LSE

THE LONDON SCHOOL
OF ECONOMICS AND
POLITICAL SCIENCE

Advances on a Methodology of Design and Engineering in Economics and Political Science

A thesis submitted to the Department of Philosophy, Logic, and Scientific Method of the London School of Economics and Political Science for the Degree of Doctor of Philosophy, 2014, London.

To Paulina, Amélie and Ulrike

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. This thesis may not be reproduced without the prior written consent of the author.

I warrant that this authorization does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of 85,504 words.

Fernando Morett

Abstract

This thesis consists of five chapters: 1.The Mechanical View, 2.Social Machines, 3.The FCC Auction Machine, 4.Self-Interested Knaves, and 5.Self-Interested but Sympathetic. In the first three chapters, I advance a methodological account of current design and engineering in economics and political science, which I call methodological mechanicism. It is not ontological or literal; it relies on a technological metaphor by describing market and state institutions as machines, and the human mind as consisting of a number of mechanisms.

I introduce the Mechanical view on scientific theories as distinct from the Syntactic and the Semantic views. The electromagnetic theories from the nineteenth century are used to illustrate this view as well as the use of minimal and maximal analogies in model-building in normal and revolutionary science. The Mechanical view is extended to the social sciences, particularly to mechanism design theory and institutional design, using the International Monetary Fund, the NHS internal markets and the FCC auction as examples. Their blueprints are evaluated using criteria such as shielding and power for calculating joint effects as well as libertarian, dirigiste, egalitarian and inegalitarian properties; and the holistic and piecemeal engineering they adopt. Experimental parameter variation is introduced as a method complementing design.

Any design assumes a particular moral psychology, so in chapters four and five I argue that the moral psychology of universal self-interest from Bernard Mandeville, and the related ideas on design and engineering, should be chosen over the moral psychology of self-interest, sympathy and sentiments of humanity from David Hume. Hume finds no solution for knavery in politics and civil society. He accepts egalitarianism as useful and consistent with utilitarian principles; however he rejects it because of some difficulties with its implementation. I show how those difficulties may be overcome, and I explain why his objections are unbalanced and not sufficiently justified.

Acknowledgments

First and foremost, I would like to thank Professor Nancy Cartwright for her support and dedication as a supervisor. Her contribution to the philosophy of science has been a great inspiration to me.

I also would like to thank Professor Rom Harré and Professor John Dupré, who very kindly accepted the offer of being the examiners of this thesis. It has been a great honour to have their wisdom and experience in the evaluation and improvement of my work.

I am very grateful to the department of Philosophy, Logic and Scientific Method at the London School of Economics and Political Science (LSE), and the Mexican Council for Science and Technology (CONACYT) for their financial support throughout my PhD.

I would like to thank the Directors of the Doctoral Programme and the Heads of the Department of Philosophy, Logic and Scientific Method at the LSE for their invaluable support and advice throughout my studies, research and teaching. I am also very grateful to all Professors and Lecturers at the Department of Philosophy, Logic and Scientific Method for their advice and teaching, which helped me to improve my knowledge and skills. I am very grateful to Alex Voorhoeve for his advice and recommendations on an early draft of chapter five.

I also would like to thank my PhD fellows from whom I learnt so much, and whose companionship and cheerful conversations I always enjoyed.

I thank all my students with whom I have shared the daily experience of doing philosophy, I learned many things from them.

It has been a great honour and a pleasure to be part of the vibrant LSE community that is so full of knowledge, new ideas and people from all over the world.

CONTENTS

‘The curious task of economics is to demonstrate to men how little
they really know what they can imagine they can design’

Friedrich von Hayek

The Fatal Conceit

Introduction

The aim of this dissertation is to present the results of my research on the methods of design and engineering in the social sciences and the philosophy about them. Design and engineering have experienced a new interest and growth in the social sciences, particularly in economics and political science. Branches such as institutional design, mechanism design theory, experimental economics and analytical sociology are examples of this methodological stance, which I call *methodological mechanicism*. This stance must be distinguished from methodological naturalism in the social sciences, which is based on methods taken from the natural sciences such as functionalism, which is adopted from biology. Functionalism is based on an organic metaphor, while mechanicism is based on a technological metaphor.

Methodological mechanicism is taken from engineering and it relies on the *machine metaphor,* which describes institutions as machines made of different mechanisms. Efficiency and reliability are defining properties of machines, and therefore they become defining properties of institutions. Such metaphor and method are justified because of the good results they can lead to, and they can be critically revised and even abandoned if they cease to be useful. Because it is methodological and metaphorical, this stance carries no ontological commitment trying to turn any person and institution into real machines. Such a possibility is open to a monistic materialism or physicalism, and the reductionism needed to reduce the psychological and biological to the physical.

Methodological mechanicism and the machine metaphor are pervasive in economics, political science and sociology, and even in psychology where behaviour is explained through mechanisms operating in the mind. Criticisms have been made of this view because it can turn any person into an automaton and the whole society into a collection of them. This criticism is fair only to the extent that an alternative view is provided or considered, where a concern with efficiency and reliability can be resolved, considerably reduced or abandoned. Otherwise, such a method and metaphor can continue having a place in science and philosophy. This criticism may be unfair if it attributes to this stance the kind of ontological commitment just described. Therefore, my argument and defence of this methodological stance and metaphor do not adopt this commitment; my argument and defence are conditioned to the effectiveness and fruitfulness of this stance, and any criticism welcome.

Besides scientists, philosophers too have adopted this stance. The Mechanical philosophy of the seventeenth century is an example. Currently, philosophers such as Nancy Cartwright and Jon Elster adopt this view. Outside the work these philosophers have published on social machines and social mechanisms, there is hardly any substantive further work published by philosophers on these topics. Therefore, the results I present in this dissertation were developed from a rather small body of literature.

The dissertation consists of five chapters, which can be separated into two parts. The first part covers the first three chapters, and the second part the last two chapters. The second part covers the topic of moral psychology, and the first one covers methodological aspects. In the methodological part, the Mechanical view is introduced and applied to both natural science and social science, particularly to the electromagnetic theories from the nineteenth century, analytical sociology, institutional design and mechanism design theory, which is illustrated with the case of the multiple-round simultaneous ascending auction.

Because any design in economics and political science assumes a determined moral psychology, the last two chapters remain related to the first chapters, in particular to chapters two and three where mechanism design theory is discussed. There is a vast body of literature on current moral psychology, so I decided to work instead on the moral psychology from the eighteenth century where new contributions can be made, and also because during this period the foundations of the current debate were established.

In chapter one, I introduce the Mechanical view in opposition to the Syntactic and Semantic views. I illustrate the Mechanical view with the case of electromagnetic theories from the nineteenth century. The role of metaphor and analogy in models is discussed, and a rule on minimal analogy is also introduced which can be applied to type-hierarchies. Type-hierarchies are representations of natural and social kinds ordered according to their level of generality forming a pyramid or a three-like classification. Type-hierarchies, metaphor and analogy are all part of the Mechanical view.

In chapter two, I review the current application of the Mechanical view to the social sciences, in particular to mechanism design theory, analytical sociology and institutional design. I concentrate on the production of blueprints for new institutions and the methodological principles which can regulate such a production. I discuss the principles advanced by Nancy Cartwright for the design of sociological machines. A further methodological evaluation of blueprints is made by distinguishing between holistic and piecemeal engineering, and also by distinguishing between libertarian and dirigiste designs.

Chapter three examines the multiple-round simultaneous ascending auction, which was a new kind of auction designed and built by mechanism design theorists and experimental economists. This new kind of auction was used for the allocation of licences to telecommunication firms for the use and the exploitation of the electromagnetic spectrum. I describe and evaluate the blueprint for this auction using the types of design and principles discussed in chapter two, and the rule of minimal analogy and type-hierarchies from chapter one. I also introduce the method of experimental parameter variation from aeronautical engineering to account for the experiments performed as part of the design and implementation of the multiple-round simultaneous ascending auction.

In chapter four, I discuss the moral psychology of self-interest from Bernard Mandeville. The aim is to carry out an epistemological evaluation of such moral psychology using the standards from the eighteenth century. I also discuss the method used by Mandeville and his refutation of the moral psychology of Lord Shaftesbury. Mandeville's definition of vice is explained as a case of functional explanation in contrast to those explanations attributing a specific moral philosophy to him. In the last part, I discuss Mandeville's ideas on design, which emerge from his psychology of self-interest. In particular, his ideas for the prevention of knavish behaviour in politics and his blueprint for a commercial precapitalist society.

In chapter five, I discuss the moral psychology of self-interest and sympathy from David Hume, and his refutation of Mandeville's moral psychology. I discuss Hume as an early modern scientist, so I also examine his descriptive sociology of utilitarian morality, and the ideas on design which can be extracted from his work. In particular, his criticisms and rejection of egalitarian distributive justice and his own ideas for the prevention of knavish behaviour in politics. The quotations from the works of David Hume, Bernard Mandeville and other authors from the same period have been made keeping the old spelling of words used in the critical editions, so the reader may see words such as 'mony' instead of 'money' or 'controuling' instead of 'controlling', and so on.

Chapter 1

The Mechanical View

1.0. Introduction

In this chapter I argue for the following four theses: The introduction and characterisation of the Mechanical view on theories in opposition to the Syntactic and Semantic views. The definition of analogy as a class and the definition of a class as a collection of analogies, this property of interdefinability is added to the argument on the logical validity of an inference from analogy. The role and importance of intensional criteria in establishing any analogy, class and type-hierarchy, such criteria underpin the descriptive elasticity needed for the creation of diverse analogies and systems of classification over the same ontology of properties and causal structures. The introduction of minimal analogy supported on a methodological rule advocating minimal similarity for the construction of models and type-hierarchies, this rule and type of analogy are defined in contrast to the opposite rule and the corresponding maximal analogy, which is commonly presented as the only relevant analogy.

These theses are put forward through the discussion of five topics organised in seven sections. These five topics are the distinction between the Syntactic, Semantic and the Mechanical views of theories; the definition of analogy; the validity of the inference from analogy; the use of metaphorical language in science; and type-hierarchies as a solution to some problems related to the use of analogies and metaphors.

In section 1.1., the differences between the Syntactic and the Semantic views are discussed through the work of Rudolf Carnap and Bas van Fraassen using geometry as an example. Two main aims justify this choice. The first one consists of making a contrast between the Syntactic and the Semantic views and the Mechanical view. The Syntactic view argues for the elimination of geometrical shapes and any graphic model, the Semantic view keeps those shapes and models but only as means for identifying structures. In the Mechanical view shapes, models and knowledge of mechanisms constitute the foundations of science. The second aim emerges from an interest in blueprints from design and engineering in the natural and the social sciences. Geometrical shapes are graphic models for axioms, theorems and equations. A relationship could therefore be established between graphic models and blueprints, and analogical inference and other kinds of inference performed with them. This aim is only partly achieved in this dissertation because I concentrate on the social sciences, where no blueprints in graphic format are used. The application to design and engineering in the natural sciences waits for a separate work.

In section 1.2., I introduce and characterise the Mechanical view as a third view on scientific theories besides the Syntactic and the Semantic views. Currently, there is no such term and category which unifyies the work philosophers of science such as Norman Campbell, Mary Hesse, Rom Harré, Nancy Cartwright, and Ronald Giere. Currently the work of these and other related philosophers is either placed as part of the Semantic view, or it remains an orphan with no family and no generic name or characterisation. Each philosopher is therefore treated separately, or is regarded as unrelated or weakly related to others. The introduction of the Mechanical view as a comprehensive position within the philosophy of science has at least three advantages: First, it unifies apparently dissimilar and unrelated positions economising and enhancing both analysis and understanding as well as helping the reappraisal of the work done by forerunners. Second, it helps to correct the wrong classification of the work from philosophers like Ronald Giere, whose work is placed as part of the Semantic view. Third, along with the Syntactic and the Semantic views, the Mechanical view exhausts virtually all philosophical research done on models, and in other areas in the philosophy of science. A unified characterisation can bring benefits to the Mechanical view itself by systematising and empowering its own view and future research.

In section 1.3., I discuss Mary Hesse's definition of analogy as a dyadic relation between two objects with similar and dissimilar properties divided into positive and negative analogies respectively. In contrast, I put forward a definition of an analogy as a class by showing that a class relies on an unacknowledged analogy by gathering items which are not identical to each other but similar and dissimilar. An analogy, therefore, becomes the small unit of class by relating two objects. The relevant difference between a class and analogy is the larger size of the negative analogy in the latter. The *interdefinability* of class and analogy shows a continuous line between the two, and it complements Hesse's argument on the validity of the inference from analogy. The size of the positive and the negative analogies can be modified using different intensional criteria, which have an important effect on the formation of any analogy and any class. Analogy and class can be turned into one or the other by enlarging the size of the negative and positive analogies. I call this property *descriptive elasticity*, which also has important effects on the formation of type-hierarchies as representations of natural kinds.

In section 1.4., Hesse's argument on the validity of the inference from analogy is discussed in the three different frameworks that she considers, namely the logical interpretation of probability, the method of falsification and Bayesianism. The validity or, more precisely, the justification of an inference from analogy relies on a rule prescribing the selection of the hypothesis and model more similar than others to the event to be explained or predicted. This selection is made before any test. The justification from the logical interpretation and falsification fail, so Hesse found the answer in Bayesianism, where a high subjective prior probability can be given to the more similar hypothesis and model supported on relevant causal knowledge. Besides subjective probabilities, such a Bayesian argument relies on two other components, namely a cluster postulate and the property of exchangeability of probability values given to individual events within a class. As a complement to these two components, I am adding the interdefinability of analogy and class, so that the exchangeability of the prior probabilities given to individual events can be applied.

The definition of analogy and the argument on validity given by Hesse fall into the kind of logicism criticised by Rom Harré. I made the choice of engaging with Hesse's logical arguments whilst aware of the criticism because I believe it has historical value, and also because it is still relevant in current logic and Bayesianism. At this stage of my research, the choice made does not mean the rejection of other possible justifications of an inference from analogy to be provided from cognitivism, pragmatism or other positions in philosophy.

Section 1.5., contains a discussion on the use of metaphor in science from the seminal work Rom Harré published on this topic. He identifies two accounts of metaphor, the comparative view and the interactive view. The first one explains metaphoric meaning with reference to an original source of literal meaning, the second one holds that the introduction of a metaphor also affects the original source by changing its initial meaning. Hesse and Harré support the interactive view, however Harré has two criticisms. The first one is about the lack of explanation as to why some metaphors are selected while others are dismissed; he calls this the problem of principled filtering, which also affects analogy. The second one is about establishing the semantic and logical priority of literal and metaphorical language. Hesse believes metaphor is logically prior, while Harré believes there is no fixed foundation. He argues instead for a historic explanation, where both metaphoric and literal languages are constantly shifting places, or are superseded due to shifts in the meaning of central scientific terms. A third related problem he identifies is the risk of establishing trivial analogies and metaphors unless criteria on relevance are provided. He offers a unified solution to these three problems through the use of type-hierarchies, which I also discuss adding the problem of inference from analogy.

In Section 1.6., type-hierarchies are characterised and analysed against the four problems mentioned above. Type-hierarchies are used by Eileen Way as graphic tree-like representations of natural kinds; they are on the side of the mind, while natural kinds are on the side of the world. Types are organised hierarchically in a pyramidal shape having at the top the most general types called supertypes; types and subtypes are placed below and tokens lie at the bottom. There are two important semantic properties of type-hierarchies: Their masking effects and meaning shifts. Each type is built from a semantic mask, which hides and exposes different aspects of natural kinds, while a meaning shift causes the reshaping of the hierarchy through the introduction of a new supertype. Supertypes cause the largest meaning shifts; smaller shifts also take place in the lower levels. Because of the constant meaning shifts, literal and metaphoric languages have no logical priority over one another.

When type-hierarchies are large enough, they can prevent trivial similarities as well as reduce the use of ad hoc criteria by guiding the selection of relevant similarities from the properties that are inherited from a supertype or a type to any token. Such an inheritance of properties also supports an inference from analogy, which can even be deductive since tokens just need to be placed under the right supertype resembling inferences made with the covering-law models of explanation and prediction. This can only happened when the hierarchy is almost fully built; with a half-built hierarchy ad hoc criteria are used instead. This process captures how type-hierarchies work in normal science, it does not tell us how new supertypes are introduced and their hierarchies built in revolutionary periods. I introduce *minimal analogy* as a solution to this problem.

In the last section 1.7., I argue for a mixed methodology containing minimal and maximal analogies, which I claim is more robust and has a larger scope. Two scenarios are distinguished. The first one, when the next type above or a supertype is already available; and a second one when there is no such a type or supertype. The second case implies a meaning shift and a new semantic mask, and it therefore corresponds to a revolutionary stage. Mary Hesse's analogy is a maximal analogy, and it therefore has better prospects in normal science. I define *minimal analogy* as an analogy with a larger negative analogy, and any analogy with a larger positive analogy is called a *maximal analogy*. Hesse did not consider the case of inferences made with minimal analogies; her work was exclusively concerned with cases of maximal analogy.

In normal science a minimal analogy is still an option chosen as the means for speeding up the construction of a type-hierarchy, however in revolutionary science it is necessary. A methodology with one rule only prescribing maximise similarity is too conservative, and in some cases it can be recessive and even regressive affecting scientific progress. A rule on minimal similarity encourages progress but implies a greater risk of failing. Therefore, a mixed methodology can offer both protection and progress. I illustrate this by comparing the models of Michael Faraday and William Thomson on magnetic fields with the models of James Maxwell on the electromagnetic and gravitational aethers. The latter are fully mechanical models, which maximised similarity with the dominant Newtonian paradigm, while the former models minimised similarity by introducing a new supertype, namely a field ontologically distinct from matter. The models from Faraday and Thomson actually anticipate important aspects of the revolution introduced by Albert Einstein nearly a century later.

The ideas of minimal and maximal analogy, type hierarchies and mechanisms as a metaphor and as part of the Mechanical view applied to design and engineering in the social sciences, are presented in chapters two and three through the discussion on blueprints and the design of the FCC auction.

1.1. Syntactic and semantic geometry

Carnap distinguishes among three scientific ‘word-languages’: arithmetic, axiomatic and physical. He uses geometry to illustrate the differences between them.[1] The use of geometry is particularly relevant because, if there is a place where the importance of graphics and graphic reasoning should be acknowledged, it is in geometry. Carnap highly praised the metamathematical method of arithmetisation developed by Kurt Gödel. With it, Gödel intended to exhibit the structure and order of mathematical propositions using natural numbers as a language of translation, by establishing a one-to-one correspondence between natural numbers

[1] R. Carnap (1934), p. 78-82, §25.

and those mathematical propositions. René Descartes proceeded in a similar fashion by using pairs of numbers on a Euclidian plane as an algebraic translation of any geometrical shape. In an important basic sense, arithmetisation is a syntactical translation—an *explication* in Carnap's own terms—which serves as a method of *logical proof* when mathematical expressions can be deduced from the so constructed metalanguage of natural numbers.

In the case of geometry, all shapes are arithmetised by assigning ordered triads of real numbers, and the linear equations constructed with them: 'a point is interpreted in the usual way as a triad of real numbers, a plane as a class of such triads which satisfy a linear equation, and so on.'[2] By doing this, all shapes in geometry disappear by being arithmetised through the assignment of ordered triads of real numbers, and the linear equations constructed with them: 'a point is interpreted in the usual way as a triad of real numbers, a plane as a class of such triads which satisfy a linear equation, and so on.'[3] Therefore, arithmetisation becomes an *eliminative method*, where all shapes and graphic models disappear. The graphical proof of Pythagoras theorem or the law of cotangents using triangles, square and circles is replaced with a syntactical proof produced using natural numbers as a metalanguage.[4]

Physical geometry comprises the set of 'definite synthetic sentences which state the empirical (namely the geometrical or graphical) properties of certain physical objects', for instance, 'these three objects A, B, C are light-rays in a vacuum each one of which intersects the other two at different points.'[5] Carnap argues that besides producing physical descriptions, scientists must also axiomatise their own theories. In the case of Euclidian geometry, such axioms were produced by David Hilbert, i.e. axiom of parallels, axiom of continuity and so on. Hilbert's axiomatic geometry contains twenty-one axioms, which any physical sentence can be related to by using 'correlative definitions'. The philosophical task is again syntactic and logical,

[2] *Ibid.*, p. 274, §71e.

[3] *Ibid.*, p. 274, §71e.

[4] Carnap explains that unlike Wittgenstein he wanted to do more than just *showing* the syntax of scientific language; he wanted to *express* it using a formal language. Arithmetisation, therefore, becomes an *explication* of the syntax of in geometry; see R. Carnap (1934), p. 53, §18,; and (1962), pp. 1-18.

[5] R. Carnap, (1934) p. 81, §25.

which consists of *explicating* the order and kinds of words of physical or empirical sentences, equations and axioms by using a metalanguage, and *proving* the deductive order of sentences, equations, axioms, theorems, and any other scientific proposition.[6]

For reasons of exactness, clarity and simplicity; axioms were selected by Carnap as the standard canonical way of expressing the terms and propositions contained in scientific theories. Following Gottlob Frege,[7] he criticised the inexact and often hazardous expression of scientific terms and propositions published by scientists in articles and books. Hence, his aim was to render these concepts and propositions exact and closed under the relation of logical consequence, having axioms as a foundation. Inexact physical descriptions with loose ends were and still are common in science. In contrast, axioms are scarcely used to express the basic terms and propositions of scientific theories.

The logical explication consists of making explicit the syntax of three different sets of scientific propositions, namely equations, axioms and definite empirical sentences by identifying features such as extension: existential or universal; size: atomic or molecular; composition: conjunctive, disjunctive or conditional as well as the sequences of reasoning performed with these elements, leading to normative patterns with the form of a modus tollens, a destructive dilemma and so on.

By doing this, philosophy becomes concerned only with sentences and their logical syntax. Any geometrical shape is reduced to triads of real numbers and equations for each physical dimension. Geometry, a basic candidate for graphic reasoning, vanishes by being reduced to sentential descriptions. Inference from graphics, a cognitive activity so crucial to geometricians, simply disappears. The same *eliminative method* could, in principle, be extended to any model and any other graphic means used in science such as diagrams, photos, engravings and blueprints.

[6] Carnap explains that besides first-order predicate logic, arithmetisation is also needed in some cases, so it must be considered as an explication method, see R. Carnap (1934), pp. 57–58, §19.

[7] See G. Frege (1879), pp. 5–8, and (1979), pp. 12-13.

Bas van Fraassen offers an alternative to Carnap's syntactic geometry. Following Alfred Tarski, he argues for *models* as the standard for expressing the content and truth-value of scientific theories, with the ultimate task of identifying isomorphic structural relations among those models and data from the world. Within the Semantic view, models comprise both set-theoretic mathematical and graphic models such as Niels Bohr's model of the atom. Accordingly, van Fraassen uses a Fano plane, also called Seven Point Geometry, as a model for the following four axioms:

A1. For any two lines, there is at most one point that lies on both.

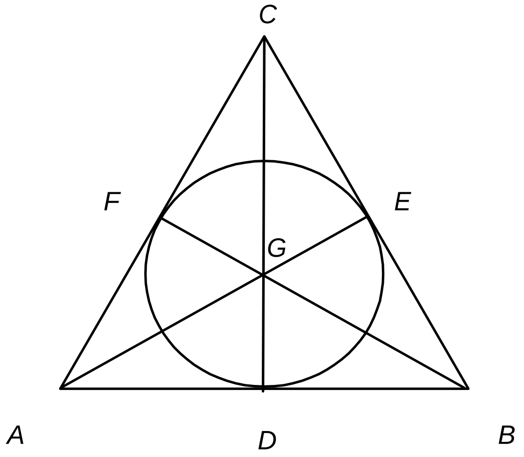
A2. For any two points, there is exactly one line that lies on both.

A3. On every line there lie at least two points.

A4. There are only finitely many points.

Van Fraassen argues that 'logical claims, formulated in purely syntactical terms, can nevertheless often be demonstrated more simply by a detour via a look at models',[8] therefore the four axioms can be proven not by using a logical metalanguage but *by reasoning from a graphic model*, namely the Fano plane below, which consists of a geometry of the seven points *A* to *D*.

Figure 1.1. Fano plane.



[8] Van Fraassen (1980), p. 43; Seven Point Geometry in p. 42.

Such a visual demonstration, however, still requires the help of the following set of sentences for the interpretation of the image: "In this structure only seven things are called 'points', namely A, B, C, D, E, F, G. And equally, there are only seven 'lines', namely, the three sides of the triangle, the three perpendiculars, and the inscribed circle. The first four axioms are easily seen to be true of this structure: the line DEF (i.e. the inscribed circle) has exactly three points on it, namely D, E, and F; the points F and E have exactly one line lying on both, namely DEF; lines DEF and BEC have exactly one point in common, namely E; and so forth."[9]

Unlike Carnap's syntax of word-languages, van Fraassen's semantics keeps geometrical shapes as models for demonstrating axioms. Philosophically, this is a very important choice. First, because it lays out some common grounds with the Mechanical view, where graphic models are taken as fundamental in science. Second, because it supports graphic reasoning, that is, it accepts that scientific inference can be based on models and other graphics means. By doing this, philosophical research is not anymore constrained to word-languages. This is a very important step for a methodology of design and engineering, where blueprints are fundamental.

Despite its prominence in science, inference from models has received scarce attention from philosophers of science and, more specifically, from logicians. Most of the philosophical research has been concerned with ontological and metaphysical aspects of models as well as their function as suppliers of truth conditions and further empirical content of scientific theories. Despite its interest in models, the Semantical view is not in a better position because virtually no further attention has been paid to inference from models. Because of the main interest of this thesis on blueprints, I concentrate on graphic models depicting mechanisms. Therefore, I do not discuss mathematical model-theory or any graphic means used in mathematics such as Euler or Venn diagrams or any Cartesian plane.

[9] *Ibid.,* p. 43.

1.2. The Mechanical view

Besides the Syntactic and the Semantic views there is the *Mechanical view*. This is a term and a description I am introducing covering a number of contemporary philosophers with closely related arguments and proposals. I place the physicist and philosopher of science Norman Robert Campbell as the founder of this view. Besides Campbell, the Mechanical view encompasses the work of Rom Harré, Mary Hesse, Nancy Cartwright, and Ronald Giere among others. This view emerged with a more defined shape in 1960s through the work of Mary Hesse and Rom Harré, who were inspired by the work of Campbell.

Norman Campbell argued against the methodological reduction of physics to mathematics as it had been pursued by scientists such as Ernest March and Henry Poincaré, who 'were primarily mathematicians and not experimenters.' Campbell drew a distinction between 'mechanical theories' and 'mathematical theories' in physics rejecting 'the view that theories of the second kind are in any manner superior in value or certainty to those of the first […] it is simply asserted that such [mechanical] theories alone can attain the ultimate end of science and give perfect intellectual satisfaction.'[10] This was his main thesis; he wanted to restore the value of mechanical theories in physics, which he claimed are supported on models depicting analogies between events from different domains.

Currently, entries and articles on models in encyclopaedias of philosophy and edited volumes do not register the Mechanical view as a unifying position, and they do not use either any other term identifying this position in the philosophy of science. Usually, the Syntactic and Semantic views are discussed as the only systematic unified positions, and then a number of main authors and problems are listed separately and discussed as unrelated, or as weakly or randomly related with one another, which all belong to the Mechanical view as I present it here. Moreover, from those female and male philosophers belonging to the Mechanical view, there are comparatively fewer systematic books with a comprehensive treatment than in the Syntactic

[10] N. Campbell (1920), pp. 8, 154-155.

and the Semantic views. The explications and discussions in main sources of reference such encyclopaedias and handbooks are no doubt relevant and philosophically rigorous, but they become too dispersed and somehow cumbersome, when they addressed the work of philosophers belonging to this view. See for instance the entries on models in the Stanford and the Rutledge Encyclopaedias of Philosophy, the volume edited by Mary Morgan and Margaret Morrison, and the comprehensive survey on models written by Daniela Bailer-Jones.[11]

Back in the early twentieth century, Pierre Duhem drew a methodological distinction between the 'abstract mind' of French and German scientists, and the 'visualising mind' of the English scientists. The abstract mind produces axioms and equations associated to perfect geometrical shapes representing real objects, and it performs all inferences through rigorous deductive steps.[12] In contrast, the visualising mind relies on mechanical models picturing imperfect real objects: axioms are not required while equations often have an instrumental role by being epistemically less important than graphic models. Models do the ultimate and more fundamental epistemic job by exhibiting and demonstrating the mechanisms through which nature operates. Duhem points out that rigorous deduction is replaced with 'rough analogies', which are 'a regular feature of the English treatises on physics. Here it is a book intended to expound the modern theories of electricity and to expound a new theory. In it there are nothing but strings which move around pulleys, which roll around drums, which go through pearl beads, which carry weights; and tubes which pump water while others swell and contract; toothed wheels which are geared to one another and engage hooks. We thought we were entering the tranquil neatly ordered abode of reason, but we find ourselves in a factory.'[13]

Indeed, we enter into a factory not only by opening that book from the nineteenth century English physicist Oliver Lodge, but we also do by opening the books from current

[11] M. Morgan and M. Morrison (1999), D. Bailer-Jones (2009); see also R. Frigg (2006a).
[12] A representative criticism from the Mechanical view on deductive rigour and formalisation in economic models can be read in N. Cartwright 'The Vanity of rigour in Economics: Theoretical models and Galilean experiments', in her (2007) *Hunting Causes and Using Them*.
[13] P. Duhem (1906), pp. 70-71, 56-57; the book Duhem is referring to is by Oliver Lodge (1889) *Modern Views of Electricity*.

philosophers of science such as Rom Harré, Nancy Cartwright or Ronald Giere, where images, diagrams, and other graphic means play a main role.

The introduction of the Mechanical view as a comprehensive position within the philosophy of science has at least three advantages. First, it unifies apparently dissimilar and unrelated positions economising and enhancing both analysis and understanding, as well as helping the reappraisal of the work done by forerunners.[14] That is, it allows the reappraisal and unification of the early work from Norman Campbell, Mary Hesse and Rom Harré with the most recent one from Nancy Cartwright, Ronald Giere, Margaret Morrison, Nancy Nersessian, David Gooding and others. Second, it helps to correct the wrong classification of the work from philosophers like Ronald Giere, whose work is placed as part of the Semantic view.[15] Third, along with the Syntactic and the Semantic views, the Mechanical view exhausts virtually all philosophical research done on models, and in other areas in the philosophy of science.

Among the female and male philosophers and historians just named as part of the Mechanical view there are of course differences. For instance, for some induction and logic play a crucial part, while for others reasoning from analogy and cognition are a fundamental part of science. In spite on these and other differences, the prominent place given by all of them to mechanical model is, I believe, strong enough to support this classification. In sum, I argue that the addition of the Mechanical view is insightful and general enough by allowing a quick and comprehensive look into the current debate on models, and more generally, in the philosophy of science.

Against the Syntactic view,[16] the Mechanical view rejects the elimination of models and causal powers, and it also rejects the idea that scientific language provides a literal description of

[14] Unlike the Syntactic and the Semantic views, the Mechanical view did not have a continuous and more cohesive and systematic development; some aspects and authors from this view are discussed D. Bailer-Jones (2009).

[15] See R. Frigg (2006b), p. 52; N. da Costa and S. Frech (2000), p. S119; and M. Morgan and M. Morrison (1999), p. 3-4.

[16] See C. Hempel (1965), pp. 433-447, and R. Carnap (1939), who argues that when 'Maxwell's equations of electromagnetism, were proposed as new axioms, physicists endeavoured to make them "intuitive" by

the world. It argues instead for the use of models, especially those depicting theoretical mechanisms and entities, which involve the vindication of causal powers. It also highlights the constitutive role of analogy and metaphor in those models, and the explanations and predictions made with them. Its own defence of inference from analogy is supported on single cases,[17] in contrast to a large number of cases, which is typical of induction and laws as defined by the Syntactic view. Because of its defence of mechanisms, causal powers and theoretical models, the Mechanical view is largely realist in opposition to the empiricism of the Syntactic and the Semantic views.

Graphic models like the Fano Plane are a common ground for the Semantic and Mechanical views; this explains why the work of philosophers such as Ronald Giere is mistakenly placed as being part of the Semantic view. Unlike this view, models in the Mechanical view are not used as means for establishing isomorphic structures among models and data from the world, nor for the interpretation of axioms or any other formalisation in a scientific theory. In the Mechanical view, knowledge of mechanisms is placed at the core of scientific models and scientific labour, such knowledge is the ultimate aim of science. In this view models are graphic representations of causal mechanisms; they are the means to expose those mechanisms. A mechanism is a cohesive arrangement of causes regularly producing an effect. Within this view, models are used for at least three outstanding purposes:

- As means for justifying new theories as well as for expanding and refining current ones
- As means for rendering scientific claims true
- As means for improving scientific and technological intervention in the world.

constructing a "model"… It is important to realize that the discovery of a model has no more than an aesthetic value or didactic or at best a heuristic value, but it is not at all essential for a successful application of the physical theory', pp. 67-68.

[17] See N. Cartwright (1989), p. 56ff; and (1992), p. 51.

With the term 'models as mediators' Morgan and Morrison tried to grasp and summarise much of the work done by philosophers working in the Mechanical view since 1980s. Such mediation between theories and the world is exposed mainly in two ways. The first one concerning the truth-value of scientific claims; the second one concerning scientific intervention into the world.

In the first one, models are the real providers of any empirical content in science, that is to say, when laws and theories are taken at face value 'they lie'—to use Cartwright's phrase—only models tell us the truth. Particularly, what she calls 'representative models', which contain a detailed description of the empirical domain of concern, often described as 'target system'. Cartwright asserts that 'theories in physics do not generally represent what happens in the world; only models represent in this way, and the models that do so are not already part of any theory.' [18] Morgan and Morrison hold almost the same thesis by criticising the conception of models as mere derivations from theories, or as simplifications of them. They argue that 'models should no longer be treated as subordinate to theory and data in the production of knowledge' but as independent and autonomous.[19] Models are autonomous because they actually help produce new causal explanations and new measurements, which cannot be derived from the theory or the data themselves.[20]

The centrality of models is also held by Ronald Giere, who claims that scientific theories comprise 'a population of models' and 'various hypotheses linking those models with systems in the real world'.[21] Such models are not set-theoretic but they are mechanical models. His preference for graphic mechanical models clearly places him into the mechanical tradition, and away from the Semantic view, which he actually criticises. He rejects isomorphism as the hypothesis explaining the relationship between scientific models and the world, and he argues instead for a relation of similarity. Also, against the Semantic view, he rejects van Fraassen's

[18] In M. Morgan and M. Morrison (1999), p. 242.

[19] *Ibid.*, p. 36.

[20] *Ibid.*, pp. 13, 21; also there see article by M. Suarez in pp. 168-196.

[21] R. Giere (1988), p. 85,

empiricism, arguing instead for a variety of realism.[22] A realist position is also shared by Harré and Cartwright.

The second aspect concerning scientific and technological intervention is one of the most recent developments within the Mechanical view. Nancy Cartwright has produced the first work and analysis with a clear focus on the implementation of social and economic policies. In particular, she has focused on blueprints regarded as a particular type of model. Her work on blueprints is discussed in the next chapter.

The pioneering work of Mary Hesse and Rom Harré on models is largely addressed to the production and justification of new scientific theories. Instead of using terms like 'normal' and 'revolutionary science', or 'progressive' and 'degenerative research programmes', Harré and Hesse use the term 'theory construction' as a description covering the creation of the new theories, their refinement and expansion. Such a term was a response to the distinction made by Logical Positivist philosophers between the contexts of discovery and justification. The term theory construction is also associated to the cognitive foundations of science adopted by the Mechanical view in contrast to the logical foundations pursued by Logical Positivism. Philosophers like Rom Harré and Ronald Giere explicitly state their methodological commitment to the cognitive approach, while others like Morgan and Morrison use the term 'learning' instead.

The Mechanical view can be summarised in the following six components:

i. Graphic models as central to science

ii. Vindication of causal powers and mechanisms

iii. Key role of single case inference with and without analogy

iv. Realism predominates

v. Metaphorical terms as important part of scientific language

vi. A concern with the use of models for intervention

[22] *Ibid.*, p. 80-82, 92-106.

The first four are the most widely shared aspects, while the last two are less widespread. In this chapter, I only discuss numbers one, three and five.

1.3. What is an analogy?

In philosophy of science, one of the easiest and common ways of making a normative recommendation on scientific inference consists of appealing to deduction and demanding complete information. Deduction is, of course, the safest inference for getting true conclusions. Induction can be considered as the second best option just because it can yield false conclusions. It is a widely spread habit in philosophy to describe as 'heuristics' almost any other type of reasoning outside deduction and induction. Such a general and indiscriminate use of the term heuristics actually creates negative effects preventing the work philosophers and logicians should do, explicating and evaluating with due care and enough detail the diversity of inferences. There is far more literature published and research done on deductive and inductive logic than in any other kind of inference, which creates a significant disadvantage for the remaining inferential diversity. In philosophy, the challenge persists on explaining and producing norms for other kinds of inference outside induction and deduction. Among others, psychologists such as Gerd Gigerenzer and Daniel Kahneman have being doing the job instead.[23]

Inference from analogy is an example of a non-deductive and non-inductive type of inference, which Carl Hempel thought could only provide 'heurist guidance' as part of 'the pragmatic-psychological aspects of explanation', but could not have a 'logic-systematic role in scientific theorising'.[24] There are a number of classical works on analogy such those of John

[23] See G. Gigerenzer (1999), and D. Kahneman, P. Slovic, A. Tversky (1982).

[24] C. Hempel (1965), pp. 441, 443.

Maynard Keynes, Rudolf Carnap and Mary Hesse; and some recent ones from Paul Bartha and those compiled by David Helman.[25] In this chapter, I discuss the work of Mary Hesse.

There are two basic philosophical questions Hesse asks about inference from analogy, namely *what is an analogy?* And, *when is an argument from analogy valid?* She explains that no proper answer has been given to these questions in modern logic. A common response given to the first question states that the answer is 'obvious or unanalyzable', while the response to the second one concludes that the validity of the analogy is 'highly problematic'.[26] In this section I discuss the answer to the first question, and in the next section I discuss the answer to the second one.

First, Hesse provides a basic definition of analogy as a *relation* between two objects; then she asks us to compare the earth and the moon identifying similarities and dissimilarities. For instance, both are large, solid, opaque and spherical, they revolve on their own axes and gravitate towards other bodies. In contrast, the moon is smaller and more volcanic with no atmosphere and no water. She calls the first set of common properties 'positive analogy', and the second set 'negative analogy', and she adds that 'there generally will be properties of the model about which we do not know yet if they are positive or negative analogies; these are interesting properties because they allow us to make new predictions'[27]. This last set is called 'neutral analogy'. A relation between the properties of two objects is a basic form of relation. Hence, a first definition of analogy (a) can be written as follows:

(i.a) An analogy is a dyadic relation between two objects consisting of positive, negative and neutral properties.

[25] J. M. Keynes (1921), Carnap (1980), D. Helman (1988), and P. Bartha (2010).

[26] M. Hesse (1966), p. 57.

[27] *Ibid.*, p. 8. Earlier, John Maynard Keynes (1921, pp. 217-232) had introduced the distinction between positive, negative and neutral analogy.

After providing this definition, Hesse concludes that 'the question of what the analogy is in this case is fully answered by pointing to the positive and negative analogies, and the discussion passes immediately to the second question.'[28] While this definition of analogy as a dyadic relation clearly captures essential aspects, it does not capture and exploit another important aspect, namely, the classes or kinds being related by an analogy. In principle her definition does not have to exploit other aspects unless certain benefits can be expected from trying an alternative definition. So, before passing on to the second question as Hesse is asking, I will put forward a definition of analogy shifting the attention from properties to classes, that is to say, I will define analogy as a class, and a class will be defined as a collection of analogies.

I use 'class' instead of 'kind' or 'type' because it fits better with the logical argument that Hesse writes in defence of the inference from analogy. Such a choice is subject to the criticisms Harré makes on classes and logicism. However, Hesse did work largely within the logicist framework. In later years she recognised the need for a cognitive approach but she did not write the respective cognitive response to the questions on the nature and validity of inference from analogy.[29]

By defining analogy as class I am trying to show at least two benefits. First, that such a definition can complement and simplify the answer on validity given by Hesse, in particular by showing how class and analogy are interdefinable. Second, it can enrich Rom Harre's argument on type-hierarchies by showing the important role of intensionality. The overall discussion on analogy will have a third benefit to be shown in the last chapter, where *design by analogy* is discussed in relation to artefactual kinds.

To motivate the definition of analogy as a class, I rely on the work of W. V. Quine on natural kinds, specifically on his discussion on similarity, which is equally fundamental for defining both class and analogy. Similarity is also fundamental for the type-hierarchies used by

[28] *Ibid.*, p. 58.
[29] M. Hesse (1988), p. 317.

Rom Harré. It is so essential that Quine actually claims that 'the notion of a kind and the notion of similarity seem to be substantially one notion'.[30] Furthermore, similarity is as basic as other logical notions such as identity and negation, and yet 'there is something logically repugnant to it'[31] due to the difficulties found trying to define it. For instance, a basic comparative use of similarity such as '*a* is more similar to *b* than to *c*' fails, when it is defined using set theory by explaining that *a* and *b* are more similar because jointly they belong to more sets than *a* and *c* do. Quine explains that such a definition fails because combinations in set theory are 'random'.[32] That is to say, the freedom allowed for assigning properties and making combinations can make similarity relations almost arbitrary, and therefore inadequate for the natural sciences where the determination of kinds face more constraints.

In logic a definition of similarity also fails, at least as it was attempted by Carnap, because his definition of a kind does not prevent the case where a collection of items cannot be a kind despite meeting the criteria. This occurs because Carnap's definition states that 'a set is a kind if all its members are more similar to one another than they all are to any one thing outside the set'.[33] Quine explains that a rather disparate set containing all red round things, red wooden things and round wooden things meets Carnap's definition without being a kind. This happens because such a set excludes plausible members. For instance, it excludes yellow rubber balls while it accepts yellow crocket balls and red rubber balls. More importantly for the definition of analogy, it allows great dissimilarity among members, for instance, by allowing one to place in the same class red cherries, cart-wheels and red wooden boxes.

A solution to these problems consists of making the similarity relation more precise and restrictive by requiring that all members share at least one property.[34] This criterion prevents

[30] .W. V. Quine (1969), p. 119.

[31] *Ibid.*, p. 117.

[32] *Ibid.*, p. 118.

[33] R. Carnap (1928), pp. 129–131, §80; 180-181, §111.

[34] A. Hausman (1979) and Rodriguez-Pereyra (1999) discuss the same solution to Carnap's definition of a kind. Nelson Goodman first identified the problem with Carnap's definition, which he called 'the

cases of disparate sets like that one used by Quine, while it also allows a large degree of dissimilarity, which is important for any analogy. Such a criterion for a class (*c*) provides the grounds for a first definition:

(i.c) A class is a collection of items with at least one common property.

Armed with this definition of a class, let us now return to the question on what is an analogy. Hesse defines a positive analogy as the set of common properties between two items. Such common properties constitute the similarity among those two items. Taking two items in isolation, namely the earth and the moon, they already form a class with two objects or a subclass within a larger class, namely the larger class of massive rotating bodies in the solar systems with elliptical trajectories. In this subclass, unshared properties such as having water or an atmosphere are the negative analogy; they constitute the dissimilarity among the two. All other properties constitute the neutral analogy. Because members in any class are not identical to each other, they all have unshared properties or common properties with variations in degree, size or aspect. This provides the basis for extending the idea of positive, negative and neutral analogy to any class. That is to say, any collection of items organised in a class relies on an *unacknowledged* analogy. Members in a class are not identical but similar and dissimilar; they are analogous to each other on some properties, and disanalogous on others. From this a second definition of a class follows:

(ii.c) A class is a collection of items with positive, negative and neutral analogies.

problem of the imperfect community' (1977, pp. 119-126). Carnap comments on this problem can be found in P. A. Schilpp (1963), pp. 946-947.

Because the positive analogy refers to common properties, this definition is consistent with the first definition (i.c.). And because this new definition uses analogy as the *definiens* for a class, the following question arises: how can a class be distinguished from an analogy? The number of common properties could make the difference. This is because commonly the number of those properties among all members in a class is presented as larger than in any pair of items forming an analogy. The distinction between a class and an analogy, therefore, could rely in the proportion of shared and unshared properties. With this consideration on the proportionality of properties, a third definition of a class and a new definition of an analogy can be advanced:

> (iii.c) A class is collection of items, where the positive analogy is larger than the negative analogy.
>
> (ii.a) An analogy is a class with two or more items where the negative analogy is larger than the positive analogy.

The reference to the neutral analogy can be omitted in order to have a simpler definition, without affecting it because its neutral character assumes no known similarity or dissimilarity. The terms *uniform class* and *analogical class* can be used as short descriptions for definitions (iii.c) and (ii.a) respectively. From the earlier discussion, it follows that any uniform class can be decomposed into a number of analogical subclasses by pairing up any two individual members and, conversely, any number of analogical classes can be composed into one uniform class by conjoining each analogy with the next one. This equivalence between a class and an analogy shows that class and analogy are *mutually definable*. In other words, *a class is a collection of analogies,* and *an analogy is a class*. I call this property the *interdefinability* of class and analogy. This is important because it creates a *continuous* line between the two, which will complement the half-

Bayesian argument Hesse makes on the validity of analogical inference. This will be shown in the next section.

If the difference between analogical and uniform classes is only a matter of proportion between the positive and the negative analogies, a further question arises on how to establish the size of each. For instance, we know that the earth and the moon are both members of a uniform class, namely the class of massive rotating bodies in the solar system with elliptical trajectories. But this only happens using this description because both the earth and the moon can also be presented or, more precisely, they can be redescribed as forming an analogical set by enlarging the size of the negative analogy in order to make them look different from each other, i.e. one is a planet and the other is a satellite with all the differences this implies.

There are cases where the negative analogy is initially small and yet a uniform set or class splits into two. Let us take the example used by Quine,[35] where mice were split into two separate infraclasses placing the marsupial mice apart. Phenotypically, both marsupial and placental mice still look very similar except from the pouch and a shorter pregnancy, among other few differences. The genome of both marsupial and placental mice is not available yet, and we don't know if their genotypes will be specific enough in order to allow us to differentiate, not only any placental mammal from a marsupial one but also, and far more specifically, the genus *Mus* (mice) from the species *Brown antechinus* (marsupial mice). Furthermore, any phenotypic description in biology, or any other taxonomic description in science made from observable characteristics, can actually be enlarged almost indefinitely, so that the positive and the negative analogy can be shifted to a very small or a very large size. For instance, in spite of the few obvious differences observed among the ordinary mice and marsupial mice, the size of such differences can easily be enlarged adding eating, sleeping and mating habits, food preferences, life span, frequency of eye colour, variety of skin colour, degree of cooperation and frequency of conflict, most frequent illnesses and so on. There is no rule or

[35] *Ibid.*, p. 128.

criterion that can establish one description as the unique objective or acceptable one. Any number of different orderings can be made all based on a causal structure, which explains the phenotypical features selected for the description. Let us call this *descriptive elasticity* the possibility of enlarging positive and negative analogies effecting the shape, size and number of classes and analogies. Intensional diversity leads to descriptive elasticity.

The philosopher John Dupré holds similar views in his analysis of natural kinds when he explains that 'what appears to be missing is not so much reality—or even existence independent of our classifying activity, since presumably things do have weights whether or not we know it—but significance.'[36] My previous discussion has shown that extensional criteria of size and proportion applied to properties are not sufficient; and because of the scope and diversity of other available properties intensional criteria have to be added, for instance prioritizing some phenotypical features over others. But intensionality is a source of significance as Dupré is asking. This leads us to the last definition of class and analogy:

> (iv.c) A class is a collection of analogies, where the number of positive analogies is larger or more significant than the number of negative analogies.

> (iii.a) An analogy is a class with two or more members, where the number of negative analogy is larger or more significant than the number of positive analogy.

Rom Harré also believes that extensionality as a criterion is, by itself, insufficient because two or more classes can have the same extension with clearly different intensional criteria and he also criticises intensionality. Harré sees a problem when 'classes corresponding to the two intensions

[36] J. Dupré (1993) p. 18; in p. 94 Dupré explains how in the absence of materialist monism and success in reductionism, Quine's argument on ontological commitment leads to an ontological pluralism.

in the hierarchy will have to be constituted by the same members',[37] which also implies membership to more than one higher order class or type. For instance, if the class of all things with shape is coextensive with the class of all things which are coloured, a red item such a red cherry can be a member to both classes and also to the genus *Prunus*, to the class of all organic things, and to the class of all sweet edible fruits.

Unlike Harré, I believe multiple membership to classes or types is not a problem but a virtue, because it provides the grounds for the growth and diversification of knowledge, when different orderings and links within the same set of items are produced. This is consistent with Harré's support of 'multifarious systems of classification'; he points out 'that there are many orders of natural kinds [...] what type-hierarchy we choose to work with may depend on the type of problem we are trying solve'; the only constraint is 'the causal structure of the world'.[38] In other words, different orderings are allowed as long as they are supported on causal laws and causal explanations. Being guided by different intensional criteria, such orderings can produce any partial and complete extensional overlaps. This is as another expression of the descriptive elasticity, which has two items related in an analogy as the foundation from which larger ordering and hierarchies are built.

Causal structures actually are not a strong constraint; they allow great intensional diversity in the construction of any ordering of natural kinds. An almost indefinite multiplicity of classifications can be built upon causal structures, for instance, different orderings based on phenotypical criteria, aesthetic criteria, market demand and consumption can all be shown to have a causal base. Moreover, causal explanations in science are often contested when more than one cause or set of causes is presented as the true explanation, and such disagreements can last for a long period. Hence, intensional diversity and descriptive elasticity meet this constraint.

[37] R. Harré; J. L. Aronson, E. C. Way (1995), p. 31.
[38] *Ibid.*, p. 42-43.

Unlike Hesse, Harré rejects the use of classes using types instead. I have relied on classes because Hesse does and furthermore I do it as means for showing the interdefinability of class and analogy as well as the important role intensional criteria and descriptive elasticity play in the creation of different orderings, which can be presented as type-hierarchies or classes. Descriptive elasticity has important effects on analogy, and analogy is fundamental in the arguments and positions Harré and Hesse hold. Any isolated item or token and any number of them can be placed under different types or classes by enlarging or reducing the size of the negative or the positive analogy. This has important effects on the shape, size, overlaps and multiplication of types and classes. Harre's hierarchies of types are discussed two sections below.

1.4. The logical problem of analogy

Hesse draws a distinction between formal and material analogies. Unlike material analogies, formal analogies such as the mathematical proportion 9:3 :: 15:5, the anatomical and physical relations wing:bird :: fin:fish; pitch:sounds :: colour:light, do not allow any prediction. Material analogies include causal relations, which provide the grounds for any prediction. We know correlations also enable predictions, so Hesse explains that 'analogical argument presupposes a stronger causal relation than mere co-occurrence.'[39] In the case of the material analogy between the earth and the moon, the aim is to make a prediction on the existence of life in the moon based on existence of life on the earth. Such an inferential leap is larger than that of enumerative induction, and therefore it involves a greater risk. Like induction, any inference from analogy is also a case of ampliative inference, and as such it faces the similar challenges on justification and validity. Hesse calls this challenge 'the logical problem of analogy', which consists of justifying any likelihood assigned to any prediction or explanation based on an analogy:

[39] M. Hesse (1966), pp. 83-84.

> given two analogues *x* and *y* which resemble each other in a number of characters *B, … Bm,* and are dissimilar in that *x* also has *A… An* and *y* has not, and *y* has also C, ... *Cl* and *x* has not, we want to know whether another character *D* of *x* is likely to belong also to *y.*[40]

She argues that the degree of similarity in the analogy can justify the likelihood of the prediction. The prediction will be more or less likely depending on how large the similarity is between *x* and *y*. That is to say, the size of the positive analogy between *x* and *y*. She claims that, 'we generally should have more confidence in a hypothesis based on a model very similar to the explicandum than a model much less similar'[41] She calls this 'a common-sense assumption', which states that the larger the similarity, the greater the likelihood. I call this *the rule of maximal similarity*, which we can rewrite as follows: 'Between two or more analogies holding a hypothesis, the analogy with greater similarity should be chosen'. But how can such a rule be justified? She revises three possible philosophical justifications, namely the logical interpretation of probability, the falsification method, and Bayesianism.

She explains that this search for a justification is not concerned with methods for testing hypotheses but with 'a method of hypothesis-selection', whose aim is to help the selection of 'more reasonable' hypotheses. [42] In this case the hypothesis consists of an analogy built into a model. Because of the concern on validity and justification, the sense of 'more reasonable' can be better understood as 'more likely'. Testing methods such as the hypothetico-deductive method does not provide any means for a pre-selection of hypotheses as Hesse needs it. Such a pre-selection of hypotheses before any actual test is actually performed would, of course, be considered as part of the psychology of science and the context of discovery by positivistic philosophers.

[40] *Ibid.*, pp..72-73.

[41] *Ibid.*, p. 118.

[42] *Ibid.*, p. 118; see also Hesse (1953) pp. 198-204.

I argue that the interest on a pre-selection of hypotheses and models before any test is performed is methodologically sound and justified. This stage was considered to be part of the context of discovery, and therefore no rules could be given. However, the search for rules or methods for pre-selecting hypotheses is justified, particularly where many hypotheses have been produced, which is not an uncommon situation in science. From the moment hypotheses are produced, there is a constant concern on how to select the most likely one. We can call this *the problem of hypothesis choice*. From a cognitive perspective, and even from a logical one, no prejudice should be imposed on any rule, routine or procedure, which could speed up the whole production and justification of scientific knowledge by finding a method for the discrimination and selection of hypotheses before any test.

She presents the following three cases, against which those three possible justifications are considered:

i. The choice between a hypothesis with a model and hypothesis without a model.

ii. The choice between two models, where one is more similar to the event to be explained called explanandum.

iii. The choice between three or more models, where the first model is more similar than the second one, and less similar than the third one to the event to be explained, which is called explanandum.[43]

A model is 'any system, whether buildable, picturable, imaginable, or none of these, which has the characteristic of making a theory predictive'.[44] As an example of a model she refers to the use of the billiard balls in random motion as an analogy for gas molecules also moving in a random fashion as postulated by the new kinetic theory of gases in the nineteenth century. Her interest on models applies only to theoretical entities and their mechanism of operation. Against

[43] M. Hesse (1966), pp. 19, 113.

[44] *Ibid*., p. 19.

Pierre Duhem and others who hold the view that models are accessory and an aid to science, she claims that 'models are logically essential for theories.'[45]

Not surprisingly, Hesse finds that neither the logical interpretation nor falsification can help make a choice among the three cases listed. The main problem with the logical interpretation of probability from Carnap or Keynes[46] is that 'any group of three characters is *a priori* as likely to occur as any other [...] so the evidence of the model, that the group *ABD* occurs, can have no effect on the chance of *BD* containing *D* rather than *X*'[47], where each capital letter stands for a known property, and *X* represents an unknown property. The model is *ABD*, and the explanandum is *BCD*. This has a negative effect on case (i) because no choice can be made between models with known or unknown properties. This happens because in the logical interpretation equal probability distributions were common due to the principle of indifference, which establishes that in the absence of any reason for giving more weight to one event or another, all events should be considered equally probable.[48] However, Carnap realised that some propositions can have *a priori* unequal probabilities. For instance, a disjunctive proposition is *a priori* more likely than a conjunctive one; he called this property the 'logical width'. Specifically on analogy, he actually claims that 'the known similarity between *b* and *c* is the greater the stronger the property M1 [positive analogy], hence the smaller its width'.[49] That is to say, the larger conjunction of properties, the smaller the probability.

Hesse does not discuss 'logical width', which actually has negative effects for cases (ii) and (iii) because in both of them if a model with greater similarity is pre-selected, its likelihood would always be smaller. In other words, the greater the similarity, the smaller the likelihood. Instead, she looked for means to make the likelihood of a model 'stronger-than-chance'; that is to say, greater than any other equally probable competing model or explanation. One way of

[45] *Ibid.*, p. 19, see P. Duhem (1906), pp. 69-75; see also C. Hempel (1965), pp. 433-447, and R. Carnap (1939) p. 67-68.
[46] R. Carnap (1950); J. M. Keynes (1921).
[47] M. Hesse (1966), p. 119-120.
[48] J. M. Keynes (1921), p. 42.
[49] R. Carnap (1945), p. 84, 87-88.

doing this within the logical interpretation consists of placing the properties *B* and *D* as an 'instance of a generalisation', or a law establishing that 'All *B*'s are *D*', but such a law is introduced as a postulate, and it is therefore arbitrary. Moreover, any negative analogy must be considered as 'irrelevant to the confirmation' in order to prevent any predictive or explanatory failure. But this is 'quite implausible, since it is reasonable to assume that the occurrence of differences [...] reduces the confirmation.'[50] So she concludes that the logical interpretation of probability does not help make a choice among (i), (ii) and (iii).

With falsification the inference from analogy also faces serious problems. For case (i), Hesse constructs a scenario with the model 'All *AB* are *D*', and two competing hypotheses, namely H1: 'All *B*'s are *D*' and H2: 'All *C*'s are *X*' with an explanandum '*BC*'. As it is required from the method of falsification, the model and the hypotheses have a universal form. Both hypotheses are equally similar to the event to be explained because they share one property in common with it. If the model 'All *AB*'s are *D*' is attached to each hypotheses, H1 should be chosen because 'the set of potential falsifiers of H1 includes and is larger than the set of potential falsifiers of H2'[51] This choice is made because Hesse constructs the model *ad hoc* by making it almost the same as H1 except from the addition of property *A*. In this way, H1 becomes a subset of M, and therefore, it includes the potential falsifiers of H1 plus its own falsifiers. This does not happen with H2.

Hesse sees the falsificationist account of analogy as plausible and satisfactory, however the solution is not an interesting one. By making the model more similar to the explanandum, analogical inference is at risk of becoming unchallenging and uninteresting. Case (i) is actually turned into case (ii). This is an important problem for Hesse's rule of maximal similarity. Unlike her, I argue that unexpected and surprising analogies with less similarity can actually produce greater advances in science. I defend this claim in the next section.

[50] M. Hesse (1966) p. 130; and (1964), p. 323-323.

[51] *Ibid.*, p. 132.

For cases (ii) and (iii) Hesse follows a different strategy. She presents the explanandum *B1*, *B2*, *C*; and two models, namely model M1: *A1*, *B1*, *D1*, and model M2: *A2*, *B1*, *B2*, *D2*. Despite being more similar to the explanandum, M2 cannot be chosen because the size of potential falsifiers has to be defined and compared to model M1. Because of the different properties each model has, none of them can be a subset of the other, so the size of the set with the potential falsifiers for each model cannot be compared, and therefore a choice cannot be made, 'we are left with hypotheses which are incomparable on both falsifiability and corroboration criteria.'[52] Individually, each model is falsifiable but together they are incomparable. The problem extends to any choice with two or more analogies and their models.

Despite these failures with the logical interpretation of probability and the method of falsification, Hesse still holds that 'material analogues', i.e. models based on material analogies depicting causal mechanisms are 'strongly predictive',[53] and she claims that the choice is clear in cases (ii) and (iii), where models more similar to the explanandum should be chosen. Maximal similarity should, in principle, lead to predictive success. Inference from analogy is therefore left without a full justification.

Later, Hesse turned her attention to Bayesianism, which seemed to be able to accommodate a pre-selection of hypotheses based on analogies and models. In subjective Bayesianism there is no prejudice on the sources and type of information used for assigning prior probabilities; this liberality looks appealing because it can provide a justification for a high likelihood given to the analogy and model more similar to the explanandum. Hesse believes that with certain modifications, 'a unified theory of Bayesian inference' can justify 'three types of inductive inference', namely 'enumerative induction', 'theoretical inference from analogies and models', and inference based by 'simplicity'.[54] The modifications she suggests emerge from the criticism she makes on the Bayes theorem as a transformation rule used for updating beliefs,

[52] M. Hesse (1966), p. 140.
[53] *Ibid*, p. 143-144.
[54] M. Hesse (1975), p. 91.

which she claims is 'vacuous as a method of induction', unless 'normative constraints are put upon belief distribution the updating of beliefs other than coherence'.[55] It is of course controversial to continue using the term 'Bayesianism' after constraints are placed on prior probabilities, which has given rise to objective versions of it. Hesse's own version should perhaps be considered as objective. Although, she makes extensive criticisms on different aspects of Bayesianism, and even considers the possibility of devising a different rule of transformation.

In Bayesianism any hypothesis based on an analogy with or without a model can be pre-selected giving to it a certain degree of belief, that is, a prior probability can be given to it before any prediction or test. She argues that the introduction of a causal model, i.e. a material analogue, prevents the posterior probability from being 'arbitrary' or an 'accident'. If coherence is the only norm, any statistical generalisation like 'Most crows are black' is supported by a long run of crows, if the prior probability of the generalisation is high enough.[56] But from case (i) above we also want a hypothesis with a model to have a higher prior probability than a hypothesis without a model, and from cases (ii) and (iii) we want more similar models to have higher prior probabilities than less similar ones. Because Hesse is arguing for a 'sufficient (though not a necessary)' method,[57] she needs to show at least one example where such values are obtained.

Unfortunately, Hesse did not produce such an example showing how Bayes's theorem can be applied to an inference from analogy, she only makes a generic reference to cases like 'prediction of the properties of a new chemical compound from knowledge of its elements and of other compounds; prediction of the properties of a full-sized object from experiments on a replica in, say, a wind tunnel; and prediction of the properties of micro-objects and events from

[55] *Ibid.*, p. 91.
[56] *Ibid*, p. 66.
[57] *Ibid.*, p. 78.

those of macroscopic models'[58]. I am therefore producing an example intended only as illustration by relying on an example from one of her earlier papers.[59] The example used there shows how an inference can be made from an analogy between objects belonging to different classes, which I apply to planets, bricks, marbles and anvils. These objects are solid, massive and opaque but vary in size, shape, chemical composition and so on. Galileo produced the law of falling bodies, Kepler produced the laws of the planetary motion, and Newton the laws of universal gravitation from which those of Galileo and Kepler can actually be derived.

Hesse claims that a new prediction can be made without a law connecting any two different classes of objects such as bricks and marbles, by identifying the relevant similarities between two classes. If there is evidence from many trials on the falling speed of marbles dropped from the top of a tower, by analogy a new prediction can be made on a first brick on trial falling with the same speed. The same prediction can be made dropping an anvil. By the same procedure, if the force of gravitation is known from several planets attracting each other, by analogy a new valid prediction can be made on the earth attracting any brick, marbles, cherry or anvil falling from the top of a tower. Here it is a simple form of Bayes' theorem:

$$P(h/e)= \frac{P(e/h)\ P(h)}{P(e)}$$

With such a theorem it is possible to obtain the posterior probability of a causal hypothesis made with the help of analogy from the prior probability of a prediction, and the prior probability of that causal hypothesis with and without assuming that the prediction was

[58] *Ibid.*, p. 96.

[59] Hesse (1970), pp. 166-170

successful. If the falling speed of bricks is well established from many trials, by analogy a first new prediction can be made on an individual item belonging to dissimilar classes such as anvils, marbles or cherries with the following values:

> P(h): is the prior probability of the hypothesis. This is the probability of the causal mechanism described in the positive analogy being real before any prediction. That is to say, it is the probability of a common cause affecting in the same way the falling speed of bricks and anvils. Because of the risk involved in the negative and neutral analogies a cautious value of 0.75 can be assigned to this prior.
>
> P(e): is the prior probability of the evidence or observation. This is the probability of the very first anvil actually falling at the same speed to that of the bricks. Because no analogy is considered here, a moderate value of 0.80 can be given to this prior.
>
> P(e/h): is the likelihood, which is the probability of the evidence or observation assuming that the hypothesis is true. This is the probability described in the previous paragraph, whose value is modified by assuming that the common cause or property presumed in the positive analogy is real. The value of the likelihoods should be high, namely 0.95

With these values, the posterior probability P(h/e) can be obtained:

$$P(h/e) = \frac{P(e/h)\ P(h)}{P(e)} = \frac{0.95 \text{ x } 0.75}{0.80} = \frac{0.7125}{0.80} = .89$$

As it can be appreciated, the probability of the hypothesis raises from 0.75 to 0.89, which is what Hesse was expecting from using Bayes' theorem. Following the rule of maximal similarity, she argues 'that the confidence we have in the prediction […] is due to the relation of analogy […] which is constituted by the repetition of predicates G[solid] and Q[massive]', so we regard the analogy as 'confirmed because the bodies described […] are sufficiently similar in some respects […] to justify the inference'[60]

Having made an illustration, we can now discuss the core of her logical argument, which consists of a postulate and a property, namely the clustering postulate and the exchangeability of single-test probabilities. From the last section, I am adding the interdefinability of class and analogy as a property, which complements both exchangeability and the postulate:

Clustering postulate:

'If groups of properties are present together in a number of instances, in the absence of other evidence it is more probable than not that they will be present together or absent together in further instances.'[61]

Exchangeability:

'The initial probability of a given outcome of a single test of a P-object for Q-ness or for not-Q-ness will respectively be the same for each individual P-object, that is, p(e1) = p(e2) = … = p(er) = p(~e1), and so on.'[62]

Interdefinability:

A class is a collection of analogies and an analogy is a class with two or more members, where the size of positive analogies is larger, or more significant, than the size of negative analogies

[60] *Ibid.*, p. 167.

[61] M. Hesse (1975), p. 94.

[62] *Ibid.*, p. 93.

The clustering postulate is analogous to the principle of the uniformity of nature used to justify induction, it is an adaption of this principle for the case of analogy. Hesse actually explains that the postulate is a different name for the 'assumption of homogeneity', where 'in these matters of induction it must assumed that God is not a coin-tosser'.[63] Therefore, the postulate is crucial because it justifies the projection of the same probability value for all future predictions of the same class of events.

Here and in other works, Hesse presents her argument on analogy as an inductive one, which I believe can be misleading. The cases she discusses actually correspond to single case inferences, which need to be distinguished from inferences based on many cases. For instance, in her discussion on the logical interpretation of probability and on falsification, single case probabilities are not considered. It should be noted that due to the challenges it poses, single case inference is treated separately from the typical enumerative induction supported on a large amount of instances. The literature on single case probabilities and recent works on experimental induction based on a few cases, offer a more suitable framework for the analysis and discussion of the inference from analogy.[64] The interesting challenge consists of keeping the inference *minimally inductive* by relying on one or as few cases as possible making scientific predictions and explanations. The conception of analogy as a single case inference is crucial for Hesse's general argument and, more particularly, for the exchangeability of initial probabilities.

The exchangeability of initial probabilities is a crucial new property, which actually does the job the logical interpretation and falsification could not do. She drew this property from the work from Bruno De Finetti, where he argues for the independent subjective probabilities, which can be separately assigned to events of the same type or with analogous characteristics.[65] From independence and separability Hesse derived exchangeability, which crucially relies on the

[63] *Ibid.*, p. 94.

[64] D. Gilles (2000) discusses single case probabilities under the different interpretations of probability; P. Galison (1987) presents historical cases of inferences in experimental physics based on single cases and on a few cases only; and N. Cartwright (1989) presents a compelling argument on single case causal explanation.

[65] De Finetti (1937), p. 120.

equal prior probability, which can be assigned to any member in the analogical class. Such conditions justify giving a high degree of confirmation to the very first member in that class, if the prediction is successful. That is, inductively scientists do not have to wait for a long series of successful tests or predictions to be justified after, say, the twentieth prediction. A single case can do the job. In other words, equality and exchangeability in the analogical class allow bringing back to the front the confirmation value of the twentieth case. By doing this, the standard inductivist rationale is inverted.

The equality of prior probabilities by itself does not allow or imply bringing back such a confirmation value. On the contrary, De Finetti is arguing for taking each event as separate from any other similar one or of the same type. Recall that in the standard account, an analogy is not a *class* of events of the same type but a *relation* of two dissimilar objects belonging to different classes. Exchangeability, therefore, remains unwarranted and arbitrary, unless an argument is provided where all relevant analogies are brought together into one class. The definition of an analogy as a class does this, thanks the property of interdefinability.

Homogeneity, exchangeability and interdefinability complement each other; together they make a cohesive argument for the inference from analogy. The high prior probability given to such inference is justified because of the rule of maximal similarity and the related causal knowledge. Then, a successful prediction produces a higher posterior probability as it has been shown in the example above on the falling speed of dissimilar objects, namely planets, anvils, marbles and cherries. The idea of exchangeable events within a class, particularly of analogous events, is very important because it allows placing an analogy as the first event forming a class. This step is fundamental in Hesse's new and stronger argument on analogy, which places inference from analogy not as supplementary or as a mere heuristic device, but as an inference which can itself be the foundation of any scientific law or theory.

This argument completes the different views Hesse held on analogy from the early work she published on this topic in 1963 to the last development and position she held adopting some aspects from Bayesianism. She kept a long battle with Positivism, and she was also dissatisfied with Falsificationism. In 1966, she was still claiming that 'analogical argument is necessary only in situations where it has not been possible to observe or to produce experimentally a large number of instances in which sets of characters are differently associated […] analogical argument is "weaker" than inductive, but on the other hand it has the advantage of being applicable where straightforward generalization is not.'[66] Nine years later, she finally reached a distinct new view:

> 'The universality which is usually held to be an essential constituent of theories is seen in this view as rather *a convenient method of summarizing case-by-case analogy judgments* […] predictions of next instances of universal generalizations are *elementary special cases* of this kind in which the notion of "analogy" between a model whose behaviour is known in domain i_1 and the predictions in domain i_2, reduces to the notions of exchangeability and clustering of instances in the different initial conditions i_1, i_2.'[67] (italics added)

This is the strongest argument and conclusion within the Mechanical view on analogy I am aware of, it is cogent and distinctive by placing single case inference from analogy at the foundations of laws and theories, which now take a derivative and secondary status. In contrast, the Syntactic and Semantic views place laws and generalisations at the foundations of any theory. Hesse's argument and justification of analogical inference provides the Mechanical view with the new grounds needed since the time Norman Campbell advanced his views on analogy, models and scientific theories.

[66] M. Hesse (1966), p. 76.

[67] M. Hesse (1975), p. 99, 97.

1.5. Metaphor

The knowledge and postulation of mechanisms are distinctive of the Mechanical view. All mechanical models theoretical or observable describe a causal mechanism responsible for certain effects. There are two fundamental features of mechanisms as they are introduced by the Mechanical view. The first one requires continuous physical contact between all the entities and effects involved. That is to say, action at distance is avoided. Newtonian mechanics is a canonical example of this, where theoretical entities and mechanisms such as the luminiferous aether and the corpuscular theory of light were postulated. The second one is a widely shared realist belief in causal powers.

The Mechanical view is itself a metaphor, which has expanded into scientific and philosophical domains where causes and mechanisms are used metaphorically. The work of Donald Davidson in the philosophy of mind and action, and that of Daniel Little in explanation in the social sciences are examples.[68] In the social sciences, mechanism design theory, an important branch in game theory, constitutes another outstanding example, where the term 'mechanism' has been introduced with a clear metaphorical sense. Mechanism designers devise specific rules, incentives and penalties, which together bring about certain behaviour. However, it is not entirely clear the kind of physical interaction existing between the presumed causes, the mind and the observed behaviour. In spite of this, the work of Davidson and mechanism design theory is a clear example of the success of metaphors in philosophy of mind and science.

Other terms such as inflation, deflation, depression and boom used in economics also have a metaphorical meaning. Besides mechanisms, functionalism is an example of another successful metaphor widely used in anthropology and sociology. Evolutionary game theory represents another well-established twofold metaphor in the social sciences, where among others terms like 'dove' and 'hawk' are widely used describing the profiles of different individuals portrayed as players. These examples show that metaphors are not a few only having

[68] D. Davidson (2001); D. Little (1991).

an accessory character; there are many of them playing a fundamental role also in the social sciences.

In contrast, in the natural sciences the use of mechanisms is often considered to be literal and real. It is believed that nature is composed of mechanisms, that is, of causes responsible for all things we see happening. This seems obvious and in principle difficult to challenge; the many successes of science predicting and intervening in nature seem to prove the reality of causes and mechanisms as well as the literality of the related descriptions. However, even here metaphors can be found in some of the most fundamental concepts.

The use and the role of metaphor in science has been a very important contribution made by Rom Harré to the Mechanical view. He explains that models, metaphors and analogies are needed when 'we have reached the limits of discernible mechanisms'.[69] While some analogies and models can be built using literal language, metaphorical terms are often required when no adequate concept or description is available. Thus metaphorical terms and analogies meet in a model at the borders of scientific discovery, conceptual change and scientific revolutions. James Maxwell's vortex-idle wheel model of magnetic force, and the billiard balls model of gas molecules are examples of such models.

In Positivism, the meaning of any theoretical terms could only be decided upon by the observable effects; no speculation on the specific nature and inner workings of unobservable mechanism and entities was otherwise acceptable. Harré demonstrates how the observational language accepted by Positivism actually contains metaphorical terms, whose meaning ultimately relies on the terms and procedures taken from another scientific branch. For example the term 'current' in electro-dynamics pictured as a flow of electrons cannot be fully defined with reference to the different readings observed on an ammeter from a simple circuit, 'because as it is used in electro-dynamics it carries with it an accretion of meaning derived from its use in hydro-dynamics, where it could be effectively taught before a flowing or running stream. Hence

[69] R. Harré (1960), p. 105.

the term 'current' is metaphorical carrying with it into the description of the phenomena encountered in electrical circuits some of the force it had in its original p.c.p.'[70]

Besides the term 'current' other fundamental terms in physics are also metaphorical such as 'force', 'field', 'repulsion', 'conductor', 'wave', 'packing fraction' and 'strangeness'. Generally, the metaphorical meaning of scientific terms goes unnoticed because 'the tradition in philosophy of language and science is that language is intrinsically literal in nature. Literal meaning is considered to be the standard and normal use of words, and it is the meaning that words possess independently of when and how they are used.'[71] This is an important observation, without it the widely shared belief that science provides literal descriptions of nature and society would persist and remain unchallenged. In science metaphorical terms 'are *picture-carrying expressions*. When we describe an electrical discharge ('discharge' is an M-term too) in a gas as the passage from a current, we are inviting ourselves to picture something flowing of which incandescence, for instance, is an effect.'[72] Therefore, figurative language is not anymore exclusive to art but it also is a systematic component of science.

The *comparison view* of metaphor explains figurative meaning by relating it to a primary literal meaning. For instance, the term 'electrical current' is metaphorical because it can be related to the literal description of clusters of molecules of a fluid like water moving along a canal. Harré criticises Norman Campbell and Ronald Giere for implicitly holding this view, when they use analogy and similarity in their philosophical accounts of models and scientific theories. He argues instead for the *interactive view* put forward by Max Black with an application to language in general, that is to say, without a special focus on science. Unlike the comparative view, this view does not assume that literal meaning remains as the fixed foundation upon which metaphor is explained. The introduction of metaphors rather shakes those foundations

[70] R. Harré (1960), p.112; he explains that 'a term has been defined with reference to a *paradigm case* (p.c.) if it *could have* been introduced by ostension. The paradigm case will be that to which we could have pointed in introducing the term, and the whole method of introduction I shall call a *paradigm-case procedure* (p.c.p.)', p. 111.

[71] R. Harré, J. L. Aronson, E. C. Way (1995), p. 96.

[72] R. Harré (1960), p. 112.

by creating new meanings, which affect any related literal meaning; Black points out that 'it would be more illuminating… to say that the metaphor creates the similarity than to say that it formulates some similarity antecedently existing.'[73]

Mary Hesse also criticised the comparison view and adopted the interactive view of metaphor, applying it to science. She explains that the interactive view accounts for the mutual affectation of both literal and the metaphorical language producing a 'shift in meaning', and a 'post-metaphoric' sense. For instance, with the metaphor 'Man is a wolf', 'men are seen to be more like wolves after the wolf-metaphor is used, and wolves seem to be more human.' And with any mechanical metaphor 'nature becomes more like a machine in the mechanical philosophy, and actual concrete machines themselves are seen as if stripped down to their essential qualities of mass and motion.'[74]

Harré agrees with this mutual affectation and believes similarity is created by choosing to relate two or more objects, rather than being there preceding the metaphor. However, he holds that the comparative view remains 'vague', at least in its application to scientific language, because 'it is not clear how the interaction or filtering is to occur, nor how similarity can be created where none was seen to exist before.'[75] He calls this 'the problem of principled filtering of positive from negative analogies.' And he also rejects Hesse's thesis on the logical priority of metaphor, which states that 'metaphor properly understood has a logical priority over the literal, and hence that natural language is fundamentally metaphorical, with the "literal" occurring as a kind of limiting case'[76] In other words, she inverts the order by placing metaphor as a more fundamental form of speech.

Besides these two problems, Harré also identifies another problem with the use of bare similarity as the kind of relationship models hold with the world, and as the criterion to be used

[73] M. Black (1962), p. 37; see also (1993), p. 35.
[74] M. Hesse (1965), p. 254.
[75] R. Harré, J. L. Aronson, E. C. Way (1995), pp. 105, 96-97.
[76] M. Hesse (1993), p. 56.

for defining metaphor. Ronald Giere places bare similarity as the criterion needed for evaluating the empirical significance of models by claiming that 'the notion of similarity between models and the real system provides a much needed resource for understanding approximation in science. For one thing, it eliminates the need for a bastard semantical relationship—approximate true.'[77] Giere says that such a basic notion could be refined by adding 'degrees' and 'respects' of similarity, however he does elaborate this claim further showing how this can actually be done. Harré believes this notion of similarity is too basic for models because it does not tell us if it is a symmetric or a transitive, and also because it 'is not rich enough to give us a ranking of models in terms of which are better approximations [...] The notion of similarity is doing too much of the work in Giere's theory; and similarity is too complex and difficult a notion to leave as unanalysed primitive.'[78]

In sum, Harré identifies three outstanding related problems on metaphor and analogy; and I am adding the fourth on the list, which is the logical problem of analogy discussed in the previous section:

1) *Priority*: the problem of establishing the logical priority of metaphorical or literal language.
2) *Salience*: the problem of filtering positive from negative analogies
3) *Triviality*: the problem of distinguishing trivial from non-trivial analogies
4) *Inference*: the problem of justifying the likelihood of a prediction or an explanation based on an analogy.

Harré argues that an ontology of types organised in hierarchies can provide a solution to the first three problems, and I am also evaluating such hierarchies against the inference from analogy.

[77] Giere (1988), p. 106

[78] R. Harré, J. L. Aronson, E. C. Way (1995), pp. 94-95.

1.6. Type-hierarchies

A type is a representation of a natural kind and any individual member of a natural kind is a token in virtue of the representation created by the type. Eileen Way, a co-author with Rom Harré, explains that 'types are on the side of the mind, kinds are on the side of the world.'[79] As a representation, a type 'is a set of individuals each of which has certain properties which are numerically identical with those in other sets of higher type'. Because types have a nominal status, the relationship they hold with their tokens cannot be that of 'qualitative identity', which only holds 'between the relevant concrete properties of each particular', [80] so numerical identity does the job of establishing the relationship needed between tokens and the types. An individual whale is a token whose properties are numerically related to those contained in the type mammal, which is a nominal representation. Types are ordered according to their level of generality, so for instance mammals are more general or higher up in the hierarchy than placental mammals, and the family *Felidae* is below these two.

Types do not mirror the world. This is because more than one type-hierarchy can be built upon any natural kind, but this does not mean types are a mere construction built solely upon convention. Linguistic conventions are accepted as part of types and hierarchies but they cannot fully account for their formation because 'real structures of natural kinds'[81] set constraints on them. Natural kinds therefore have a metaphysical status. Harré argues for a realism of natural kinds and their causal structures upon which a representational pluralism is built using types hierarchically organised.

Type-hierarchies are graphic representations; they are not presented using sentential descriptions. This is consistent with Harré's defence of graphic models, which he calls 'iconic models', in opposition to the 'statement view' of theories and scientific knowledge held by the

[79] *Ibid*, p. 27

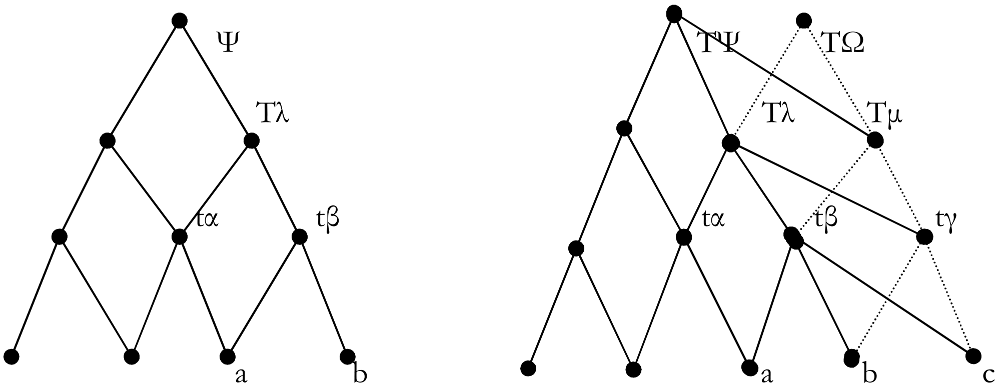
[80] *Ibid*, pp. 15-16. More details on Harre's argument on nominalism can be found in Harré and R. H. Madden (1975), pp. 16ff.

[81] R. Harré, J. L. Aronson, E. C. Way (1995), p. 17.

Syntactic view. An interest in the development of artificial intelligence programmes also motivates a graphic representation and the actual choice for types, which help simplify inferences under computational constraints.[82] Their shape and operation of type-hierarchies is analogous to Truth Trees in logic.

The figure below is an adaptation of the graphic model of type-hierarchies used by Eileen Way.[83] Because of their increasing level of generality of types the model has a pyramidal shape. Literal and metaphorical descriptions are called *semantic masks*.

Figure 1. 3. General model of type-hierarchies



Literal mask

Metaphorical mask

Where TΨ and TΩ are called supertypes; Tλ and Tμ are called types; tα, tβ and tγ are subtypes; and *a*, *b* and *c* are tokens. Supertype TΩ, type Tμ, subtype tγ and token *c* only emerge with the metaphorical mask. So, for instance, *a* and *b* are tokens of subtype tα in the literal mask, *c* becomes a token of tα and *b* a token tγ only under the metaphorical mask, while tβ becomes a subtype of Tμ, and TΩ becomes a supertype for all also under the metaphorical mask.

[82] S. E. Fahlman (1979) uses type-hierarchies in artificial intelligence programming; he explains that it is not enough to retrieve 'isolated facts with no deduction' from a knowledge base, while the other hand 'we cannot expect to deduce quickly *everything* that could be deduced', so the inheritance of properties through type-hierarchies offers 'an intermediate level of deductive ability', p. 15.

[83] R. Harré, J. L. Aronson, E. C. Way (1995), p. 102.

***1. Logical priority*.** Neither metaphoric nor literal language can be established as logically prior because the distinction between the two is highly dependent on the wider semantic context, and also because such a distinction is subject to fundamental meaning shifts. Eileen Way describes these wider semantic contexts as *masks* because they can hide or expose certain properties from any natural kind, as well as certain vertical and horizontal relationships in the hierarchy. Each semantic mask contains different literal terms and different metaphoric ones, which can reveal or hide new relationships among kinds along the transverse, horizontal and vertical axes in the hierarchy. The introduction of a new metaphorical term leads to a meaning shift in the literal terms, causing the reshaping of the hierarchy. Hence, literal does not mean a definitive true description; it only 'denotes the presently accepted classification of natural kinds and species'.[84]

The successive theories of gases between the seventeenth and nineteenth centuries are an example of this shift in meaning and reshaping of hierarchies: From the Newtonian theory postulating particles surrounded by an elastic fluid called 'caloric' to the kinetic theory, which rejects caloric and repulsion relying only on the random motion of particles colliding with each other. Each theory functions as a semantic mask. The caloric theory of gases was developed by Pierre-Simon Laplace in analogy with Newtonian physics, where forces of attraction and repulsion carry the explanatory power. Because the small particles from gases would be too far from each other to be able to exert any attraction, the fluid caloric was introduced in order to hold them together. It was also assumed that the gas molecules attracted the caloric. Latent caloric formed a material core from which free caloric was released filling in the space between the molecules.[85]

The caloric theory actually retained part of its explanatory power on some observable effects on temperature and volume. Nonetheless, it became an outmoded metaphor after the kinetic theory was accepted as the new literal description. As a semantic mask, the caloric theory

[84] Ibid., p. 102.
[85] See R. Fox (1971), pp. 68ff; and H. Chang (2004), pp. 69-75.

exposed gas molecules as a natural kind, while it hid the possibility of accounting for their movement without any fluid. A fundamental meaning shift occurred with the emergence of the kinetic theory, which explained the same events and regularities observed between volume and temperature by replacing 'attraction' and 'repulsion' with 'collision' and 'random motion'. No fluid like caloric or any other was needed. The effects of this meaning shift were not only local because at the time caloric was also used in other theories, so structural changes were forced into all related natural kinds and the type hierarchies built upon them. A vast number of observable events had to be explained and re-described using the new theory.

The same conclusions can be extended to Hesse's example discussed in section 1.4. A shift in meaning and the reshaping of types also occurred when the literal terms 'free falling' and 'constant speed' from the Galilean semantic mask, were replaced with 'attraction' and 'gravitational force' under the Newtonian mask.

The argument on meaning shifts and the reshaping of type hierarchies relies on a normative judgement, which accepts the postulation of unobservable entities and mechanisms for scientific explanation. Strict empiricist normative standards reject such postulation of theoretical entities and mechanisms, while realist standards accept it. Against Ernest Mach and the Copenhagen interpretation in physics, Harré argues for the following principle, which he calls P1: 'If you don't know why certain things happen then invent a mechanism (in accordance with the view you take of how the world works)–but it is better still if you find out how nature really works.'[86] This principle actually provides Campbell and Hesse's work of theoretical models with a systematic foundation, and it also anticipates the defence of the inference to the best explanation as normative principle. For instance, Gilbert Harman writes, 'the inference to the best explanation corresponds approximately to what others have called "abduction", "the method of hypothesis", "hypothetic inference" […] and "theoretical inference" […] In making this inference one infers, from the fact that a certain hypothesis would explain the evidence, to

[86] R. Harré (1960), p. 101.

the truth of that hypothesis.'[87] The next step would be to reduce inference to the best explanation to a case of inference from analogy in order to obtain a fully Campbellian view of theories. Unlike Campbell though, Eileen Way makes analogy to depend upon type-hierarchies, which in her view provides the ultimate systematic and methodological grounds for scientific models and scientific inference.

2. & 3.Salience and Triviality. On a one-to-one basis, an analogy relies on bare similarity and *ad hoc* criteria, which can lead to trivial relations of both similarity and dissimilarity. Bare similarity as a criterion for building analogies is insufficient because many trivial or scientifically uninteresting relations of similarity can easily be established. For instance, similarity relations between layers of bricks on a wall and layers of cells on a human tissue, or between racing cars running on an elliptical race and planets rotating around the sun. As Eileen Way points out, 'clearly some properties are more important or *salient* than others for the model; how are these determined? Why don't we consider that electrons may have an analogy with the moons craters or an atmosphere; or that the nucleus may have a gaseous and turbulent structure?'[88] Lack of salience leads to triviality, criteria on salience could be created *ad hoc* on one-by-one basis but this is not good enough when a few general criteria can be produced reducing normative requirements and increasing scope.

Way explains that triviality is a fundamental problem for the Semantic view, while *ad hoc* similarity is a fundamental problem for Hesse and Giere's view on models. This happens because structural isomorphism between models 'is not a powerful enough relation', since 'there will be an endless number of systems that exhibit the requisite mapping'. [89] While bare similarity remains on the 'bottom most-layer' of any ontology by relating only one token to another. In other words, in the Mechanical view as developed by Hesse and Giere, similarity does too little

[87] G. Harman (1965), pp. 88-89; see also P. Lipton (2004)

[88] R. Harré, J. L. Aronson, E. C. Way (1995), p. 91.

[89] *Ibid.*, p. 92.

by relating two or a few tokens only, while in the Semantic view similarity relations among those tokens can be trivial.

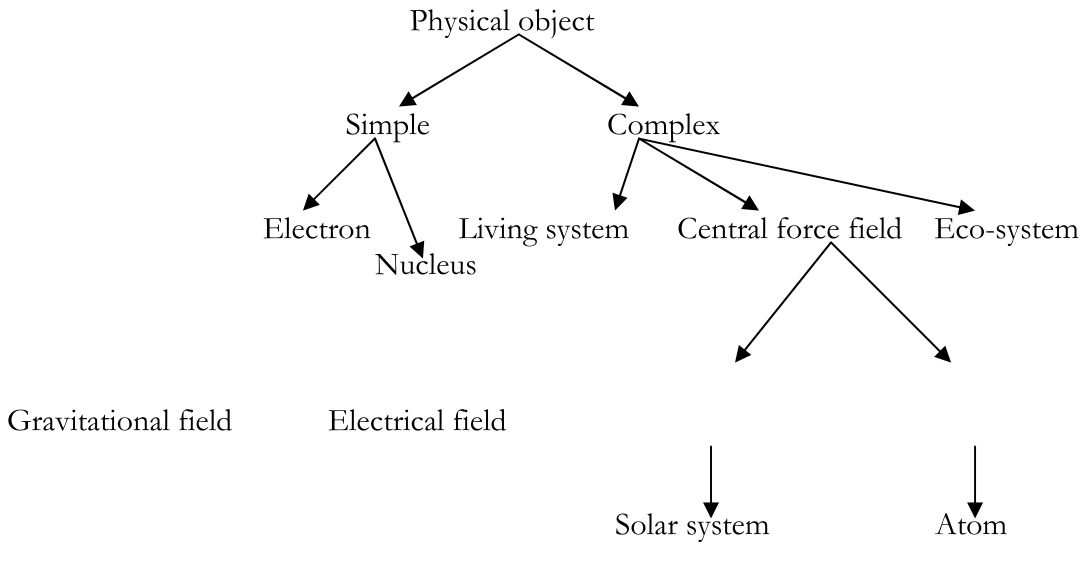
Eileen Way argues that a hierarchy of types can provide a solution to these two problems. She claims that types are not static but dynamic representations of different relationships among objects all belonging to natural kinds, so 'the type-hierarchy *generates* salience and similarity through inheritance and the empirically determined ordering of these kinds.'[90] The knowledge of causal and non-causal properties of types, and the inheritance of these properties from one type to another constitute the solution to the problems of salience and triviality.

Supertypes at the top of hierarchy denoted by TΨ and TΩ, and types Tλ and Tμ as well as subtypes tα, tβ, tγ at the lower levels can provide such a solution through relations of inheritance reaching any token. Any successful analogy following the logical standards established by Hesse is actually laying a bridge between two types through one or more common properties, it has not just been guided by a type or a supertype as suggested by Way but by *ad hoc* criteria. For instance, the case in section 1.4., where gravitation is a property common to dissimilar tokens such as a marble, an anvil, and a planet is an example of how properties in different types and subtypes are inherited through a supertype. Such a Newtonian supertype, TΩ in Figure 1.3., consisted of all massive, solid and opaque objects regardless of their location, size, shape, chemical composition and other dissimilarities constituting a negative analogy. TΨ can be described as a Galilean supertype consisting of all massive and solid terrestrial objects regardless of their location, size, shape, chemical composition and many other elements comprising the negative analogy. The semantic shift from the Galiliean mask to the Newtonian one implied the reshaping of types and hierarchies of all terrestrial and celestial objects

[90] *Ibid.*, p. 112.

Besides providing similarity with systematic grounds, inheritance relations from supertypes and types also prevent triviality by guiding the exclusion of some common properties from any positive analogy, just because they can be irrelevant from the natural kinds perspective. The isomorphism of the elliptical paths followed by a planet, a racing car and a termite flying around a light bulb can be therefore excluded as grounds for a natural kind. While the isomorphism found in the paths of planets and electrons can be retained as part of the positive analogy because it is an inherited property from a supertype, namely the central force field. Therefore, Way explains that 'there is no need to rule out the negative analogies *ad hoc*, because the common supertypes will generate only positive analogies with the systems.' [91] Here is the type-hierarchy for the atom and the solar system:

Figure 1.4. A partial type-hierarchy for the atom and the solar system:[92]



In this way, type-hierarchies organise knowledge of properties and relationships, which otherwise would remain separate and unrelated, and it also economises on such knowledge because properties are gathered into one supertype from which they are inherited by types,

[91] *Ibid.*, p. 108.

[92] *Ibid.*, p. 109.

subtypes and tokens. A scientific ontology of types organised in hierarchies becomes a powerful methodological tool, which could also solve the logical problem of analogy as stated by Hesse and discussed in section 1.4. Inheritance of properties is important not only ontologically but it also has an impact in logic. For instance, the uncertainty involved in the inferential leap taken from one token to another may disappear because any new property can be inferred with the support provided by the supertype.

4. Inference. How can a type-hierarchy help solve the logical problem of analogy? As we saw earlier, this problem consists of justifying any likelihood given to a prediction or explanation based on an analogy, which is a single case inference. Way argues that any new property in a token, such as 'having wings' for an animal, can be deduced from a supertype. This is because 'the hierarchy makes it easy to deduce these facts by a form of modus ponens: Bob is a canary, canaries are birds, and therefore, Bob *is a* bird. This is called an *inheritance hierarchy*, or sometimes an *isa*-hierarchy […] Thus, in order to determine if Bob has a certain property, just trace up the *isa*-hierarchy and assume that any of the meta-properties asserted about higher nodes can be considered assertions about lower nodes as well.'[93] Indeed, if the type hierarchy is large enough and if most or all of its properties and relationships have been empirically determined, then any inference on properties can be deductive.

Because the inference on properties is deductive, true conclusion are warranted, and therefore any prediction on new properties is certain *a priori*, which solves the logical problem of analogy. For instance, a few phenotypical features could be identified in an animal or a plant, from which a positive analogy can be made. Once the right positive analogy is established, the inference on properties follows in the fashion of a Hempelian covering-law model of prediction and explanation but based on supertypes, so it works as a 'covering-type model'. Way actually

[93] *Ibid.*, p.37.

calls it ‘subsumption’, he explains that it is an inductive inference that ‘involves the subsumption of the entity in question under a type in the type-hierarchy’[94]

From the discussion on Mary Hesse’s work we have learnt that the alternative to covering laws is a single case inference. If Hesse’s half-Bayesian argument on single case inferences is used for type-hierarchies, two important problems arise:

1) The role of the maximal similarity rule when a type is available.

2) The role of the maximal similarity rule when a type is not available.

As we know from Mary Hesse, any analogy must be supported on causal knowledge and a positive and negative analogy. The rule of maximal similarity prescribes the selection of the model with the larger positive analogy, that is to say, similarity must be maximised in order to protect the inference from the risk involved in the negative analogy. Is this rule still adequate adding a type-hierarchy? It is not adequate anymore for it can delay, and even prevent the progress of scientific knowledge. I argue instead that a combined strategy using also the opposite rule prescribing the selection of models and analogies with a larger negative analogy is methodologically more robust and more adequate. This rule can be called *the rule of minimal similarity*.

In a type-hierarchy, similarity increases horizontally and decreases vertically, the closer types are to each other the larger the similarity and the positive analogy are, the farer they are the larger the dissimilarity and the negative analogy. Therefore, proximity and position in horizontal and vertical axes of the hierarchy are logically relevant properties. If natural kinds are a jigsaw puzzle whose final shape is unknown, inferences relating distant pieces can speed up the final solution. That is to say, inferences guided by the rule of minimal similarity, i.e. based on larger negative analogies, can relate tokens from distant types as well as discover new types accelerating in this way the construction and expansion of type-hierarchies. Like what happens

[94] *Ibid.*, pp. 198-199.

with the rule of maximal similarity, for the rule of minimal similarity there are also two cases: (1) when the next type above or a supertype is already available; or (2) when there is no such type or supertype. The second case implies a meaning shift and a new semantic mask. The two cases can also be described as cases of normal science and revolutionary science, particularly when a new supertype is introduced displacing an old one.

A methodology with one rule only prescribing to maximise similarity is too conservative, and in some cases it can be detrimental both methodologically and epistemically because it can undermine the growth of scientific knowledge. By maximising similarity we are reducing dissimilarity, and therefore the construction of any type-hierarchy slows down. Hence, a successful inference with a larger negative analogy has both greater epistemic and methodological value but it also carries greater risk. The success of an inference based on a larger negative analogy depends on how much causal knowledge is available, how many accumulated anomalies there are in the present type-hierarchy, and how sharp and well-endowed is the mind of scientists for creating new supertypes.

I call an analogy with a larger negative analogy *a minimal analogy*, and any analogy with a larger positive analogy *maximal analogy*. Mary Hesse did not consider the case of inferences made with minimal analogies; her work was exclusively concerned with cases of maximal analogy. In normal science a minimal analogy is still an option chosen as the means for speeding up the construction of a type-hierarchy, however in revolutionary science it is necessary.

To illustrate how a minimal analogy works, in the next section I discuss the introduction of the new supertype 'force field' in the nineteenth century through the work Michael Faraday and James Maxwell did on magnetism, which eventually lead to a scientific revolution in the early twentieth century.

1.7. Minimal analogy

On one side, there can be almost fully built type-hierarchies from which many properties can be deduced on new tokens, on the other side there are cases where no supertype or type is available for a certain new token, and therefore no properties can be deduced. In this case, single case inferences are needed in order to establish any new property. In the middle of these two cases, there are half-built type-hierarchies with different degrees of development. Harré takes the first case with a largely built hierarchy as the basis for his argument.

The inference of properties based on a large and highly defined type-hierarchy is subject to the old problem of induction just because any empirical universal proposition can be falsified. For instance, inferences made on any new token, say a mouse, which were based on the supertype placental mammal were reliable for a long time until the new properties of a token challenge the hierarchy. When the first marsupial mice were found, a new type had to be introduced splitting the class into two infraclasses.

In the process of building up and updating a type-hierarchy, Harré accepts a certain degree of ad hoc-ness, he explains that 'when attempting to explain an unknown system in terms of a known system, we may try many different locations in our ontological scheme (or type-hierarchy). At one time it was suggested that the structure of the atom might best resemble that of plum pudding—a sponge-like solid with denser matter (raising) scattered throughout […] Whether a particular model is a good one or not depends on how well the unknown system can inherit the laws and properties of the relevant supertypes.'[95] The problem with this idea is that in many cases no type or supertype is available, and therefore no property can be inferred as an inherited feature. In these cases, scientists face a significant epistemological and methodological challenge trying to establish a new supertype, which is likely to produce a shift in meaning and a new semantic mask. The research pursued by Michael Faraday on the magnetic lines of force during the nineteenth century is an example of this kind of challenge

[95] *Ibid.*, p. 107.

and shift in meaning. In these cases, the process of establishing a new supertype is largely *ad hoc*, and it is not deductive or inductive either; the challenge lies mainly in the interpretation of experimental results rather than on their replication.

I argue that the opposition between the supertypes '*mechanical aether*' and '*force field*' in nineteenth century physics, illustrate the contrast between minimal and maximal analogies as rules guiding scientific research. I claim that minimal analogies represent a necessary and progressive method needed for building up type-hierarchies, and I also hold the view that maximal analogies are conservative, and that they can even have recessive or regressive effects in scientific progress. The differences between maximal and minimal analogies and their effects, are illustrated with the models of James Maxwell and Michael Faraday on the magnetic lines of force.

In 1852, Michael Faraday published his strongest defence of the separate ontological status of the magnetic lines of force as continuous physical entities distinct from matter.[96] His argument challenged the idea of action at distance by arguing instead for a non-mechanical and physical continuum as the explanation for the magnetic forces of attraction and repulsion. Following the Newtonian paradigm, James Maxwell wanted instead to produce a mechanical explanation of such an unobservable physical continuum: 'I propose now to examine magnetic phenomena from a mechanical point of view, and to determine what tensions in, or motions of, a medium are capable of producing the mechanical phenomena observed.'[97] The leading idea for such an explanation was that of long vortices parallel to each other created by small particles revolving on their axes. The position and direction of such vortices coincided with those of the lines of force observed around a magnet. Hence, the lines of magnetic force observed on the iron powder scattered around a magnet, were explained as the observable effect of such vortices.

[96] M. Faraday (1852) 'On the physical character of the lines of magnetic force'; the same year Faraday published a second article complementing this one with the title, 'On the Lines of Magnetic Force: Their definite character; and their distribution within a magnet and through space'.
[97] J. C. Maxwell (1861-1862), p. 162.

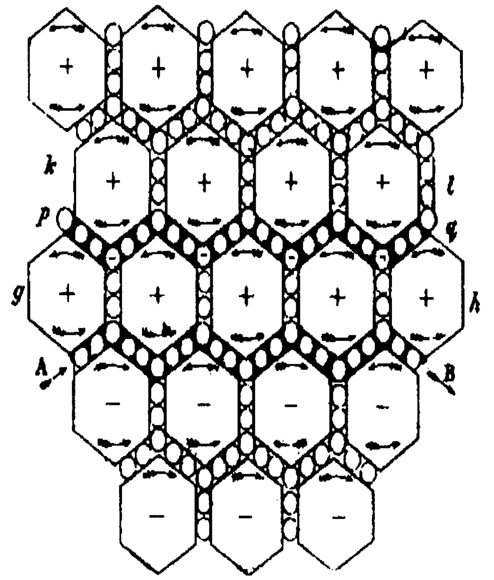
The creation of a full mechanical model was not an easy task. An important problem was to think of a mechanism which could allow all vortices to move in the same direction when an electrical current is induced. If we imagine vortices as pipes placed next to each other, they all would get stuck and stop if each of them moves in the same direction. This is how Maxwell explains the solution to this important problem:

> 'I have found great difficulty in conceiving of the existence of vortices in a medium, side by side, revolving in the same direction about parallel axes. The contiguous portions of consecutive vortices must be moving in opposite directions; and it is difficult to understand how the motion of one part of the medium can coexist with, and even produce, an opposite motion of a part in contact with it. The only conception which has at all aided me in conceiving of this kind of motion is that of the vortices separated by a layer of particles, revolving each on its own axis in the opposite direction to that of the vortices, so that the contiguous surfaces of the particles and of the vortices have the same motion. In mechanism, when two wheels are intended to revolve in the same direction, a wheel is placed between them so as to be in gear with both, and this wheel is called an "idle wheel".' [98]

The postulation of some kind of particle functioning as an idle wheel was a clever mechanical solution to the problem of how to make both electricity and magnetism work together. It combines mechanics of fluids and the mechanics of solids with an analogy and a metaphor taken from natural phenomena, like cyclones or tornados and metallic wheels as they operate in a machine. Maxwell's model relies on a mechanical analogy from the action of natural phenomena and the mechanics of a machine creating a full mechanical explanation, which turns into a maximal analogy within the dominant Newtonian view. This is the graphic model he produced of such a mixed mechanism:

[98] J. C. Maxwell (1861-1862), p. 283.

Figure 1.5. Maxwell's vortex-idle wheel model:[99]



'Let the current from left to right commence in AB. The row of vortices *kl* still at rest, then the layer of particles between these rows will be acted on by the row *gh* on their lower sides and will be at rest above. If they are free to move, they will rotate in the negative direction, and will at the same time move from right to left, or in the opposite direction from the current, and also form and induced electric current.' [100]

The model actually resembles the schematic diagram of a mechanism inside a machine. If we magnified the image, or if we relate it to an actual physical macroscopic model, we can actually appreciate the metaphor in its full dimension. By magnifying it, we can obtain an even more mechanical impression similar to that of tornadoes in an electrical storm, or an image of a hybrid machine such as a hydroelectric power plant, which combines technology with the mechanical force of a natural phenomenon such as a river. This was the kind of model Duhem criticised as distinctive of the English mind, in which one feels like entering into a factory with 'tubes which pump water while others swell and contract; toothed wheels which are geared to one another and engage hooks'.[101] This model almost works as a form of figurative language.

Maxwell created this model following the *method of physical analogy*, which anticipates the work Norman Campbell and Mary Hesse did on the topic. Maxwell borrowed this method from the physicist William Thomson, who had produced successful analogies between different

[99] For more on the explicit use on this analogy see also M. Hesse (1961), pp. 206-212; and N. Nersessian (1984), pp. 69-93.

[100] J. C. Maxwell (1861-62), p, 291.

[101] P. Duhem (1906), pp. 70-71, 56-57

observable phenomena and their theoretical explanations, for the purpose of developing common mathematical solutions. For instance, he drew a fruitful analogy between the electric and magnetic forces by arguing that both were 'distortions' caused by 'the absolute displacement' and 'the angular displacement' of a particle.[102]

Maxwell explains that 'by a physical analogy I mean the partial similarity between the laws of one science and those of another which makes each of them illustrate the other [...] [a] method of investigation which allows the mind at every step to lay hold at a clear physical conception, without being committed to any theory founded on the physical science from which that conception is borrowed, so that it is neither drawn aside from the object in pursuit of analytical subtleties, nor carried beyond the truth by a favourite hypothesis.' [103] Note that a physical analogy is not necessarily false; there is just no definite answer yet on its truth-value.

Instead of using only the terms 'force' or 'energy' in his analogy, Maxwell used the term 'aether' as a description for the unobservable magnetic fluid depicted in his model. The aether was still matter just of a subtle kind. Over two centuries the aether was a well-established natural kind in physics, which can be described as a supertype with several types and subtypes such as the luminiferous aether introduced by Newton, the stationary and gravitational aether postulated by Christian Huygens, the elastic and solid aether suggested by George Stokes, and the electromagnetic aether depicted in Maxwell's model. By maximising similarity with the predominant supertype, the electromagnetic aether simply became another subtype in the Newtonian semantic mask, where all types of aether were mechanical. Once a new subtype is added, properties are just inferred as inherited traits. There is no meaning shift; the mask virtually covers all aspects inheriting properties from a supertype to different types and subtypes. The main scientific task consists only of figuring out how a new mechanism would look like and how it would operate, which is what Maxwell did following the rule of maximal similarity.

[102] W. Thomson (1847), p. 62.

[103] J.C. Maxwell (1855-1856), p. 156.

Because of this, Maxwell's model was methodologically conservative, and it later became recessive and regressive. Ontologically there was no big leap, no significant gain for nearly a century, until Albert Einstein in 1905 and 1920 rejected the need for an aether and established the concept of a field.[104] In contrast, Faraday throughout his investigations and in his exchange with Thomson was reluctant to accept a mechanical explanation of the lines of force; he explicitly wanted to de-mechanise them.

For more than three decades, Faraday tried different analogies and theoretical explanations of magnetism and electricity, which finally led him in 1855 to the postulation of a magnetic force field distinct from matter.[105] This ontological distinction anticipated the current distinction we draw between the two supertypes energy and matter. The whole discovery was an *ad hoc* process, during which different hypotheses were entertained by Faraday, who increasingly became aware of the limitations of the dominant Newtonian paradigm. His research and findings show he was working at the semantic boundaries of the Newtonian paradigm trying to make sense of phenomena such as diamagnetism, which remained anomalous within the mechanical view.

Faraday's search for an explanation of the magnetic lines of force started in 1820, when he rejected André Marie Ampère's hypothesis of an undulating fluid with two electric effluvia as the explanation of magnetism. Ampère believed magnetism was not a new phenomenon but mere electricity in motion. In 1830, Faraday studied Augustin-Jean Fresnel's undulatory theory of light, which didn't need Ampere's electric effluvia, and rested instead on an analogy between the vibrations of the sound and the waves of light. Fresnel rejected the idea of aether as a fluid, and postulated instead an elastic solid aether able to transport both longitudinal and transverse waves. Faraday used this idea of an elastic solid aether, and he placed the locus of magnetic

[104] A. Einstein (1905), p. 2; (1920), pp. 13, 16; see P. M. Brown (2002) for the differences between Einstein's concept of a field and current views, which Brown claims are closer to those of Faraday than Einstein's

[105] Historical accounts with different explanations of Faraday's creation of the concept of a magnetic force field can be found in B. G. Doran (1975), and D. Gooding (1980).

action in the ‘inductive lines of force’. Then in 1845 he met William Thomson; the exchange between the two gave rise to the non-Newtonian concept of a magnetic field.

Thomson’s main interest was to produce a mathematical theory of magnetism with a method based on metaphors and analogies that he created by relating different phenomena. He first suggested an analogy between heat and magnetism assuming that the inductive lines of force acted like heat waves. Faraday had rejected action at distance as an explanation of magnetism, so his main challenge was to find a satisfactory explanation of the continuity of magnetism in space. The analogy with the waves provided a model for such continuous action. A constant problem Faraday saw with this and other analogies and models, was the need for a surrounding substance—an aether—which would serve as the medium allowing the travel and action of magnetic forces. This implied an ontology with three elements: magnetism, matter and aether. The alternative hypothesis consisted of eliminating the aether by assuming an empty space, but he just could not make full sense of the lines of magnetic force acting in a vacuum. This was a problem that persisted for a century in the theories of James Maxwell, Hendrik Lorentz and Albert Einstein.[106]

Stimulated by Thomson’s analogy, Faraday developed in 1846 a new model where forces form a plenum filling up all space such that no aether was needed. This plenum was made up by atoms acting as the centres of forces around them; he explains that ‘the point intended to be set forth for consideration of the hearers was whether it was not possible that the vibrations, which in a certain theory are assumed to account for radiation and radiant phenomena, may not occur in the lines of force which connect particles, and consequently masses of matter together; a notion which as far as it is admitted, will dispense with the aether, which, in another view is

[106] Further historical details of this problem from Faraday and Einstein can be found in N. Nersessian (1984).

supposed to be the medium in which these vibrations take place.'[107] A model with atoms and forces was only closer to the current conception of fields derived from the work of Einstein.

But there was no lineal progress in Faraday's search for the best model and hypothesis explaining the nature and operation of the magnetism. By 1850 he abandoned the dualism atoms-forces by reconsidering again aether as a medium. This time as a fluid whose action was described with the analogy of a stretched spring transmitting the magnetic forces. He acknowledges that the idea of the lines of force acting in an empty space without a medium 'is difficult to comprehend according to the Ampere theory […] or with any other generally acknowledged, or even any proposed view or even any trial speculation that I am aware of.'[108] One year later he goes back to an explanation with no aether: 'we have to consider the true character and relation of space free from any material substance. Though one cannot procure a space perfectly free from matter, one can make a close approximation to it in a carefully prepared Torricellian vacuum […] Mere space cannot act as matter acts, even though the utmost latitude be allowed to the hypothesis of an ether; and admitting that hypothesis, it would be a large additional assumption to suppose that the lines of magnetic force are vibrations carried on by it.'[109]

By 1851 new doubts and hesitation appeared, when he writes that 'how the magnetic forces is transferred through bodies or through space we know not; whether the result is merely action at a distance, as in the case of gravity; or by some intermediate agency, as in the case of light, heat, the electric current, and (as I believe) electric static action.'[110] In 1852, he finally converted to the field concept Thomson had originally suggested it to him. Faraday explains that 'I conceive that when a magnet is in free space, there is such a medium (magnetically speaking) around it. That a vacuum has its own magnetic relations of attraction and repulsion is

[107] Faraday (1846) 'Thoughts on ray-vibrations'; the idea of centre-atoms and forces is similar to that of R. J. Boscovich, whose work was known to Faraday, although it is controversial the extent to which Faraday took this idea from him; see B. G. Doran (1975), p. 166.
[108] In Martin, T. (1932-1936), Vol. V, #10834; see also B. G. Doran (1975), p. 174.
[109] M. Faraday (1851), p. 194; #2787.
[110] M. Faraday (1852), p. 330; #3075.

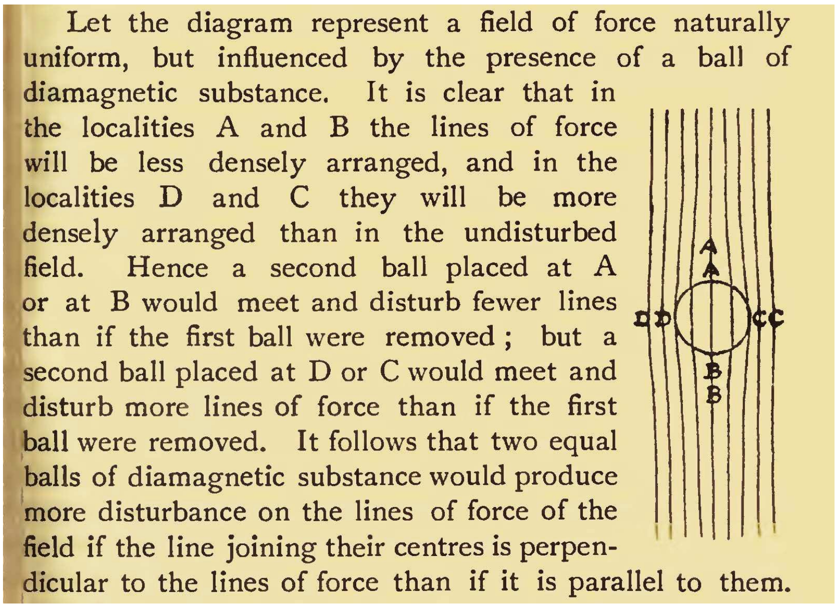
manifest from former experimental results; and these place the vacuum in relation to material bodies, not at either extremity of the list, but in the *midst* of them [...] What that surrounding magnetic medium, deprived of all material substance, may be, I cannot tell, perhaps the aether.'[111] In his last statement from 1855, he fully abandons the hypothesis of an aether, which he now considers to be inadequate and old:

> My physico-hypothetical notion [...] views these lines as *physical* lines of power [...] Those who entertain in any degree the aether notion might consider these lines as currents, or progressive vibrations, or as stationary undulations, or as a state of tension [...]It was always my intention to *avoid* substituting anything in place of these fluids or currents, that the mind might be delivered from the bondage of preconceived notions; but for those who desire an idea to rest upon, there is the old principle of the aethers.[112]

As we know, a few years later Maxwell would go back to the 'old principle of the aethers' with his vortex-idle wheel model. Jointly Faraday and Thomson produced the concept of a force field after ten years of collaboration. Like Faraday, Thomson also thought that magnetism was distinct from matter by claiming that 'this imaginary substance possesses none of the primary qualities of ordinary matter, and it would be wrong to call it either a solid, or the "magnetic fluid", or "fluids"'[113] Although, he was more interested in developing a mathematical theory than investigating the 'physical nature of magnetism', he nonetheless produced the idea of a 'field of force' supported on a basic graphic model, which he communicated to Faraday for the first time in a letter from 19th June 1849:

[111] *Ibid.*, p. 425; #3277.
[112] M. Faraday (1855), pp. 529-530; #3301-3302.
[113] W. Thomson (1851), p. 251.

Figure 1.6. First basic model of a magnetic field:[114]



Let the diagram represent a field of force naturally uniform, but influenced by the presence of a ball of diamagnetic substance. It is clear that in the localities A and B the lines of force will be less densely arranged, and in the localities D and C they will be more densely arranged than in the undisturbed field. Hence a second ball placed at A or at B would meet and disturb fewer lines than if the first ball were removed; but a second ball placed at D or C would meet and disturb more lines of force than if the first ball were removed. It follows that two equal balls of diamagnetic substance would produce more disturbance on the lines of force of the field if the line joining their centres is perpendicular to the lines of force than if it is parallel to them.

Thomson represented the magnetic field as naturally uniform affected by a ball of diamagnetic matter. In his later work he refined this basic model showing the different effects different spherical bodies produced, namely a ball with no intrinsic magnetism, and a ball inductively magnetised. Such models were the support of the sophisticated mathematics he developed with a number of equations, values and descriptions of regular effects. Some of those values and graphic sophistication can be appreciated in the following three models:

[114] S. P. Thompson (1910), p. 215; see also B. G. Doran (1975), p. 175.

Figure 1.7. Model of a magnetic field 'with an inductively magnetized globe'[115]

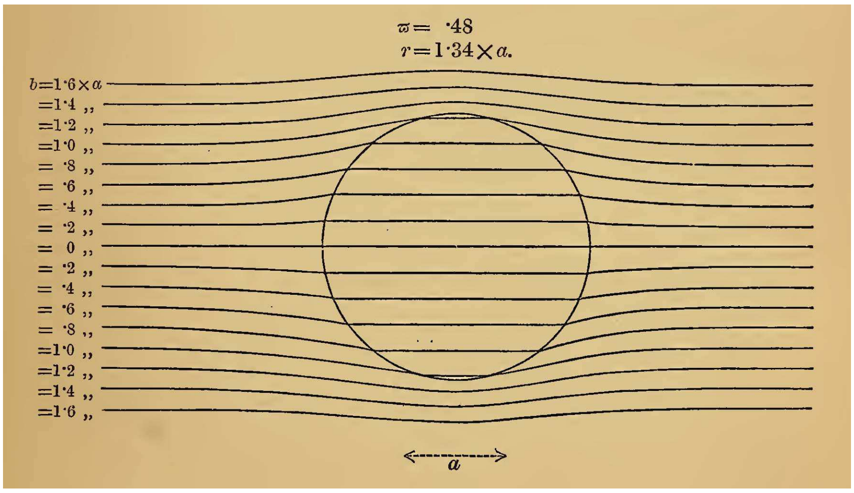
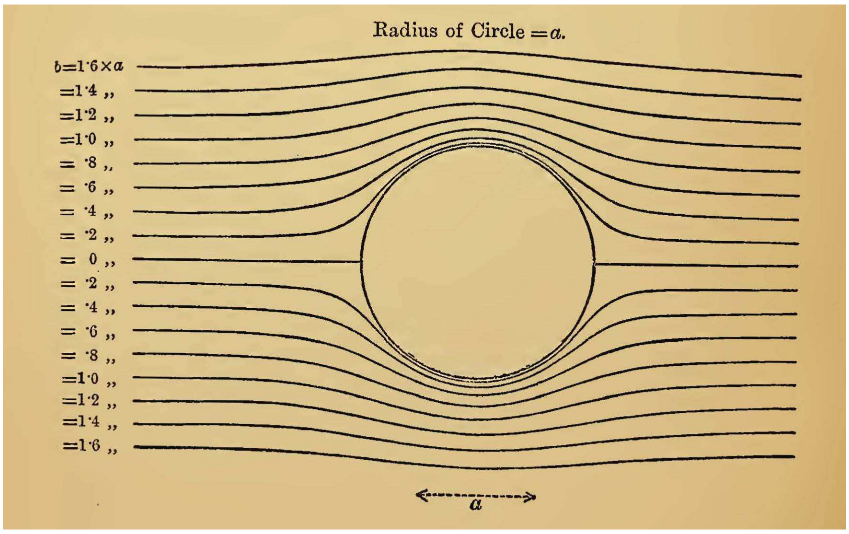
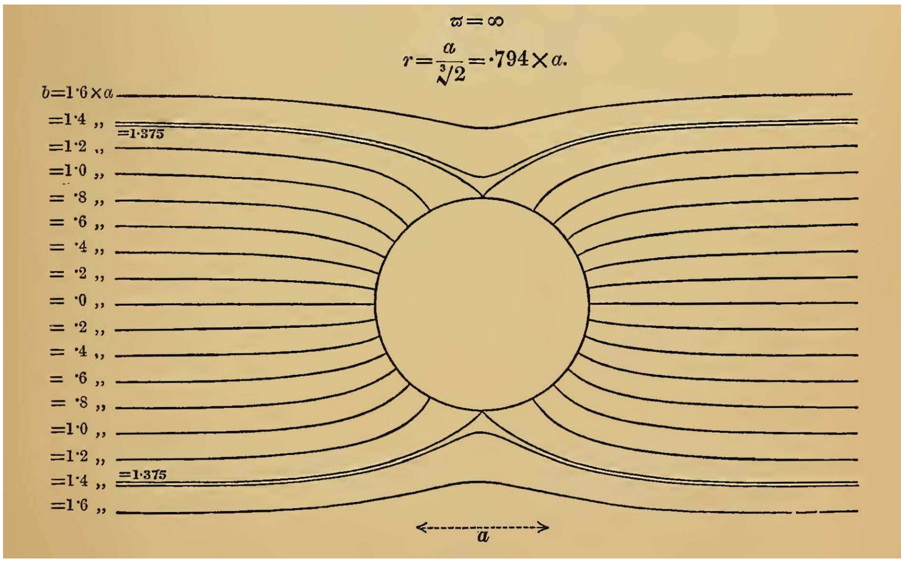


Figure 1.8. Model of a magnetic field representing 'the lines of magnetic force in the neighbourhood of a solid globe of any ferromagnetic or diamagnetic homogeneous material destitute of intrinsic magnetism, put into a uniform magnetic field':[116]



[115] W. Thomson (1872), p. 493.
[116] *Ibid.*, p. 491.

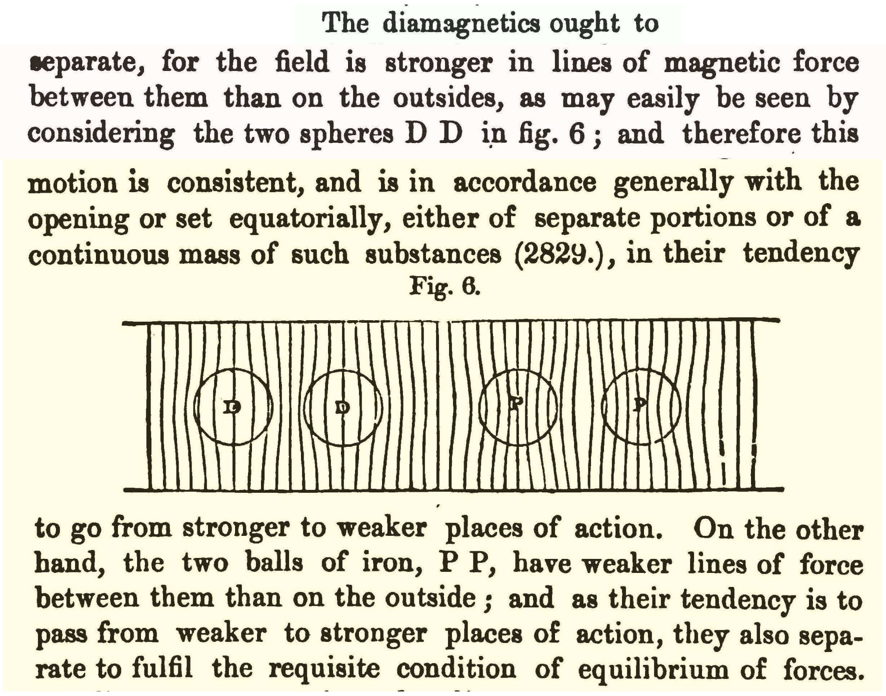
Figure 1.9. Model of a magnetic field representing 'the lines of magnetic force in the neighbourhood of a globe of soft iron in a uniform magnetic field'[117]



By 1850, Faraday was using the same graphic model for representing similar magnetic phenomena, namely the opposite effects diamagnetic and iron balls have on a magnetic field as it can be appreciated in the figure below.

[117] *Ibid.*, p. 491.

Figure 1.10. Model of a magnetic field affected by iron and diamagnetic ball[118]



The diamagnetics ought to separate, for the field is stronger in lines of magnetic force between them than on the outsides, as may easily be seen by considering the two spheres D D in fig. 6; and therefore this motion is consistent, and is in accordance generally with the opening or set equatorially, either of separate portions or of a continuous mass of such substances (2829.), in their tendency to go from stronger to weaker places of action. On the other hand, the two balls of iron, P P, have weaker lines of force between them than on the outside; and as their tendency is to pass from weaker to stronger places of action, they also separate to fulfil the requisite condition of equilibrium of forces.

Fig. 6.

By comparing Maxwell's vortex-idle wheel model (Figure 1.5.) to the magnetic field models of Faraday and Thomson, it is possible to appreciate a sharp and clear meaning shift from a semantics of contiguous action based on the mechanical action of subtle matter, to a semantics of contiguous action based on the non-mechanical action of force fields. Faraday was aware of this for he expressed how difficult it was to make sense of distinct nature of the lines of force, and how they would act without a medium.

Faraday and Thomson's models are examples of *minimal analogies*, where the similarities with the Newtonian mask are minor; they relied on a minimal mechanical analogy represented mainly by presence of balls of different kind affecting the field. The remaining part of the models is non-mechanical, and therefore it constitutes a disanalogy. The minimal analogy was a road to scientific progress in the construction of a new supertype and its respective hierarchy with types and subtypes. In contrast, Maxwell's model is mechanical it all its details, and

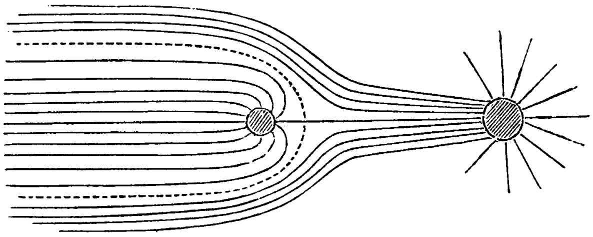
[118] M. Faraday (1851), pp. 211-212; #2831; see also p. 208; #2831; p.204; #2807 for more examples of the same kind of model.

therefore it exemplifies a maximal analogy. In spite of the mathematical progress Maxwell made using the vortex-idle wheel model, it was ontologically regressive because it relied on an ontology of aethers already superseded by Faraday. Maxwell knew Faraday's work but he decided to continue working within the Newtonian paradigm, and he actually tried to reconcile the magnetic lines of force with the action of a gravitational aether.

On 9th November 1857, Maxwell wrote a letter to Faraday, where he put forward a definition of gravitational force as a 'pushing force' stemming from the sun and from each planet. The crucial difference between the two was the status of force fields as extended non-mechanical separate entities, where massive bodies are *placed into* versus extended non-separate mechanical entities being *emitted* by those bodies. In his letter, Maxwell drafted the following basic graphic model:

Figure 1.11. Lines of force of gravitational aether:[119]

The lines of Force from the Sun spread out from him, and when they come near a planet *curve out from it*, so that every planet diverts a number depending on its mass from their course, and substitutes a system of its own so as to become something like a comet, *if lines of force were visible*.



The lines of the planet are separated from those of the Sun by the dotted line. Now conceive every one of these lines (which never interfere but proceed from sun and planet to infinity) to have a *pushing* force instead of a *pulling* one, and then sun and planet will be pushed together with a force which comes out as it ought, proportional to the product of the masses and the inverse square of the distance.

The difference between this case and that of the dipolar forces is, that instead of each body catching the lines of force from the rest, all the lines keep as clear of other bodies as they can, and go off to the infinite sphere against which I have supposed then to push.

[119] P. M. Harman (1990), pp. 548-552. In Queries 21 and 22 (*Opticks*, 1717, pp. 325-327), Isaac Newton had speculated on the composition and operation of the gravitational aether, which he thought was made of small particles; the impulses of a stream of these particles bombarding the planets would cause gravitation. This gravitational aether would be denser in empty space than in the vicinity of planets or any other massive body. Hence, the Earth moves towards the Sun under the pressure of the aether, like a cork rising from the depths of the sea.

Compare this model to Faraday's model above (Figure 1.10). The lines of force in both act in a similar fashion by expanding and contracting, but the explanation of such effects and the nature of those lines, makes the difference between a Newtonian model and Faraday's model. Faraday responded rapidly to Maxwell, first in a letter written on 13th November, and later in an addendum he published in June 1858, where he criticised him for turning magnetism into a 'mechanical force'.[120] He makes a clear statement writing that 'I do not use the word "force" as you define it, "the tendency of a body to pass from one place to another" […] such a thought, if accepted, pledged them [experimental physicists] to a very limited and probably erroneous view of the cause of the force, and to ask them to consider whether they should not look (for a time, at least), to a source in part external to the particles.'[121]

Maxwell's model of the lines of gravitational force and his vortex idle-wheel model actually complement each other. The first model makes the lines of gravitational force visible by zooming into the actual shape and pathways followed by those lines, the second model zooms in even further to make the actual micro-composition and operation of the lines of magnetic force visible. In both cases mechanisms described with different degrees of detail are offered as explanations of gravitational and magnetic forces. We can assume that a Maxwellian microscopic model of the gravitational aether would be similar to the vortex idle-wheel model, perhaps also with wheels and vortices or similar mechanical parts.

The contrast shown between the models of Faraday and Thomson, and those of Maxwell demonstrates the need for a mixed methodology with both kinds of analogy, namely minimal and maximal analogies. Hesse's inference from analogy is a case of maximal analogy because it prescribes a choice for models with greater similarity; this type of inference can therefore be renamed as *inference from maximal analogy*. I am arguing for a second type, which can be called *inference from minimal analogy*. The same half-Bayesian justification Hesse produced for the inference from maximal analogy could be used for the inference from minimal analogy,

[120] M. Faraday (1858), p. 460.

[121] B. Jones (1870), pp. 390-391; letter from Faraday to Maxell from 13th November 1857.

which is best represented here by the Faraday's models of force fields and the ontology underpinning them.

I argue that mixed methodology responds better to the demands from type-hierarchies and meaning shifts as they are advanced by Eileen Way. On the one hand, a methodology relying on maximal analogies like that of Hesse is at risk of becoming not only conservative but also regressive, or at least recessive, like Maxwell's models of the electromagnetic and gravitational aether show. A methodology that also includes inference from minimal analogies provides the grounds for scientific progress as it has shown with Faraday and Thomson's models.

On the other hand, there is a greater risk of failing with any inference from a minimal analogy; progressive rules often carry greater risk. Conservative inferences from maximal analogies are less risky. Hence, only the use of both analogies along different scientific communities or individuals provides both protection against failure building up a type-hierarchy and protection against ontological regression, where new semantic masks and new supertypes are not developed further and more rapidly. The exclusive use of one kind of analogy would be a methodological mistake just as it would also be a mistake to use both undermining the advance of one of them; the right science policy should ensure opportunities of equal progress.

For nearly a century, the scientific labour and the theories produced by Faraday, Thomson and Maxwell show how *de facto* scientists on the whole were following a mixed methodology pursuing minimal and maximal analogies. This thesis can be extended to the work of Lorentz and Einstein. A philosophical justification provides such labour and its products with *de jure* grounds, not only to episodes from the past but also to current scientific research. It meets the needs for the construction of type-hierarchies both in normal and revolutionary science. Only the justification of a mixed strategy can provide both protection and progress as well as guidance on science policy.

—O—

Chapter 2

Social Machines

2.0. Introduction

Currently, there is only a thin bridge connecting the Mechanical view with the social sciences, and there is no comprehensive account either on how this view can be applied to those sciences. The main aim of this chapter is to enlarge the bridge and lay some initial grounds for such an account. I do this in two steps. The first step covers sections 2.1., 2.2. and 2.3, where I introduce the machine metaphor and discuss five methodological principles of blueprint-making as well as other philosophical aspects of design and engineering in the social sciences. The second step includes sections 2.4., 2.5., and 2.6., where I do a selective review of some of the main features of the three branches in the social sciences concerned with design, engineering and social mechanisms, namely mechanism theory, institutional design and analytical sociology. Along the discussion in all sections, I identify a number of relevant problems which set up an agenda for further research, such as the underdetermination of theories and the iteration of metaphor both relevant for the explanation by social mechanisms.

In this chapter, I argue for the following four theses. Blueprints from game theory are Galilean blueprints that must necessarily be completed by the engineering methods and knowledge from experimental economics. Holistic engineering is feasible and successful, while piecemeal engineering can fail and can also be unfeasible; a choice between the two must rely on the amount and reliability of social technological knowledge. A wider methodology of design and engineering can be produced by detaching piecemeal engineering from its association with liberal capitalist societies, and by detaching holistic engineering from its association with socialism. Operant conditioning from behaviourist psychology is consistent with the design and engineering institutions and policies in representative democracies. These four theses are put forward through the discussion organised in the following six sections.

In section 2.1., I introduce the machine metaphor by relying on the work from Nancy Cartwright on socioeconomic machines. I use the term ‘social machines’ instead of ‘socioeconomic machines’ for referring to any state institution, firm or farm. I show how the machine metaphor is an escalation from a mechanical metaphor based on natural forces to a mechanical metaphor based on artefacts, which implies an ontology of artefactual institutions and artefactual behaviour brought about by design and engineering. This is in contrast to an ontology of traditional institutions and traditional or customary behaviour. Besides this distinction, three methodological principles of blueprint making are discussed as well as two related ontological theses on realism of capacities and individualism. Such principles and theses belong to the work from Cartwright on socioeconomic machines; they are illustrated with a game theory model on debt contracts produced by the economists Oliver Hart and John Moore. The main aim of this section is to introduce and build up an insightful and fruitful discussion of the machine metaphor to be used as the foundation for a methodology of design and engineering in the social sciences.

In section 2.2., I discuss two further principles of blueprint-making, which are concerned with the shielding of social machines, and with how to get these machines running. All together these two principles and the three principles discussed section in 2.1. constitute the method for blueprint-making advanced by her. A relation is established between these five principles and those on policy-making also advanced by Cartwright. The analysis of the model from Hart and Moore shows that blueprints from game theory offer poor information on shielding, and no information on how to get social machines running. I argue that this occurs because the blueprints produced by game theorists are Galilean blueprints, which despite creating novel mechanisms for solving important problems, they remain highly idealised, leaving important gaps on crucial aspects about shielding, construction and operation. The identification of these gaps in the design of social machines prepares the discussion for chapter three, where the work of experimental economists is discussed showing how it fills those gaps.

In section 2.3., three important distinctions are made relevant to design and engineering in social machines. The first one is an ontological distinction between artefactual institutions and traditional institutions related to the constructivist and evolutionary views in the social sciences; the International Monetary Fund and the International Gold Standard are used as examples of each view. The second one is a methodological distinction between two types of design, namely dirigiste and libertarian; the blueprints from John Maynard Keynes and Friedrich Hayek on a new international monetary system are used as examples of these two types. I show how the final design of the International Monetary Fund was the product of a hybrid blueprint that combines dirigiste and libertarian aspects.

The third one is a distinction between two types of social engineering, namely holistic and piecemeal. Against Popper, I argue that holistic engineering is feasible and that piecemeal engineering, by itself, does not offer a safe method for it can also fail. I hold that a choice between holistic and piecemeal engineering depends on the amount and reliability of the social technological knowledge available. Piecemeal engineering is generally associated with liberal capitalist societies, while holistic engineering is generally associated with socialist ones. Against this association, I also argue for a methodology of design and engineering, which can be detached from ideological and historical biases. I do all this by referring to the cases of successful holistic engineering in Russia, East Europe and Chile. I also make reference to the failures of piecemeal engineering implementing new education and nutrition policies in California and Bangladesh, and the successful cases of piecemeal engineering in China and Vietnam, implementing special economic zones, and in the United Kingdom with the design and implementation of the internal markets in the National Health Service.

In section 2.4., I present analytical sociology as the social science that more comprehensively adopts the machine metaphor and the realist views from Rom Harré and

Nancy Cartwright with a strong focus on social mechanisms and methodological individualism. Analytical sociology illustrates how the Mechanical view is used in social explanation by making a clear contrast between the positivistic covering-law model of explanation and the explanation by mechanisms, and also between grand theories and middle-range theories. Social mechanisms are explained by individual decisions based on beliefs, intentions and desires, so consistently with the realist view on unobservables and causality that intentional explanations become cases of causal explanation.

Because a material view of the mind is also adopted, I identify two challenges for the explanation by mechanisms. The first one is the underdetermination problem, which arises because of the different neurological theories about the localisation of brain functions, and the connection and communication among nerve cells. The second one consists of the constant iteration of metaphor, where terms referring to macroscopic observable events such as 'currents', 'force', 'field', 'repulsion', 'conductor' and 'wave' are metaphorically used for describing microscopic unobservable events.

In section 2.5., I discuss institutional design in political science, which emerged as a reaction to the methodological individualism from behaviouralism in political science and rational choice theory. Methodologically, institutional design lies between mechanism design theory, where new mechanisms are created, and analytical sociology where actual mechanisms are studied. Such a method consists of comparative studies on the positive and negative effects that different sets of rules, incentives and penalties can have from current and past institutions. The results of these comparative studies are used for assisting the choice over alternative institutional structures to be implemented with some adaptations in a new domain, expecting the same or similar effects. It is also described as political engineering concerned mainly with designs on different forms of government, electoral systems and constitutions.

Like analytical sociology and mechanism design theory, institutional design also relies on the machine metaphor by describing political institutions as machines moved by the 'engines of

Bentham', which shape political behaviour through 'punishment' and 'reward', that is to say, by using incentives and penalties. I argue that operant conditioning from behaviouristic psychology is the updated scientific version of the engines of Bentham, which shapes behaviour through the regulation of reinforcement and control over contingencies. B.F. Skinner extended these ideas for a technology of behaviour and cultural design, which are consistent with the dirigiste, piecemeal and holistic engineering performed in representative democracies with free markets and welfare state. Only alternative libertarian designs, which also foster economic equality, can produce a more substantive reduction of control and dirigisme.

The last section 2.6., consists of an analysis and discussion of mechanism design theory, it further illustrates and supports the discussion and conclusions from sections 2.1. and 2.2. on the limits of Galilean blueprints from game theory. I discuss the work from Leonid Hurwicz, one of the founders of mechanism design theory, who discovered an inconsistency in the design of free markets, which requires true information on preferences and other aspects, while at the same time it creates incentives for lying on those preferences. This is a very important problem because it leads to an inefficient allocation of resources with suboptimal equilibria, and a potential for large social losses. This discovery gave rise to a principle of design called 'incentive compatibility'. I illustrate the work on design produced by mechanism design theorists, with the 'multiple auction by sealed bids' devised by William Vickrey. The design of this new auction shows both the design power of game theory and the significant shortcomings it shows on how to shield the auction, and how to get it running. These conclusions show that knowledge of design necessarily requires knowledge of engineering from experimental economics. Together both kinds of knowledge constitute the social technological knowledge available from the science of economics. Decisions on feasible and unfeasible blueprints for new social machines are also necessarily subject to the advancement of such technological knowledge.

2.1. The machine metaphor

James Maxwell produced a fully mechanical model of magnetic force mainly based on natural forces adding one component only from a machine, namely an idle wheel. Nancy Cartwright extends this view creating a *machine metaphor* of both nature and society. Nature and society are seen as an array of steady machines producing regular outcomes, and each of these machines consists of an array of separate parts assembled into mechanisms under the guidance of a blueprint.

Maxwell described the solar system as fully mechanical with no fields but with gravitation conceived as a pushing force, whose microscopic model would have contained parts similar to the wheels and vortices of the electromagnetic aether. Natural mechanical forces largely define his models and, in spite of having an important role, artefactual mechanical effects are small in proportion. Cartwright also sees the solar system as mechanical but she escalates the Mechanical view by creating a metaphor entirely based on artefacts, that is to say, on machines, which she calls *nomological machines*. A nomological machines is 'a fixed (enough) arrangement of components, or factors, with stable (enough) capacities that in the right sort of stable (enough) environment will, with repeated operation, give rise to the kind of regular behaviour that we represent in our scientific laws.'[122] Using the laws of Kepler, she explains how the nomological machine metaphor works.

Based on the astronomical data on Mars gathered by Tycho Brahe, Johannes Kepler established the following three laws of planetary motion: i.) The orbit of every planet is an ellipse with the Sun at one of the two foci, ii.) The line joining a planet and the Sun sweeps out equal areas during equal intervals of time, and iii.) The square of the orbital period of a planet is proportional to the cube of the semimajor axis of its orbit. Later, Isaac Newton postulated a gravitational force and established the magnitude of such force required to keep a planet in such elliptical orbit with a constant speed. Generally, the laws of Kepler and Newton are presented as examples of regularities with no further explanation on how they arise. The

[122] N. Cartwright (1999), p. 50.

machine metaphor provides an answer to this question by postulating *capacities*. This is done by figuring out 'the nomological machine that is responsible for Kepler's laws—with the added assumption that the operation of the machine depends entirely on the mechanical features and their capacities. This means that we have to establish the arrangement and capacities of mechanical elements, and the right shielding conditions that keep the machines running properly, so that it gives rise to the Kepler regularities.'[123]

Hence, the machines that give rise to natural laws like those of Kepler consist of three main parts, namely capacities, the specific assembling of them, and the provision of a shield for protection. More specifically, this means a realist belief in gravitational force as a capacity or causal power existing in each planet and other massive bodies in the solar system; knowledge of the joint effects of this capacity from massive bodies of different size placed in different positions; and knowledge of events which can affect or prevent isolated or joint effects of the gravitational forces in operation. The philosophical choice for capacities constitutes a radical departure from empiricist standards, which ultimately relies on the cogency of a realist argument.[124]

The joint effects of gravitation for any set of known planets and massive bodies can be calculated reliably by using Newton's laws and equations. Knowledge of the presence of new planets or potential colliding objects such as asteroids and comets, which can affect the running of the solar system as a machine, can only be obtained gradually and normally *a posteriori* when a distortion has already been observed. This affects the scope and power of the shielding conditions. Cartwright accepts this limitation explaining how the discovery of a new planet as an 'observed irregularity points to a failure of description of the specific circumstances that characterise the Newtonian planetary machine. The discovery of Neptune results from a

[123] *Ibid.*, p. 50.

[124] In *Nature's Capacities and Their Measurement* (1989), Nancy Cartwright has produced such a realist argument for capacities.

revision of the shielding conditions that are necessary to ensure the stability of the original Newtonian machine.'[125]

In this way, the nomological machine metaphor is employed also for philosophical purposes. It works as a mask, exhibiting new features of scientific theories and scientific explanation, which remain hidden under the Syntactic view. Under this technological metaphor, any scientific laws only holds relative to the operation of a nomological machine, which comprises a number of parts assembled under the right plan or blueprint as well as a protective shield, and further ceteris paribus conditions. All these elements remain unnoticed under the regularity view of scientific laws. With the Mechanical view, Kepler's laws and any other natural law arise as the product of different nomological machines. Scientific explanation ceases to be guided by the covering-law model, and theories become collections of models of nomological machines. Nature consists of a big array of nomological machines.

The metaphor also extends to the state, markets and society. Economic and political institutions as well as contracts among individuals are also seeing as technological artefacts. Society as a whole becomes an array of nomological machines, which Cartwright calls *socioeconomic machines*, while theories in the social sciences become collections of models on those socioeconomic machines. This is the Mechanical view escalated from natural mechanical forces to artefactual ones now being extended to society and theories in the social sciences.

As Nancy Cartwright advances it, the Mechanical view applied to the social sciences consists of five explicit methodological principles, and three ontological theses. In this section, only the first three principles and the first two theses are discussed, the remaining two principles and single thesis are discussed in the next section. The first three principles establish that any model of a socioeconomic machine must show:[126]

[125] N. Cartwright (1999), pp. 52-53.
[126] *Ibid.*, p. 146.

i) The parts that make up the machine, their properties and the separate capacities.

ii) How the parts are to be assembled.

iii) The rules for calculating the outcome from the joint operation of the assembled parts.

To illustrate these methodological principles, Cartwright uses an example from game theory applied to long-term debt contracts. In particular, the model of a 'repudiation-proof contract' produced by the economists Oliver Hart and John Moore. [127] Seen as socioeconomic machines, investment contacts must function steadily by producing regular outcomes, which depend on the knowledge game theorists have on the individual players and their capacities as well as knowledge of the different expected outcomes from their mutual interaction. In this case, the regular expected outcome consists of a timely delivery of credits from the investor, and the accomplishment of business targets by the entrepreneur until the full completion of the project.

In the model, Hart and Moore describe the parts of the machine and the capacities of those parts, namely two individual players an investor and an entrepreneur, both displaying specific psychological capacities. These consists of self-interest, greed, perfect and costless calculation, and full rationality. It is also assumed that the entrepreneur has a special capacity consisting of particular skills relevant to the project, which are not easily and costlessly replaceable. Because of this, he enjoys greater bargaining power. Other parts are structural or external to both players such as identical discount rates, certainty in all operations, rules for renegotiation, and the existence of a frictionless second-hand market for the physical assets of the project.[128] The structural parts and the players are assembled in one game in two main stages, one with an initial negotiation and agreement on a certain distribution of the surplus, and a second one when repudiation of the contract occurs and the surplus is now divided in equal parts of 50% each.

[127] O. Hart and J. Moore (1994); the analysis from Cartwright is based on an earlier version of this article published in 1991 as Discussion Paper No. 129 by the LSE Financial Markets Group.

[128] O. Hart and J. Moore (1994) p. 861.

Long-term debt contracts pose particular challenges. One of these challenges arises from the opposite repayment preferences between the investor, who prefers a fast repayment, and the entrepreneur, who prefers a slow repayment. This tension increases when opportunities for outside investment of capital or skills exceed the returns of the current project. This leads to greed, self-interest and defection from each player a real possibility. From a social perspective, Hart and Moore wanted to prevent these contracts from failing because of the social losses and inefficiencies that failure creates. The challenge consisted of reversing the repudiation of the contract by devising a set of new rules, which would create opportunities for negotiations available to both players, so that the project is not abandoned but completed. Easy and costless defection must be prevented, while the conditions for renegotiation must keep returns attractive to both players.

They devised a mechanism by relying on the assumptions of certainty and a continuum of optimal points during the renegotiation period, they explain that 'the assumption of perfect certainty, combined with that of renegotiation, implies that there is a continuum of optimal debt contracts', which implies that ' the parties can write a succession of short-term contracts that are renegotiated, or a long-term contract that is never renegotiated along the equilibrium path', and therefore 'a debt contract can be agreed to such that in equilibrium *D* [debtor/ entrepreneur] never repudiates.[129] Recall that in the model the entrepreneur enjoys greater bargaining power. With those two assumptions, the calculation of the joint effects after repudiation is made by using equilibrium theory using specific rules for renegotiation, and by relaxing the assumption of a common discount rate, while the capacities of self-interest, greed and rationality remain the same for each player. In this way, Hart and Moore's repudiation-proof contact illustrates the three methodological principles any model of a socioeconomic machine should follow.

Besides those three principles, Cartwright adds two important ontological theses on socioeconomic machines:

[129] *Ibid.*, pp. 842, 849.

i) Realism of capacities.

ii) Ontology of individuals.

Against empiricist standards, Cartwright argues for a realist belief in unobservable capacities, which she also called 'natures' following Aristotle. Natures or capacities of individuals cannot be reduced to the constant conjunction of two or more episodes of observable behaviour. We should also add that they should not either be considered as having the instrumental status of convenient fictions used only for explaining observable behaviour, nor should they be considered as the product of an inference to the best explanation of observable behaviour in the absence of alternative better explanation. The realist thesis is stronger than instrumentalism and the inference to the best explanation because it holds 'natures as primary and behaviours, even very regular behaviours, as derivative. '[130]

Although, the realism of capacities or natures enjoys better prospects in experimental and behavioural economics, Cartwright argues for it using models and examples from game theory. The contrast between realist and antirealist standards in the social sciences can be clearly observed in the controversy between cognitive psychology and behaviouristic psychology, and between utility theory and preference revealed theory in economics. Adopted as a thesis for socioeconomic machines, the realism of capacities justifies and prescribes the use of psychological capacities as the ultimate explanation for any expected or any observed behaviour.

Against a holism of social facts or social structures, Cartwright argues for individuals and their capacities as the ultimate grounds for explanation in the social sciences. Using the science of economics as an example, she explains that this thesis 'is based on the hope that we can understand aspects of the economy separately and then piece the lessons together at a second

[130] N. Cartwright (1999), p. 149; earlier (1989, p. 9) she chose the term 'capacities' over 'causal powers', currently she believes 'natures' is a better term: 'most of my arguments about capacities could have been put in terms of natures had I recognised soon enough how similar capacities, as I see them, are to Aristotelian natures.' (1999, p. 85); see also N. Cartwright and J. Pemberton (2013).

stage.'[131] This thesis is both ontological and methodological for she explains that 'the analytic method works in physics: to understand what happens in the world, we take things apart into their fundamental pieces, to control a situation we reassemble the pieces, we reorder them so they will work together to make things happen as we will.'[132]

Ontologically and methodologically, individualism is widely accepted, and used in economics and all branches of game theory including mechanism design theory. In contrast, individualism has been abandoned in political science, particularly in institutional design, while it has been strongly vindicated in analytical sociology. Mechanism design theory, institutional design and analytical sociology are discussed in the last three sections of this chapter.

The machine metaphor helps to meet two important scientific tasks, namely the explanation of actual states of the world and the design of new ones. The work of Cartwright addresses both: first through the ontological description of the components of actual socioeconomic machines, and second through the establishment of methodological principles for the blueprints of those machines. The machine metaphor implies a transition from natural systems, natural laws and traditional institutions to constructed laws, systems and institutions. Thus, the solar system, the Roman Senate and the International Monetary Fund become machines just like a bulldozer, a microprocessor or a blender. Natural laws like those of Kepler and economic relations of trade are seen as artefactual just as the flow of electrical currents in a microprocessor. Cartwright writes, 'here it is my strong claim: look at any case where there is a regularity in the world (whether natural or constructed) that we judge to be highly reliable and which we feel that we understand [...] what you fill find is a nomological machine.'[133]

Therefore, the three principles and the two ontological theses, which have just been discussed, apply to both traditional and constructed institutions as well as traditional and constructed social relations. Game theory models can be models of any traditional institution or

[131] N. Cartwright (1999), pp. 149-150.
[132] *Ibid.*, p. 83.
[133] *Ibid.*, p. 58.

social relationship but they can also be models of constructed institutions and social interactions. Unlike Cartwright, I use the term 'constructed' exclusively for artefacts produced with the help from scientific designers and engineers, and I use the term 'traditional' instead of the term 'natural' for any institution or social relation, where no scientific design or engineering has be used. Unlike the term 'natural', the term 'traditional' in the social sciences seems to be accurate, and it also creates a sharper contrast with 'constructed' or 'designed'.

The model from Hart and Moore belongs to those models describing a constructed regularity, that is to say, the model is a blueprint for replacing a traditional or customary type of behaviour, namely the repudiation of debt contacts with a new constructed or artefactual behaviour, namely the ability to renegotiate contracts until the completion of a project. In this way the metaphor of the socioeconomic machine, and the related principles and ontological theses, apply to constructed or designed contracts and institutions. In contrast, debt contracts with no design rely on trade traditions inherited through generations of bankers and traders, so the rules of those contracts are the product of learning across generations without the help from game theorists or social scientists in general.

The repudiation of contracts certainly is an important social problem, and a lasting efficient solution that can benefit all parties involved without creating social losses is not easy to find. Traders and bankers can continue relying of their own means and experience for solving the problem but they can also seek help from social scientists. The use of science is what distinguishes tradition from construction, traditional from designed and natural from artefactual. More precisely, the science to be used is a science of design, whose main task is the production of blueprints.

2.2. Blueprints for social machines

Blueprints are a fundamental and distinctive part of any science of design just like models are also fundamental to the natural and the social sciences. An important distinction must be made between models describing parts of the actual world, and blueprints projecting parts of possible worlds. Because of this basic ontological difference, a *science of design* should be distinguished from what can be described as a *science of facts*. Such a distinction has also been made using terms like basic science in contrast to applied science, and natural and social science as distinct from engineering and technology. The economist Herbert Simon distinguishes natural sciences such as physics and biology from sciences of the artificial; he explains that 'the engineer, and more generally the designer, is concerned with how things ought to be—how they ought to be in order to attain goals and to function. Hence a science of the artificial will be closely akin to a science of engineering', while 'natural sciences are concerned with how things are.'[134]

In economics, Leonid Hurwicz makes a similar distinction, when he writes that 'traditionally, economic analysis treats the economics system as one of the givens. The term "design" in the title [of the article] is meant to stress that the structure of the economic system is to be regarded as an unknown. An unknown of what problem? Typically, that of finding a system that would be [...] superior to the exiting one.'[135] The distinction between positive and normative economics serves the same purpose as he also explains that 'the study of economic systems can be approached either in the spirit of "positive" science ("what is") or "normative" science ("What should be")'.[136]

In economics, blueprints are actually also called 'models', which is generically used without making a distinction between models of the actual social world and models of possible social worlds, or parts of them. In spite of being useful and important, modal distinctions between actual and possible worlds are nonetheless insufficient and partly inadequate for a

[134] H. Simon (1996), pp. 4-5, 114-115.
[135] L Hurwickz (1973), p. 1.
[136] L Hurwickz (1972), p. 425.

science of design.[137] This is because possible worlds in philosophy have been treated mostly formally, and with no interest in how they are part of engineering and design. As a consequence virtually no attention has been paid in philosophy to how the work and products of engineering and design affect the content and normative standards currently used in the possible worlds literature. A further distinction is, therefore, needed between possible and feasible worlds. The latter are the subject matter of a philosophy of the science of design and the blueprints it produces.

Ontologically, feasible worlds are a subset of possible worlds, which can initially be distinguished by criteria obtained from the advancement of *technological knowledge*. That is to say, feasible worlds are a function of the scope, power and reliability of technological knowledge. In a basic sense, technological knowledge consists of justified practices and propositions concerning the design, construction, operation and functioning of social and physical artefacts.[138] Engineering and design produce and preserve this kind of knowledge, so the science of design complements the science of engineering. The success of any new socioeconomic machine and any new policy rely on both knowledge of design and knowledge of engineering. Designers produce blueprints, engineers build the machines, and only technologically feasible blueprints must be selected for building social machines.

In spite of their central role and impact on nature and society, blueprints have received scarce attention from philosophers of science. Among the few works available, there are those from Nancy Cartwright and Francesco Guala.[139] With the Syntactic and the Semantic views, theories became the fundamental units of analysis in the philosophy of science. For the semantic view set-theoretical and physical models became essential as the means for providing

[137] See D. Lewis (1986); S. Kripke (1980).

[138] On technological knowledge see M. De Vries, S. Hansson, A. Meijers (2012), pp. 55-64; J. Pitt (2001), and E. Layton (1987); a body of literature on 'knowledge-how' has been produced in epistemology, which requires a separate research to establish the relationship between both discussions with an interest in design and engineering, see G. Ryle (1946), J. Stanley and T. Williamson (2001), J. Stanley (2011), K. Hawley (2003), and J. Bengson and M. Moffett (2011).

[139] F. Guala (2005), pp. 161-183.

scientific theories with an interpretation, which is crucial for understanding the claims the theory is making as well as for giving minimal empirical content to it. Idealised models such as that of a pendulum with a massless bob, a body in motion on a frictionless plain, and a perfectly rational individual partly fulfilled that purpose. Among others, the work of Nancy Cartwright on realistic models, also called representative models as opposed to interpretative models,[140] represents a shift in the philosophy of physics and the natural sciences towards the Mechanical view. Her more recent interest and work on the social sciences represent the same view; it is the extension of the Mechanical view to blueprint-making and policy-making methods.

Both blueprint-making methods and policy-making methods rely on the use of social mechanisms, while they differ on the scale of the projected changes. Policy changes are considered to be comparative smaller than those considering large state institutions or bigger markets. Cartwright discusses blueprints and policy-making separately, although the methodological principles and ontology she argues for are the same, namely capacities, mechanisms and causal models.

With the model from Hart and Moore, the three methodological principles (i), (ii), (iii), and two ontological theses (i) and (ii) were illustrated and discussed. Two more principles are discussed in this section, while the third ontological thesis is discussed in the next one. They all make a total of five principles and three theses, which all together constitute the Mechanical view from Nancy Cartwright on blueprints and socioeconomic machines:

- Five principles for blueprint-making:

  i) The parts that make up the machine, their properties and the separate capacities.
  ii) How the parts are to be assembled.
  iii) The rules for calculating the outcome from the joint operation of the assembled parts.
  iv) What counts as shielding.
  v) How the machine is set to run.

[140] In M. Morgan and M. Morrison (1999), p. 242.

- Three ontological theses on socioeconomic machines:
    i) Realism of capacities.
    ii) Ontology of individuals.
    iii) Rejection of evolutionary change.

From being prescriptive on how to build models of actual socioeconomic machines, these principles and theses now become prescriptive on how to build blueprints for feasible socioeconomic machines. That is to say, models describe actual machines; blueprints project feasible ones. These principles represent two complementary sides of the machine metaphor, and the methodological argument derived from it. The principles listed are therefore described now as *principles for blueprint-making*.

Game theory models like that of Hart and Moore and others devised for fixing problems such defection, imperfect equilibria or free riding can all be categorised as blueprints. More precisely, these models belong to mechanism design theory. As we saw earlier, Cartwright is fairly optimistic about how informative Hart and Moore's blueprint on long-term debt contracts regarding the first three principles is. In particular, this blueprint defines the individual and the structural parts of the socioeconomic machine, namely the type of individuals participating and their capacities as well as the same discount rates for both, certainty in all operations, rules for renegotiation, and the existence of a frictionless second-hand market for the physical assets of the project. With the help of equilibrium theory outcomes can be calculated from the interaction between investor and entrepreneur. In contrast, she is sceptical on how informative the blueprint is regarding the last two principles, namely how to shield the machine, and how it should get running.

To make the contract enforceable for both entrepreneur and investor, Hart and Moore establish a number of shielding conditions such as no initial sunk costs, expected initial returns larger than those offered by any alternative project, and some penalties. If the entrepreneur repudiates the contract, he loses control over the project's physical assets, so the investor can liquidate them. Furthermore, grounds for continuous renegotiation are also considered, which can lead to expected returns larger than those from liquidation. This happens because in the model the entrepreneur has special skills required for the execution of the project, so the investor has a strong incentive for renegotiating despite his initial wish to defect.

These shielding conditions constitute an important theoretical progress, which become severely limited when the prospects for real application are considered. As Cartwright points out, 'what counts as shielding conditions will heavily depend on what the specific material instantiation is. This is especially true of game-theoretic models, where few clues are given about what real institutional arrangements can be taken to constitute any specific game.'[141] Limitations on shielding are certainly a problem because game theory is not a strong empirical science. Furthermore, it does not have a proper specialised branch of scientists trained with the knowledge and skills required for building, shielding and operating the games it designs, that is to say, the socioeconomic machines it creates blueprints for. Due to the increasing need and pressure for better designs, game theorists working on the design of mechanisms have been adding, in a piecemeal fashion, important internal shielding conditions as it is prescribed here in principle (iv). Among others, they have designed mechanisms against the suppression of norms by the players and against false revelation of preferences.

The demands posed by principle (v) on how to build and get the machine running represent a far more severe problem not only for game theory, but for all branches in the social sciences concerned with the design of institutions and policies. The model from Hart and Moore provides no information on how get the machine running; it only states the game will

[141] N. Cartwright (1999), p. 147.

function optimally as a 'sub-game of perfect equilibrium. Repeated running imply means playing the game again and again.'[142]

The problem extends to all blueprints from game theory, where mechanisms for improvement, reform or new institutions are projected. However, this is not a surprise because these blueprints are *Galilean blueprints* produced by theoretical scientists. The term comes from the models produced by Galileo with some highly idealised objects and conditions such as frictionless inclined planes, and massless cords holding bobs in a pendulum. Similarly, Hart and Moore's blueprint and all game theory blueprints contain different idealisations such as perfect rationality, costless calculations and perfect certainty in all operations. Cartwright explains that this kind of idealisation 'eliminates all other possible causes to learn the effect of one operating on its own'; despite being unrealistic, these models and blueprints are empirically relevant for design because they can establish 'facts about stable tendencies'[143]

Hence, the problems of how to improve shielding for socioeconomic machines and how to get them running remain unsolved. Cartwright argues that detailed causal models of the target population could help improve the shielding conditions for new policies; her methodological view on policy-making consists of the three following principles:[144]

- *Principle 1.* A good way to evaluate whether a policy will be effective for a targeted outcome is to employ a causal model comprising of:
  a) A list of causes of the targeted outcome that will be at work when the policy is implemented.
  b) A rule for calculating the resultant effect when these causes operate together.

- *Principle 2:* Causes are INUS conditions.

- *Principle 3*: Mechanisms matter

[142] *Ibid.*, p. 147.
[143] N. Cartwright (2007), pp.221, 225.
[144] N. Cartwright and J. Stegenga (2011), pp. 308, 313.

The first principle restates principles (i) to (iii) listed above for blueprint-making, while the second is a refinement within the Mechanical view stressing the complexity of causes due to their combined and separate effects. The third principle deserves special attention because it is fundamental in design, it highlights the distinctive task of designers and the main component of blueprints, namely the creation of new mechanisms for solving problems such as free riding, contract repudiation and inefficient allocation of economic resources. Cartwright quotes motherly love, fear of punishment and desire to conform as examples of social mechanisms.[145] Social mechanisms have become a major topic in sociology and political science. Jon Elster has been one of the main contributors to this topic. His work and the different applications of mechanisms are discussed in sections 2.4. and 2.6.

Cartwright explains that causal models in policy-making only help 'to estimate, if only roughly, whether, were a proposed policy to be actually implemented, a specific, identified outcome would be produced.'[146] Therefore, the questions of how to implement and operate a policy and how to get a socioeconomic machine running still need answer, which is already available from the work experimental economists have been doing testing designs and getting new social machines running.

Over the last two decades, a spontaneous division of scientific labour has emerged, so that scientists with the closest set of skills to the those required for the job of an engineer have assumed the challenge, they are the experimental economists. Their skills originally learned and developed for the purposes of testing theories and hypotheses as well as for producing new experimental findings, have been adapted for testing the rules and mechanisms projected in the blueprints produced by game theorists. An outstanding example of this is the job of the experimental economist Charles Plott, who gets new socioeconomic machines running, namely the new multiple-round ascending auction designed for the allocation of segments in the airwave spectrum to telecommunication firms in The United States. This case is discussed in

[145] *Ibid.*, p. 314.
[146] *Ibid.*, p. 289.

chapter three. There the discussion will show how the Galilean blueprints produced by game theorists remain largely undefined not only on how to get any machine running but also on all four remaining principles advanced by Cartwright.

2.3. The engineering of social machines

The final aspect from Cartwright's Mechanical view on design and blueprints is the ontological thesis (iii), where she rejects the evolutionary change of social machines. This is a very important and fundamental thesis of design, which deserves special attention. In the philosophy of the social sciences, this thesis has produced a strong division between those supporting design and those standing against it. Although the debate about social design is an old one, the contemporary expression of it emerged in the 1920s and 1930s with the debate over central planning through the work of economists such as Ludwig von Mises, Friedrich Hayek, Oskar Lange and Abba P. Lerner, and the philosophers, Karl Popper, Michael Oakeshott and Michael Polanyi.

Ontologically, the thesis of a designed order in society opposes the thesis of a spontaneous social order. Spontaneous order is defined as a stage of relative equilibrium within an evolutionary process, where there is no leading agency. The rejection of evolutionary change and the support for designed institutions stands against the thesis of a spontaneous social order, and institutions exclusively based on tradition. Cartwright explains that 'the third thesis is one about which evidence is divided. Ordinary machines do not evolve. They have to be assembled and the assembly has to be carefully engineered […] one of the most clear-cut examples of a designed institution in economics is the International Monetary Fund.'[147] In contrast, she places the International Gold Standard as 'an institution which was not designed, but which evolved gradually over the nineteenth century.'[148] The International Gold Standard is considered as an

[147] N. Cartwright (1999), p. 150.

[148] *Ibid.*, p. 150.

example of spontaneous international order formed through the decisions of individual traders in the marketplace with no agent leading the process.

The economist Carl Menger explains how money originated through a spontaneous process, driven by the increasing need for making trade operations more efficient. Commodities of greater saleability arise as the means of exchange mainly because they are available in large quantities, so by choosing them as currency the bad effects of scarcity are prevented. Other properties such as easy transportation and fitness for preservation also play a part in the selection of the commodity to be used as money. Thus, cowrie shells, cocoa beans, salt bars and some metals have all been used as currencies in different periods of human history. Later, the development of metallurgy set the conditions for a widespread use of metals as the currency across large geographical areas. Menger explains that 'the origin of money (as distinct from coin, which is only one variety of money) is entirely natural […] Money is not an invention of the state. It is not the product of a legislative act'. Instead, the creation of money is the result of the unconcerted behaviour of each individual 'led by [his economic] interest, without any agreement, without legislative compulsion, and even without regard to public interest, to give his commodities in exchange for other, more saleable, commodities'.[149]

In contrast, one crucial aspect of many designed institutions is the inclusion of a leading agent in charge of overseeing and regulating the actions of the participating individuals, in order to ensure the accomplishment of certain aims and goals as they have been defined in a blueprint. In game theory this implies the inclusion of an agent regulating the game by sanctioning the actions of the players. The regulating agent can be a representative of the state, a representative of the proprietor of a firm, or a representative of a landowner. A second crucial aspect of design is the participation of social scientists as designers, that is to say, as blueprint makers whose job is to use the best scientific knowledge available designing any new institution or policy.

[149] C. Menger (1871), pp. 260, 261.

Note that the history of traditional institutions such as commodity money going from cocoa beans and salt bars to silver or gold, shows the need for change and adaptation when faced with challenges. Individuals and institutions then face a choice between responding with the best of their accumulated traditional knowledge, means and custom, or they can respond by getting help from social scientists, which implies the development of artefactual means and artefactual behaviour. This is the crux of the dilemma between design and a spontaneous order.

The challenges and problems experienced by the International Gold Standard after the First World War called into question the use of gold as commodity money. One of the main benefits of the gold standard was the positive effect it had in keeping the exchange rates stable, which provided certainty to trade operations and jobs. However, it had an important flaw for it lacked the means for a prompt response to large fluctuations in the gold stocks, and its availability in the markets. Gold stocks and reserves are subject to the discovery and exploitation of gold mines, which can cause shortages as well as an oversupply leading to inflation, deflation and trade imbalances with the loss of jobs. The gold standard offered no mechanism for restabilising stability within a short period, so by the time stability was regained through the transfer of capitals, high social costs had been paid. These include drastic falls in trade, large unemployment and loss of purchasing power in large parts of society.

Between the two War Worlds a number of alternatives were produced by economists, which attempted to solve the problems created by the Gold Standard.[150] In 1923 the economist John Maynard Keynes made important criticisms of the Gold Standard exposing the different negative effects it had produced over the past decades.[151] He advanced a monetary plan, which would become one of two blueprints discussed at the Bretton Woods Conference, where the foundations for the International Monetary Fund were laid out. The two main components of Keynes's blueprint consisted of the creation of a supranational agency endowed with regulation powers, namely a Clearing Union; and the use of fiat money by both the Clearing Union and

[150] See F. Cesarano (2006), pp. 36-41, 68-99.

[151] J. M. Keynes (1923) *A Tract of Monetary Reform*.

national central banks. The blueprint also included the creation of an international currency to be called 'unitas' or 'bancor'. Both the Union and central banks were required to be technical and not political, they would intervene to solve two main problems created by the Gold Standard, namely the instability of international prices created by the scarcity of gold as commodity money, and the large unemployment created by employers who react against the natural tendency of wages to rise beyond the limits set by the volume of money attached to gold. This is how Keynes describes the mechanism and the leading role of the supranational currency:[152]

> The peculiar merit of the Clearing Union as a means of remedying a chronic shortage of international money is that it operates through the velocity, rather than through the volume, of circulation […] If hoarding is discouraged and if reserves against contingencies are provided by facultative overdrafts, a very small amount of actually outstanding credit might be sufficient for clearing between well-organised Central Banks […] The primary aim of an international currency scheme should be, therefore, to prevent not only those evils which result from a chronic shortage of international money due to the draining of gold into creditor countries but also those which follow from countries failing to maintain stability of domestic efficiency-costs and moving out of step with one another in their national wage policies without having at their disposal any means of orderly adjustment.

The alternative plan put forward by the economist Harry Dexter White still relied on the gold standard; it rejected the creation of an international currency and a supranational agency. It confined any change in exchange rates only to special circumstances with a fundamental disequilibrium, which were left unspecific.[153] The United States held a good part of the gold reserves and a strong power position, so White's plan was largely adopted at the Bretton Woods conference setting up a world-wide system of currencies attached to the US dollar, while the US dollar was itself attached to gold reserves.

[152] J. M. Keynes (1943), pp. 185-186.

[153] See F. Cesarano (2006), p. 133-145.

The basic rationale underpinning White's blueprint is explained by the economist Friedrich Hayek, who criticised Keynes's plan because while it 'might, indeed, be superior […] it is 'not a practical proposition' for it requires 'a wisely and impartially controlled system of managed currency for the whole world'[154] led by a supranational agency. To solve the problems of the scarcity of gold and the resulting instability with trade imbalances, inflation and deflation, Hayek suggested replacing gold with various storable raw commodities such as wheat, sugar, copper and rubber; his own blueprint considered the coordination of national policies, the use of private specialist brokers, and a minimal and almost mechanical role for a monetary agency, which would have very little or no discretionary powers. These are the basics of the blueprint for the new system:[155]

> With this system in operation an increase in the demand for liquid assets would lead to the accumulation of stocks of raw commodities of the most general usefulness. The hoarding of money, instead of causing resources to run to waste, would act as if it were an order to keep raw commodities for the hoarder's account. And as the hoarded money was again returned to circulation and demand for commodities increased, these stocks would be released to satisfy the new demand […] There would, in particular, be no need for the monetary authorities or the Government in any way directly to handle the many commodities of which the commodity unit is composed. Both the bringing together of the required assortment of warrants and the actual storing of the commodities could be safely left to private initiative. Specialist brokers would soon take care of the collecting and tendering of warrants […] In this respect the business of the monetary authority would be as mechanical as the buying and selling of gold under the gold standard […] the monetary authority shall be empowered in precisely defined circumstances to accept in place of (or substitute for) warrants for stored commodities contracts for future delivery of any commodity.

[154] F. Hayek (1943), p. 176.

[155] *Ibid.*, pp. 179-180, 182-183.

There are two important differences between the blueprints from Hayek and Keynes. The first one on the moral psychology used for design, the second one on the methodology and ontology of design.

The moral psychology adopted in design affects the scope and prospects of success for each blueprint. Hayek did not believe that government agencies, national or supranational, could be trusted with power and control over commodity money, while Keynes does not believe that proprietors of firms or landowners can be trusted to provide rapid and coordinated action for stabilising prices, or the availability of commodity money without depressing wages or creating unemployment. They both cared about efficiency and stability of capitalist societies but disagreed on the means needed for achieving those two aims.

Just as Hayek criticises Keynes's plan for being impractical by relying on technocrats, who are assumed to be wise and impartial, Hayek's own plan can be criticised for being also impractical. As much as the success of Keynes's blueprint depends on the wisdom and the principled and impartial behaviour of technocrats, the blueprint from Hayek requires, for its success, enlightened self-interested proprietors and landowners, who are willing to act in coordination and rapidly neglecting their most immediate interests and relatively safe trade options, for the risky prospects of balancing back their returns from future trade operations. These assumptions imply the absence of slow, short-sighted, reckless and fraudulent behaviour.

These are crucial components of the moral psychology assumed in the design; the truth and reliability of such assumptions define the scope and prospects of success for any blueprint. In Keynes's case, if the assumptions are true, technocrats can be trusted on the use of wide discretionary powers and full control over fiat money. Equally, in Hayek's case if the assumptions from the opposite blueprint are true, proprietors and landowners can be trusted over the full control of commodity money and enlightened, coordinated and rapid action. Each blueprint would then compete almost exclusively on the stability of prices, low inflation and low unemployment within capitalist societies. Further aspects on the moral psychology of design are discussed in chapters three and four.

Methodologically, a blueprint that relies on an agency leading the process is called dirigiste, while a blueprint relying on a minimal or no leading agency is called libertarian. Ontologically, a blueprint relying on traditional knowledge, traditional means and custom is called evolutionary. In contrast, a blueprint relying on science, artefactual means and artefactual behaviour is called constructivist. So, a blueprint based on science and artefactual behaviour is described as constructivist and artificial, while a blueprint based on traditional knowledge and custom is called evolutionary and natural.

Hayek introduced some of these descriptive terms making a contrast between the two methods, which I have been describing as blueprint-making methods. The term 'blueprint' may seem to be exclusive of design, however I argue that whenever scientific knowledge is used, as Hayek does producing a new plan, the term can be applied to both cases. That is to say, the extension of the term 'blueprint' covers all cases when science is used for creating any new blueprint, where the new behaviour is considered to be an adaptation from custom or an artefact from construction.

The blueprint from Keynes is both dirigiste and constructivist. It is dirigiste because it relies on supranational and national monetary agencies regulating the behaviour of the participants, and leading the process towards the stability of prices, low inflation and low unemployment. It is constructivist because it intentionally attains all this by creating artefacts such the International Clearing Union, a new design for Central Banks as well as international and national fiat money. The blueprint from Hayek is libertarian because it rejects the creation of any supranational monetary agency, allowing only a minimal intervention from national monetary agencies with constrained, and almost fully defined powers. It is evolutionary because the stability of prices, low inflation and low unemployment are attained as the product of individual actions relying on traditional rules, and knowledge accumulated across generations.

The final design of the International Monetary Fund included features from both Hayek and Keynes's blueprints. It retained gold as a reserve and adopted dollar as the international currency instead of artefactual fiat money, and it did not provide any means for preventing the US Treasury from printing money, which effectively meant a discretionary power for creating reserves. Amounts of gold and national currencies were transferred from the country members to the coffers of the IMF, and it was agreed that countries would lose control over 75% of those transfers. Technocrats at the IMF would have full control over those funds to be used for lending to other countries, and they would also have the power to survey national economies as a condition for lending as well as make recommendations on domestic policies.

Impartiality was compromised because most of the IMF staff positions were filled with US economists and technocrats. Exchange rates and other national monetary policies were also subject to the approval and regulation of the new supranational bureaucracy. The new discipline in policies recommended by the IMF, and other new behavioural changes required from national monetary authorities and local politicians were more than adaptions from custom, they had to be constructed from training and maintained through regular supervision, incentives and penalties which were all decided by the IMF bureaucracy and technocrats.

Because of the Keynesian aspects of the International Monetary Fund, and because her methodological advice is addressed to state agencies in charge of designing, building and implementing policies and institutions,[156]the Mechanical view from Nancy Cartwright on blueprint-making and policy-making falls into the dirigiste methods considered as a part of social engineering. Of course, the size of the intervention from leading agencies varies from holistic central planning in socialist societies to piecemeal decentralised planning and intervention in largely capitalist societies, with different sizes in the provision of welfare institutions and state-owned enterprises. Her principles on policy-making require piecemeal

[156] See N. Cartwright (2009), N. Cartwright and J. Stegenga (2011), and N. Cartwright and J. Hardie (2012). Other examples she uses are from education and nutrition policies implemented in 1990s, namely the new policy on class-size reduction in primary schools in California, and the child nutrition programme implemented in Tamil Nadu and Bangladesh.

intervention, they demand the identification of social mechanisms, and the production of comprehensive causal models, which if they are used 'in applied science and engineering, why should we expect it to be substantially different—and substantially easier—in social engineering?'[157]

Like Cartwright, Karl Popper also advocated the use of piecemeal social engineering; and like her he also used the machine metaphor, which he described as the 'technological approach'. He points out that 'just as the main task of the physical engineering is to design machines and to remodel and service them, the tasks of the piecemeal social engineer is to design social institutions, and to reconstruct and run those already in existence.'[158] Popper was critical of holistic social engineering as socialist and national socialist government practised it, so he placed piecemeal engineering as part of liberal capitalist societies and their governments.

The views from Popper and Cartwright on social engineering actually correspond with policy reforms and institutional design as they are practiced today in capitalist societies with provision of welfare institutions. For instance, Popper explains that unlike those blueprints for holistic engineering, 'blueprints for piecemeal engineering are comparatively simple. There are blueprints for single institutions, for health and unemployment insurance, for instance, or arbitration courts, or anti-depression budgeting, or educational reform.'[159]

The association of piecemeal engineering with liberal capitalist societies and Keynesian policies, and the association of holistic engineering with socialist societies have both become a common place. It has also become a common place to associate libertarian capitalism with anti-design views grounded on evolutionary arguments similar to those held by Hayek.[160] At the same time, libertarian egalitarianism has been largely ignored in design; no holistic or piecemeal engineering, and no evolutionary argument exist on how to attain economic equality and wealth

[157] N. Cartwright and J. Stegenga (2011), p. 296.
[158] K. Popper (1961), p. 59.
[159] K. Popper (1966), p. 172.
[160] M. Oakeshott (1947) and M. Polanyi (1940) held views similar to those from F. Hayek.

with a minimal state. Such neglect and associations could be explained by the origin of the respective scientific views and the circumstances surrounding them, and also because of the values and biases leading scientific research.

In opposition to those commonly held views, I argue that the application of piecemeal and holistic engineering depend on the amount and reliability of the existing social technological knowledge irrespective of economic and political beliefs. I do this by showing how piecemeal and holistic engineering methods have been applied to economic and political programmes of different size and ideology. The ultimate product of this argument and position should be a methodology of design and engineering in the social sciences, which can be detached from their current ideological and historical biases, and can therefore be made available to all and be used for making blueprints from political and economic standpoints with low or no records of design.

Whenever modern science is used for a new policy, institution or constitution design and engineering are present, including plans for reducing the state from right and left libertarianism. In contrast, design and engineering are excluded from those societies exclusively relying on accumulated traditional knowledge with no use of social modern science. Piecemeal and holistic engineering are the two extremes within a continuous line, so there is a large variation on the size of the changes considered in any design. Whatever the size and challenges, any decision of design and engineering must rely on the amount and reliability of the existing knowledge. A small reform in the electoral system or in the public health services can fail just because social scientists do not have the required knowledge, while large economic reforms can succeed just because the required knowledge is available.

The recent capitalist engineering with successful mass privatisation performed in Russia[161]within a short period of three years, and in East Europe[162] in less than a decade, proved

[161] See Boycko, Maxim *et al.* (1995),

[162] See J. Elster, C. Offe, U.Preuss (1998); J. Zielonka (2001); and C. Bjørnskov and N. Potrafke (2011).

the success of holistic engineering. The behavioural changes in the population did not emerge from custom as Hayek would argue but from the new rules, incentives, penalties and other enforcement means embodied in the new socioeconomic machines being built, namely private firms. In factories, the officials from the communist parties were replaced with representatives from the new proprietors of firms. The application of incentives, penalties and supervision on female and male Russian workers remained, just the size, content and structure changed. More significantly for the anti-design libertarian views held by Hayek was his practical support and theoretical justification of a military dictatorship that decided upon and safeguarded capitalist holistic engineering, as it occurred in Chile. [163] This case proved not that only successful free-market holistic engineering is possible, but also that it can also take place under military control and protection.

The amount of accumulated scientific knowledge on the design, construction and operation of firms and free markets is so vast that many of the positive and negative consequences can be anticipated. All this makes the design and engineering of free markets comparatively more reliable. Similarly, the positive and negative consequences of a central planning from the state are also well known. Some of these consequences can currently be observed in Venezuela, where large-scale social engineering has been taking over the last few years, expanding the control and power from the central government.[164]

By itself, piecemeal engineering is not any more reliable than holistic engineering. Piecemeal engineering of small reforms and policies can also fail, even in cases where scientific knowledge has been made available. Nancy Cartwright demonstrates this point using cases where government education and nutrition programmes have failed. In 1996, elected politicians from the House of Representative in California decided on a new education policy for reducing

[163] See Hayek (1982), Vol. 3, pp. 124-126. Friedrich Hayek, Milton Freedman and other right-libertarian economists from the Mont Pelerin Society supported the holistic capitalist reforms in Chile, they met and advised General Augusto Pinochet and the economists in charge of the reforms, see J. G. Valdés (1995).

[164] See M. Weisbrot and J. Johnston (2012).

the class size in primary schools, the new policy was motivated by the poor performance of students in previous years. The decision on the new policy was based on experimental evidence from randomised control trials performed in primary schools in Tennessee over a period of four years, where a reduction from 25 to 13 students produced a significant improvement in reading and mathematics.

In spite of the evidence the implementation in California failed, which also represented a loss of one billion dollars that was allocated for the policy. The failure was explained by identifying differences in the demographic profiles of pupils, which likely created an unequal distribution of confounding factors contained in the different populations, while problems of implementation were also found such as the hiring of teachers with low qualifications, and a structural change in the thresholds for class reduction defined in each state.[165]

Cartwright criticises randomised control trials for their inability to discriminate confounding factors and some structural ones, which also affect implementation. By using causal models instead, she explains that her own 'solution of capacities to underwrite the inference ticket from efficacy to effectiveness solves the problem of the relevance of Tennessee to California […] The logic of capacities—when applicable—thus solves all three problems in one fell swoop. Regarding problems of confounding and implementation, it accounts for the fact that we would not normally expect the same outcome outside the experimental setting as inside.'[166] She holds similar conclusions for the failure of the Bangladesh Integrated Nutrition Project inspired by the success of the Tamil Nadu Integrated Nutrition Project both funded by the World Bank.[167]

A successful case of piecemeal engineering can be found in the design, implementation and current operation of internal markets in the National Health Service (NHS) in the United Kingdom. Initial ideas on a blueprint for implementing competition within the health care

[165] See G. W. Borhnstedt and B. M. Stecher (eds.) (2002), pp. 7-9.

[166] N. Cartwright (2009), p. 132.

[167] N. Cartwright and J. Hardie (2012), pp. 76-90.

system in The United States had been advanced by the economist Alain Enthoven in the early 1980s; internal competition for budget allocation was presented as a solution to the continuous rising costs and high inefficiency of health care publicly funded. The NHS was experiencing the same problems, which prevented bringing the least well-off districts up to the level of the best-off districts as it was originally planned in 1948 under the labour government.

Planning and administration in the NHS was centralised in London and the six other regions with only a small number of decisions made by local districts. Enthoven presented the basic components of a decentralised blueprint by transferring the decisions on budget spending, control over assets and other important aspects to the districts, he explained that 'districts are now subject to many controls that are intended to satisfy the needs of central government, the Region, the medical profession, national unions, etc. but are not focussed on efficient service at the point of delivery.'[168]

His blueprint consisted of six basic rules such as full freedom for each district authority to decide upon the allocation for resources for its own patients, and the payment for emergency and non-emergency services to patients outside each district. Decisions on wages, working conditions and firing decisions would also be made locally; consultants and general practitioners would make contracts with the district authorities, and each district could buy or sell services and assets to other districts and the private sector as well as borrow money at government interest rates.

Enthoven realised that control over budget and assets as well as freedom to buy, sell, borrow and sign contracts may be themselves be insufficient, and they can actually lead to the opposite outcomes unless they were supported by a behavioural and cultural change, so he added six prerequisites such as the provision of incentives to make cost-effective decisions in the design, appointment of suitably trained managers, good cost information available, a commitment to suppress vested interests, and the development of a new culture of buying and

[168] A. Enthoven (1985), p. 38.

selling health services, which he noted did not exist, so the behavioural changes needed would therefore be artefactual.[169]

The British economist Alan Maynard put forward similar ideas on a blueprint suggesting that general practices could become fund-holders with powers similar to those of the district health authorities.[170] Besides the changes in district authorities and general practices, the final blueprint for the NHS internal markets introduced similar changes for hospitals, which would change their status, becoming trusts. The new health trusts, district authorities and general practices would have equal decision power over the allocations of funds, buying, selling and signing contracts. A successful implementation required behavioural and cultural changes; each health trust, district, and general practices would now act as private firms; committees and managers in each of them would behave like proprietors of a firm, and competition would take place among trusts, district authorities and general practices for attracting and retaining as many patients as possible, who should now behave as consumers.

The new NHS internal markets would not be completely free markets but managed markets because those trusts, districts and general practices losing in the competition would be saved by the central government ensuring that patients would not be affected. Regular universal health care would remain for all residents and legal immigrants, and emergency services would remain available to any person. With some minor reforms and improvements, internal markets remain as the main operational principle of the NHS, both the Labour Party and the Conservative Party have supported this design. The NHS internal market is a good example of a socioeconomic machine. More precisely, it is a welfare state liberal socioeconomic machine combining aspects of a right-libertarian design with some socialist ones.

While working on their own blueprints, Enthoven and Maynard rejected a right-libertarian design for the NHS, while they have also rejected the existing central planning.

[169] *Ibid.*, pp. 39-41.

[170] A. Maynard (1986).

Maynard argued that 'as in all other markets, capitalists become the enemies of capitalism. By this we mean that general practitioners would be likely to use their market power to stabilise and maximise their income and employment. Such behaviour, although it might benefit the members of the profession, would be unlikely to lead to greater efficiency or patient satisfaction.'[171] While Enthoven argued that 'markets do not protect the weak, the disadvantaged, or the unlucky. Social protections are needed and generally present in modern democratic market economies […] The challenge is to find something in the middle that captures some of the best of both central planning and market forces.'[172]

As a socioeconomic machine, the NHS has been working reliably and more efficiently than it did under central planning.[173] One of its unintended consequences was to save the Labour Party from its own centralised policies, and just as it helped Labour it could have helped state socialism. Because internal markets are a hybrid design with socialist components without privatisation, it could have used in the ex-USSR, East Germany and East Europe by extending it to all social and economic domains where central planning was used, and it could also be used now in Cuba or North Korea following the kind of piecemeal engineering successfully performed in China and Vietnam just not for internal markets but for free markets instead, implemented in the so called 'special economic zones'.[174]

The successful engineering of the special economic zones is particularly relevant because of their right-libertarian design. There is no government intervention, no minimum wages, no taxes and virtually no other barrier; capital investors and entrepreneurs almost enjoy full freedom. Female and male workers are still under the control of the entrepreneur

[171] A. Maynard, M. Marinker and D. Pereira (1986), p. 1439.

[172] A. Enthoven (1999), pp. 11, 13.

[173] See J. Le Grand, N. Mays and J. Mulligan (1998), pp. 117-143; A. Enthoven (1999), pp. 24-32; see also A. Enthoven (2002); C. Spoor and J. Munro (2003); and M. Dusheiko *et. al.* (2007) Due to the lack of information on some years, and the control of available information by the government, the measurement and comparison of efficiency rates and other aspects like quality, equity, accountability and choice are not comprehensive, and in some cases they are not fully accurate.

[174] See D. Z. Zeng (ed.) (2010); Y. Tao and L. Zhiguo (ed.) (2012); O. Weggel (2007); and M. Than and J. L. H. Tan (ed.) (1993). China implemented the first special economic zone in 1979, while Vietnam implemented the first one in 1991.

representatives but free from the control of the communist party representatives, and they earn wages that are comparatively larger at the local level, while they may remain significantly low compared with international standards. Also, working conditions can be poor and detrimental to the workers in the special economic zones. Nonetheless, all this is consistent with a right-libertarian design.[175]

The success of the internal markets in the United Kingdom, and the success of the special economic zones in China and Vietnam show that each of these socioeconomic machines can be implemented in socialist and capitalist countries following the method of piecemeal engineering, precisely as it was suggested by Karl Popper. He argued in favour of a social engineering 'confined to a factory, or to a village, or even to a district' because 'only in this way we can learn how to fit institutions into the framework of other institutions, and how to adjust them, so that they work according to our intentions. And only in this way can we make mistakes, and learn from our mistakes, without risking repercussions of a gravity that must endanger the will to future reforms.'[176] The precautionary principle implicit in his position led him to reject holistic engineering using socialism as an example. He claimed that 'of the two methods, I hold that one is possible, while the other one simply does not exist: it is impossible.'[177]

In contrast, the cases discussed above show that large scale engineering on free decentralised markets has been successfully performed because of the vast social technological knowledge available on how to build them, and because of its comparative advantages over socialist central planning. It is not so much the size of the changes as the knowledge on how to produce them. Successful holistic engineering is feasible and therefore possible just as piecemeal engineering can be unfeasible and therefore impossible, at least until the relevant knowledge is produced. No piecemeal engineering should be performed just because the size of grave

[175] See M. Zwolinski (2007).
[176] K. Popper (1966), p. 176.
[177] K. Popper (1961), p. 69; see also K. Popper (1966), p. 175.

repercussions is small, and no holistic engineering should be rejected just because it represents a large-scale change. The amount and reliability of the social technological knowledge available should be used as criteria for deciding between piecemeal and holistic engineering. These are criteria freed from some of their ideological and historical biases, namely the association of piecemeal engineering to liberal capitalist societies and Keynesian policies, and holistic engineering from state socialism. A science of design should be made available to all and, more importantly, it should be performed on designs from political standpoints with no or few cases of design, which can offer solutions to current social problems.

2.4. Analytical sociology

The machine metaphor has proven to be methodologically fruitful. Through different sources and influences, this metaphor has spread across the social sciences mainly in the search for social mechanisms to be used in explanation and design. There are three outstanding examples of this metaphor in the social sciences, namely analytical sociology, mechanism design theory and institutional design. Analytical sociology is mainly concerned with explanation, while mechanism design theory and institutional design are concerned with design. The methodologically conscious choice for mechanisms and the close attention paid to them in analytical sociology provide the grounds for an initial comparison between these three branches, and the identification of some challenges related to the use of mechanisms such as the material nature of the mind and the scope of causes external to the mind in the design and engineering of artefactual behaviour.

Analytical sociology constitutes the best example in the social sciences of the Mechanical view advanced by Nancy Cartwright and Rom Harré. It rejects empiricist standards such as the need for laws and universal theories. It places *social mechanisms* at the centre of research in sociology and some related extensions to economics, and it explicitly adopts causal realism and realism of unobservable entities such as beliefs, intentions and desires. The use of mechanisms was inspired by the work of Jon Elster, while the realist argument for causes and unobservable entities came through the influence of Rom Harré. It adopts methodological individualism, which is also part of the analytical method followed by Cartwright.[178]

Within analytical sociology, the work from the economist Thomas Shelling[179]and that from the sociologist Raymond Boudon[180]can be considered as foundational in the conscious attempt for using individual mechanisms for the explanation of large-scale social effects. More recently, the sociologists Peter Hedström and Richard Swedberg[181] have made important advances in establishing the methodological foundations of this research programme.

The standards of empiricism prescribe the use of laws, while it criticises the realist conception of causes and mechanisms and its use in scientific explanation. The deduction or induction of the event to be explained must be supported by laws, while any reference to causes or mechanisms must be replaced with a set of observable initial and boundary conditions, which are constantly conjoined to the observable effect.[182] Hedström and Swedberg criticise the covering-law explanations for being 'black-box explanations'; this is because 'they do not attempt to reveal any mechanisms that might have generated the observed relationships'[183]. The same view on the covering-law model is held by Cartwright, who uses the term 'vending machine' instead[184]. Following Jon Elster,[185] Hedström and Swedberg claimed that social mechanisms provide real explanatory power, and therefore any true explanation must be causal.

Elster has also been critical of the covering-law model of explanation; he put forward an alternative model of explanation by mechanisms in the social sciences. He places mechanisms as 'intermediate between laws and descriptions', and defines them as 'causal patterns that are

[178] N. Cartwright (1999), pp. 83. 149-150.

[179] T. Shelling (1978).

[180] R. Boudon (1974) and R. Boudon (1981).

[181] P. Hedström and R. Swedberg (ed.) (1998), P. Hedström (2005); see also P. Demeulenaere (ed.) (2011).

[182] C. Hempel (1965), pp. 351-352.

[183] P. Hedström and R. Swedberg (ed.) (1998), p. 8; see also P. Hedström and R. Swedberg (1996), p. 287.

[184] N. Cartwright (1999), pp. 58-59, 184-186.

[185] J. Elster (1989), (1999).

triggered under generally unknown conditions or with indeterminate consequences.'[186] Elster's explanation by mechanisms holds important similarities with the explanation by causal models advanced earlier by Nancy Cartwright, so the work from both consistently underpins the research programme in analytical sociology.[187]

Individual psychological mechanisms are fundamental for any explanation in analytical sociology. These mechanisms are mainly taken from rational choice theory and cognitive psychology; some examples of them are cognitive dissonance, wishful thinking, self-fulfilling prophecy, the endowment effects and utility maximisation. Recall that Cartwright also considers some cognitive mechanisms for explanation such as motherly love, fear of punishment and desire to conform.

Analytical sociology emerged not only as a reaction to the methodological standards of empiricism but also against the holism of some sociological theories, also called 'grand theories'. Theories such as structural functionalism hold a prominent focus on *social structures* as the theoretical entities with major explanatory power, which are also framed as functional explanations. Social structures are sets of relationships established among individuals, and a set of such structures constitutes a social system; kinship, caste and social class relations are examples of those structures. In these theories, individual behaviour meets the 'exigencies' demanded for 'the production, maintenance and development of cultural systems', so psychological 'motivational mechanisms of the personality must be understood and formulated relative to the functional problems of this unit.'[188] Hedström and Swedberg criticise structural functionalism as 'empty theorising' for 'it ignores the principle of individual action'[189]. In contrast, they argue that 'a focus on explanatory mechanisms helps sociology to avoid the trap

[186] J. Elster (1999), p. 1; see D. Bailer-Jones (2009, pp. 35-41) for a detailed discussion on mechanisms; see J. Dupré (2001) for an argument against mechanisms in economics.

[187]Jon Elster (1999, p. 2) explains that the work from Nancy Cartwright on explanation by causal models in physics anticipated his work on explanation by mechanisms in the social sciences.

[188] T. Parsons (1951), pp. 21, 10, see also pp. 116-125.

[189] P. Hedström and R. Swedberg (1996), p. 299.

of mindless empiricism on the one hand, and conventional and empty theorising on the other.'[190]

The philosophers Mario Bunge and Daniel Little hold similar views; they have also produced arguments in defence of social mechanisms as the micro-foundations of theories in the social sciences. Little, for instance, holds that 'social causation depends on regularities that derive from the properties of individual agents: their intentionality, their rationality, and various features of individuals motivational psychology'; and he adds that 'causal mechanisms are more fundamental than regularities of association between causal variables.'[191] Bunge argues that 'grand theories' such as 'Parsonianism' must be avoided and considered with suspicion, while knowledge of social mechanisms should be prioritised.[192] Such mechanisms should be the main components of theories with limited scope called 'middle-range theories'. The following two examples illustrate the explanation by social mechanisms, which belong to middle-range theories.

To illustrate how such mechanisms work, Raymond Boudon quotes the explanation for the lack of support socialism received within the US American working classes during the late nineteenth and early twentieth centuries. It is a contrastive explanation originally produced by the sociologist Werner Sombart, which Boudon breaks down into thirteen premises; here it is in a condensed form of six premises:[193]

1) The US American society is a stratified society. In a stratified society, people consider upward social mobility to be something desirable.
2) Upward social mobility requires an investment from each individual with varying costs and uncertain returns.

[190] *Ibid.*, p. 299.

[191] D. Little (1991), p. 18; and D. Little (1998), p. 245.

[192] M. Bunge (1999), pp. 47, 55-63.

[193] R. Boudon (1981), pp. 20-24; Werner Sombart's explanation reflects the social and political conditions of the late nineteenth century, his book *Why is There No Socialism in The United States?* was published in 1906

3) If, on average, the costs and risks do not exceed certain individual thresholds, individual strategies of upward social mobility will be chosen. If such individual costs and risks exceed the threshold, individual strategies will be dismissed, and collective strategies of upward social mobility will instead be considered. Collective strategies also carry costs and risks.
4) The costs of individual upward social mobility are greater in societies with pronounced class differences, so collective strategies reducing individual costs and risk will be chosen.
5) A socialist programme legitimises and coordinates a choice for collective strategies of upwards mobility. A collective socialist strategy will be chosen if rival collective strategies of upward mobility have greater costs and risks with fewer returns, and if a large enough number of individuals share the same belief on such greater costs and risks.
6) Class barriers are more visible in The USA than in Europe, therefore the belief in a greater upward social mobility within capitalism is more widely shared in The USA than in Europe. The individual costs and risks of upward social mobility within US American capitalism are perceived to be less than those involved in socialist collective action.

Therefore, socialism is less appealing in The USA than in Europe.

The question regarding the lack of support of a socialist programme was puzzling and the explanation was challenging because there was evidence of the low rate of upward social mobility in The USA; poverty and unemployment were also large in American society. Therefore, an explanation would not be trivial but revealing. Boudon explains that there are two key elements in the explanatory mechanism proposed by Sombart, namely 'the weaker the visibility of social barriers, the greater the belief in the possibility of crossing them', and 'the lower the cost of a strategy, the greater the chance that it will be adopted'.[194] The belief in a greater and easier upward social mobility can actually be false, that is to say, it is enough that a large enough number of individuals believe it to be true; ' it is only necessary […] a low visibility

[194] *Ibid.*, p. 23.

of symbolic barriers between classes'.[195] This was the case in the The USA, where social mobility was actually far lower than it was perceived, and lower than in some European countries. Hence, the crux of the explanation is the false, shared belief in a greater upward mobility within the US American workers both employed and unemployed. In addition, political and economic decentralisation in The USA also hampered a wider communication and coordination within the working classes at a national level.

The second example uses a game theory model to illustrate the surprising effects small variations in individual thresholds have for collective action such as riots, strikes, voting and migration. The sociologist Mark Granovetter criticises sociological explanations of collective action based on institutionalised norms, individual preferences and motives because they are insufficient for the explanation of individual decisions with effects on collective outcomes. He argues that explanation by social norms assumes 'a simple relation between collective results and individual motives', therefore a model with a mechanism explaining 'how these individual preferences interact and aggregate' is needed.[196] He found such mechanism in the variation of individual thresholds for decision making on the participation in collective actions. In his model, Granovetter assumes a crowd with one hundred individuals randomly taken from a population with different quantitative values on the number of people needed for them to join a riot: 'a distribution of riot thresholds equals to the uniform distribution: 1% has threshold 0%, 1% of the population has threshold 1%, 1% has threshold 2% … 1% of the population has threshold 99%'.[197]

Initially only the person with a threshold of 0% will participate, and her participation will activate the person with a threshold of 1%, this action will activate the person with a threshold of 2%, and so on until the person with a threshold of 99% joins completing the set. If the distribution of thresholds changes slightly, for instance, if the person with a threshold of 3% is

[195] *Ibid.*, p. 22.

[196] M. Granovetter (1978), p. 1421.

[197] *Ibid.*, p. 1431.

replaced with a person with a threshold of 4%, the collective outcome would dramatically change. There will be no riot; the collective action will end with three people only. Other changes can be modelled with the same population showing how small changes can have dramatic collective outcomes. Hence, variation in individual thresholds is a simple mechanism with surprising explanatory power, unexpected large-scale effects and large scope because of the many cases of collective action it can explained.

Threshold models and the explanation of the low appeal for socialism within the US American working classes follow the three methodological principles suggested by Hedström and Swedberg,[198]namely the principle of direct causality by identifying individual decision making and interaction among individuals, the principle of limited scope by building models, which are part of middle-range theories, and the principle of methodological individualism by explaining collective action and its aggregate effects by individual decision-making. The first two principles ask for a finer causal description with a limited scope, which is not explicitly requested in the five principles put forward by Cartwright, so in this sense these principles are a refinement within the Mechanical view. The third principle on methodological individualism is already part of this view.

Besides these three principles, analytical sociology also postulates the existence of unobserved explanatory mechanisms. Assumptions of intentions, discounting, and preferences have proven extremely useful for the analysis of individual action even though they never can be observed.'[199] Hedström and Swedberg refer to the work from Rom Harré in support of the postulation of theoretical entities such as beliefs, intentions and desires and their causal powers, particularly when they form a psychological mechanism for individual action, which on the aggregate level constitutes a social mechanism.

[198] P. Hedström and R. Swedberg (1996), p. 298.

[199] *Ibid.*, p. 290.

Such mechanisms have *generative* causal power to be distinguished from a Humean *sucessionist* explanation of causation, Harré explains that 'in the generative theory the cause is supposed to have the power to generate the effect and it is connected to it […] the world being what it is, replete with generative mechanisms located in the many things and materials that exist in nature, not every possible outcome is equally likely.'[200] This realist view on the causal power of mental states is also shared by Nancy Cartwright, with the adscription of capacities or natures discussed above in section 2.1. The influence of Harré's realist argument extends across other sociologists within analytical sociology, for instance Mohamed Cherkaoui writes that 'realist philosophy, which is mainly British, clearly bears the stamp of Harré, notwithstanding the contributions of [Roy] Bhaskar, and [Andrew] Sayer among others.'[201]

Following the work of Jon Elster, Hedström and Swedberg describe mechanisms as the 'cogs and wheels' of social explanation; Elster writes that 'mechanisms [are] –nuts and bolts, cogs and wheels–that can be used to explain quite complex social phenomena.'[202] Semantically, the machine metaphor is part of the core of the research programme in analytical sociology. One of the strongest ontological views on this metaphor is held by Elster, who argues that 'all explanation is causal'. He accepts intentional and functional explanations, however he claims that 'at the most fundamental level, though, all explanation is causal.'[203] He follows Donald Davidson's argument,[204] which turns intentions and other unobservable mental entities into causes just as they are used in physical sciences, so that intentional explanation becomes a case of causal explanation. This claim is important because the mechanical metaphor may turn into a literal description if it is accepted that intentions and other mental entities relevant for the explanation of behaviour are physical or material entities. There are two positions on this ontological thesis.

[200] R. Harré (1972), pp. 116-117; see also R. Harré (1970), pp. 39-40, 102-111. .

[201] M. Cherkaoui (2005), p. 95; see also P. Demeulenaere (ed.) (2011), pp. 18-21, 87, 188, 268.

[202] J. Elster (1989), p. 3; see P. Hedström and R. Swedberg (1996), p. 286.

[203] J. Elster (2007), pp. 7-8, 14, 30 and 271; Daniel Little (1998, p. 208) holds similar views.

[204] D. Davidson (2001), pp. 3-19.

Elster aligns with the anomalous materialism from Davidson, while Mario Bunge argues for full materialism. Bunge argues that 'an explanation by reasons is just a particular case of causal explanation', 'the internal causes of overt behaviour are mental events such as decisions motivated in turn by intentions (which are in turn processes in the frontal lobes of primates and perhaps of other higher vertebrates as well).'[205] Either version of materialism faces at least two problems. The first one is the existence of mutually inconsistent theories of unobservable entities and processes, which is a problem already discussed in chapter one with the case of the eather and field theories of electromagnetism. In neuroscience too there are mutually inconsistent theories competing for the explanation of unobservables such as the localization of brain functions versus theories of distributed brain functions, electrical versus chemical theories of synaptic transmission, and theories of nerve cell connections by cytoplasmic continuity versus connections by surface contacts.[206] This is just the problem of *underdetermination of theories* for which the rules of minimal and maximal analogy were discussed also in chapter one.

The second problem consists of the *iteration of metaphor*. As it was shown in chapter one, central terms in the vocabulary from microphysics rely on metaphors taken from macroscopic events such as 'currents', 'force', 'field', 'repulsion', 'conductor', 'wave', and 'strangeness'. This shows the large scope of metaphor importantly illustrated by Rom Harré.[207] The ubiquity and constant iteration of metaphor in science under different semantic masks justifies and strengthens the use of mechanical and technological metaphors such as that of social mechanisms in analytical sociology. This makes the possibility of having a scientific semantics made up by literal terms and descriptions unlikely. For it will be shown that not only in analytical sociology but also in mechanism design theory and institutional design the scope, explanatory power and methodological fruitfulness of the machine metaphor remain strong.

[205] M. Bunge (1999), pp. 45-46.
[206] See M. Jacobson (1993), pp. 151-228.
[207] R. Harré (1960).

2.5. Institutional design

Institutional design is a branch within political science concerned with the design of institutions such as forms of government, electoral systems and constitutions. It relies on the machine metaphor and on comparative methods of research, and it emerged from New Institutionalism. The publication of the book *Rediscovering Institutions* in 1989 by the economist James March and the political scientist Johan Olsen marked the return to the study of institutions in political science, this new trend was described as New Institutionalism in contrast to the Old Institutionalism, which predominantly had normative and legal concerns.[208] New Institutionalism came into political science as a reaction to the methodological individualism of behaviouralism in political science and rational choice theory. The focus of New Institutionalism on the study of actual institutions, and the constant need for reforming and creating new ones, led almost naturally to the use of this knowledge for the design of political reforms as well as new institutions.

The method used in institutional design consists of comparative studies of the positive and negative effects of different sets of rules, incentives and penalties from current and past institution. The results of these comparative studies are used for assisting the choice over alternative institutional structures to be implemented in a new domain expecting the same or similar effects. The degree of resemblance between the known domain of operation and the new domain plays a crucial in the choice; it is assumed that by maximising similarity the likelihood of getting the desired effects are comparatively larger. Because this method only provides a choice from actual designs past or present, further work and information are still required in order to adapt the design to any new domain. Therefore, this method lies in the

[208] See B. Guy Peters (2012), p. 1-24. In economics, a shift towards institutions also occurred in reaction to neoclassical economics and related aspects of rational choice theory, see D. North (1990); D. North, L. Aston, and T. Eggertsson (1996); O. Williamson (2000); and E. Ostrom (2005). Unlike the new institutionalism in political science, the new institutionalism in economics has almost exclusively concentrated on producing empirical studies on actual institutions.

middle of a science of design such a mechanism design theory, and a science of facts such as analytical sociology.

Within institutional design, electoral systems have received special attention because 'they are the most manipulative instrument of politics' producing one of the largest effects on the distribution of political representation and political power.[209] Extensive empirical studies have been published on the transformation of Russia and Eastern European counties into representative democracies.[210] Comparative studies on electoral engineering have also been extensively produced covering data from Papua New Guinea, Fiji, Estonia, Mexico, Denmark, Iceland, Northern Ireland, the United Kingdom and many other countries, which have been published by the political scientists such as Benjamin Reilly, Pippa Norris and Amel Ahmed.[211] New branches such as constitutional design have also emerged with an interest in supranational constitutions such as the constitution for the European Union.[212]

The political scientist Giovanni Sartori was one of the early initiators of this method of comparative design, which he applied to constitutions and party systems. He introduced the term 'political engineering' pointing out the effects it has shaping the behaviour of politicians and society, he argues that electoral engineering should be a main target because political parties are the political channels of mass societies, that is to say, the place where 'the pace and the path of mass behaviour are set' and power distributed by 'the regulation and timing of enfranchisement, districting, and the translation of votes into seats.[213] For instance, by comparing the majoritarian electoral system and proportional representation, he argues that the double ballot system is better because it is 'highly flexible' by allowing 'for both majoritarian and

[209] G. Sartori (1997), p. ix; see also B. Reilly (2001), p.12; and A. Ahmed (2013), p. 10.
[210] See J. Elster, C. Offe, U.Preuss (1998); J. Zielonka (2001); and M. Boycko *et al* (1995).
[211] B. Reilly (2001); P. Norris (2004); and A. Ahmed (2013).
[212] J. M. Buchanan and G. Brennan (1985); J. M. Buchanan, *et. al.* (1990); S. Voigt (ed.) (2002); and S. Voigt (ed.) (2013).
[213] G. Sartori (1968), pp. 273, 276.

proportional arrangements', it 'is majoritarian where there are single-member constituencies, and proportional in multiple-member constituencies.'[214]

Sartori also compares presidential and parliamentary systems, holding a similar argument by suggesting a system with intermittent presidentialism, which can come into play when a parliamentary system is failing, so that by alternating one and the other, incentives are created for a better performance from each during his time in power. He argues that 'presidentialism and parliamentarism are single-engine mechanisms', and 'far more often than not the presidential engine falters in its downward parliamentary crossings, while the parliamentary engine impairs, in its upward ascent, the governing function.'[215] The machine metaphor description can be fully appreciated, where each form of government becomes a machine, an engine producing reliable outcomes each with different effects, and such effects are produced by shaping individual and collective behaviour with the right set of rules, incentives and penalties. This is consistent with Cartwright's own metaphor of socioeconomic machines, which regularly and reliably produced certain outcomes.

Sartori further explains that by "putting the metaphor and an etymology together I come up with 'constitutional engineering'"; constitutions, electoral systems and other political institutions "are like (somewhat like) 'engines', i.e. mechanisms that must 'work' and that must have an output of sorts", they are 'unlikely to work as intended unless they employ the engines of Bentham, i.e., punishments and rewards.'[216] Jeremy Bentham wrote on the distribution and effects of punishments and rewards for the enforcement of laws, he argued that 'the greatest happiness of the greatest number ought to be the object of every legislator: for accomplishing his purposes respecting this object, he possesses two instruments—Punishment and Reward [...] the springs of that mechanism developed, whence those laws arise to which the power is

[214] G. Sartori (1997), p. 11.
[215] *Ibid.*, p. 153, see also p. 159.
[216] *Ibid.*, p.ix

attributed of executing themselves'[217] The 'engines of Bentham' are a fundamental part of comparative institutional design and mechanism design theory; I argue that they are a mechanical metaphor of *operant conditioning*, which is a scientifically updated version of those engines.

The psychologist B. F. Skinner defines *operant behaviour* as the behaviour conditioned 'upon the posterior reinforcing stimulus.'[218] It is brought about using a prompting device and a reward after the performance; both the device and the reward are designed and decided by the experimenter. Because 'reward suggests compensation *for* behaving in a given way, often in some sort of contractual arrangement', behaviourists use the term 'reinforcer' instead, which 'designates simply the strengthening of a response'.[219]

The experiments performed by Skinner with rats and pigeons were highly successful, not only shaping behaviour but also making important discoveries.[220] He worked on the extension of his findings and method to human behaviour, particularly on social matters such as education, industrial relations and politics. With behaviourism, the social world 'may be regarded as an extraordinarily complex *set* of positive and negative reinforcing contingencies'[221], and the aim is to increase the control over those contingencies using operant conditioning. On a large scale, operant conditioning should ultimately lead to a comprehensive technology of behaviour to be applied to cultural design. Skinner argued that 'a program of cultural design in the broadest sense is now within reach'[222], an 'industrialist may design a wage system that maximizes his profits, or works for the good of his employees […] A party in power may act

[217] J. Bentham (1833), p. 192.

[218] B. F. Skinner (1938), p. 22; in contrast, *respondent behaviour* 'is the result of something previously done to the organism […] the stimuli precedes the occurrence of the response', p. 22.

[219] B. F. Skinner (1963), p. 505.

[220] See B. Skinner (1948) and B. F. Skinner and C. B. Ferster (1957), in these works he presents his discoveries on superstitious behaviour in pigeons, and the effects created on animal behaviour when reinforcement is strategically scheduled.

[221] B.F. Skinner (1958), p. 57.

[222] *Ibid.*, p. 99.

primarily to keep its power, or to reinforce those it governs (who in return keep it in power), or to promote the state, as by instituting a programme of austerity.'[223]

He further argued that 'what we need is a technology of behaviour. We could solve our problems quickly enough if we could adjust the growth of the world's population as precisely as we adjust the course of a spaceship, or improve agriculture and industry with some of the confidence with which we accelerate high-energy particles, or move towards a peaceful world.'[224] This is consistent with the way many policies and institutions are currently designed, implemented and built by many democratic governments from the left, right and centre in politics. Furthermore, the evolutionary role of operant conditioning described by Skinner is consistent with evolutionary game theory; they both shared common grounds with evolutionary theory.[225]

Ontologically and methodologically, operant conditioning has been criticised for neglecting the mind and individual agency by reducing human behaviour to a mere response prompted by a specific stimuli. This was put into contrast with cognitive psychology, which has gained a solid and well-justified consensus exposing the active role of the mind. Skinner replied to his critics explaining that 'I should not want to try to prove that there are no innate rules of grammar or internal problem-solving strategies or inner record-keeping processes';[226] he rejected the existence of a human mind and accepted only the existence of the brain and the associated genetics, whose constitution and causal power generating human behaviour 'eventually neurology will tell us all we need to know'.[227]

Rom Harré criticises Skinner because he 'embraces more tightly than anyone the other two ideas that make up the basis of modern psychology, the mechanistic model of human

[223] B.F. Skinner (1971), p. 151.

[224] *Ibid.*, p. 5.

[225] See B.F. Skinner (1981).

[226] See C. Catania (ed.) (1988), p. 364.

[227] *Ibid.*, p. 301,

action, and the Humean conception of cause.'[228] He argues that this mechanistic model neglects human agency by turning any person into an automaton, and he proposes instead a metaphysics of natures and powers as the source of human agency. In contrast, Jon Elster argues that individual psychological mechanisms can be explained by consequences and natural selection just like B. F. Skinner also argues.[229] Elster explains that the consequences of recurrent behaviour 'can enter into the causes that make its occurrence on a later occasion more likely. There are two main ways in which this can happen: by *reinforcement* and by *selection*.'[230] He is concerned with how incentives and penalties shape behaviour causing the selection of some types of behaviour and the extinction of others.

Notwithstanding the continuing debate on the nature of the mind and the brain, and the scope of their causal powers and those from the environment, the causal power of stimuli supplied by the environment is widely accepted; human behaviour can largely be conditioned by the particular constitution of the environment and the consequences from past behaviour. The existing consensus on this claim constitutes the common ground the claim I make by holding that operant conditioning is a modern and scientific version of the engines of Bentham. In other words, it provides the grounds for the discussion on the use of incentives and penalties in design in the social sciences, whose precise scope may remain indeterminate but it certainly is not small or negligible.

Politically, the ideas of Skinner on cultural design and behavioural technology have been criticised as carrying potential support for full dirigisme and full social control. Nonetheless, Skinner explicitly constrained his views on social design to representative democracies with a welfare state and free markets,[231]which is consistent with the way institutions and policies are currently designed, engineered and implemented in countries of this type. Indeed, the basics of operant conditioning such as positive and negative reinforcement, as well as the strategic

[228] R. Harré and P. F. Secord (1972), p. 34; see also R. Harré (1999).
[229] B.F. Skinner (1951).
[230] J. Elster (2007), p. 271.
[231] B.F. Skinner (1971), p. 169.

distribution of them using time schedules and changes in quantity; are currently used with some modifications and with different names in institutional design and mechanism design theory.

This occurs because in representative democracies, individual voters give their representatives the power to decide for them using the scientific and legal means necessary to ensure the accomplishment of specific established objectives and goals. A similar analysis and conclusion apply to the relationships female and male workers hold with their trade union representatives, and the representatives from the proprietors of firms and landowners. The state, the firm and the farm are social machines, whose efficient functioning depends on the quality of the design, which commonly relies on the engines of Bentham, that is to say, on some form of operant conditioning. Such design is consistent with the piecemeal social engineering advocated by Karl Popper, and the blueprint-making and policy-making methods advanced by Nancy Cartwright.

In contrast, a right-libertarian design offers an alternative where dirigisme and control from the state is reduced, while it preserves economic inequality. A property-owing democratic design offers a substantive reduction in both economic inequality and dirigisme and control from the state. A left-libertarian design is more comprehensive for it offers a more substantive reduction in the control and dirigisme not only from the state but also from the firm and the farm, while it fosters economic equality.[232] The expansion of freedom and equality rely on the knowledge of reliable methods of design detached from their ideological and historical biases.

[232] From a right-libertarian position, Friedrich Hayek criticised both the dirigisme of J. M. Keynes's policies and the constructivism of the cultural design programme advocated by B. F. Skinner, see F. Hayek (1943), and F. Hayek (1978, p.6). From a left-libertarian position, Noam Chomsky agreed with cultural design but criticised the neglect and disempowerment of human agency and individual autonomy implied in Skinner's explanation of human behaviour, see N. Chomsky (1971); for recent views on left-libertarianism see P. Vallentyne and H. Steiner (ed.) (2000). John Rawls (1999, pp.xiv-xv, 242-251; 2001, pp.135-140) argues for a property-owing democracy, which offers less economic inequality and more political freedom than capitalism with a welfare-state and state socialism.

2.6. Mechanism design theory

Whereas in analytical sociology, mechanisms are studied as they are found already existing and operating in society, in mechanism design theory, they are designed and built. Mechanism design theory is a branch in game theory, which emerged from the debate over efficiency and problems of resource allocation between socialism and capitalism, that is, between centralised economic planning and decentralised free markets. One of its main founders, the economist Leonid Hurwicz explains that 'we can think of an economic system as defined by a set of institutional or behavioural rules that enable us to distinguish, for example, capitalism from socialism, pure laissez-faire from mixed economy, or perhaps perfect competition from oligopoly […] The totality of these behaviour patterns (as distinguished from environment and state descriptions) may be called the economy *mechanism*.'[233] This metaphor presents the whole economy as a grand mechanism, as a big social machine with different parts and small mechanisms assembled to produce a specific outcome.

Social machines vary in size and aims, an indefinitely large amount of mechanisms constitute a national economy, while firms and farms require a smaller number of mechanisms to function as machines. The need to solve economic and other social problems creates a constant demand from new designs including large designs such as new rules for international trade in the European Union or the NAFTA in North America, and medium and small designs such as the NHS internal markets in the United Kingdom, the new multi-round ascending auction designed for allocation of licences to telecommunication firms in The USA and Europe, and the repudiation-proof contract devised by Oliver Hart and John Moore.

The design of mechanisms for the allocation of resources within any given society to those who can make the best use of them has been one of the main challenges in economics. Private competition in the market and central planning are two well-known mechanisms for resource allocation. Hurwicz observed that private market competition as devised and used in

[233] L. Hurwickz (1972), p. 425.

all major economic theories actually creates incentives for cheating on preferences and prices, which severely undermines the possibility of reaching an efficient allocation with equilibrium closest to the optimal. The further equilibrium is from the optimal, the grater social losses are. He explains that this problem is due to 'a fundamental conflict among such mechanism attributes, [namely] the optimality of equilibria, incentive-compatibility of the rules, and the requirements of informational decentralization.'[234] In other words, parts of the design are inconsistent with the incentives it creates.

This was an important discovery in economic design, which is called the 'incentive-compatibility problem'. It gave rise to a design principle now widely followed in economics, Hurwicz explains that the 'concept of incentive-compatibility merely required that no one should find it profitable to "cheat", where cheating is defined as behaviour that can be made to look "legal".'[235] The problem is not exclusive of free markets but it also extends to central planning.

Targets and norms are set in all economies. In capitalism and socialism, 'there is a "superior" and a "subordinate", and the latter has an incentive to depress the norms when the penalty for failure to reach a target is severe.'[236] This situation applies to any two individuals involved in a market transaction or in the allocation of public goods, 'participants would "cheat" without openly violating the rules. A participant could try to "cheat" by doing what the rules would have required him to do had his characteristics been different from what they are, i.e., he could "pretend" to be poorer than he is, or less efficient, or less eager for certain goods. (It is important to understand that he would not be doing this directly by uttering false statements, but indirectly by behaving inappropriately according to the rules for his true characteristics.)'[237] After identifying the problem, Hurwicz briefly explores some of the possible solutions such as the creation of teams, which eliminates competition and creates incentives for

[234] L. Hurwicz (1973), p. 24.
[235] L. Hurwicz (1972), p. 445.
[236] L. Hurwicz (1973), p. 24.
[237] L. Hurwicz (1973), p. 23.

truthful revelation of preferences; or an increase in the share of the total output given to the workers, who then would have an incentive for work, maximising the utility of the firm or the farm.

Decentralisation, revelation of true preferences and information efficiency are some of the main areas of interest in mechanism design theory. For instance, important mechanisms for the revelation of true preferences have been designed, which can be applied to important problems such as the demand for public goods. This is the case of the sealed-bid auction designed by the economist William Vickrey.

One of the causes of market failure is imperfect competition, which can lead to undersupply and oversupply of commodities in the market. Imperfect competition may occur when buyers or sellers are too few in number to ignore the effects of their actions on the market price. It can also happen when buyers or sellers are too many, too naive or too isolated from each other to engage in any overtly or tacitly concerted action. Vickrey considers how a government agency called 'marketing agency' could intervene in the market, so that competitive equilibrium prices are attained. One of the possible solutions he contemplates is a monopolistic marketing agency to which all sales of the commodity must be made, and from which all supplies must be bought.

In principle, this could allow the agency to determine the competitive equilibrium price, however, Vickrey observes that the agency would need information coming from 'reports and actions of buyers and sellers, who would have an incentive to understate prospective demands and supplies, or to curtail their actual sales and purchases in the hope of inducing the marketing agency to change the price in their favour.'[238] Besides being expensive, this solution could lead to large inefficient outcomes because the revelation of true preferences is clearly compromised. Note that this situation is the same for any state institution holding a monopoly, such as the

[238] W. Vickrey (1961), p. 9.

NHS in The United Kingdom before the implementation of the internal markets as it was discussed above in section 2.3.

Another solution considered by Vickrey is a Dutch auction, performed by the marketing agency, where the auctioneer announces prices in descending sequence; in this case the auction ends with the first and only bid. In spite of being fast and therefore inexpensive, this auction creates prices with uncertain values with respect to an efficient equilibrium, which are likely to lead to an inefficient allocation. This occurs because as soon as the price comes down to the full value of the commodity given by the most eager bidder, the price paid implies a zero gain for him, that is to say, 'as the announced price is progressively lowered, the possibility of a gain emerges, but as the gain thus sought increases with the lowering of the point at which a bid is to be made, the probability of securing this gain diminishes.'[239]

The final design suggested by Vickrey consisted of a multiple auction by sealed bids, where multiple identical commodities are put on sale and each bidder submits a bid in a sealed envelope. In sealed bids the usual practice is to accept a certain number of bids starting from those offering the highest price, where the effective price is that one established in each individual bid. An alternative method pointed out by Vickey consists of setting the effective price at the level of the last bid accepted, which allows all successful bidders to benefit from a uniform price. This prevents discrimination in the final price available to all bidders. Vickrey introduces a final variation of this method by making the uniform price to be charged to the successful bidders equal to the first bid rejected rather than the last bid accepted, he explains that 'only in this way is it possible to insure that each bidder will be motivated to put in a bid at the full value of the article to himself, thus assuring an optimum allocation of resources […] avoiding any incentive for wasteful individual expenditure on general market research.'[240]

[239] *Ibid.*, p. 15.
[240] *Ibid.*, p. 26.

Comparatively, the design with a multiple auction by sealed bids offers the most efficient method for attaining competitive equilibrium prices of commodities put on sale by a government agency. The analysis and design produced by Vickrey was highly praised by other economists such as Edward H. Clarke and Theodore Groves,[241]who added some refinements by introducing a two-part tariff. Roger Myerson and Eric Maskin[242] also developed further auction mechanisms, general bargaining problems and bilateral trade; and they have also expanded design to problems of environmental economics. In 1990s the design of the multiple auctions attained extraordinary success with the design of the new simultaneous ascending auction used for the allocation of exploitation rights of the wave space in The USA, and in Europe a few years later, raising staggeringly large revenues for the government. The design of this auction is discussed in chapter three.

Methodologically, the same virtues and shortcomings found in the repudiation-proof contract designed by Oliver Hart and John Moore discussed above in section 2.1. also apply to the multiple auction by sealed bids designed by Vikrey, and in general to all designs from mechanism design theory. Vickrey's design tells us (i) the parts of the machine and the capacities of those parts, namely the bidders with specific psychological capacities such as self-interest and greed and an impartial marketing agency with reliable knowledge on equilibrium competitive prices. It also tells us (ii) how the parts are to be assembled by establishing the basic rules for bidding, and the procedures to be followed by the marketing agency; (iii) the rules for calculating the outcome, that is, equilibrium competitive prices are calculated using equilibrium theory with relevant information on each bidder. In the design, (iv) some information is available on shielding, for instance, on how to prevent collusion among bidders, side payment and communication or signalling.

[241] E. H. Clarke (1971); and T. Groves (1973).

[242] R. Myerson (1981) and (1979); R. Myerson and M. Satterthwaite (1983); E. Maskin and J. J. Laffont (1979); and E. Maskin and S. Baliga (2003).

Finally, (v) no information is made available on how to get the multiple auction by sealed bids running. As it was pointed out earlier, such information and knowledge is provided by experimental economists; who have developed the skills of social engineers. The knowledge of design from mechanism design theorists necessarily requires the knowledge of engineering experimental economists have. Together both kinds of knowledge constitute the technological knowledge available from the science of economics. Decisions on feasible and unfeasible social machines are necessarily subject to the advancement of social technological knowledge.

—O—

Chapter 3

The FCC Auction Machine

5.0. Introduction

The FCC auction was a new kind of auction used for the allocation of licences for the use and exploitation of the electromagnetic spectrum in The United States. This auction set a methodological standard of design and engineering in economics; its design adopted some properties from the traditional English and Dutch auctions and also add new innovative properties, such as multiple rounds where bidders can return unwanted items. Unlike the English and the Dutch auctions, the FCC auction was designed and built by social scientists. The large revenue it raised was hailed as a proof of success of mechanism design theory. This success led some European governments to hire mechanism designers for the design and implementation of similar auctions for the allocation of licences on the electromagnetic spectrum.

The success was not only due to the knowledge available from mechanism design theory but also from the practical knowledge experimental economists have, they performed the experiments testing the rules and mechanisms, which produced data crucial for the design and the implementation of the new auction. In this chapter, I present a methodological account of the FCC auction design discussing two main components of it, namely the blueprint produced by mechanism designers and the experiments performed for producing the data missing in the blueprint. I also evaluate this blueprint using the types of design and principles discussed in chapter two, and minimal analogy and type-hierarchies from chapter one.

I characterise the method used by experimental economists as the *experimental parameter variation*, which I take from aeronautical engineering. The introduction of the method of experimental parameter variation allows philosophers to pay attention to practical knowledge, or knowledge of practices, as opposed to propositional knowledge. Practical knowledge has been largely ignored in epistemology and in the philosophy of science. Science is not only the knowledge of theories, laws and inferences; there is a vast array of practices, some of them highly successful and sophisticated. Engineering and experimental methods have been mostly developed in the natural sciences, where they have been growing in size and sophistication. In economics and other social sciences these methods have been developed only recently, and there seems to be an increasing demand for more experimental and engineering knowledge in these sciences.

In section 5.1., I describe and discuss the FCC auction blueprint, which is a multiple-round simultaneous ascending auction. This blueprint was produced by three mechanism design theorists, Paul Milgrom, Robert Wilson and Preston McAfee. Using the types of design and principles discussed in chapter two, I characterise this blueprint as partly dirigiste and oligopolistic, and explain why on four of the five principles of design advanced by Nancy Cartwright, this blueprint falls below the standard by leaving some gaps in the design. Using the rules on minimal and maximal analogy and type-hierarchies discussed in chapter one, I argue that this blueprint is a case of minimal analogy, and therefore it is a progressive design within the type-hierarchy of auctions.

In section 5.2., I introduce and describe the method of experimental parameter variation from aeronautical engineering. I rely on the work from Walter Vincenti, who illustrates this method using the experimental work the mechanical engineers William F. Durand and Everett P. Lesley did in the 1920s, when they tested a large number of new air propeller prototypes using a wind tunnel. The data obtained were crucial for the manufacturing of propellers ready to be assembled in a new model of aircrafts superior to those available at the time.

In section 5.3., I show how the method of experimental parameter variation can be extended to experimental economics, and in particular to the experiments performed by Charles Plott and his team searching for data crucial for the successful implementation of the FCC auction. The experimental work done by Plott and the data obtained filled the gap left in the blueprint submitted by Paul Milgrom, Robert Wilson and Preston McAfee.

5.1. The FCC blueprint

Multiple-round simultaneous auctions are a new kind of auction designed and implemented by the mechanism design theorists and experimental economists. The creation of this new kind of auction came as a product of a call made by the US Federal Communications Commission (FCC) in 1993 for a new more efficient mechanism to be used for the allocation of licences to telecommunication firms for the use and exploitation of the electromagnetic spectrum.

A multiple-round simultaneous auction is a social machine consisting of three main mechanisms, namely a simultaneous market, ascending biding and multiple rounds. In a multiple-round simultaneous auction, several markets are open at the same time, so any bidder can place any number of bids in different markets. The markets run in rounds and remain open until the bidders have accomplished the best purchase by selling back some items and buying new ones. These properties of the auction allow a highly efficient allocation of licences and the maximisation of revenue for the auctioneer, which in this case was a government institution. The design of this new auction relied on the pioneering work of William Vickrey discussed in chapter two, section 2.6. Vickrey designed an auction of multiple items with a sealed bid where the auctioneer is also a government agency just like the case of the FCC where multiples licences are auctioned. With this design, Vickrey was trying to solve the problem of imperfect competition in free markets, which can lead to undersupply and oversupply of commodities. An auction of multiple items with a sealed bid provides the blueprint of a social machine, whose mechanisms could attain competitive equilibrium prices of commodities.

The design and successful implementation of the first multiple-round simultaneous was hailed as an outstanding achievement almost exclusively due to game theory, which clouded the important and distinctive engineering work done by experimental economics. The philosopher Francesco Guala made a significant advancement showing the crucial contribution made by experimental economists; he presents the case mainly as a problem of logic, where inferences made in the laboratory have to be extended to the outside world.[243] Unlike Guala, I present the case as a methodological problem concerned with design and blueprint-making methods. In particular, I argue that the method of experimental parameter variation was used by experimental economists in order to produce data essential for the design and implementation of the FCC auction.

As part of the decentralising trend of public assets and services in 1980s, the US Congress decided to look for a new and more efficient mechanism for the allocation of licences for the use and exploitation of the airwave space, which would lead to the provision of mobile communication with cellular telephones and radio systems, and the transmission of data with fax machines. Until 1982 these licences were allocated using an administrative hearing process known as the 'beauty contest', in which each applicant had to persuade the FCC of the benefits of adjudicating a licence to them. This allocation procedure was slow, opaque and highly bureaucratic. A first attempt at replacing the beauty contest was made by introducing a lottery where licences were randomly allocated to the applicants. This new mechanism was fast, transparent and simpler; however it created strong inefficiencies by allocating licences to applicants who have no real interest in exploiting the licence. This created a secondary market where licences were sold and resold creating large profits for private individuals, and a loss in revenue for the government.

The US Congress was aware of the disastrous experience in New Zealand and Australia in the early 1990s, where licences were allocated using first-price and second-price sealed-bid auctions. These auctions were chosen without asking for scientific advice; they produced large losses in the government's revenue, and they also prompted strong criticism from the public

[243] F. Guala (2005), pp. 178-181, 194-199.

and rival political parties.[244] The US government looked for scientific advice issuing in 1993 a 'Notice of Proposed Rule Making', where the FCC advanced an initial design of an auction in two stages, expecting replies and comments mainly from economists and game theorists. In order to prevent an oligopolistic distribution and promote economic equality, the original policy set by the Congress considered a distribution of licences to minority-owned and women-owned companies, small businesses, and rural telephone companies. However, the final design excluded these groups by allocating the licences to those bidders holding the highest bids, which led to an oligopolistic distribution with an increase of inequality.

Game theorists model auctions as non-cooperative games played by self-interested utility-maximising bidders. This game assumes a solution under Nash equilibrium, namely that given everyone's moves, no player can be better off than she currently is by shifting to a different strategy. There were two important problems mechanism design theorists faced in designing the new FCC auction. The first one was related to the complementary character of licences in contiguous regions of the spectrum. The second one was related to the existence of perfect substitutes in different portions of spectrum. Given these two properties, the value of any package of licences would vary according to number and combination of contiguous and non-contiguous portions of the spectrum. Moreover, a number of further conditions such as affordable technologies and operation costs had to be considered in the design. These further conditions added to the perfect substitution and complementary values produced an excessively large number of packages with almost each of them having a different value.

Generally, auction models assume a common value of the items, that is to say, the value of the auctioned item is assumed to be the same for every bidder but unknown to all. The design of auctions where items have different values for different bidders was in an early stage. The economic theory available at the time did not provide the means for estimating the different outcomes of an auction where the items have different values. Some insights pointed

[244] See J. McMillan (1994).

to the highly problematic nature of items with complementary properties, whose unstable value produces different Nash equilibria with no clear indication as to which of them is optimal. Therefore, the design of the FCC auction represented an important challenge due to the lack of data on important aspects which no theory could provide. The situation is the same to that of the design of the new air propellers to be discussed in the next section, where data which the blade element theory could not supply were lacking.

The FCC hired the economist John McMillan, who suggested an auction in two stages. In the first stage, the licences would be auctioned in packages using a sealed bid, and in the second stage only individual licences would be auctioned. This mechanism seemed to solve the complementarity problem since those bidders who value packages over individual licences would place high bids in order to get more than one licence. In the second stage, bidders with a preference for individual licences would equally place high bids. In both cases, an auction with two stages seemed to be efficient by allocating licences to bidders who could maximise their use and exploitation based on their willingness to pay more for them. This design was supported by the National Telecommunications and Information Administration (NTIA), a public institution advising the government and the FCC, which had also suggested package-bidding after getting the advice from the economist John Ledyard, who had worked on the design of combinatorial auctions.[245] Unlike the beauty contest and the lottery, this design was scientifically supported. Because this design was fully controlled by FCC and the NTIA, and because these two government agencies decide the combination of licences in each package, the design is dirigiste, that is to say, it contains some properties of central planning discussed in chapter two, section 2.3.

Some telecommunication firms were critical of package-bidding as it was not competitive enough because for it prevented some bidders from purchasing some licences, which created an unfair advantage for those who may be allocated with a large part of the

[245] See J. Ledyard et. al. (1997).

spectrum; they thought that an open bid could provide equal bidding opportunities to all. Telecommunication firms realised that a bad design could actually affect their own interests by creating unfair and inefficient allocation, and so they decided to hire their own scientific advisors. The economists Paul Milgrom, Robert Wilson and Charles Plott were hired by Pacific Bell; Jeremy Bulow and Barry Nalebuff by Bell Atlantic; Preston McAfee by Airtouch Communications, Robert Weber by Telephone and Data Systems; Mark Isaac by the Cellular Telecommunications Industry Association; Robert Harris and Michael Katz by Nynex, Daniel Vincent by American Personal Communications, Peter Cramton by MCI; and John Ledyard and David Porter by the National Telecommunications and Information Administration.[246]

Paul Milgrom and Robert Wilson put forward a new design which they called 'simultaneous ascending-bid auction'. Separately, Preston McAfee put forward a similar design. A simultaneous open auction constituted the answer to the concerns voiced by private firms on package-bidding with a sealed bid, and it also represented an improvement on the two stages considered in the FCC initial design.

In a simultaneous open auction several markets are open at the same time and bidders can participate in all of them at once. This was a true innovation in auction design. Unlike a sealed bid, an open simultaneous auction allows each bidder to monitor the behaviour of other bidders. This valuable information enables her to assess her chances of buying the combination of items she prefers. During the auction, bidders can move freely from one combination to another by selling back to the market those items over which their preference has changed, until they accomplish a combination with the highest value. Another important advantage of this new design over a sealed bid is that it helps prevent the winner's curse, that is to say, the possibility of overbidding. This can be prevented because bidders can monitor the pricing behaviour of others.

[246] See J. McMillan (1994); F. Guala (2005), pp. 167-168.

Besides the open character of the new auction, simultaneous bidding on several markets all opened at the same time was also another important innovation. In the traditional English ascending auction, items are auctioned one by one starting with a low price, and bidders continue making offers until the market is closed, which usually occurs when no new offer is put forward. Therefore, the possibility of getting a combination of items is not directly made available. This could only occur if a second market is open where items are resold but not all items may be there, and prices would also increase because of the costs and time involved in opening a second market.

In the traditional Dutch descending auction time is fixed and items are sold in packages starting with a high price, which prevents other bidders from purchasing individual items they have a strong preference for. Again, a secondary resale market could be open but the same problems of time and cost rising would appear. Therefore, a direct sale in one single market represents a more efficient design. Because in the Dutch auction prices start high and time is limited, demand may be prematurely terminated affecting prices and efficiency in the allocation of items. A simultaneous ascending auction prevents this situation by allowing more time holding a long round until no new bid is put forward. It also prevents a resale in expensive secondary market by providing different rounds, where bidders can sell back to the market any number of items as well as buy new ones until they are satisfied with a package.

The final blueprint was prepared and submitted by Paul Milgrom, Robert Wilson and Preston McAfee. It contained the descriptions of the three new mechanisms, namely a simultaneous market, ascending biding and multiple rounds. This blueprint can be evaluated using the types of design discussed in chapter two, namely libertarian and dirigiste, and the five principles of design and engineering advanced by Nancy Cartwright. Also, a further evaluation can be made using the distinction between minimal and maximal analogy, and by constructing a type-hierarchy as it was done in chapter one with the magnetic force models from James Maxwell and William Thompson.

Because the electromagnetic spectrum is controlled and fully regulated by the state through the FCC and the NTIA,[247] and because these two agencies still controlled part of the design, this blueprint retained some aspects of central planning. The blueprint is oligopolistic because by allocating the licences to those holding the highest bids, it excludes minority-owned, women-owned companies, small businesses and rural telephone companies, so such a design fosters the domination of the market by a small number of firms.

The contrast between the traditional English and Dutch auctions and the new FCC auction with multiple-rounds, simultaneous markets and ascending bidding provides a further case and illustration of the distinction between traditional and artefactual institutions discussed in chapter two, section 2.3. There, the contrast was made between the International Gold Standard and the International Monetary Fund. Like the Gold Standard and other cases of commodity money, the Dutch and the English auctions were also created without the help from scientists, that is, without using mechanism design theory and neoclassical economics. In contrast, the FCC auction is the product of scientific design, it is a social machine made up of three main mechanisms assembled to create a whole new machine. Friedrich Hayek argued against the creation of an international monetary institution endowed with the power to dictate national economic policies and produce fiat money, as it had been suggested in the blueprint put forward by John Maynard Keynes. This was only a case of a general argument Hayek made against design and engineering—which he described as 'constructivist'—and against dirigisme, that is, against central planning and control.

The first design of the auction in two stages where the FCC and the NTIA decided on the combination of licences in each package was a case of dirigisme with central planning. Such dirigisme was prevented by the action from telecommunications firms who hired scientists to

[247] The design of FCC auction was made under the USA Communications Act of 1934, which defined the electromagnetic spectrum as publicly-owned resource and prohibited any private ownership of it; those granted with a licence were defines as 'public trustees'. The law rapidly changed in 1996 after the first FCC auctions were run extending the rights of the licence holders, who could now hold the licence almost permanently; see K. Corbett (1996) for details.

produce designs where their own interest were fostered and protected. Therefore, the final blueprint became partly libertarian by giving those firms the power to decide how to form their own licence packages. A full right-libertarian blueprint would have considered giving private firms the control and ownership of the electromagnetic spectrum instead of just giving them a licence. This would have led to the extinction of the FCC and the NTIA or the reduction of them to agencies supervising the quality standards of the telecommunication services. In contrast, a blueprint which includes licences for minority-owned, women-owned companies, small businesses, and rural telephone companies as it was originally planned would have been at least partly egalitarian, although still dirigiste.

A sharper contrast can be made with the blueprints from left-libertarianism and a property-owing democracy, where direct widespread ownership of the electromagnetic spectrum among the unemployed, low-income families and other worst-off groups could be considered. In this case, without having to wait for the distribution of the revenue raised by the FCC auction and taxes through welfare institutions under the blueprint submitted by Milgrom, Wilson and McAfee. Additionally, the size of the welfare state would be reduced and also the power and size of central government, which in this case is represented by the FCC and the NTIA. The contrast with left-libertarianism and a property-owing democracy can only be generic because blueprints from these positions are virtually inexistent.[248] Mechanism design theory and experimental economics are dominated by neoclassical economics and welfare economics. This is why in chapter two, I argue for a methodology of design and engineering in the social sciences, which can be detached from their current ideological and historical biases, and can therefore be made available to other positions; particularly those where design and engineering are poor or inexistent.

[248] For recent views on left-libertarianism see P. Vallentyne and H. Steiner (ed.) (2000). John Rawls (1999, pp. xiv-xv, 242-251; 2001, pp.135-140) argues for a property-owing democracy.

A second evaluation can be made by using the rules on minimal and maximal analogy and type-hierarchies discussed in chapter one, sections 1.6. and 1.7. The magnetic force models from James Maxwell and William were presented as examples of maximal and minimal analogies. The model from Thompson was more progressive because by describing the magnetic force as a field it minimised the analogy with the mechanical Newtonian paradigm, while the model from Maxwell maximised such an analogy. This analogy was further appreciated by building a type-hierarchy. In a similar way, minimal and maximal analogies can be applied to blueprints also building a type-hierarchy.

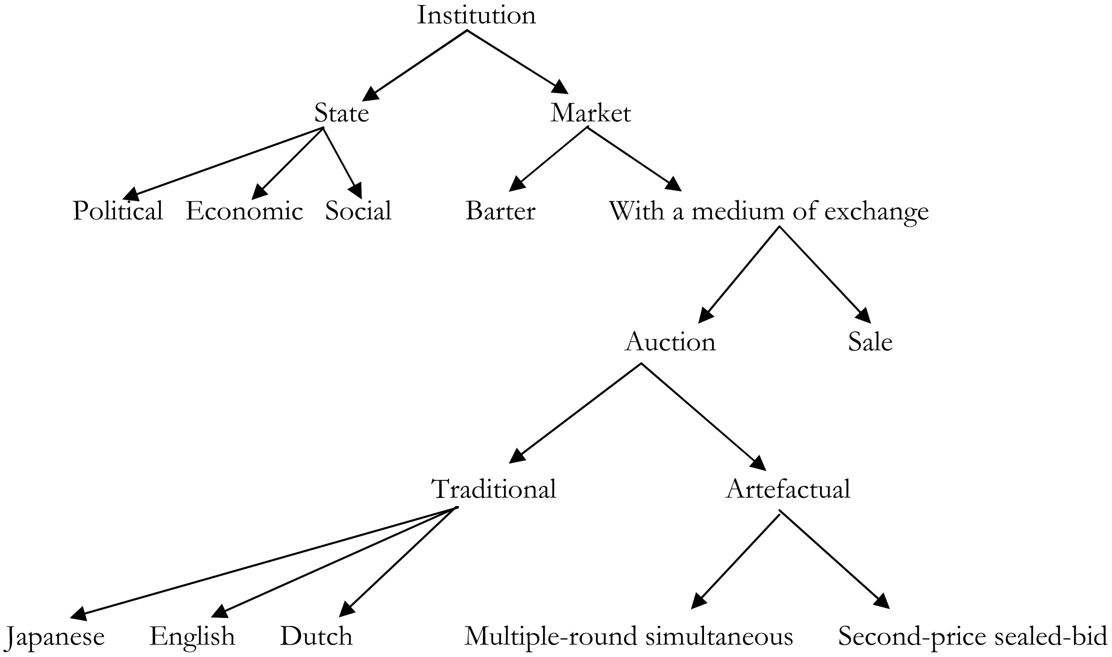
Eileen Way defines a type as 'a set of individuals each of which has certain properties which are numerically identical with those in other sets of higher type'. Because types have a nominal status, the relationship they hold with their tokens cannot be that of 'qualitative identity', which only holds 'between the relevant concrete properties of each particular'[249]; numerical identity does the job of establishing the relationship needed between tokens and the types. For instance, a 'gold coin' is a token whose properties are numerically related to those contained in the 'commodity money', which is a nominal representation. Types are ordered according to their level of generality forming a pyramid or a three-like classification. Because commodity money is a traditional kind of money distinct from fiat money, is it necessary to distinguish between traditional social kinds and artefactual social kinds. Traditional kinds rely on custom and knowledge accumulated across different generations without the intervention of science, while artefactual kinds are a product of science, design and engineering. The same distinction can be applied in the natural sciences, for instance in chemistry where natural and synthetic elements are distinguished, or in synthetic biology where a distinction is made between natural and synthetic DNA.

In chapter one, type-hierarchies were graphically presented using a tree-like shape placing at the top the type with the largest extension, which is called a supertype. The same

[249] R. Harré, J. L. Aronson, E. C. Way (1995), pp. 15-16.

graphic presentation can be made for the multiple-round simultaneous auction placing 'institution' as the supertype, and also by distinguishing traditional from artefactual auctions:[250]

Figure 3.1. A partial type-hierarchy of multiple-round simultaneous auction:



Although, there is no paradigm shift in the design of the multiple-round simultaneous auction, significant progress was made in the design of artefactual auctions, which started with the work of William Vickrey, who designed the second-price sealed-bid auction. The multiple-round simultaneous auction is an artefactual auction which combines aspects of the English and the Dutch auctions, namely ascending bidding and the combination of items in packages, adding to them multiple rounds, the return of any unwanted licences, and bid increments decided by the auctioneer. The similarities with the English and the Dutch auction constitute the positive analogy, and multiple rounds, the return of unwanted licences and bid increments constitute the negative analogy. Because the size of the negative analogy is larger, the blueprint is a case of

[250] Different criteria can be used for classifying auctions and there are further types of them, see P. Klemperer (2004), and P. Milgrom (2004).

minimal analogy, and therefore it is a progressive design within the type-hierarchy. Design by analogy does not exist as part of the methods in mechanism design theory, it is a topic to be developed both in philosophy and the social sciences.

A third evaluation of the Milgrom-Wilson-McAfee blueprint can be made using the *five principles for blueprint making* from Nancy Cartwright discussed in chapter two, section 2.2., namely (i) The parts that make up the machine, their properties and the separate capacities, (ii) How the parts are to be assembled, (iii) The rules for calculating the outcome from the joint operation of the assembled parts, (iv) What counts as shielding, and (v) How the machine is set to run

Cartwright used the blueprint of a repudiation-proof contract from Oliver Hart and John Moore to illustrate how these principles work and how their demands should be met. With the help of equilibrium theory and the rules for renegotiation designed by Hart and Moore an optimal equilibrium can be accomplished by decisions made by the players, which solves the inefficiency created when the contract is repudiated. Hart and Moore's blueprint only meets the requirements from the first three principles because it describes the parts of the machine, namely two individual players the investor and the entrepreneur both displaying specific psychological capacities: self-interest, greed, perfect and costless calculation, and full rationality. Other parts are structural or external to both players such as the same discount rates, certainty in all operations, rules for renegotiation, and the existence of a frictionless second-hand market for the physical assets of the project. The structural parts and the players are assembled in a single game with two stages, one with an initial negotiation and agreement on a certain distribution of the surplus, and a second one when repudiation of the contract occurs and the surplus is now divided in equal parts. However, the blueprint does not provide information on how to shield the new contract and how to implement it.

The evaluation of the multiple-round simultaneous auction blueprint is less positive. The parts of the machine were known, namely self-interested telecommunications firms with high purchasing power and the FCC as a greedy government agency wanting to maximise the revenue. The structural parts were also known, which consisted of rules defining the three main mechanisms, namely a simultaneous market, ascending biding and multiple rounds. Although, Milgrom, Wilson, McAfee and others were confident that the auction would work, there was no knowledge on how to put all the different parts together and how to set the whole auction running; and there were no means either for getting a reliable calculation on the outcome. There were concerns about collusion among the bidders and attempts from them to outwit the rules, however no precise shielding against these possibilities was part of the blueprint. McAfee and Milgrom actually explain that 'the spectrum sale is more complicated than anything in auction theory. No theorem exists–or can be expected to develop–that specifies the optimal auction form.'[251]

Two of the main problems were complementarity and perfect substitution of the licences, and a solution using Nash equilibrium was not feasible. For instance, because licences packages would be formed, the existence of complementary values means 'that market-clearing prices may not exist. Equilibrium is likely to exist if the buyers have similar views about how the goods should be aggregated, whereas it may not if they disagree about what constitutes good aggregations.'[252] The solution to this and other problems was provided by the experimental economist Charles Plott and his team, who devised the experiments which produced the data needed using the method of experimental parameter variation. Milgrom himself recognises this when he writes that 'much of what is known about multi-unit auctions with interdependencies comes from experiments.'[253]

In the next section, I introduce the method of experimental parameter variation, which is taken from aeronautical engineering.

[251] P. McAfee and J. Milgrom (1996), p. 171.
[252] *Ibid.*, p. 172.
[253] J. McMillan (1994), p. 151.

5.2. Experimental parameter variation

The engineer Walter G. Vincenti has produced a methodological account of aeronautical engineering, where he surveys different historical episodes of engineering research and design to illustrate a number of methodological practices. One of the most suggestive methodological practices he identifies in this survey is Experimental Parameter Variation (EPV), which he defines as: 'the procedure of repeatedly determining the performance of some material, process, or device while systematically varying the parameters that define the object of interest or its conditions of operation.'[254] He explains that this method is distinctive of engineering in contrast to scientific theories:

> Experimental parameter variation is used in engineering (and only in engineering) to produce the data needed *to bypass the absence of a useful quantitative theory*, that is, to get on with the engineering job when no accurate or convenient theoretical knowledge is available. This is perhaps the most important statement about the role of parameter variation in engineering.[255]

Vincenti illustrates this method by discussing the work from the mechanical engineers William F. Durand and Everett P. Lesley, who performed extensive experimental research between 1916 and 1926 with the purpose of designing and producing new fixed-pitch air propellers superior to those available in Europe. Prior to the development of variable-pitch propellers in the 1930s, only fixed-pitch propellers were used in aircrafts. Since the shape of a fixed-pitch propeller could not be changed during different flight conditions, they were optimised for cruise, climb or take-off depending on which one was most critical for the airplane mission. Choices were also made selecting a propeller which could attain a compromise general performance, where no aspect was optimised.

[254] W. Vincenti (1990), p. 139.

[255] *Ibid.*, pp. 161-162.

In the United States no significant research had been done since the pioneering achievements of the Wright brothers in the first decade of the twentieth century. Although, some information on air propellers was available at the time, no systematic data existed which could support a new design. Only a few results were available from the experimental work done by Gustave Eiffel, a French engineer who had developed a new type of wind tunnel for experimenting with three families of different propellers, with each family containing four types of propellers. Experimental engineering research work on air propellers began in England, France and Germany around 1910. By 1913 in England comparisons were made between previous theoretical work and experimental data showing that theory was only useful for the general qualitative aspects of design. Accordingly, the quantitative part would have to be developed from data to be obtained in the laboratory.

In contrast, the amount of systematic data on marine propellers was significantly larger. By 1905 William Durand had produced experimental results on forty-nine different prototypes using the method of experimental parameter variation. By 1908 in England, Robert Froude had reported results on thirty-six marine propellers. In the United Stated this was followed by a hundred-and-twenty more results reported by David Taylor in 1910. Because of the availability of data on marine propellers, Durand and Lesley relied on them for their research on air propellers.

In addition to the existing experimental data on marine propellers, the blade element theory from Stefan Drzewiecki was also available. This theory divides the blade of a propeller into a large number elements at different radii, and each element is modelled as a small aerofoil moving in a straight line with a velocity determined by three components, namely the forward speed of the propeller, the tangential speed of the rotating element, and a secondary speed of flow induced by the aerodynamic action of the propeller itself. Then, the forces of each element at its appropriate velocity are estimated from experimental aerofoil data. Finally, the performance of the propeller is determined by summing all those forces.

One of the main problems Durand and Lesley faced was the calculation of the secondary flow induced by the action of the propeller. They used the blade element theory neglecting this secondary force and other complicating effects. By doing this they were able to calculate the performance of eighty two-blade air propellers by varying the parameters in a theoretical fashion. They compared these results to those obtained through experimentation finding that the general trend was same, while the quantitative values were substantially different and erratic.[256] This discrepancy between theory and experiment is very important because it shows the limits of theoretical knowledge for purposes of design. Theoretical knowledge is frequently insufficient for design; no reliable and efficient design can exclusively rely on it. The theoretical calculus of trends in the performance of air propellers made by Durand and Lesley is analogous to calculus of behavioural tendencies made by experimental economists, who also produced experimental data for design which a theory cannot provide. This is shown in the next section with the design of the multiple-round simultaneous auction.

Durand and Lesley produce new data by testing different prototypes of propellers made of different materials and with different shapes by systematically varying the parameters within the range of practical concern, defined mainly by a set of foreseeable flight requirements and conditions. They define the performance of a fixed-pitch air propeller as the function of two different sets of parameters, namely the conditions of operation and the geometrical properties of the propeller. The former includes the forward speed V and the revolutions per unit time n; the latter includes the diameter D and a number of ratios r_1, r_2, … etc. which contain information on the geometrical shape. The propeller performance P is determined by the following equation: $P = F(V, n, D, r_1, r_2, \ldots)$. The description in the equation is approximate because it leaves out complicated secondary effects from viscosity, compressibility of the air and the elastic bending of the propeller. Given the aim of the design, these effects can be neglected. Once the value range of concern has been fixed and the list of particular values has been

[256] *Ibid.*, p. 155.

established, 'parameter variation for the propeller consists of systematically varying the values of the parameters within the parentheses and measuring the resulting variation of propeller performance.'[257]

Because of the crucial role of the geometrical shape of the propeller, the ratios became the relevant parameters to be tested. After some preliminary tests, Durand and Lesley selected a diameter of three feet for all the small-scale prototypes and they established five parameters of relevance defined by ratios r_1 to r_5. The most important parameter was the mean pitch ratio, which is a measure of the angular orientation at some standard representative radius relative to the plane of propeller rotation of the blade section. This parameter is particularly important because the larger the mean pitch ratio, the higher is the angular orientation of all blade sections. The other four parameters contained information on the distribution of the pitch ratio along the blade and the type of blade section. They chose three equally spaced values of mean pitch ratio and two values of each of the other four parameters. Using all possible combinations of values, Durand and Lesley obtained forty-eight different propeller models, which were distributed in a representative way over the field of design. Using a wind tunnel, each model was tested using a prototype through a series of values of rotational speed n at distinct values of forward speed V to determine its performance P. Those with the highest value were selected.

This was the initial and fundamental stage of the research, where the method of experimental parameter variation was crucial for obtaining data needed in further stages until the completion of the full design, construction and final test of the new propellers. The research continued until Durand and Lesley built and tested a full-scale prototype. Vincenti explains how they used laws of similitude and dimensional analysis to proceed from the data obtained on the forty-eight small-scale prototypes to the construction and testing of small-scale models and full-scale prototypes.[258] Once the full-scale prototypes successfully passed all necessary tests, the

[257] *Ibid.*, p. 148.

[258] *Ibid.,* pp. 159-166.

engineering research phase was followed by the manufacturing of propellers ready to be assembled into the aircrafts. Propellers only work in combination with the right engine and airframe, so new airplanes were designed with engines and airframes adequate to the selected propeller. In this way, the vast amount data provided by Durand and Lesley using experimental parameter variation became crucial for the design of new superior aircrafts, which had been the ultimate aim of the research.

Their work set a new standard in engineering research and design. Their case demonstrates the essential role experimental parameter variation plays in engineering research and the limits of theoretical knowledge, in this case the blade element theory. Within a short period experimental parameter variation spread and became an established method that encompassed the early work from William Durand in the United States, Robert E. Froude in Britain, and Karl Schaffran in Germany.[259]

In the next section, I show how the method of experimental parameter can be extended to design and engineering in economics.

5.3. The engineering of the FCC auction

The philosopher Francesco Guala characterises the FCC auction as a case of economic engineering. He is mainly concerned with the problem of external validity. In particular, he is concerned with the kind of inferences which extend internally valid propositional knowledge produced in the laboratory into the outside world. The problem is philosophically relevant because those true and reliable inferences made predicting and explaining behaviour in the laboratory are not obviously true and reliable when new markets and state institutions are to be built. He argues that the combination of inferences by analogy, eliminative inferences and the

[259] *Ibid.,* p. 294; see also D. W. Taylor (1924).

reproduction of real world conditions in the laboratory explain the success of the FCC auction.[260]

While the propositional knowledge engineers have is certainly essential, the practical knowledge they have for the construction of social machines seems to be more distinctive of engineering. Such a practical knowledge from engineering actually starts in the laboratory, where new mechanisms are tested. I argue that experimental parameter variation is an example of this practical knowledge. Guala himself is aware of the existing gap in the philosophical research on this kind of knowledge, which actually explains how while new markets and state institutions are built. He acknowledges this in the replies he gives to Anna Alexandrova and Frank Hindriks.

Alexandrova and Hindriks are both critical of the explanation Guala provides on the role experiments have in producing knowledge which lies outside theories and blueprints. They actually do not use the term 'blueprints', they use the term 'models' instead. Alexandrova is mainly concerned with the limitations blueprints have on the behaviour and other relevant conditions to be found in the outside world; when a new kind of auction is implemented; she explains that when 'economic models and experiments are used for engineering institutions such as spectrum auctions […] sometimes it is simply not known whether or not some assumption essential for deriving a particular effect in the model can be satisfied by the target system economists are constructing.'[261] Hindriks makes a general criticism to theoretical economists who are sceptical or neglect the contributions experimental economists could make creating new knowledge, and he criticises Guala for not making wider and stronger case in favour of experimental economics beyond inference and external validity. He explains that 'except for a few scattered remarks, however, Guala does not directly address the scepticism that economists display about experiments.'[262]

[260] F. Guala (2005), pp. 184-202.
[261] A. Alexandrova (2008), pp. 199-200.
[262] F. Hindricks (2008), p. 217.

In his reply, Guala highlights the good job experimental economists do testing the hypotheses contained in the blueprints, while at the same time recognises that 'the story is very different for experiments that are closer to application ('testbed' experiments). Here Alexandrova is right –no standards account of modelling does a good job at explaining what is going on.'[263] In his reply to Hindriks he explains that 'as he correctly points out, MEE [Guala's book *Methodology of Experimental Economics*] is quite bold in making prescriptive claims about experimental inference but relatively modest the role of experimentation in economics as a whole.'[264]

My argument on experimental parameter variation as a method of experimental economics answers the concerns expressed by Alexandrova and Hindriks. The use of experimental parameter variation shows the distinctive contribution experimental economists make to the design and engineering in economics. Moreover, the scope of experimental parameter variation could be extended to experiments performed in other social sciences.

The blueprint submitted by Milgrom, Wilson and McAfee represented a good solution to important problems such as complementary values, perfect substitution and preference maximisation on package-bidding. Nonetheless, its implementation represented a great challenge, the joint functioning of the three main mechanisms looked too complicated. Mechanism design theorists were no able to create a reliable expectation on how it all would work. Besides the right functioning, there were also concerns on how to prevent collusion and cheating. Unlike the other kinds of auction such as the Dutch and English auctions, multiple-round simultaneous auctions had never been tried before.

Rules constitute a fundamental part of mechanisms, and it the case of the FCC auction blueprint 'the most important – and debated – rules concerned increments, withdrawals,

[263] F. Guala (2008), p. 229.

[264] *Ibid.*, p. 227.

eligibility, waivers and activity.'[265] The auction would not be continuous but split into rounds with no pre-fixed number of total rounds, that is, the rounds would continue until no offer is put forward, and the winner is satisfied with the licences she has purchased. To ensure a maximal satisfaction of preferences, withdrawals were an important part of the rules. It was also important to prevent unnecessary delays speeding up the action without prematurely terminating demand, so rules on bid increments and an eligibility based on a deposit were considered in the blueprint.

As part of the activity rules the eligibility of bidders was important because some of them may want to slow down the bidding process by following a 'wait and see' strategy. Such a delay could cause significant inefficiencies, and it would also increase the costs of the auction. Therefore, the eligibility of any bidder would be subject to an initial deposit called 'initial eligibility', which would also set a limit to the number of markets the bidder could participate in. This rule of eligibility also prescribed the regular use of such a deposit by spending parts of it in each bid. A refusal to do this would affect the eligibility of the bidder by reducing the number of bids she could make in the next round. Neither game theory nor auction theory provide information on how long an auction with multiple rounds could last, so with the eligibility rule, the auctioneer would be able to speed up the auction by enforcing an early commitment from all bidders. This rule would also help identify bidders who lacked any real interest in acquiring the licences, which was a problem auctions in New Zealand and Australia faced where uninterested bidders caused significant delays.

Three key data were missing on these rules, which no theory or previous knowledge on mechanisms could provide information on, namely:

1) Optimal bid increment.
2) Estimate of the total number of rounds.
3) Length cycles produced when licences are sold back

[265] F. Guala (2005), p. 175.

Without reliable data on these three aspects, the efficiency and smooth running of the auction would be compromise, and its full implementation could actually fail. The FCC hired the economist Charles Plott and asked him to perform experiments on these and other aspects of the auction. Guala provides a rich description of the experiments performed by Plott, however he does not draw a systematic methodological lesson from it. This is also pointed out by Alexandrova.[266] Charles Plott also provides a detailed description of the experiments he and his team in Caltech performed calling them 'testbeds', which he defines as 'a simple working prototype of a process that is going to be employed in a complex environment. The creation of the prototype and the study of its operation provide a joining of theory, observation, and the practical aspects of implementation.'[267]

The idea of a 'working prototype' is insightful and it actually corresponds to the term used in engineering, however the definition on the whole is poor and uninformative for any scientist who would like to have a clear and simplified understanding of the crux of the method. There is no abstraction made from the descriptive details, which would enable any scientist to see in a simplified manner the nature and systemic side of those practices. This is why I argue that by extending the method experimental parameter variation to the design of the FCC auction, we draw and extend methodological lessons which otherwise would remain lost in the rich description provided. Let us recall that experimental parameter variation consists of determining via experimentation the optimal performance of materials, processes or devices by varying the parameters of their operation.

The most comprehensive report of the experimental practices performed in preparation for the implementation of the FCC auction is provided by Plott. However, parts of the report are insufficient for producing a richer and more detailed methodological description. Another problem is the small number of experiments he performed. Unlike Durand and Lesley,

[266] *Ibid.,* p. 197.
[267] *Ibid.,* p. 607.

who carried out comprehensive tests of propellers with a great range of variation, Plott and his team only conducted a small number of experiments due to the deadline and time and budget constraints set by the FCC. He explains that 'pressures of time and money substantially limited the amounts of experimental data that could be collected', therefore 'the strategy was to select certain key aspects of the parameter/theory space and collect such data as one could.'[268] Only two parameters were subject to variation, namely the total number of licences and the number of those with complementary values. In one case, seven licences were auctioned with two collections of three licences each having complimentary values; in the second case nine licences were auctioned with all of them having complimentary values. The experiments had two aims. The first one was to compare the efficiency of the multiple-round simultaneous auction allocating licences to bidders who value them most against a Japanese auction. The second one was to provide information on optimal and estimate values of the activity rules from the multiple-round simultaneous auction.

(1) *Optimal bid increment*. As an auctioneer, the FCC had an interest in identifying the winners rapidly, so that the auction could finish as soon as possible without negatively affecting the demand. For this purpose, the blueprint considered a bid increment every round. The auctioneer would do this by identifying the highest standing bid at the end of each round introducing an increment for the minimal bid in the next round. On the one hand, an excessive increment could deter potential bidders, causing demand-killing and the reduction of eligibility. On the other hand, a too small of an increment would not speed up the auction enough. Therefore, the discovery of the optimal increment became an important problem of design.

During the variation of increments performed the laboratory, Plott and his team observed that large increments above the highest standing bid effectively eliminated bidders too quickly placing at risk the inefficient allocation of licences. Without specifying the number and values of the variations, Plott explains that 'experiments had also produced evidence of the

[268] *Ibid*. p. 614.

capacity of large increments to be demand-killing: A bidder failing to bid because of a large increment could lose eligibility.'[269] The FCC reports that an increment of ten to twenty percent above the highest standing bid was found to be the optimal range.[270] This was enough to speed up the auction but not too big to cause demand-killing and inefficiency.

(2) *Estimate of the total number of rounds*. The second data to be obtained was an estimate on the total number of rounds. The FCC was concerned about the operation costs if the auction extended for a long time. Plott considered different aspects of the behaviour from the bidders and the auctioneer, which could compromise the efficiency of the auction. On the one hand, there was the strategic interest bidders may have in slowing down the auction. On the other, too much pressure on the bidders could also lead to overbidding. A further concern emerged from the allowance the blueprint made for the bidders to have time off for revising their strategies and budgets; the rule prescribed a stop after a number of rounds starting again the next day. This rule also helps prevent the winner's curse saving the FCC from expensive mistakes by preventing a legal case in court. Milgrom explained the case noting that 'sales of major companies take a long time. There are billions of dollars at stake here, and there is no reason to rush it when we are talking about permanently affecting the structure of a new industry.'[271] Therefore, getting an estimate of the total number of rounds and intervals became crucial data of design with important political, economic and legal implications.

Hence, time between rounds would allow bidders to put forward more sensible bids, and it would also help prevent overbidding. At the same time, it was also important to reduce the number of rounds and intervals as much as possible to save on operation costs. In the experiments performed, Plott observed that the total time of the auction was mainly dependent on the number of rounds, rather on the intervals between them. He explains that 'many of the early experiments that were allowed to terminate naturally involved continuous-time processes

[269] *Ibid.*, p. 633.

[270] 'Smoothing Methodology Fact Sheet, 31th March 2003, FCC Experiments, Papers & Studies, electronic source: http://wireless.fcc.gov/auctions/default.htm?job=papers_studies

[271] P. Milgrom (1994), p.11.

without stages. Examination of these data suggested that the FCC auction could go through as many as a hundred rounds. The more rapid the rounds, the sooner would be the termination.'[272] This estimate of a hundred rounds was good enough because it allowed the FCC to calculate the operation costs and consider the need for an adjustment on the activity rules.

(3) *Length of cycles*. The Milgrom-Wilson-McAfee blueprint also included a rule allowing withdrawals because the winner decided that the price was too high, or because she just changed her preferences. The rule established that licences could be sold back to the market but the bidder returning them would have to pay the price difference, if the final price was lower than her own bid. Theoretically, it was expected that withdrawals could lead to 'cycles' where licences returned to auctioneer would have to be sent back to the market more than once, until one of the bidders becomes satisfied with the price. Although this possibility was envisaged, there was no way to calculate how long cycles might be.

Therefore, the production of experimental data on the occurrence and length of cycles was another important task which, along with the estimate of the total number of rounds, was relevant for estimating the total time of the auction. Too many cycles might significantly delay the termination of the auction, or even prevent the auction from ending. The experiments showed that a licence package may be released up to three times with the last holder losing money. Plott reported that 'since the new price of the item is above the average value of the marginal person, the new holder lost money. Panel B shows that releases can occur more than once during an auction. As can be seen in that experiment, the item was released two times, leading to a cycle of length three.[273] Hence, cycles were short but overpricing was likely to occur.

These data on cycles and those on bid increments and the total number of rounds were crucial for the final design and implementation of the FCC auction, which presumably led to an

[272] Plott, 1997, p. 633.
[273] *Ibid.,* p. 625.

efficient allocation of telecommunication licences.[274] The revenue from the first round with nine auctions run between 1994 and 1996 was of twenty-three billion dollars, a large amount that has been hailed as a proof of the efficiency of the auction, and the power of game theory for design. However, the same credit should be given to experimental economists whose contribution was decisive for the final design and the successful implementation of the auction. The success of the FCC auction led governments in Europe to the implementation of auctions also for allocation of exploitation rights of the electromagnetic spectrum.[275]

In philosophy of science, design and engineering methods are often neglected by the excessive attention paid to theories and the methods associated to them. The use of theories for the design of blueprints has led some to argue that the success of the simultaneous ascending auction was due the advancement game theory and auction theory. While one can recognise the use of theories in both cases the fixed-pitch air propeller and the FCC auction, it would be a mistake to attribute the successful design and implementation of them exclusively to those theories.

By reducing the explication of such success to the derivation of knowledge from theories, theory-testing experiments and externally valid interferences, philosophers of science are overlooking the distinctive features of experimental and engineering methods. My aim in this chapter has been to show the distinctive epistemic and methodological character of these practices and the knowledge they produce. Without a set of systematic practices producing data for design, engineers and policy-makers would be left only with a set of abstract models and predictions on some tendencies.

Experimental parameter variation is a good example of practical knowledge which produces data theories cannot provide. It is also an example for philosophers and scientists on how to get a systematic and insightful interpretation of some of the practices performed by

[274] The efficiency of the first round of FCC auctions has been a matter of controversy; see C. Plott (1997), p. 637; and P. Cramton (1997).

[275] See K. Binmore, and P. Klemperer (2002).

experimental social scientists, which otherwise would remain implicit or lost within long and detailed descriptions published in articles and books. Experimental parameter variation has a normative force analogous to any form of argument and inference studied in logic using rules such as *Modus Ponens,* a *Celarent* syllogism and Bayes' theorem. These rules provide instructions on how to perform inferences, experimental parameter variation provides rules on how to perform practices. Philosophers of science have excessively focused on inferential rules and propositional knowledge from theories and abstract models; by doing this they have overlooked and dismissed the role of scientific practices and the knowledge they produce. A comprehensive philosophy of design and engineering in the natural and the social sciences is needed. The subjects discussed and the arguments put forward in this dissertation are presented as an advance towards such a philosophy.

—O—

Chapter 4

Self-Interested Knaves

4.0. Introduction

Psychological assumptions constitute an essential component of the design of rules, policies and institutions such as electoral systems, public health services, constitutions, contracts and auctions. These assumptions are used by theoretical and experimental social scientists for making blueprints as well as for the construction and implementation of those rules, policies and institutions. Moral psychology constitutes a subset of such assumptions concerned with the explanation of moral behaviour. Whereas some assumptions give prominence to attitudes towards risk, cooperation and learning; moral assumptions focus on attitudes towards norms and values such as free-riding, truth-telling, promise keeping, fairness, envy and knavery. In game theory, for instance, aspects of such moral behaviour are sometimes synthesised using analogies with animals such as dove-like and hawk-like behaviour.

Moral psychology assumptions help regulate the expectations of new policies and new institutions by setting up some initial limits on possible behavioural changes. The success or failure of any design in the social sciences crucially depends on the accuracy and reliability of these assumptions. This is why a good science of facts is needed; in particular for psychology, sociology, and some aspects of evolutionary theory. A science of actual moral behaviour must be able to measure, explain and predict as well as regulate the expectations of possible behaviour brought about by design and engineering. In this chapter and the next, I discuss moral psychology and some basic aspects of moral sociology.

Moral psychology is a branch of psychology, and therefore it should not be considered as a philosophical branch. This seems to be an obvious and unnecessary remark. Nonetheless, moral psychology is often ambiguously discussed and taught as a philosophical branch. This happens in part because moral psychology is not yet a fully developed branch in psychology. I argue that this ambiguous status in philosophy has undermined the progress of moral psychology as a scientific branch. Moral psychology should be considered only as a branch of psychology, and therefore it should be developed by scientists and evaluated by philosophers of science with the methods and criteria used in philosophy of science.

The philosopher Jay Wallace, for instance, defines moral psychology as the study of 'the psychological conditions for the possibility of binding norms of action; the ways in which moral and other such norms can be internalized and complied with in the lives of agents; and a range of psychological conditions and formations that have implications for the normative assessment of agents and their lives.'[276] Currently, the study of all those psychological conditions can only be performed by psychologists and neurologists, who have the methods and training needed for producing this kind of knowledge.

I agree with philosophers such as John Doris, Stephen Stich and others when they argue that ethics should be 'richly informed by relevant empirical considerations' such as field and experimental evidence. However, I disagree when they argue that in order to keep a 'competitive advantage' ethics, i.e. moral philosophy, must make 'empirical claims with enough substance to be seriously tested by the empirical evidence'.[277] Ethics provide normative arguments as to why a set of values should be preferred over a different set. The empirical justification or refutation of the definitions and explanations of the moral norms and values followed by an individual or a social group is a scientific task, not a philosophical one. An individual or a population can *de facto* act according to certain values and norms, which they may not provide an argument for. That is to say, they might act according to certain values and norms without being able to explain why those values should be chosen. Until such a normative argument is provided there is no moral philosophy.

[276] In F. Jackson and M. Smith (eds.) (2005), p. 87.
[277] *Ibid.*, p. 119; see also J. M. Doris (ed) (2010).

Any moral philosophy presupposes a certain moral psychology and a moral sociology.[278] Calls for an empirically informed ethics make sense and are welcome, because often moral philosophers write their argument paying little or no attention to the psychological and sociological assumptions they use. They also do it without enquiring into the scientific status of those assumptions, by asking whether they have enough empirical support or whether there is any scientific controversy over them. Moral philosophy and design in the social sciences must be based on a good science of facts, namely a psychology and sociology of moral behaviour.

One side of the scientific task consists of establishing certain aspects of actual facts, that is, of actual individual and social behaviour. The other side consists of establishing the feasible set of the behaviour projected in moral arguments or in economic and political blueprints. Therefore, this knowledge of design and knowledge of engineering must be added as part of the demand for an empirically informed moral philosophy.

In this chapter and the next one, I analyse and discuss moral psychology as a branch of psychology. I do this by evaluating the grounds on which knowledge claims are justified as well as the choices made over rival theories. In other words, I treat moral psychology as an ordinary case of science. In particular, I discuss and compare the moral psychologies of Bernard Mandeville and David Hume, whom I consider to be the early modern founders of moral psychology. Because these are cases of early modern science I use, as much as possible, the epistemological standards of that period.

The term 'moral psychology' did not exist in the eighteenth century, so Mandeville describes his work as 'moral anatomy' by creating a methodological analogy with anatomy in medicine. He was trained and graduated as a physician, writing a dissertation and a treatise on health subjects.[279] He had previously graduated in philosophy, where he wrote a dissertation

[278] For a survey on moral sociology see S. Hitlin, and S. Vaisey (eds.) (2010); Émile Durkheim (1887) laid some foundations for a sociology of morality.

[279] B. Mandeville (1691) *Disputatio Medica lnaugralis de Chyosi Vitiata*. Leyden: Elzevier; and B. Mandeville (1711) *A Treatise on the Hypochondriack and Hysterick Passions*. London: D. Leach.

supporting the Cartesian thesis, which neglects that non-human animals have a soul.[280] Hume also described himself as an anatomist; they wanted to produce a science of moral subjects using the methods from the natural sciences.

It is a challenge to select criteria for evaluating Mandeville's moral psychology as a science, mainly because the normative standards of the time do not fit with the purpose of a building a moral psychology. Specifically, there were two main rival standards at that time: the Aristotelian and the Cartesian. A third less prominent standard came from Galileo, Gassendi, and Bacon. The Galilean and the Cartesian standards share some common grounds on the fundamental role of experimentation and systematic observation. However, they disagree on the foundational place of metaphysical principles, from which some deductions on physical laws and properties could be made. Gassendi criticises the Cartesian method as well as the experimental basis shared by Galileo and Bacon.[281] The experimental methods and systematic observation championed by Galileo, Gassendi and Bacon are the closest to the method followed by Mandeville. However, even these standards are not entirely appropriate because they were developed from astronomy and physics. In the eighteenth century there was no philosophy of medicine and no philosophy of psychology or the social sciences.

Because of these reasons, I can only partially meet the requirements for an evaluation of the methods and knowledge claims from Mandeville's theory. Such an evaluation consists of two parts. The first one in this chapter covers sections 4.1., 4.2. and 4.3.1., which include a description of the methods and inferences performed by Mandeville as well as his refutation of the moral psychology used by the moral philosopher Lord Shaftesbury. The second part is presented in the next chapter, where I discuss David Hume's refutation to the moral psychology of Bernard Mandeville.

[280] B. Mandeville (1689) *Disputatio Philosophica de Brutorum Operationibus*. Leiden: Elzevier. Later, he changed his views by arguing instead for the similarities and continuation between humans and other animals.

[281] See J. Losee (2001), pp. 46-71.

Besides the methods and theory of moral psychology produced by Mandeville, I also discuss his sociology, which complements his moral psychology. In section 4.4., I discuss different accounts which characterise Mandeville's work as a case of moral philosophy. Against these accounts, in section 4.5. I argue that his work is an early modern example of functionalism applied to social aspects. The functionalist explanations he provides of the economy and society most likely stem from his knowledge as a physician on human anatomy.

In the last sections 4.6., 4.7. and 4.8, I discuss the ideas Mandeville had on design, which are consistent with his moral psychology of universal natural self-interest. Specifically, I examine his ideas on the prevention of knavery in politics, and the behavioural changes required for the transition from an agrarian society to a precapitalist commercial society.

4.1. Refuting romantic moral psychology

In Augustan England, the recent increase of trade and the introduction of public credit and national debt for financing government were viewed with concern, mainly because besides producing large revenues, they also produced big losses for the whole society. Frequently, condemnation of material progress came as reaction to what people perceived as a threat from a new set of moral values based mainly on greed. In spite of bringing economic prosperity, these new values were also considered as fostering corruption in both society and government. The rise and accumulation of new wealth was seen as a political threat, because it increased the crown's capacity for patronage. Nonetheless, the exaltation of Christian values like charity, frugality and good faith was still common. Such values were presented as the moral foundation of a society concerned with securing public benefits like social cohesion, trust and generalised welfare. Charity schools became the epitome of these social concerns inspired by Christian compassion. Politically, it was also argued that if these values were embraced by politicians in parliament and the crown, they would produce good governance and public benefits, including respect for civil liberties and the making of new socially beneficial laws.

Such was the civic spirit in the humanism from Petrarch, and Ancient Greek and Roman philosophers like Horace, Xenophon, and Epictetus, which Anthony Ashley Cooper, third Earl of Shaftesbury, was well acquainted with. Lord Shaftesbury was a leading moral philosopher during the period, who was widely read in Britain and in continental Europe. He was a leading voice that opposed Thomas Hobbes's psychological egoism as the basis for organising the state and the society. He is also considered by many as the founder of the moral sense theory. His most important work, *An Inquiry Concerning Virtue, or Merit,*[282] assumes a moral psychology criticised by Bernard Mandeville as false and romantic.

Shaftesbury constructs a natural teleology, which he extends to society and moral behaviour. His moral psychology, as well as his distinction between virtue and vice, is both built to meet the two main tenets of such a teleology, namely happiness and harmony of the whole. Holistically, he describes nature and society as a system where balance and harmony are to be maintained despite any disorder or imbalance. The main emphasis is placed on harmony and balance because they lead to the continuation and survival of the whole. Such balance is obtained when each member in a society fulfils their function adequately by behaving virtuously just like 'any Organ, Part or Member of an Animal-Body, or mere Vegetable, to work in its known Course, and regular way of Growth. 'Tis not more natural for the Stomach to digest, the Lungs to breathe, the Glands to separate Juices, or other entrails to perform their several Offices.'[283] Virtues such as justice, honesty, modesty and moderate self-interest as well as benevolence and love towards the whole humanity fulfil a function, which helps to maintain harmony and the preservation of society.

[282] The *Enquiry* was first published in 1699, with revised editions published in 1711, 1714, and 1732. Currently, Shaftesbury's work is hardly discussed, however, during the eighteenth century he was widely read and discussed. A compilation of his work with a revised version of the *Enquiry* was published for the first time in 1711 with the title *Characteristicks of Men, Manners, Opinions, Times*, with further editions printed in 1714 and 1732. Along with J. Locke's *Second Treatise on Civil Government*, *Characteristicks* was the most reprinted book in English language during that century with translations to French and German from 1738 to 1779. In English, his influence is notorious in the moral philosophies of Francis Hutcheson and Adam Smith.

[283] Shaftesbury (1732), vol. 2, pp.77-79.

Shaftesbury was highly critical of Hobbes's psychological egoism.[284] Throughout the *Inquiry,* he tries to disprove it by showing how public interest, benevolence and other kinds of social love are actually widely observed in society.[285] He holds that that self-interest and virtue are not opposed by arguing that moderate self-interest is actually virtuous because it contributes to the healthy balance of the whole:

> 'The Affection towards private or Self-good, however *selfish* it may be esteem'd, is in reality not only consistent with publick Good, but in some measure contributing to it; if it be such, perhaps, as for the good of the Species in general, every Individual ought to share; 'tis so far from being ill, or blameable in any sense, that it must be acknowledg'd absolutely necessary to constitute a Creature *Good*.'[286]

According to Shaftesbury, only a narrow, immoderate self-interest is a vice because it harms the 'Interest of the Species' by negatively affecting the balance and harmony of society. Virtue and vice are therefore defined using holistic and teleological criteria. Vice occurs when any passion or affection becomes immoderate. Vices such as ambition, avarice, vanity and love for luxury are all 'ill' or 'unnatural' passions because they harm society.[287]

Besides being moderate and causing no harm to the society, any action has to satisfy two further requirements in order to be virtuous*: i*) It has to be grounded on full affection, *ii*) and on knowledge of the effects it has on society. According to Shaftesbury, any 'Partial Affection, or social Love *in part*, without regard to a compleat Society or *Whole*, is in it-self an Inconsistency,

[284] The attribution of psychological egoism to Hobbes has become increasingly controversial, see B. Gert (1967); G. Kavka (1986); and D. Boonin-Vail (1994).

[285] See Shaftesbury (1732), vol. 2, pp. 80-81; and Shaftesbury (1732), vol. 1, pp. 118-119, where he criticises Hobbes's assertion on universal selfishness: 'Thus Civility, Hospitality, Humanity towards Strangers or People in distress, is only *a more deliberate Selfishness*. An honest Heart is only *a more cunning one:* and Honesty and Good-Nature, *a more deliberate,* or *better-regulated Self-Love* [...] And thus *Love of one's Country,* and *Love of Mankind,* must also be *Self-Love*. *Magnanimity* and *Courage,* no doubt, are Modifications of this universal *Self-Love!*'

[286] Shaftesbury (1732), vol. 2, p. 23; see also pp. 16, 66-67 and 161-162.

[287] *Ibid.*, pp. 163, 166-169.

and implies an absolute Contradiction.' Partial and unstable passions and affections in turn rely on 'capriciousness and humour '.[288] Because they are disorderly and unstable, they harm society and reduce individual joy and generalised happiness, which is inconsistent with the holistic teleological principles set as the ultimate criteria for virtue and vice. Only a full and stable affection in the individual can lead to true virtuous action. Moreover, any action with positive effects on the whole cannot be virtuous, unless individuals knew in advance the effects it could bring. Shaftesbury explains that 'we call any Creature *worthy* or *virtuous,* when it can have the Notion of a publick Interest, and can attain the Speculation or Science of what is morally good or ill, admirable or blameable, right or wrong.'[289]

From a factual perspective, Shaftesbury recognises that there is vice and virtue, and believes too that punishment and reward can reduce vice in the society and the family. The main point of contention is the existence of true virtue with no expected reward or punishment, where individuals themselves naturally and voluntarily constrain their own self-interest, and act instead in the public interest and, more generally, in the general 'Interest of the Species'[290], that is to say, the whole humanity.

Mandeville argues that the psychological assumptions in Shaftesbury's argument are false, namely his holistic claim on the balance and harmony of a society founded on moderate self-interest and other virtues such as justice, modesty and honesty. Mandeville argues that self-interest is by nature universal and unrestrained, since no action is performed because of public interest. In contrast, he claims that immoderate passions are actually necessary for creating a thriving economy and keeping society in a good balance and harmony. 'Dormant' passions keep individuals 'in a State of slothful Easy and Stupid Innocence', where no great vices are expected but no great virtues either. In contrast, vices such as ambition, luxury, vanity, envy, and pride are all necessary: 'it would be utterly impossible, either to raise any Multitudes into a Populous,

[288] *Ibid.,* p. 110.
[289] *Ibid.,* p. 31.
[290] *Ibid.,* pp. 23, 17.

Rich, and Flourishing Nation, or when so rais'd, to keep and maintain them in that Condition, without the assistance of what we call Evil both Natural and Moral.'[291] Mandeville regards the moral psychology adopted by Shaftesbury as 'romantick' and his moral philosophy as mistaken in trying to establish 'Heathen Virtue on the ruins of Christianity.'[292]

In his essay, 'A Search Into the Nature of Society' (1723), Mandeville states that Shaftesbury had completely misunderstood the true nature of society claiming that ''two Systems cannot be more opposite than his Lordship's and mine. His notions I confess are generous and refined [...] What a Pity it is that they are not true', adding that he has 'demonstrated in almost every page of this Treatise [*The Fable of Bees*, Vol. I], that the Solidity of them is inconsistent with our daily experience.'[293] The first volume of *The Fable of the Bees* is long and detailed as well as rich in discussions of examples and authors. Because it tries to prove the universality of self-interest, the argument and examples there challenge Shaftesbury's moral psychology, although it explicitly refers to Shaftesbury only a few times. It is in the 'The First Dialogue' of the second volume of *The Fable of the Bees*, where Mandeville actually explicitly states that he is refuting Shaftesbury's argument by reducing it to the absurd.[294]

Trying to gather comprehensive empirical support for his claims, Mandeville discusses seven representative cases of groups from different social classes, jobs and professions. These seven groups are: laborious poor classes, physicians, lawyers, clergymen, tradesmen, cardinals and members of the court. His style is satirical but that was a common and acceptable format for an argument also used by Shaftesbury and many others during that period. At the time, both satires and essays constituted the corpus of an argument in moral philosophy and politics.

[291] B. Mandeville (1732), vol. I,. pp. 199-200, 373. All page numbers quoted from this book correspond to the original edition, which are normally added in square brackets in all editions.

[292] B. Mandeville (1732), vol. II, p. 432.

[293] B. Mandeville (1732), vol. I, p. 372. In *Letter to Dion* (1732a, p. 34), Mandeville writes 'I differ from My Lord Shaftesbury entirely, as to the Certainty of the *Pulchrum & Honestum*, abstract from Mode and Custom: I do the same about the Origin of Society, and in many other Things, especially the Reasons why Man is a Sociable Creature, beyond other Animals.'

[294] B. Mandeville (1732), vol. II, p. 25.

The first case is of a poor laborious woman, who goes through hardship and pain saving forty schillings to pay for getting her six-year old son into an apprenticeship as a chimney sweep. These actions lead to important social benefits in a large number of households, by providing them with the conditions needed for clean cooking as well as for keeping the rooms warm, in this way preventing illnesses and death during cold periods. Because of the sacrifices the woman makes in order to save money, and also by placing her young son into a highly risky job, which bring many benefits to the society her actions should regarded as virtuous. However, because Shaftesbury's definition of virtue requires intentionality by demanding knowledge of the expected effects into the wider society, the actions of the poor woman are not virtuous despite the important benefits they bring to society.

Therefore, a mere external correlation between individual action and the social benefits it creates can be misleading. A further criterion is needed in order to be able to establish if there was some knowledge of the effects. Shaftesbury does not provide such a criterion, and does not even seem to be aware of the problem. In contrast, Mandeville pays close attention to the need for reliable knowledge of psychological motivations in order to qualify any action as virtuous or vicious.[295] I discuss his own criterion and solution to the problem in the next section.

The case of the industrious poor woman leads to absurd untenable consequences because despite of the public benefits she creates she cannot be virtuous; she is 'an indigent thoughtless Wretch, without Sense or Education'[296], who cannot 'act from such generous Principles.' Mandeville extends the consequences of this case to all the poor, illiterate and uneducated working classes, 'the labouring Poor, which are by far the greatest part of the Nation'.[297] The extension is important because it challenges the scope of Shaftesbury's claims.

The next case Mandeville discusses is that of people from 'higher stations', who are literate and professionally educated such as physicians and lawyers. These people are literate and

[295] B. Mandeville (1732) vol. I, pp. 34, 42 and 467.
[296] B. Mandeville (1732), vol. II, p. 23.
[297] B. Mandeville (1732), vol. II, p. 24.

their work brings many benefits to society by curing the sick and by helping those who face detention and potential imprisonment. In spite of the public benefits created, the actions of physicians and lawyers are not virtuous either, this is because the observed 'Patience', 'Assiduity', 'Labour' and 'Fatigue' are all motivated by a search for 'Fame, Wealth, and Greatness'. This becomes evident because none of them would undergo even 'a quarter' of such a fatigue and effort, if there would be no money and no enhanced reputation in return. 'Therefore, when Ambition and the Love of Money are the avow'd Principles Men act from, it is very silly to ascribe virtues to them.'[298] For the same reasons, no virtue can be attributed to tradesmen who bring many social benefits to both the extravagant rich and the poor craftsmen by buying all kinds of toys and gadgets from the latter to be sold to the former. They know the public benefits brought by their actions but they were not motivated by such benefits.

Mandeville draws the same conclusions for cardinals and other clergymen, who enjoy large fees, housing, food and other comforts for their service. These cases are highly relevant because individuals holding such positions are supposed to care for the welfare of the whole society, and therefore should be considered among the best candidates for virtuous actions following Shaftesbury's definition of virtue. Mandeville points out the College of Cardinals in Rome as 'the best School to learn the Art of Calling', where 'each Member, besides the Gratification of his own Passions, has nothing at Heart but the Interest of this Party.'[299] Similar conclusions extend to members of the Royal Courts, who despite being named and employed to serve the public interest, they 'rob the Publick' instead and are dominated by 'excess of Vanity and hurtful Ambition unknown among the poor'. Furthermore, 'Envy, Detraction and the Spirit of Revenge, are more ranging and mischievous in Courts that they are in Cottages'.[300]

[298] B. Mandeville (1732) vol. II, p. 26.
[299] *Ibid.*, pp. 34 and 35.
[300] *Ibid.*, p. 42.

If there are social groups where 'the Speculation or Science of what is morally good or ill, admirable or blameable, right or wrong'[301] is to be found, those groups should be the clergy and the courts. Again, like in the case of lawyers, physicians and tradesmen, the actions from the cardinals, other clergymen and members of the Royal Court cannot be virtues because there is no intentionality in the public benefits they create.[302]

In this way, Shaftesbury's moral psychology is reduced to the absurd because of all the untenable consequences it leads to. On the one hand, the poor laborious classes are not virtuous because they have no knowledge of the benefits they bring to society. On the other hand, those who have the knowledge such as the members of the royal courts, cardinals and physicians are not virtuous either because they do not have in mind those benefits when they act. All the seven types considered are exclusively motivated by self-interest, and therefore none of them are virtuous.

After refuting the moral psychology of natural restrained self-interest from Shaftesbury, Mandeville challenges any eventual critic of such refutation to put forward any evidence against his own opposite claim of a natural universal unrestrained self-interest. Until such evidence is provided his claim stands as a *best* or *true* psychological explanation of moral behaviour. In the middle of a dialogue, Horatio, a supporter of Shaftesbury's moral philosophy, accepts the existence of unrestrained selfishness, but he believes it is not universal: 'I don't conclude from Selfishness in some, that there is no Virtue in others'[303]. However, Cleomenes, a supporter of Mandeville's views, provides further evidence to finally persuade Horatio of the universality of natural immoderate self-interest.

Lord Shaftesbury died in 1713, so he had no opportunity of replying to Mandeville, who first criticised him in 1723. David Hume responded to Mandeville three decades later in the 1750s. I argue that the controversy between Shaftesbury, Hume and Mandeville is a problem of

[301] Shaftesbury (1732), vol. 2, p. 31.
[302] B. Mandeville (1732), vol. II, p. 39.
[303] *Ibid.*, p. 28.

theory choice, which can be tackled by using epistemic criteria. To avoid as much as possible any epistemological anachronism, epistemic criteria from the same period, whenever they exist, must be used. Because Hume explicitly says that he is emulating Isaac Newton's science and method, in the next chapter I compare his moral psychology to that of Mandeville using criteria from the eighteenth century, namely those from the *vera causa*, which are analogous to those of the inference to the best explanation in current philosophy of science.

4.2. Erecting moral psychology

Moral psychology as a separate specialised branch of knowledge did not exist in the eighteenth century. I argue that the work of Bernard Mandeville represents the first early modern case of a psychology of morality, which he produced in great detail and with a large scope. One of the main aims of my discussion of Mandeville's work is to vindicate him as an early modern scientist, in contrast to those who vindicate him as a moral philosopher. Mandeville's moral psychology also provides important insights on important problems of design such as the distinction between *natural* and *artificial* or *artefactual* behaviour. Mandeville's work in moral psychology is part of his research into a comprehensive science of human nature, which covers subjects such as politics, socialisation, the origin of laughter, good manners, language, suicide and the function of religion in war.

Mandeville studied medicine, and worked as a physician, specialising himself in nervous diseases and digestion. His knowledge of medicine, the anatomy of the human body and the brain provided him with the grounds needed for erecting a moral psychology using naturalistic methods. Although physiology did not exist as a specialised science during Mandeville's lifetime, many of his explanations are functional explanations similar to those of the human organs. The main subject matter of his investigations was the functions and relationships of passions and affections as causes of moral behaviour.

Shaftesbury had already called attention to the need for producing an 'anatomy of the mind' in support of moral philosophy. As we have seen, his moral argument was mostly constructed on a deductive method by drawing the consequences from a set of basic premises about the holistic and teleological character of nature and society. The psychological assertions he makes on the existence of a natural concern for the public interest are not supported on examples or representative cases, therefore his moral argument is almost empirically empty. Rather, he takes those empirical assertions as being obviously true. Once the crucial assumption has been made of the existence of moderate self-interest, it all becomes an almost exclusively analytical argument by expanding and elaborating on teleological and holistic consequences. Nonetheless, such a method should not be considered as entirely mistaken, because Shaftesbury was mainly concerned with writing a philosophical argument not a scientific theory. He was aware of the empirical shortcomings acknowledging the need for an 'Anatomy' with the aim of determining the 'Fabrick of the Mind':

> The Parts and Proportions of *the Mind,* their mutual Relation and Dependency, the Connexion and Frame of those Passions which constitute the Soul or Temper, may easily be understood by any-one who thinks it worth his while to study this inward Anatomy. 'Tis certain that the Order or Symmetry of this *inward Part* is, in it-self, no less real and exact, than that of the *Body.* However, 'tis apparent that few of us endeavour to become *Anatomists* of this sort. Nor is any-one asham'd of the deepest Ignorance in such a Subject.[304]

Shaftesbury extends his aesthetic principles of order and symmetry to the constitution of the mind following an analytic *a priori* reasoning, which Mandeville criticises. He instead defends *a posteriori* synthetic reasoning based on trained observation. He explains that just like 'those that study the Anatomy of Dead Carcases may see, that the chief Organs and nicest Springs more immediately required to continue the Motion of our Machine, are not hard Bones, strong

[304] Shaftesbury (1732), vol. 2, p. 83

Muscles and Nerves, nor the smooth white Skin that so beautifully covers them, but small trifling Films and little Pipes that are either over-look'd, or else seem inconsiderable to Vulgar Eyes'[305]

Small fine parts with large causal effects can therefore escape the untrained observer, who may focus instead on the largest and most obvious parts. Only a scientifically trained observer can find those distinctive causes that make a difference for the performance of any action. Hence, the *moral anatomist* has to be able to see the small hidden springs responsible for moral action. In this context, the main problem consists of establishing the scope of self-interest as the true or best explanation for moral behaviour. The rival explanation relies on altruism or public interest as the true motivation. Mandeville uses the term 'self-denial' to describe actions where self-interest is voluntarily restrained because of the influence of *custom* and *education* with no expectation of reward; it is to 'prefer the good of others to their own, if at the same time he had not shew'd them an Equivalent to be enjoy'd as a Reward for the Violence, which by so doing they of necessity must commit upon themselves.'[306] The most important contention in his argument is the rejection of naturally occurring self-denial complemented with the claim of universal natural self-interest.

In producing such a moral anatomy, Mandeville was critical of the overemphasis moral philosophers place on norms, paying little or no attention to the actual causes of behaviour, 'Writers are always teaching Men what they should be, and hardly ever trouble their Heads with telling them what really are.'[307] This important oversight, which he finds in Shaftesbury's moral philosophy and other moral philosophies such as Stoicism, explains why 'the Theory of Virtue is so well understood, and the Practice of it so rarely to be met with.'[308]

[305] B. Mandeville (1732), vol. I, p. iii.
[306] *Ibid.*, p. 28.
[307] *Ibid.*, p. 25.
[308] *Ibid.*, p. 180.

In order to prove the universal scope of self-interest as an empirical claim, Mandeville uses again representative cases from different social groups, giving special weight to cases where self-denial is expected, and it should therefore be likely to find evidence supporting it. Besides the seven groups discussed in the precedent section, I am now discussing eight more selected from the many cases found in his books. The new groups are mothers, soldiers, nuns, friars, beggars, mendicant orders, kings and ministers. The empirical support is enlarged with the interesting single cases on Lucrecia in Ancient Rome, the Spanish and the Dutch societies of the sixteenth and seventeenth centuries, and charity schools in England. In the examples, Mandeville also identifies different forms of knavish and corrupted behaviour motivated by self-interest.

Like any burglar, lawyers survey the law in order to find gaps, which can help them to advance their own interests and those of their clients by harming those of other individuals. Soldiers do not risk their lives for others, they act motivated by their own self-interest, trying to avoid public shame if they hide or run away, while at the same time they also seek personal glory and immortality. When priests and nuns provide emotional help and comfort to the poor and the rich, they also act motived by self-interest for they want to ensure a place in heaven and veneration on earth. Moreover, nuns and friars frequently breach the vow of celibacy by having children, which they then abort or hide after being born. Mothers love their children, however they love themselves and their own preferences even more. The same woman who can neglect, give away or even kill her bastard child because of public shame or burden, can be tender, caring and sweet to a child born in proper marriage. By giving excessive care and protection to their offspring mothers care more about their own preferences even by ruining their own children, who then become spoiled and dependent.[309]

Ministers affect the public interest by engaging in bribery and corruption in order to advance their own self-interest or that of their own party. Kings send their subjects to death and

[309] *Ibid.*, lawyers, p. 4; soldiers, p. 50; nuns and friars, p. 164, mothers, p. 67.

impoverish their own countries because of a stubborn self-interested motivation in continuing with a lost war, or because of excessive spending caused by their love for pomp and luxury. Because of self-interest, tradesmen cheat in order to take an unfair advantage over their competitors having no concern for the harm they can inflict on them. Beggars and mendicant orders deceive people with a pitiful voice to get the money they will later spend indulging in their own appetites in the company of their friends.[310]

Mandeville claims that human nature 'has always been the same, and that the Strength and Frailties of it have ever been conspicuous in one part of the Globe or other, without any Regard to Ages, Climates, or Religion'.[311] His moral psychology of universal natural self-interest, knavish behaviour and widespread vice became known to the wider public with the publication of the 'Essay on Charity and Charity Schools' in the 1723 edition of *The Fable of Bees*. This essay was strongly rejected by some members of the political establishment in England, who took Mandeville to the Grand Jury of Middlesex accusing him causing public nuisance. In this essay, as in all previous ones, Mandeville presents a sharp and revealing account of the true motivations of apparent virtuous behaviour motivated presumably by compassion based on Christian beliefs. Specifically, he argues that charity, one of the most important social pillars in England, was not motivated by compassion and public interest but by self-interest.

Mandeville holds that the rich and famous help the poor and needy to pride themselves on charitable behaviour, while they are also eager to be flattered, honoured and praised by the public. He explains that charity is given to hospitals, orphanages and universities because 'they are the best Markets to buy Immortality at with little Merit.'[312] Charity is a highly relevant case in moral psychology and moral philosophy because it was, and still is considered by some, as definitive evidence on the existence of true unselfish behaviour, so any successful refutation of

[310] *Ibid.*, ministers, p. 114; King, p. 179; traders, p. 53, beggars, p. 165.
[311] *Ibid.*, pp. 255-256; see cases of Lucretia in pp. 231-232, and Dutch and Spanish in pp. 201-205.
[312] *Ibid.*, p. 300.

such evidence would have important consequences for the respective moral philosophy and the moral psychology supporting it.

Furthermore, in the same essay Mandeville also calls into question the education policy associated with charity schools by advancing a *functionalist explanation* of the role that the poor working classes have in the national income. Mandeville realised that the existence of a large number of poor uneducated individuals ready to work with low wages had unintended consequences for an increase in exports and gross domestic product granting time and wealth for the pleasure and ease of a small minority:

> It is impossible that a Society can long subsist, and suffer many of its Members to live in Idleness, and enjoy all the ease and Pleasure they can invent, without having at the same time great Multitude of People that to make good this Defect will condescend to be quite the reverse, and by the use and patience inure their Bodies to work for other and themselves besides. The Plenty and Cheapness of Provisions depends in a great measure of the Price and Value that is set upon this Labour.[313]

By identifying the function of the poor working classes, Mandeville also refutes the social policies based on a moral philosophy inconsistent with the economic system. An extended policy on charity schools would cause a dysfunction by spoiling the adequate economic functioning of the society as a whole by reducing the number of people willing to accept low wages, which would, in time, reduce exports and gross domestic product. On the normative side, the actual association of poverty to economic growth and self-interest challenges prevailing moral philosophies built upon the influence of Christian values, such as that of Shaftesbury and David Hume. On the positive side, a functional explanation of poverty, self-interest, and their unintended consequences was needed to solve the Mandevillian paradox of how widespread

[313] *Ibid*., p. 326, see also 328.

vice can actually lead to economic growth and political stability instead of creating decay and severe social conflict.

Mandeville explains that such an outcome is not natural, that is to say, it is not the result of a spontaneous order as Friedrich Hayek and others[314] would argue but the product of a design from 'the dextrous Management of skilful politicians'[315], who can turn private vices into public benefits by devising laws and policies, which curb self-interest and vice in the right way, and keep the right number of poor workers who can provide cheap labour. His work shows the importance moral psychology has as a test and foundation of any moral philosophy, and of any design of policies and institutions in economics and political science.[316] To succeed both design and moral philosophy need accurate and reliable knowledge of the desires, intentions and other mental states causing moral, political and economic behaviour.

4.3. The method of moral psychology

Mandeville himself explains the method he has been using for producing his psychological theory, he calls it the 'Method of reasoning from Facts *à posteriori*'[317], which consists of reasoning from observation and experience only. He recognises reasoning from experimentation in chemistry also as an *a posteriori* method, although he explains that experiments cannot be performed on the brain. Alluding to Descartes, he rejects *a priori* reasoning by claiming that 'all our knowledge comes *à posteriori*, it is imprudent to reason otherwise than from Facts.'[318] He also makes an important distinction between 'conjecture' and 'knowledge', explaining that the latter provides certainty while the former does not.[319]

[314] F. Hayek (1967); F. Hayek (1982); and B. Norman (1982).
[315] *Ibid.*, p. 428.
[316] *Ibid.*, p. 160.
[317] B. Mandeville (1732), vol. II, p. 193, see also and p. 177.
[318] *Ibid.*, p. 304, see also p. 207.
[319] *Ibid.*, pp. 163, 167, 263.

In this case, ‘certainty’ becomes a decisive epistemic criterion, because a psychological theory consists of knowledge of unobservable processes and entities taking place inside the mind. Certainty has also been a long-standing problem across the social sciences, where an important distinction is made between the inside and outside of any individual or collective action. This distinction has produced a methodological divide between naturalism and hermeneutics, while in psychology a similar distinction created a divide between behaviouristic and cognitive psychology. Mandeville falls into the category of those scientists who try to get knowledge of the inside of any action using naturalist methods, that is to say, knowledge of the mental entities and processes causing actions. Therefore, the methodological challenge consists of producing knowledge of the mind that is certain and reliable.

Mandeville explicates his own *method of inference* to get this knowledge of the mind by using an analogy with knowledge of the inner parts and functioning of a watch:

> I don’t believe there is a Man in the World of that Sagacity, if he was wholly unacquainted with the Nature of a Spring-Watch, that he would ever find out by dint of Penetration the Cause of its Motion, if he was never to see the Inside: But every middling Capacity may be certain, by seeing only the Outside, that its pointing at the Hour, and keeping to Time, proceed from the Exactness of some curious Workmanship that’s hid; and that the Motion of the Hands, what number of Resorts soever it is communicated by, is originally owing to something else that first moves within. In the same manner we are sure that, as the Effects of Thought upon the Body are palpable, several Motions are produced by it, by contact, and consequently mechanically.[320]

Note that according to Mandeville only a ‘middling certainty’ can be attained for any knowledge about the mind. He explains why full certainty cannot be achieved by pointing to two important methodological constraints related to the anatomy of the brain. First, he explains that the anatomist can only have access to the brain when it is already dead, so the main ‘spring of life’ is

[320] *Ibid.*, p. 177.

gone, and therefore a full understanding of the inner functioning is not possible. The second constraint is set by the limited scope of macroscopic observation, which is constrained to large parts and organs such as nerves, blood vessels, folds and windings, while millions of small cells remain unobserved. Because of these two methodological constraints, the psychologist cannot gather the information needed for producing a theory with a higher degree of certainty. Therefore, the 'best Naturalist must acknowledge that he can only 'give any tolerable Guesses', or actually admit 'as to the mysterious Structure of the Brain itself, and the Oeconomy of it, that he knows nothing.'[321]

With those limitations, the inference to inner live processes and entities proceeds exclusively from observable behaviour. He is aware the epistemic challenge this implies by explaining that 'it is impossible to judge of a Man's Performance, unless we are thoroughly acquainted with the Principle and Motive from which he acts.'[322] By becoming thoroughly acquainted with information available on the subjects, the scientist can reach a 'middling' or reasonable degree of certainty.

Another means for raising the degree of certainty on a theory consists of refuting rival theories. Mandeville does this by refuting the moral psychology of Lord Shaftesbury. If it is believed that the theories in dispute are exclusive and exhaustive, the refutation of one of them confirms the other. In this case self-interest and self-denial are mutually exclusive and exhaustive, at least if no gradation between the two is considered. Because only a middle certainty is available, the moral psychology of universal natural self-interest from Mandeville can obtain a higher degree of certainty or confirmation through eliminative reasoning in the form of a disjunctive syllogism between both rival theories.

Any refutation relies on the amount and quality of the knowledge available from the rival theories, which is produced using a specific method. The rules of an early modern method

[321] *Ibid.*, pp. 179-180.
[322] B. Mandeville (1732) vol. I, p. 42.

of psychology can extracted from the previous discussion on Mandeville's methodological ideas, they are the following six rules:

- Reason from experience only.
- Gather as much information as possible on the subjects.
- Trained observation skills are indispensable.
- Look for unexpected motivations hidden by self-deception or interest.
- Test rival claims by trying to refute them.
- The epistemic aim of these rules is reasonable certainty.

Mandeville does not discuss further or explicate these rules, so they may look poor according to current philosophic and scientific standards. However, as an early modern psychologist, he was not in a worse position than others. Indeed, as an early modern psychologist David Hume faced similar methodological and epistemic challenges when he was also trying to produce a comprehensive science of human nature. Moreover, certainty on the knowledge of the inside of any action, that is to say, on mental entities and processes remains as a key problem in psychology and the social sciences.[323]

An evaluation of Mandeville's method and its products is presented in the next chapter by comparing his theory of universal natural self-interest to David Hume's theory of natural sympathy and self-interest; such a comparison becomes a problem of theory choice, which is solved by using the *vera causa* criteria from the eighteenth century.

[323] See R. G. Collingwood (1946); M. Weber (1922); C. Geertz (1973); and A. Giddens (1984).

4.4. Defining virtue and vice

According to Mandeville, self-denial is a decisive psychological feature for the definition of virtue. As it was quoted earlier, Mandeville defines self-denial as the restraint of self-interest through custom and education.[324] That is to say, virtue is not natural but artefactual. Self-denial is important because it is a necessary condition for any action to be considered as motivated by public interest or altruism.

Under different names and technical terms, the discussion on self-denial is still relevant today in economics and political science. The definition of true 'self-denial' addresses similar concerns to those expressed by Amartya Sen with his own defence of 'counterpreferential choice' based on social commitment.[325] Ken Binmore has recently argued that the theory of revealed preference is superior to psychological theories arguing for universal self-interest because revealed preferences leave open the possibility of unselfish behaviour.[326] Another example is the definition of altruism in sociobiology as 'reduction of individual fitness', and the response to it given by Gary Becker, who argues that altruism actually increases individual fitness, if 'indirect' unintended effects are also considered.[327] The philosopher Elliot Sober tries to provide an ultimate answer by arguing that both selfish and unselfish psychological motivations cause behaviour.[328] My interest in Mandeville and Hume's psychological theories is therefore historical and contemporary at the same time, for I looked into the modern origin of the debate between moral psychologies of universal self-interest, and those arguing for altruism, sympathy, counter-preferential choice and similar terms requiring unselfish motivations, that is, some form of self-denial.

Mandeville defines virtue as 'every Performance, by which Man, contrary to the impulse of Nature, should endeavour the Benefit of others, or the Conquest of his own Passions out of

[324] B. Mandeville (1732) vol. I., p. 28.

[325] A. Sen (1977).

[326] K. Binmore (2009).

[327] G. Becker (1976), pp. 282-294.

[328] E. Sober, and D.S. Wilson (1999), pp. 329-338.

a Rational Ambition of being Good.' Vice occurs when a person satisfying her or his own passions or appetites 'might either be injurious to any of the Society, or ever render himself less serviceable to others.'[329] Therefore, rational motivation for action distinguishes virtue from vice.

In spite of bringing public benefits some passions such as greed, envy, vanity and love for luxury are not virtuous but the basis of vice because they are motivated by self-interest, and also because 'skilful politicians' do the job implementing the laws and policies necessary for creating those benefits.

A more interesting case of vice are the actions caused by passions or emotions such as pity, compassion and parental love, which can bring public benefits but can also harm society. This is because actions caused by those emotions obey 'an Impulse of Nature, that consults neither the publick Interest nor our own Reason, it may produce Evil as well as Good. It has help'd to destroy the Honour of Virgins, and corrupted the Integrity of Judges'. Further, 'no Pity does more Mischief in the World than what is excited by the Tenderness of Parents, and hinders them from managing their Children as their rational Love to them would require.' [330] Hence, there is no virtue in such tender passions. First, because there is no reasoning helping discriminate between those actions which can bring public benefits from those that can harm the society. Second, because even if they do bring some public benefits, they are motivated by some expected reward such as fame, adoration and flattery.

Any definition of virtue and vice has two sides, positive and normative. A moral psychologist should be mainly concerned with the positive side. The psychologist of moral behaviour should engage with the normative debate only for purposes of clarification or feedback. To make this distinction may seem obvious and unnecessary, nonetheless both aspects are frequently conflated. The distinction between these two aspects in moral psychology

[329] B. Mandeville (1732), vol. I, p. 34, see also pp. 256, 294; and vol. II, pp. 106-107. Mandeville recognises that a full conquest of the passions is unfeasible, so the reduction of pride to a small expression of it can be enough for attributing self-denial, see vol. I, p. 43.

[330] *Ibid.*, p. 42 and 294, see also p. 68.

constitutes a main claim in this chapter. Moral psychology should be fully recognised as a separate scientific branch to be researched and developed by scientists with the use of scientific methods. It should not be researched or developed by philosophers, who should instead investigate and evaluate such knowledge with the tools and methods available in the philosophy of science. Otherwise, confusion can grow further spoiling the advances of scientific knowledge of moral behaviour.

For instance, Frederick Kaye in his comprehensive and still influential study on Mandeville finds tension, and even contradiction, between his definition of virtue and the refutation of public interest as the motivation for moral action. This happens because if no virtuous action exists where public interest is realised through self-denial, there seems to be no sense in keeping a definition of virtue which precisely relies on such self-denial. What is then the sense in keeping self-denial as criterion if self-interest is universal? For Kaye, the introduction of reason and self-denial is actually 'arbitrary' and 'superficial', a 'final twist given to his thought after it has been worked out in harmony with the opposite or empiric view'.[331] Besides Kaye, three other scholars also try to make sense of the contradiction they find between Mandeville definition of virtue and his psychology of universal natural self-interest. They do this by trying to establish the 'true' moral philosophy Mandeville held.

Like Kaye, Maurice Goldsmith also recognises the impossibility of distinguishing between vice and virtue once self-interest becomes universal. He explains that 'Mandeville's view destroyed the possibility of distinguishing between virtuous proportionate consumption and vicious overconsumption, and thereby undermined the notion that there was a settled social order such that what was proper and what was luxury in every rank could be determined.'[332] Once such a distinction collapses, there seems to be no purpose in holding a definition of virtue.

[331] F. B. Kaye (1924) 'Commentary Critical, Historical, and Explanatory', in *The Fable of the Bees*, Vol. I. p. liii.

[332] M. M. Goldsmith (2001), p. 135; see also p. 139.

Goldsmith explains that 'Mandeville presents the alternative of virtue as a logical rather than a practical possibility […] there is no logical inconsistency in holding Mandeville's strict definition of virtue and denying that it is ever exemplified in practice.'[333] Indeed, the alleged inconsistency disappears if virtue is declared inexistent. He argues that such a 'strict definition of virtue' was used rather rhetorically to prove the inconsistency between the predominant views in moral philosophy and public opinion.

Hector Monro also believes that virtue is 'empty' and that its discussion only carries aspirational purposes. He explains that 'since [virtuous] men do not in fact exist, the best we can hope for is that men will control their desires, so as to prevent them from conflicting with the good of others.'[334] Another scholar, Martin Scott-Taggart also accepts that there are no instances of virtuous behaviour but, unlike Kaye, he finds no contradiction in this because he argues that Mandeville's definition of virtue should be understood a 'regulative notion' just like Kant's regulative ideals.[335]

All four authors place their conclusions on virtue and vice as part of their arguments on the kind of moral philosophy Mandeville presumably supported and argued for. Thus, they use terms like 'strict', 'rigoristic', and 'ascetic' to describe the definition of virtue. I argue that the normative connotations of these terms, and the moral philosophies associated to them, prevent any understanding and evaluation of Mandeville's work as scientific, and therefore as positive. They all try to demonstrate that Mandeville was a moralist but none of them considers him as an early modern scientist. Consequently, they do not discuss any epistemic standards, methodological rules, or any other descriptive and evaluative tools from current philosophy of science or any relevant philosophy from the eighteenth century.

Kaye argues that Mandeville is both a moral anarchist and a utilitarian pioneer, for he holds relativist views 'in theory', while 'in practice' actual moral behaviour was judged as

[333] *Ibid*, pp. 57, 148-149.
[334] H. Monro (1975), p. 237, see also p. 135, 189.
[335] J. M. Scott-Taggart (1966), p. 232.

virtuous by its effects on society, i.e. whether they increased or decreased general happiness. There is, therefore, a contradiction between theory and practice, and because Kaye believes that the morally rigoristic definition of virtue is arbitrary, he solves the problem by arguing for the retention of 'useful vice' and the rejection of 'harmful vice'. Once this is done, we can conclude that Mandeville was 'practically, if not always theoretically, a utilitarian'[336]

Monro agrees with Kaye, although he goes further by including both vicious passionate action and virtuous 'dispassionate concern for the public good' as 'devices' in a utilitarian moral philosophy.[337] Goldsmith also believes that Mandeville remained on both sides and served utilitarian purposes by 'retaining definitions of virtue which required acting without a consequentialist motive'.[338] Against Kaye, Scott-Taggart argues that Mandeville was not a utilitarian moralist but a moralist holding mild ascetic views. He claims that if individuals want to be consistent, and if they hold an abstract ideal of moral good, they would learn from their mistakes and try to meet this ideal by shifting from vice to virtue, that is, from hedonism to a mild asceticism.[339]

I will not discuss the views from these authors any further because I do not believe there is enough evidence to vindicate Mandeville as a moral philosopher, and I do not think that it is necessary either. Mandeville did not argue for a moral philosophy in any substantive manner, nor did he develop an independent argument that favours some moral philosophy. I believe that any attempt at trying to establish a Mandevillian normative ethics as utilitarian, ascetic or combining both is mistaken, and it is also alien to the motivations and purposes stated by Mandeville himself. In his work, there is an unambiguous commitment to explaining moral behaviour by producing a psychological theory. Evidence from both volumes of *The Fable* and *An Enquiry Into the Origin of Honour and the Usefulness of Christianity in War*, show that he fully adopted a scientific attitude and used a scientific method for his research into moral psychology

[336] *Op. cit.*, p. lxi.
[337] *Ibid.*, pp. 238, 247.
[338] *Ibid.*, p. 148
[339] *Ibid.*, pp. 224-228

and some related social aspects, which led him to a functional explanation of vice and an explanation of virtue by design.

4.5. Functional vice

Kaye, Goldsmith, and Monro do not discuss the existence of a large number of poor workers with low wages and no education, which Mandeville recognises as necessary for a booming commercial economy based on large exports and an increasing domestic demand for products from the rich and middle classes. This omission causes a significant selection bias in their arguments because they accept regulated vice as useful, and therefore as consistent with utilitarianism, without supplying also a utilitarian argument for the justification of low wages and poor education. For the same reasons, the consideration of an ascetic morality by Goldsmith and Scott-Taggard is also mistaken. Just like an increase in the number of charity schools for the poor is inconsistent with a booming commercial economy, a moral conversion of the rich and the middle classes from hedonism and consumerism to asceticism and frugality is equally inconsistent, because it would lead to a drastic reduction in the local demand.

Mandeville explains how self-interest and vice curbed by the government, poor education and low income are all *functional* to a thriving commercial society. Greed, envy, vanity, and love for luxury are all examples of vice. Because vice is *useful* it is turned into a virtue within utilitarian moral philosophy. The study of the consequences of individual or collective moral behaviour on the society as a whole is often associated with utilitarian arguments in philosophy, and related areas of policy-making in economics and political science. However, there is no necessary connection between the two. Although utilitarianism relies on functionalism, utilitarianism and functionalism are not equivalent. The first one is a moral philosophy; the second is a method in the social sciences. Mandeville's extensive and detailed study of unintended consequences must be philosophically analysed and evaluated as a case functional explanation in the social sciences. As it was discussed earlier, Mandeville built moral psychology by drawing an analogy with the anatomy and physiology of the human body. This is important because explanations in physiology are functional explanations, and because they are taken from medicine and biology, functional explanations are a case of methodological naturalism in the social sciences.

Unlike the inferences made by Mandeville on unobservable entities and processes in the mind, which can be evaluated using normative standards from the eighteenth century, namely those from the *vera causa*; functionalist explanations cannot be evaluated using criteria from the same period because there are none that are appropriate for the task. After the publication in 1543 of Andreas Vesalius's book on anatomy *De Humani Corporis Fabrica* the distinction between anatomy and function was not very clear yet. Later, in 1548 the publication of William Harvey's book *Exercitatio Anatomica de Motu Cordis et Sanguinis in Animalibus* on the heart and blood circulation laid the initial grounds for the separation between anatomy and function, which were fully developed and assimilated decades later. In 1641, Nicolaes Tulp published his influential book on anatomy *Observationes Medicae*; Tulp studied medicine at the University of Leiden, where he later became a professor. Encouraged by his father, who was a reputed physician in Rotterdam, Mandeville went to Leiden to study medicine and philosophy eleven years after Tulp's death in 1674.[340]

In the *The Fable of the Bees* and other works, Mandeville actually produced functional explanations without using any explicit functionalist or physiological vocabulary. This can be explained because only a century later physiology emerged as a separate branch in medicine with the work of Claude Bernard. Nonetheless, already in the sixteenth, seventeenth and eighteenth centuries functional explanations were part of medicine.[341]

Due to the lack of methodological standards from that period on functional explanations, parts of current criteria can be used. Functionalism in the social sciences has

[340] See T. V. N. Persaud (1997); J. F. Fulton (1966); K.E. Rothschuh (1973).

[341] See M. Foster (1901).

grown in size and sophistication since the publication of the works from Bronislaw Malinowski and Alfred Redcliffe-Brown, and more recently those from Talcott Parsons, Robert Merton and others. Many of the methodological standards set by these authors surpass any standard applicable to Mandeville's work, nonetheless the basic but powerful distinction made between latent and manifest functions and the basic model of a functionalist explanation can both be used.

Carl Hempel identified four basic components in a functional explanation: a whole or system (*S*), the internal and external conditions around the system (*C*), a persistent trait in that whole (*T*), and the effects from such a trait which meet a need (*N*) in that whole or system. If the system functions adequately, it is inferred that the effects of the trait have met a certain need or condition necessary for the adequate continuous functioning of the system.[342] This is the basic logical form of functional explanations:

Explanans:

• At time *t*, *S* functions adequately in conditions *C*.

• *S* functions adequately in conditions C only if condition *N* is satisfied.

• If trait *T* were present in S then, as an effect, condition *N* would be satisfied.

Explanandum: At time *t*, trait *T* is present in *S*.

Using this model, Mandeville's explanation of vice in England can presented now as a functional explanation:

[342] C. Hempel (1965), p. 306-310; see also D. Little (1991), p. 94.

Explanans:

- In 1723, the English economy was booming having a large number of poor uneducated individuals and skilful politicians in government, while foreign demand for English woollens and manufactured goods was rapidly increasing.

- The English economy is booming having a large number of poor uneducated individuals and skilful politicians in government, while foreign demand for English woollens and manufactured goods is rapidly increasing; only if employment, wages consumption and exports are having a large increase.

- If every person in England were a self-interested and vice were widespread then, as an effect, employment, wages, consumption and exports would increase.

Explanandum: In 1723, every person in England was a self-interested and vice was widespread.

In this functional explanation, *S* is 'The English economy is booming', and *C* is 'To have a large number of poor uneducated individuals, skilful politicians in government and a rapidly increasing foreign demand for English woollens and manufactured goods'. *N* is 'Employment, wages, consumption and exports have a large increase', and *T* is 'Every person is a self-interested and vice is widespread'. Hempel identifies a number of problems in functional explanations such as the fallacy of affirming the consequent, and he makes a number of recommendations on how to fix these problems, for instance by prescribing the use of general laws and functional equivalents. It is not necessary to describe these recommendations in more detail at this point. All that is epistemically needed in this case is to demonstrate that a persistent trait such as self-interest and widespread vide produced unintended effects or consequences, which become functional because they satisfy vital needs of the whole society, namely the creation of jobs and a rise in wages, consumption and exports.

A further appreciation of the functional character of Mandeville's explanations can be reached by comparing it to the functional explanation made by Robert Merton on the ceremonial rain dance performed by the Hopi:[343]

Explanans:

- In 1947 the Hopi tribe remains strong, having an economy based on farming and a government with representatives from each village forming a council, while it is surrounded by Navajo tribes.
- The Hopi tribe remains strong, having an economy based on farming and a government with representatives from each village forming a council, while it is surrounded by Navajo tribes, only if social cohesion and group identity are promoted.
- If a large number of individuals from the Hopi tribe people were to participate in the rain dance, then social cohesion and group identity would be promoted.

Explanandum: In 1947, the Hopi regularly performed the ceremonial rain dance.

Both the rain dance performed by the Hopi and the self-interested actions with widespread vice performed by the English have unintended consequences, which are beneficial to both societies. Both explanations produce knowledge of functional properties certain traits have, which are not obvious but unexpected and surprising. Merton distinguishes between obvious or manifest functions and latent ones. He explains that '*Manifest* functions are those objective consequences contributing to the adjustment or adaptation of the system which are intended and recognized by participants in the system', and '*Latent* functions, correlatively, being those which are neither intended nor recognized.'[344] The manifest function of the Hopi dance is to bring rain, while the latent function is to promote social cohesion and group identity. The manifest function of self-interest and vice is to satisfy individual needs, desires and preferences, while the latent one is the creation of jobs and a rise in wages, consumption and exports.

[343] R. Merton (1968). pp. 110, 118-121.

[344] *Ibid.*, p. 105; see also R. Merton (1936).

Merton traces the identification of unintended consequences in the social sciences back to Adam Smith and others after him without acknowledging Bernard Mandeville. Such an omission is not uncommon, neither Smith nor others gave credit to Mandeville for his scientific finding, namely the unexpected functional relationship between public benefits and private vices.

Twenty years later, David Hume produced the first utilitarian definition of virtue based on how useful, agreeable or pleasing actions are; this definition set the initial grounds for later developments of a utilitarian moral philosophy.[345] By a simple definitional change, the whole balance between vice and virtue changes since everything Mandeville defined as vice becomes virtue.[346] Hume did not dispute the overall inner workings of passions and the need for constant reward as described by Mandeville, and he did not challenge either the functional relationship between public benefits and private vices and self-interest. He only gave a new different interpretation to the same set of empirical observations and findings made by Mandeville.

With his findings on moral psychology and functional explanations, Mandeville forced the utilitarian solution to the paradox, which gave rise to a new moral philosophy. Without such empirical findings and explanations, it would not have perhaps been possible for Hume and others to contend that usefulness or utility are morally good and self-interest is a virtue. Similarly, it would not have perhaps been possible for Adam Smith to produce the metaphor of the invisible hand leading self-interest towards the creation of public benefits. Mandeville was an early modern forerunner of both functionalism and moral psychology; upon his scientific work, utilitarian moral philosophy and the related economic theory were built.

[345] D. Hume (1772), *SBN*268; see also A. Smith (1790), pp. 178, 188-189; and H. Sidgwick (1907), pp. 423-426.

[346] See A. Atkinson (1998) and A. Sen (1981) for examples on the significant effect different definitions have on measurement in the social sciences.

4.6. Virtue by design

It was discussed earlier, all four authors, Kaye, Goldsmith, Monro and Scott-Taggard hold that virtue as defined by Mandeville is either inexistent or a mere rhetorical device. Unlike Kaye, Goldsmith and Monro leave open the possibility of having cases of virtue open but without explaining how such cases can be found. Scott-Taggard argues that reflection guided by a regulative ideal would eventually lead to virtue. These ideas were advanced as a solution to the negative consequences created by the definition virtue as self-denial.

Evidence from Mandeville's work shows that he did not believe virtue was inexistent. The opposite conclusion reached by the aforementioned authors can be explained by a failure to notice the role of education as a rudimentary form of social engineering. The distinction between *natural* and *artificial* or *artefactual* behaviour is crucial for this. Mandeville distinguishes between 'taught' and 'untaught nature', that is to say, between natural and artefactual. He explains that a moral psychology of universal natural self-interest and passions driving human behaviour implies the artefactual character of society and its constitution as a polity. Mandeville uses the term 'passions' instead of the terms 'emotions' or 'sentiments', which currently are more common, he explains that passions such as pride, shame and fear are continuously used for shaping human behaviour. For instance, good manners, willingness to fight in a war, business partnerships, modesty and restraint of women, marriage and the ability to conceal certain passions or emotions are not natural but artefactual, and they all are produced and maintained by causing intense fear and shame, and a strong sense of pride and worthiness. It all is an intense process of rudimentary social engineering, which starts very early in infancy with the use of flattery, threats, punishments and rewards as technologies for shaping human behaviour.[347]

A more challenging task consists of restraining self-interest and the passions by using rational discussion as a technology, which anticipates one of tenets of Enlightenment in

[347] B. Mandeville (1732), vol I. pp. 54-59, 63, 69, 252-153, 172, 200.

continental Europe. In *An Enquiry into the Origin of Honour and the Usefulness of Christianity*, Mandeville explains:

> I am willing to allow, that Men may contract a Habit of Virtue, so as to practice it, without being sensible of Self-denial, and even that they may take pleasure in Actions that would be impracticable to the vicious: But then it is manifest, that this Habit is the Work of Art, Education and Custom; and it is never acquired, where the Conquest over the Passions had not being already made.[348]

Already in the second volume of the *The Fable of Bees*, Mandeville explains that individuals can be virtuous 'by Reason and Experience; and not by Nature, I mean, not by untaught nature.'[349] His definition of virtue is consistent with his refutation of Lord Shaftesbury's moral psychology because Mandeville was refuting the claim that individuals perform actions in the public interest *by nature*. In other words, he was refuting the claims that self-denial is natural. The distinction between 'untaught nature' and 'taught nature', that is, between natural and artefactual behaviour, dissolves any possible inconsistency or puzzle between the definitions of virtue and vice. Mandeville did not state virtue was inexistent and he did not appeal to logic to solve the problem but to education. When he explicates that virtue is 'the Work of Art, Education, and Custom', he is highlighting the great challenge skilful politicians or other leaders face designing and enforcing any law or moral rule either by relying on passions such as shame or pride, or by relying on rational discussion. Virtue as self-denial is an artefact; it is the product of art and education.

Mandeville discusses one case only where virtue as self-denial is realised. He finds this case in religion, particularly in the Roman Catholics from the eighteenth century and before, whose behaviour exhibits a comparatively larger proportion of self-denial than Protestants or Muslim Turks. He explains that 'all *Roman Catholicks* are brought up in the firm Belief of the

[348] B. Mandeville (1732b), pp. x-xi.

[349] B. Mandeville (1732) vol. II. p.106.

Necessity there is of Self-denial. They are strictly forbid to eat Flesh on Fridays; and Pains are taken to inspire them from the very Childhood with a Horror against the breaking of Commandment. It is incredible, what Force such a Precept is of, and how the Influence of it sticks to Men.' Mandeville explains this successful case of self-denial by referring to the skills and knowledge Catholic cardinals and priests had of what can be described as an accurate and successful folk moral psychology, which despite not being scientific was highly effective. The 'Architects of the Church of *Rome*' created the moral rules 'most difficult to comply with', and were able to get large multitudes to actually follow those rules, 'not only by Words and in Theory, but by Practice and Example.'[350]

Mandeville acknowledges the existence of vice among Catholics but still finds the achievements of the Catholic religious leaders impressive for curbing strong passions such as sex, hunger and craving for meet. For analogous reasons, he criticised the moral philosophies of Stoicism and that of Lord Shaftesbury, which lack the knowledge of the social technologies needed for producing and maintaining the virtuous behaviour they argue for. Because of the great success Catholic leaders had in realising the moral philosophy of Roman Catholicism, he attributes to them the best knowledge and skills needed for shaping the moral behaviour of large multitudes. For instance, he points out to the power of 'the Fear of an invisible Cause'[351] has shaping the behaviour of large multitudes represented by the devil in different forms, and through different means of invisible evil actions. The use the devil was supplemented with the skilful design of further devices such as the introduction of a large number of saints people could pray asking for help and protection, and a number of rituals such as the confirmation, the first communion, the anointing of the sick and many others.

The case of Catholic behaviour in the eighteenth century and before could be criticised as positive evidence of virtue as self-denial; the lack of further evidence could also be criticised

[350] B. Mandeville (1732b), pp. 110-112.
[351] B. Mandeville (1732), vol. II, p. 235-237.

but not the lack of any evidence. Kaye, Goldsmith, Monro and Scott-Taggard mistakenly denied the existence of any positive evidence put forward by Mandeville. This negative conclusion justified the search for the purpose and the moral philosophy that could make sense of the choice of the definition of virtue as self-denial. The discussion above shows that such a choice was justified and motivated by the idea of making a clear contrast between two Christian moral standards, namely the new emerging utilitarian moral philosophy associated to Protestantism, and the old moral philosophy of Roman Catholicism.

Because of the success Catholic religious leaders had shaping behaviour by creating all those invisible causes of evil, worship figures and rituals, they proved to have reliable knowledge of moral psychology and related social aspects. This success is analogous to that of the skilful English politicians in the design of laws, taxes, incentives and penalties during the eighteenth century. In both cases self-interest and the passions were successfully shaped in different proportion and through different means with the purpose of creating a booming economy based on self-interest, and a widespread religion based on self-denial. Both virtue as self-denial and virtue as self-interest curbed with sympathy are artefacts. English politicians and Catholic leaders did not have scientific knowledge as we have it now; nonetheless their skills and outcomes can be considered as a successful case of folk design and social engineering.

4.7. Design for self-interested knaves

Besides erecting moral psychology and producing important functional explanations with an emphasis on the unintended consequences of moral behaviour, Bernard Mandeville also advanced some ideas on design. He did this by arguing for the knaves principle, and by discussing some general ideas on a very basic blueprint for a commercial precapitalist society. Both the knaves principle and his ideas on such a blueprint are based a moral psychology of self-interested passion-driven individuals, who are also knaves. 'Knavery' is generic term Mandeville uses for referring to any kind of behaviour which remains legal despite being dishonest or corrupt. This kind of behaviour was discussed in chapter two as one of the main concerns in current mechanism design theory, there the term 'cheating' was used instead also with a generic sense. For instance, the economist Leonid Hurwicz explains that the rules in any design concerned with an efficient allocation of economic resources 'requires that no one should find profitable to "cheat", where cheating is defined as behaviour that can be made look "legal".'[352]

Because self-interest and cheating or knavery are universal, it is important to design laws and policies that also regulate the behaviour of politicians, this is what the knaves principle does. Later, this principle was also discussed and adopted by David Hume, and more recently it has also adopted in public choice theory. The origin of the knaves principle can be traced back to 1720, when Mandeville published his essay 'Of Government'. In this essay, he criticises the power kings and their close relatives had to rule in an absolutist fashion, and argues instead for a constrained monarchy where the rule of law is applicable not only to all subjects but also to all three branches of government, namely the Crown, the Lords and the Commons. Mandeville explains how the three forms of government (monarchic, aristocratic and democratic) coexist and are actually mixed in the British government. Analysing this mixed form of government, he pays special attention to the design of constitutional laws for the prevention of tyranny by removing prerogatives of the crown, and by allocating more power to parliament.

Abuse in the use of royal prerogatives such as the arbitrary provision of protection and privileges to individuals by exempting them from the rule of law, must be prevented in any constitutional contract, this is because 'all Persons are accountable for their own Actions, and that no Order of the King, how plain or express soever, tho' produced in writing, and corroborated with his Sign Manual, can extenuate a Man's guilt, much less exempt him from it,

[352] L. Hurwicz (1972), p. 445.

if in executing that Order he has acted against the Law.'[353] Mandeville wanted to prevent cases like that that of James II in England, whose dictatorial policies abolished the right to vote for members of parliament as well as the right to hold Protestant beliefs. These actions led to the Glorious Revolution in 1688, and the eventual coronation of William of Orange and his wife Mary as King William III and Queen Mary II.

The reign of James II came after the restoration of monarchy in England in the aftermath of the English Civil War, and the short-lived republican commonwealth under the protectorate of Oliver Cromwell. It was a time with continuous political turmoil and prolonged social unrest with all events taking place within a short period of five decades. It constitutes the historical background in which Mandeville argued for the knaves principle and published the essays where he defends individual liberties, religious tolerance and the empowerment of parliament, particularly of the Commons and of non-rich Lords. As a result, the knaves principle was advanced as a principle for constitutional design:

> I have often heard well-meaning People say, that would every Body be honest ours is the best Constitution in the World. But this is no Encomium, where every Body will be honest and do their Duty, all Governments are good alike. That is the best Constitution which provides against the Worst contingencies, that is armed against Knavery, Treachery, Deceit, and all the wicked Wiles of humane Cunning, and preserves itself firm and remains unshaken, though most Men should prove Knaves. It is that which can bear most Fatigues without being disorder'd, and last the longest in Health, is the best.[354]

Mandeville points out kings and queens as the most conspicuous self-interested knaves, who use any prerogative for trying to maximise their own preferences in the social order and the allocation of burdens and benefits to specific groups. Queen and kings may engage in matters of

[353] B. Mandeville (1720), p. 302; all page numbers cited from this book correspond the page numbers from the original publication, which are placed in square brackets in the 2001 edition by Irwin Primer.
[354] *Ibid.*, p. 297.

public interest and they may eventually play by the rules only because of considerations of self-interest. Mandeville believed that their knavish and absolutist political tendencies may become beneficial to the wider population, if appropriate constitutional constraints are designed and implemented for a more equal distribution of political power and land.

While some prerogatives may be considered necessary and therefore may be granted to the king, 'he has not one that can make him Tyrant, or his Subjects Slaves. The Rights and Privileges of Parliament, and the Liberty of the People are as Sacred Branches of the Constitution as any thing the King can claim.'[355] Echoing James Harrington,[356] Mandeville also claims that 'Dominion always follows Property, and that, where the one is wanting, it will ever be impracticable for any long Continuance to enforce the other.'[357] Therefore, when new constitutional reforms are designed, excessive property and excessive power in any of the three branches of government must be prevented. Mandeville highlighted the increasing number of Lords joining the parliament who were not rich, and estimated that at the time three quarters of the total land was owned by the Commons. Such a distribution would prevent any coalition between the King and the Lords from overpowering the Commons both economically and politically.

It is impossible for designers to anticipate and prevent all kinds of knavish behaviour when they are making a blueprint for a new law or a new constitution. As a generic principle for design, the knaves principle meets the requirements of large scope and good knowledge by spotting an important problem and suggesting how to solve it. The design of supplementary rules and laws is therefore required to deal with specific cases. For instance, on a more specific level, Mandeville emphasises the risk of not removing the veto power granted to the Crown as part of its prerogatives: 'Representatives of the People are come on a very foolish Erran[d], if there is another Power upon Earth, that without their Consent can make void, and with

[355] *Ibid*, p. 305.
[356] J. Harrington (1656), p. 91, see also pp. 287-193.
[357] *Ibid*, p. 314.

impunity annul, perhaps the next Day, what they have been enacting with so much Solemnity, and after so mature a Deliberation.'[358]

The knaves principle is a response to historical events in England and parts of Europe, and it also reflects Whig values. [359] The case of tyranny and absolute power exerted by the monarchies and the subsequent constitutional changes made leading to parliamentary and republican systems, show how the anticipation and effective prevention of undesirable scenarios devising new rules are largely determined by past and related experiences as well as on the ability to learn from them. Since the time it was published, the knaves principle has been highly praised by philosophers and scientists alike as a fundamental principle of design in the social sciences.

Currently, the knaves principle still captures our basic understanding and fears on political abuse and corruption in representative democracies with large bureaucracies. This is one of the reasons why this principle still draws the attention of economists, political scientists, and philosophers. The knaves principle has been incorporated into public choice theory and more widely into game theory, behavioural economics and political theory by economists and political theorists such as Ken Binmore[360], James M. Buchanan and Geoffrey Brennan[361], Alan Hamlin[362], and Bruno Frey;[363] and also in public policy by Julian Le Grand[364]; and in philosophy by Philip Pettit[365], Daniel Hausman[366] and David Gauthier.[367] The current impact and influence of this principle requires separate research, the current purpose has only been to trace its origin and application to design back to the eighteenth century.

[358] *Ibid*, p. 301.

[359] See M. M. Goldsmith (2001), pp. 99-100, 112.

[360] K. Binmore (1998), p. 272; and K. Binmore (2005), p. 136.

[361] J. M. Buchanan and G. Brennan (1985), pp. 67-68; and J. M. Buchanan and G. Brennan (1983).

[362] G. Brennan, and A. Hamlin (2000), pp. 6-10, 61-66,

[363] B. Frey (1997); and B. Frey, A. Stutzer, and M. Benz (2001).

[364] J. Le Grand (2003).

[365] P. Pettit (2002), pp. 275-307; and P. Pettit (1997), pp. 212-230.

[366] D. Hausman (1998).

[367] D. Gauthier (1982); and D. Gauthier (1992).

4.8. Blueprint for a precapitalist commercial society

In the eighteenth century England was a precapitalist commercial society. Besides studying the moral psychology of England during that period, Mandeville also identified some of its main structural features, which can be described as a basic blueprint of this type of economy and society. Because this economy emerges from a feudal one the blueprint describes two stages. In the first stage self-interest and other relevant passions remain 'dormant', so the first task consists of activating them. In the second stage they are curbed preventing any excess that can be harmful to the economy and the society.

Mandeville compares two types of society: one is a lumpish machine; a frugal and honest society that is idle, innocent and ignorant with an opulent and rich society, where the economy is booming and the arts and sciences flourish. He explains that without precapitalist reforms society remain 'poor, ignorant, and almost destitute of what we call the Comforts of Life, and all the Cardinal Virtues together won't so much as procure a tolerable Coat or a Porridge-Pot among them: For in this State of slothful Ease and stupid Innocence, as you need not fear great Vices, so you must not expect any considerable Virtues. Man never exerts himself but when he is rous'd by his Desires: While they lie dormant […] the lumpish Machine, without the influence of his Passions, may be justly compar'd to a huge wind-mill without a breath of Air.'[368] Hence, the challenge for the designer consists of producing a blueprint with new laws and moral rules that can activate those dormant passions and desires.

It is the task of skilful politicians as designers to create and pass those new laws and policies necessary to turn a feudal society into a precapitalist commercial one. Because Mandeville argues that sloth and greed are both self-interested passions, self-interest exists in both types of societies. While self-interest, other passions and desires in a feudal society remain to a certain extent dormant, in a commercial they are stimulated and turned into greed, vanity,

[368] B. Mandeville (1732) vol. I, pp. 199-200.

love for luxury and a stronger eagerness for work. This the basic blueprint produced by Mandeville for achieving this transformation:

> Would you render a Society of Men strong and powerful, you must touch their Passions. Divide the land, tho' there is never so much to spare, and their Possessions will make them Covetous: Rouse them, tho' but in Jest, from their Idleness with Praises, and Pride will set them to work in earnest: Teach them Trades and Handicrafts, and you'll bring Envy and Emulation among them: To increase their Numbers, set up a Variety of Manufactures, and leave no Ground uncultivated; Let Property be inviolably secured, and Privileges equal to all Men; Suffer no body to act but what is lawful, and every body to think what he pleases […] Would you have them bold and Warlike, turn to Military Discipline, make good use of their Fear, and flatter their Vanity with Art and Assiduity: But would you moreover render them an opulent, knowing and polite Nation, teach 'em Commerce with Foreign Countries, and if possible get into the Sea, which to compass spare no Labour nor Industry, and let no Difficulty deter you from it: Then promote Navigation, cherish the Merchant, and encourage Trade in every Branch of it; this will bring Riches, and where they are, Arts and Sciences will soon follow.[369]

Reforms on property rights, exploitation of land and crop production, diversification of manufacturing, and expansion of domestic and foreign trade as well as readiness for war, they all rely on a folk or scientific social engineering, which can successfully make the transition to precapitalist society by reshaping self-interest and other passions and desires. With this programme of reforms Mandeville sketches a blueprint for building a rich, warlike, commercial society inspired mainly by the Dutch and the English societies of the seventeenth and eighteenth centuries. These reforms are still fashionable; they are almost the same to those implemented in Russia, some East European countries and Chile as it was discussed in chapter two.

[369] *Ibid.*, pp. 200-201.

Societies play different games under different social contracts. Different social contracts rely on moralities of different kind. The contrast made by Mandeville between a frugal and ignorant agrarian society and a lavish and rich commercial society pose a challenge to the institutional designer, who has to devise and implement the necessary reforms to transition from one social contract to the other preventing chaos and a social collapse. The ultimate aim is to have a well-ordered commercial precapitalist society, a rich but vile machine which runs steadily and fairly efficiently; a machine whose 'Vileness of the Ingredients that all together compose the wholesome Mixture of a well-order'd society'.[370] Such ingredients are described in analogy with a 'Bowl of Punch':

> Avarice should be the Souring and Prodigality the Sweetening of it. The Water I would call the Ignorance, Folly and Credulity of the floating insipid Multitude; while the Wisdom, Honour, Fortitude and the rest of the Sublime Qualities of Men, which separated by Art from the Dregs of Nature the fire of Glory has exalted and refin'd into a Spiritual Essence, should be an Equivalent to Brandy.'[371]

Satirical writing in Mandeville becomes the means for communicating a scientific stance on human psychology, which is dispossessed of any romanticism or concession. This is important because he criticises that 'most Writers are always teaching Men what they should be, and hardly ever trouble their Heads with telling them what they really are.'[372] Mistakenly, some social scientists call describe this position as 'realist',[373] when it should simply be described as 'positive', at least as positive as science can be considering the effect of values. Such a 'realist' or positive stance has been adopted by the economist James M. Buchanan, founder of public choice theory, who calls economists and political scientists to do science 'without romance'. Buchanan explains that he has produced this kind of science thanks to the 'rediscovery' of 'the

[370] *Ibid.,* p. vi.
[371] *Ibid.*, pp. vi, and 106.
[372] *Ibid.*, p. 25.
[373] See E. H. Carr (1945).

methods of the eighteenth century philosophers such as Mandeville, Hume, and particularly Adam Smith.'[374]

Part of the satirical or 'realist' attitude of Mandeville may be explained because of his Whig position in politics. Ken Binmore, a contemporary Whig economist, shares a similar attitude and moral psychology. Binmore explains that a Whig is a leveller who avoids the naivety of the Left and the recurrent crisis of the markets by introducing some planning and continuous reform. He adopts self-interest as a fundamental moral assumption in game theory and warns designers on the naivety and self-deception of those designs that do not rely on penalties for enforcing compliance with the rules. He argues that 'cattle prods' must be used by the social designer preventing excessive deviation and defection. Like Mandeville, Binmore also believes that anger, pride and envy are essential elements of human sociality. [375]

From a positive perspective, Mandeville made the first important and long lasting contributions to moral psychology, which he placed against mistaken 'romantic' psychologies like that of Lord Shaftesbury and Stoicism, and he also established important functional relationships between individual behaviour and large scale social effects. He exposed the functional need for poverty and ignorance in a rich commercial precapitalist society, and also the need for producing widespread fear as the means for keeping the population in a permanent state of alert in preparation for any eventual war. In the eighteenth century, Mandeville and his moral psychology were described by David Hume, Adam Smith and others as 'malignant philosopher', 'evil' and lacking 'feelings of humanity'.[376] This description can be extended to current design in economics and political science, which virtually use the same moral psychology as it was shown in chapter two. Large parts of current design and the moral psychology supporting it are a Mandevillian legacy. Because of the positive nature of moral

[374] G. Brennan and J. M. Buchanan (1981), p. 163; see also J. M. Buchanan (1979).

[375] K. Binmore (1994), p. 145; K. Binmore (1998), pp. 47, 120., 378; and K. Binmore (2004), pp. 186-187. In his first works Binmore assumes enlightened self-interest or just self-interest; recently in K. Binmore (2009, pp. 19-22) he is critical of those psychological assumptions and has now adopted the method of revealed preference theory.

[376] D. Hume (1772), pp. SBN215, SBN243, SBN302; A. Smith (1790), p. 308.

psychology and design, designers in the social sciences can work for any god or any demon; they are highly qualified technocrats who can work for Left and the Right, for the Libertarian and the Totalitarian, and for any intermediate position between them.

—O—

Chapter 5

Self-Interested but Sympathetic

5.0. Introduction

David Hume has been largely read as a philosopher but not as a scientist. In this chapter I discuss his work exclusively as a case of science; in particular as a case of early modern science. I examine how the moral psychology of self-interest, sympathy and sentiments of humanity he argues for fits with his descriptive sociology of the utilitarian morality in Britain during the eighteenth century. I compare the moral psychology from Hume to the moral psychology from Mandeville by presenting it as a case of theory choice. I present the ideas on design and engineering, which can be extracted from Hume's work, and I discuss the objections he made to the egalitarian distributive justice advanced by James Harrington and by the Levellers.

David Hume, Adam Smith and Francis Hutcheson regarded the psychology of self-interest advanced by Thomas Hobbes and Bernard Mandeville as the rival theory to be defeated; it was a theory 'making so much noise in the world', Smith reports[377]. Hume positively praises Mandeville's theory in the first pages of the *Treatise*. This volume pays more attention to self-interest than *An Enquiry Concerning the Principles of Morals*, in which Hume discusses the disinterested passions of humanity and benevolence in more detail. Both Hume and Smith were highly critical of Mandeville's theory, which they considered to be 'wholly pernicious' because it leaves no grounds for 'feelings of humanity'. They actually allude to Hobbes and Mandeville with epithets such as 'sportive sceptic' and 'superficial reasoner', who created a 'malignant philosophy' and whose reasoning is 'ingenious sophistry'.[378]

Hume strongly criticises the self-interested individuals described by Mandeville who are 'monsters' 'unconcerned, either for the public good of a community or the private utility of

[377] A. Smith (1790), p. 313.

[378] A. Smith (1790), p. 308; D. Hume (1772), SBN215, SBN243, SBN254 and SBN302.

others'[379]; they are replicas of Ebenezer Scrooge who even at Christmas shows no humanity, no concern for others. In contrast, Hume describes a polite, sympathetic and utilitarian individual who, despite being self-interested, is capable of performing acts of disinterested benevolence and humanity, that is, a Scrooge who is morally reformed by secularised Christian values. Hume not only claims that true disinterested charity and beneficence exists grounded in the natural sentiments of humanity, but he also claims that these sentiments can 'overpower' and 'over-balance' self-interest.[380]

The debate between the psychology of self-interested knaves and the psychology of self-interested sympathetic individuals is about *true causes* or *true motives* of moral behaviour. This debate is concerned with the 'metaphysical' part of moral psychology. Using the same naturalistic analogy from Mandeville, Hume explains that the moralist is a painter who is concerned with the beauty of moral behaviour, portraying it with 'the most graceful and engaging airs', whereas the moral anatomist is concerned with 'the most hideous and disagreeable' parts analogous to the 'the inward structure of the human body, the position of the muscles' and 'the fabric of the bones'.[381] Hume did not consider himself to be a moralist but a moral anatomist; he did not write any substantive normative moral argument. This is why the moral philosopher Francis Hutcheson criticised the *Treatise* because of its lack of 'Warmth in the Cause of Virtue'.[382]

Because the different passions and sentiments and their mutual operations cannot be observed, such a moral anatomy becomes metaphysics in search for the 'hidden truths' and 'the secret springs and principles' of the inward parts of human nature, which can only be discovered by 'painful' and 'abstruse' enquiry.[383] Methodologically, the production of this new science represented a great challenge because of the difficulty of producing accurate and reliable

[379] D. Hume (1772), SBN235.

[380] *Ibid.,* SBN276; see also D. Hume (1739-40), SB487.

[381] D. Hume (1748), SBN10; D. Hume (1739-40), SB620-621.

[382] Greig, J. Y. T. (1932), Letter 13, 17th September 1739.

[383] D. Hume (1748), SBN6, SBN10.

knowledge of unobservable entities and processes in the mind by relying on observable behaviour. Indeed, Hume wanted to build this new 'science of man' as a 'true metaphysics' of human nature based on 'experience and observation.'[384]

In the introduction to the *Treatise* and the opening section of the *Enquiry*, Hume states his commitment to the observational and experimental method with an explicit reference to Francis Bacon, and also by quoting Isaac Newton. He considers the introduction of the experimental method as a key methodological innovation in the study of morality, which would allow one to treat 'passions, motives, volitions, and thoughts' as 'matters of fact' existing 'in the mind' just as it is done in physics with properties such as 'sounds, colours, heat, and cold', so that 'discovery in morals, like that other in physics, is to be regarded as a considerable advancement of the speculative sciences'.[385] In line with this methodological commitment, he explicitly appeals to three epistemic criteria in his attempt to refute Mandeville's psychological theory, namely inductive support, *experimentum crucis* and simplicity. All this seems to provide enough evidence for evaluating Hume's work as case of early modern science. Therefore, the new science of man he was erecting must be evaluated by looking into its epistemic merits and methodological grounds.

In section 5.1., I present the controversy between the moral psychologies from Hume and Mandeville as a problem of theory choice. I use four criteria for comparing them, namely the *vera causa* principle, inductive support, *experimentum crucis* and simplicity. On all four criteria the moral psychology of universal self-interest from Mandeville performs better than the moral psychology of self-interest and sentiments of humanity from Hume, so the first one should be chosen as the true or best supported theory. A main problem for Hume's theory is the lack of a refutation of possible self-interested motivations that explain the cases of disinterested humanity and generosity. He relies on a simple enumerative induction with a few cases only put

[384] D. Hume (1748), SBN12; and D. Hume (1939-40), SBNxv-SBNxvi.

[385] D. Hume (1739-40), SBN469.

forward as evidence. Another important problem is the folk psychology he relies on for building his own theory.

Section 5.2. consists of two main claims. First, I argue that Hume produced a basic descriptive moral sociology of utilitarian morality. This is against the characterisation of his work as moral philosophy. I do this by showing his explicit commitment to building a moral science, and his rejection of the normative method from moral philosophy. Such a descriptive moral sociology consists of two major empirical claims, namely the existence of the new moral principle of utility and a psychosocial mechanism supporting its implementation. This mechanism adjusts individual behaviour to the demands placed by the environment, a utilitarian one in this case. Second, I argue that charity, beneficence, clemency, industry and perseverance are all artificial virtues, that it to say, they are brought about by incentives and penalties supplied by the environment. This is important because Hume argues that these virtues constitute the natural support of utilitarian morality.

In section 5.3., I present and discuss the ideas on design and engineering which can be extracted from Hume's work. Hume was aware of the ontological gap between the 'is' and the 'ought', that is, between actual moral behaviour and the possible behaviour depicted in any moral philosophy. A moral psychology of self-interest, sentiments of humanity and the ability to sympathise does not naturally lead to a utilitarian morality. Therefore, design and engineering are needed in order to implement the utilitarian morality closing the ontological gap. I summarised the ideas Hume had on such implementation into four principles of design and engineering, namely redirection of self-interest, excitement and restraint of sympathy and the related sentiments of humanity, propaganda and reinforcement, and rational discussion and reflexion.

Section 5.4. consists of two parts. In the first one, I discuss the case of justice as an artificial virtue; there is no instinct, no passion, no affection or any other element of human nature which naturally ensures allegiance to justice. Any commitment to it is a product of education and good design and engineering applied to the redirection of self-interest and excitement, and restraint of sympathy and the related sentiments of humanity. In the second part, I discuss Hume's rejection of egalitarian distributive justice; he accepted that equality was useful and therefore consistent with utilitarian morality, nevertheless he rejected the egalitarian ideas advanced by James Harrington and by the Levellers. Hume argues that equality is impracticable because people would conceal the real value and size of their wealth and property, and pernicious because it can lead to widespread indigence, tyranny and the extinction of authority and subordination. I show how these potential problems could be solved rendering egalitarianism consistent with a utilitarian morality and with a psychology of self-interest and sentiments of humanity.

In section 5.5., I discuss the case of greedy self-interested individuals who subvert utilitarian morality violating the rules of justice; they are the self-interested knaves. In politics they act openly and insensibly; in civil society they are act secretly and sensibly. Their behaviour compromises the adequate functioning and survival of society, and it also challenges the allegiance of the honest and self-interested to those rules. The aim is to evaluate Hume's psychological theory with respect to the means it offers to the designer and engineer for turning self-interested political knaves into patriots with a public interest, and self-interested civilian knaves into honest members of society.

5.1. Refuting the selfish theory

As it was explained in the previous chapter, current epistemic criteria used for theory choice such as novel predictions, falsifiability or ontological heterogeneity can be inappropriate for a choice between moral psychologies of the eighteenth century. Therefore, I use the three criteria used by Hume himself, namely inductive support, *experimentum crucis* and simplicity adding those from the *vera causa* principle, which in the late eighteenth and early nineteenth century became the systematised expression of the rules Isaac Newton advanced in the *Principia*. The *vera causa* principle addresses epistemic concerns analogous to those of Hume and Mandeville concerning the knowledge of 'passions, motives, volitions and thoughts.'[386] By using these four criteria, the controversy between the psychology of universal natural self-interest from Mandeville, and the psychology of natural self-interest and sympathy from Hume, becomes a standard case of theory choice.

This early modern controversy between these two psychologies not only has historical value because the debate about self-interest and unselfish behaviour or altruism continues. In recent years, rational choice theory and neoclassical economics have been a main battle ground for this controversy. The controversy is also relevant today because some of the criteria used such as inductive support and simplicity are currently used in theory choice. Today, the inference to the best explanation addresses almost the same problems the *vera causa* principle was trying to solve.[387]

Here it is the list with the four criteria to be discussed:

- *Vera causa*
- Inductive support
- *Experimentum crucis*
- Simplicity

Vera causa. To get true and reliable knowledge of unobserved entities and processes causally responsible for observable effects became a major challenge in the eighteenth century. Hume's main concern, and indeed the main problem for Newton and other natural scientists at the time, was about the standards for accepting an explanation based on unobserved entities. The rules of reasoning advanced by Newton constitute a response to this concern. Mandeville's methodological analogy with the inference to the inner pieces and functioning of a spring-watch

[386] D. Hume (1739-40), SB468-469.

[387] See P. R. Thagard (1978); P. Lipton (2004); and N. Cartwright (1983), pp. 6, 87-99.

also reflects the same concern and awareness of the problem. The epistemic justification of the existence and causal efficacy of gravitation, self-interest and sympathy was a main scientific challenge. In the late eighteenth century this led to the development of the *vera causa* principle by Thomas Reid, John Hershel and Charles Lyell. This method was later used by Charles Darwin in his defence of genetic variation across very long periods of time as the true cause for the origin of new species, and migration as the true cause for the existence of colonies of the same species found in distant places.[388]

The first rule of natural philosophy as stated by Newton dictates that '*No more causes of natural things should be admitted than are both true and sufficient to explain their phenomena*. As the philosophers say: Nature does nothing in vain, and more causes are in vain when fewer suffice. For nature is simple and does not indulge in the luxury of superfluous causes.'[389] By appealing to this methodological rule, Newton was trying to prove the existence of gravitation as the *vera causa* of the attraction between celestial bodies against the vortex theory advanced by Descartes, which, by multiplying causes unnecessarily, depicted nature as superfluous and idly complex.[390] Therefore, theoretical simplicity was not an instrumental principle but a realist one justifying the choice for theories with fewer unobservable entities.

Like Newton, Thomas Reid also defines the *vera causa* principle by using the two criteria of truth and causal sufficiency; he writes that 'when men pretend to account for any of the operations of nature, the causes assigned by them ought, as Sir Isaac Newton taught us, to have two conditions, otherwise they are good for nothing. *First*, they ought to be true, to have a real existence, and not to be barely conjectured to exist without proof. *Secondly*, they ought to be sufficient to produce the effect.'[391] John Herschel explained that 'Newton has applied the term *verae causae*; that is, causes recognized as having real existence in nature, and not being mere

[388] See M. Ruse (1976); and R Laudan (1982).

[389] I. Newton (1726), p. 794.

[390] See I. Newton (1726), Book II, proposition 53 and Scholium, and Book III, General Scholium.

[391] T. Reid (1785), p. 80.

hypotheses or figments of the mind.'[392] As we know, Mandeville draws a similar distinction between 'conjectures' and 'knowledge', arguing that only 'middling certainty' can be attained in the knowledge of the passions and the mind.

Just as the controversy between Newton and Descartes was about the true causes of the same set of phenomena, namely the motion of the planets, the controversy between Hume and Mandeville was about the true causes of the same domain of human behaviour. Newton introduced the first rule to prove that his theory had only used the sufficient number of causes, whereas Descartes used more than a sufficient number of them. In spite of stating his commitment to Newton's method for the creation of a new moral science, Hume does not mention nor discuss Newton's first rule. Nonetheless, I believe the use of this rule for evaluating his moral psychology is both justified and adequate. To explain the same domain of human behaviour, Mandeville uses one cause or motive only, i.e. self-interest, whereas Hume uses two, i.e. self-interest and the sentiment of humanity and benevolence.

The actions from Greek and Roman characters such as Pericles, Marcius Berea Soranus, Publius Thrasea Paetus and King Henry IV of France are presented as examples of unselfish acts of patriotism, statesmanship and friendship motivated by sentiments of humanity and benevolence. The main problem with them is the lack of consideration Hume gives to the possible existence of self-interested motivations, which he could then refute. In contrast, Mandeville considers and refutes possible unselfish motivations for precisely the same kinds of actions Hume is using in support of his own theory. Because of this refutation Mandeville's theory must be chosen as the simpler one, while Hume's theory loses the contest standing as a theory 'indulging in the luxury of superfluous causes' just like Descartes's vortex theory did against Newton's theory of gravitation. This is shown in the paragraphs below where the cases of Pericles, Marcius Berea Soranus, Publius Thrasea Paetus and King Henry IV of France are discussed.

[392] J. Herschel (1831), p. 144.

Inductive support. Newton's rule number four explains the epistemic power of induction as follows: '*In experimental philosophy, propositions gathered from phenomena by induction should be considered either exactly or very nearly true notwithstanding any contrary hypotheses, until yet other phenomena make such propositions either more exact or liable to exceptions.* This rule should be followed so that arguments based on induction may not be nullified by hypotheses.'[393] The refutation of Mandeville's psychological theory relies on the existence of acts of disinterested benevolence and humanity motivated by sympathy, which Hume claims is a natural component of the human mind. Because Mandeville's theory is universal, one case of disinterested benevolence is enough for refuting it. Hume, however, wants to do more than that because he is seeking to reduce the scope of self-interest by enlarging the scope of sympathy, producing as many cases as possible of disinterested benevolence and humanity.

The psychological capacity humans have for sympathising with others is the main foundation for any benevolent and humanitarian action. Hume defines sympathy as the 'communication' of the 'inclinations and sentiments' of others 'however different from, or even contrary to our own', so 'hatred, resentment, esteem, love, courage, mirth and melancholy; all these passions I feel more from communication than from my own natural temper and disposition', 'and 'tis certain that we may feel sickness and pain from the mere act of imagination, and make a malady real by often thinking of it.'[394] Within the different passions and inclinations sympathy may elicit, Hume selects only those related to humanitarian and benevolent actions.

There are hardly any cases of disinterested benevolence and humanity discussed in the *Treatise*, so specific cases can only be found in the *Enquiry*. There, Hume quotes statesmanship, patriotism, motherly love, friendship and love relationships as strong evidence for disinterested actions, which can prove the existence of humanitarian motives.

[393] *Ibid.,* p. 796.

[394] All quotations in this paragraph are taken from D. Hume (1739-40), SBN316-319.

The first particular case he discusses is that of Pericles, the 'great Athenian statesman and general', who in his death-bed stopped his friends from paying tribute to him by citing all his great achievements as a statesman. He described them as 'vulgar advantages' in comparison to the 'the most eminent' of his accomplishments, namely that 'no citizen has ever yet worne mourning on my account.' Further cases include Marcius Berea Soranus, roman proconsul in Asia, and Publius Thrasea Paetus, roman senator and republican leader, who rebelled against the emperor Nero; they were 'intrepid in their fate, and only moved by the melting sorrows of their friends and kindred. What sympathy then touches every human heart!' Then, he quotes the case of the mother who 'loses her health by assiduous attendance on her sick child, and afterwards languishes, and dies of grief, when freed, by its death, from the slavery of that attendance'. And also friendship and love relationships when people love and care for others even at the expense of being hurt, like King Henry IV of France whose 'amours and attachments' 'during the civil wars of the league, frequently hurt his interest and his cause.'[395] Hume argues that in all these cases, a sympathetic sentiment prompts humanitarian and unselfish actions, which benefit the citizens within a country, children within a family and friends and lovers within a close circle. No self-interested motivation is considered, self-interest plays no role.

Unlike Mandeville, Hume does not consider if statesmen, mothers, friends and lovers act expecting to be flattered, adored and glorified. This is important because it would allow Hume to advance an argument against any self-interested motivation. His method instead is that of simple enumerative induction, which was criticised by Francis Bacon, who explained that 'the induction which proceeds by simple enumeration is childish; its conclusions are precarious and exposed to peril from a contradictory instance.'[396] Both in the *Treatise* and the *Enquiry*, Hume highly praised the work of Newton and Bacon, whose methods he claimed to be following. The lack of consideration to 'contradictory instances', that is, to self-interested motives is an

[395] D. Hume (1772), Pericles, SBN176; Marcius and Publius, SBN223; mother, SBN300; Henry IV, SBN258.

[396] F. Bacon (1620), p. 97.

important flaw in Hume's method, who takes the obvious as true making no further enquiry. An expectation for glory and public tribute might be the motivation for Pericles, Marcius and Publius, whereas overindulgence in sexual passions and fun might be the motivation for Henry IV, even at the expense of losing a war. Similar considerations apply to the devoted mother, who might be motivated by the veneration she gets from her child. Because Hume does not consider opposite motivations, the cases he presents provide poor support to his theory leaving it vulnerable to refutation.

Another important weakness of Hume's argument is the very small number of cases he presents of disinterested actions in both the *Treatise* and the *Enquiry*. He was aware of this because he finishes his defence of disinterested benevolence by claiming that 'these and a thousand of other distances are marks of general benevolence in human nature',[397] which is a poor justification for not providing further evidence. The cases discussed above are cases of 'particular benevolence' delivered to individuals we have a close connection with. In contrast, 'general benevolence' refers to all those individuals outside this close group, for instance those living in other countries and any distant place. Hume gives no example of this type of general benevolence; in a footnote he only writes: 'I assume it is as real, from general experience, without any other proof'.[398]

Hume's inductive evidence of disinterested benevolent and humanitarian actions is clearly small, and it is subject to easy dismissal because he does not consider the existence of self-interested motivations and how they could be contested. As it was shown in the previous chapter, the cases discussed by Mandeville are numerous and diverse, ranging from the behaviour of businessmen, lawyers and soldiers to motherly care, suicide and moral behaviour in other societies such as Spain and Holland. Besides the two volumes of *The Fable*, there is also *An Enquiry into the Origin of Honour and the Usefulness of Christianity in War* as well as several other essays and books, where Mandeville refined his psychological theory and enlarged the number

[397] D. Hume (1772), SBN300.
[398] *Ibid.,* SBN298.

of cases supporting it. The amount of evidence he provides largely surpasses the evidence supplied by Hume in the *Treatise* and the *Enquiry*. In consequence, Hume's argument is poor and hardly compelling, and therefore Mandeville's theory remains stronger and more convincing.

***Experimentum crucis*.** If inductive support is considered insufficient to decide the controversy between Hume and Mandeville, it could perhaps be settled by presenting a successful *limiting case* in the critical region of the domain. The *experimentum crucis* could do this. Hume explains that 'it is easy to attain what natural philosophers, after Lord Bacon, have affected to call the *experimentum crucis*, or that experiment, which points out the right way in any doubt or ambiguity.'[399] The experiment under consideration is a case of *benevolence to enemies*. Because of its exceptional features within the domain of benevolent actions, it becomes a limiting case, even though it actually is not an experiment but rather a case of 'cautious observation' and 'experience' from records in history.[400]

Hume describes how Demosthenes, a prominent Greek politician of the fourth century B.C., helped his long-standing enemy Aeschines, who was leaving the city after being sent out to political exile. Demosthenes 'secretly followed, offering him money for his support during the exile, and soothing him with topics of consolation in his misfortunes "Alas!" cries the banished statesman, "with what regret must I leave my friends in this city, where even my enemies are so generous!"'[401] The case is presented as proving the existence of true generous sympathetic passions causing benevolent acts, which overpower selfish sentiments of hatred and revenge. Hume concludes that 'compelled by these instances, we must renounce the theory, which

[399] *Ibid.*, SBN219.

[400] Hume was aware of the impossibility of performing experiments in moral science as they were performed at the time in physics. In the introduction to the *Treatise* (SBxix).he explains that 'we must therefore glean up our experiments in this [moral] science from a cautious observation of human life, and take them as they appear in the common course of the world, by men's behaviour in company, in affairs, and in their pleasures.'

[401] D. Hume (1772), SBN10.

accounts for every moral sentiment by the principle of self-love.'[402] According to him universal self-interest has thus been refuted and, because benevolence to enemies is a limiting case, the likelihood of finding disinterested benevolence in the middle regions increases.

Compared to the Mandevillian theory, Hume's conclusions seem naïve and deceptive, or at least incomplete. This is because he does not consider testing his own explanation against the existence of self-interested motivations by asking how much admiration, social tribute, reputation and power or personal relief from remorse Demosthenes would get by giving money to his enemy. An explanation based on a humanitarian act motivated by true sympathy can raise the degree of belief of a true concern for relieving Aeschines's emotional pain and economic hardship, by showing how the rival explanation based on a self-interested passion could be dismissed. This omission undermines the confirmation value of the evidence Hume is presenting.

Hume should have considered at least as equally likely a self-interested motivation, that apparent benevolence to enemies could be a calculated act of self-promotion. Indeed, because his theory holds that both self-interest and sympathy cause moral behaviour regardless of the external aspect of it, both of them should in principle be considered as equally likely. Then, a test should be performed or further evidence provided for choosing one cause over the other, which could then become an *experimentum crucis.* Methodologically and epistemologically, a theory with two or more causes is more challenging because the number of tests and the need for evidence as well as the overall uncertainty increases proportionally to the size of the causal set. These are the consequences of multiplying unnecessarily the number of unobserved causes, in this case self-interest and sympathy, which is a violation of the first rule from Newton discussed above. A further and potentially more damaging problem arises from Hume's decision to stop at the most obvious explanations of moral behaviour which are accepted in

[402] *Ibid.,* SBN219.

'common life', that is, folk psychology explanations. This problem is discussed in the next paragraphs.

Simplicity. Hume argues that a theory which holds both self-interest and disinterested benevolence is simpler than a theory based only on self-interest. His argument consists of two parts, the first one proceeds by a direct comparison between self-interested and disinterested actions; the second part relies on an analogy with secondary self-interested passions.

First part. Hume actually criticises simplicity as a criterion by expressing doubts on the prospects for accomplishing in moral psychology the 'perfect simplicity' observed in physics. In spite of these reservations, he still insists that a theory with two fundamental motives is simpler or, more precisely, imperfectly simpler.[403] The 'selfish theory', he claims, is more complex because it uses 'very intricate and refined reflections', where 'metaphysical' effects of self-interest are 'twisted and moulded, by a particular turn of the imagination' of the scientist, so it can explain a 'variety of appearances'. Hume accepts that people may deceive themselves with respect to the 'predominant motive or intention', that 'is indeed, frequently concealed from ourselves, when it is mingled and confounded with other motives, which the mind, from vanity or self-conceit, is desirous of supposing more prevalent.' However, he thinks that theoreticians such as Mandeville have gone too far by inferring self-interested motivations. All this makes the selfish theory more complex in the theoretical or 'imagined' descriptions it provides of the inner workings of self-interest. Because of such descriptive complexity, Hume argues that the theory is 'fallacious'[404], that is, false. By contrast, folk psychology–which he calls 'common life' psychology–offers simpler explanations, and because of this simplicity it should be considered as the probable true psychology of moral behaviour:

[403] *Ibid.*, SBN299.

[404] All quotation in this and paragraph are taken from D. Hume (1772), SBN297-SBN300.

> Many an hypothesis in nature, contrary to first appearances, has been found, on more accurate scrutiny, solid and satisfactory […] that there is a general presumption for its arising from the causes, which are the least obvious and familiar. But the presumption always lies on the other side, in all enquiries concerning the origin of our passions, and of the internal operations of the human mind. The simplest and most obvious cause, which can there be assigned for any phenomenon, is probably the true one.[405]

Hume recognises that physics has succeeded in going beyond first appearances. However, he believes that going beyond the most obvious causes is a methodological mistake in psychology, mainly because of the 'abstruseness' of the alleged motive and its functioning. He illustrates this by explaining how such 'imagined' functioning is false in the case of a rich patron who is grieving at the death of a poor man, who was also his friend: 'how can we suppose, that his passionate tenderness arises from some metaphysical regards to a self-interest, which has no foundation in reality.' The 'common life' psychology tells us that because the patron is rich, it is unlikely or false that he grieves the death of a poor friend because of self-interest. In contrast, the same folk psychology tells us that if the rich patron dies, the poor man falsely 'may flatter himself, that all his grief arises from generous sentiments, without any mixture of narrow or interested considerations.' The standards of 'common life' psychology establish that the 'simplest and most obvious cause'[406] is the *vera causa*, the true cause.

Again, following Mandeville's method, the alternative hypothesis with non-obvious concealed motivations explaining the same behaviour must also be tested with the information available. For instance, it should be considered if the same false or self-deceptive flattery Hume places on the poor man grieving a rich friend and patron should also be considered in the case of the rich man, who may flatter himself for grieving the death of a man who is actually poor, dull and ignorant.

[405] *Ibid.,* SBN299.

[406]All quotations in this paragraph are taken from D. Hume (1772), SBN299-300.

Hume is right suggesting that by taking a step into a deeper explanatory level the risk of failing increases. Nonetheless, physics had proved the success of taking this methodological step with the many scientific advances it made. Only by disregarding this success as Hume does, his own argument can gather some support. Why psychology should be different from physics? He would have to answer this question in order to gain more support for the division he makes by placing moral sciences in a separate category. The philosophical debate on such a divide between social and natural sciences remains open, and the current advances made by both behaviouristic and cognitive psychology have not solved the problem yet.

With Mandeville, Hume shares the commitment to produce a metaphysics of human nature, that is, a psychology postulating unobservable entities and processes. Because of this commitment they both share the same epistemological and methodological challenges and risks. However, the argument Hume puts forward rejecting on the one hand the metaphysical descriptions of the selfish theory, and accepting the metaphysics of folk psychology on the other is weak and unpersuasive. Concealed self-interested motives are a main challenge to his theory, they are hardly considered and the few occasions when they actually are, they are quickly and unconvincingly dismissed. Those concealed motives were the main contribution from Hobbes and Mandeville to the explanation of moral and political behaviour.

Second part. Hume presents an argument by analogy trying to prove the simplicity of his theory. He explains that 'If we consider rightly the matter, we shall find, that the hypothesis, which allows of a disinterested benevolence, distinct from self-love, has really more *simplicity* in it, and is more conformable to the analogy of nature'.[407] The analogy is set against primary appetites like hunger and thirst, which have the acts of drinking and eating as their primary ends. When drinking and eating are independently performed without feeling hunger or thirst, both their motivation and pleasure are 'secondary and interested'. Similarly, the desire for fame and power becomes independent and secondary, although it is derived from the primary

[407] *Ibid.,* SBN301.

passion of 'self-love and a desire of happiness.' In the same manner, acts of benevolence and humanity are performed even when people are not in need, such acts are also secondary. By analogy, if the love of fame and power ultimately derives its pleasure from self-interest; generosity to the prosperous person must ultimately derive its pleasure from disinterested benevolence or, as Hume writes, 'from the combined motives of benevolence and self-enjoyment'.[408] Therefore, any secondary passion, interested or disinterested, cannot be explained unless the respective primary passions exist.

The comments and remarks I have made earlier also apply to this case, i.e. secondary passions also have to be tested against self-interest. Enumerative induction is too weak, and a single case is even weaker unless supplementary support is provided. Even if the analogy is accepted as it stands now, it is still weak like most analogies are. Further tests or evidence have to be provided, so that the analogy can gain epistemic strength. However, even if the analogy becomes fully warranted, it still remains difficult to accept that a theory with two fundamental motives is simpler just because it relies on the non-abstruse, clear and easy metaphysics of folk psychology.

In sum, on all four criteria for theory choice, namely *vera causa*, inductive support, *experimentum crucis* and simplicity, the psychological theory from Hume achieves a lower score than the psychological theory from Mandeville. Consequently, there are better grounds for believing in a psychology of universal self-interest than in a psychology combining both self-interest and sentiments of humanity and generosity.

Compared to Hume, Mandeville was epistemically more cautions, more rigorous and was more aware of the uncertainty involved in making inferences to unobserved entities and processes in the mind. There are six rules in his method as well as the analogy he makes with the inference to the inner pieces and functioning of a spring-watch. For tackling the problem of rival explanations, Mandeville's own method has two important steps. The first one consists of

[408] All quotations in this paragraph are taken from D. Hume (1772) SBN301-302.

a test and a deductive inference, that is, any claim on disinterested benevolence as the motivation for action must be tested against the rival hypothesis, namely self-interested motives. This test takes the form of a refutation. If the rival hypothesis becomes refuted, then via a disjunctive syllogism the alternative one significantly increases its chances of being true. The second step supplements the first one, first by adding a detailed, penetrating and sharp description of how self-interested motives may operate in the particular case under scrutiny, and second by adding numerous cases where selfish motivations are confirmed via a simple induction leading to a generalisation from other similar cases. In contrast, Hume relies on enumerative induction and the number of cases he presents is small. Moreover, the folk psychology he relies on had already been discredited, among others, by Thomas Hobbes[409] in Britain and François de La Rochefoucauld[410] in France.

5.2. Moral sociology

In both the *Treatise* and the *Enquiry* Hume states his methodological commitment to investigating vice and virtue as a 'matter of fact' by 'uniform experience and observation', which he complements with the rejection of the rationalistic method and deductive inference in science, and the rejection of the normative method from moral philosophy.[411] He applies this factual or empirical inductive method for the explanation of moral behaviour and for establishing the moral preferences of society, which he calls 'moral taste'. The first leads to the production of a moral psychology, the second to a descriptive moral sociology. His psychological theory explains moral behaviour as caused by self-interest and sympathetic

[409] T. Hobbes (1642) and (1651).

[410] F. de La Rochefoucauld (1678).

[411] D. Hume (1739-40), SB468-469; D. Hume (1772), SBN231.

sentiments of humanity and generosity, while his descriptive sociology finds that the British society of the eighteenth century has a utilitarian flavour.[412]

Currently, in the social sciences there are a number of methods for data collection, data interpretation and measurement that were not available in the eighteenth century. Hume performs the task of establishing the moral taste of society by collecting evidence on the judgements people make approving and disapproving different kinds of moral behaviour and architectural styles from houses and buildings. That is to say, he gathers evidence on moral and aesthetic judgement to demonstrate the utilitarian flavour of society.

He realises that the tasks of a descriptive sociology can be tedious. If the metaphysical speculation on the passions and the inner working of the mind becomes 'abstruse' and 'minute', the description of the moral judgements of 'common life' becomes a 'superfluous' task. He explains that 'to prove, by any long detail, that all the qualities, useful to the possessor, are approved of, and the contrary censured, would be superfluous. The least reflection on what is every day experience in life, will be sufficient.'[413] This happens because Hume had to prove something that was already obvious to many, and also because of the character of a scientific task which consists of reporting findings on the existence, size and characteristics of a social feature. Indeed, to the reader the description of conventional judgements on the utilitarian approbation of individual and social behaviour of cases like discretion, assiduity, flowing affability and delicate modesty may seem uninteresting. However, any judgement on this should consider the positive value of the basic descriptive sociology Hume was producing.

From this sociological perspective, Hume finds that the British society of the early eighteenth century had a utilitarian flavour. With this finding he is not making a moral

[412] David Hume anticipates only some aspects of classical utilitarianism; in spite of the differences the basic criterion of judging any action as virtuous by how much it contributes to the happiness of society remains as distinctive. Henry Sidgwick explains that Hume used the term 'utility' in a 'narrower sense', and without as much precision as Jeremy Bentham did. Sidgwick (1907, p. 424) explains that 'there is a great difference between the assertion that virtue is always productive of happiness, and the assertion that the right action is under all circumstances that which will produce the greatest possible happiness on the whole.'

[413] D. Hume (1772), SBN235; see also SBN176, SBN217; D. Hume (1748), SBN7.

philosophy but a descriptive science; he does not argue in any substantive manner for utility as a norm; the purpose is only to demonstrate the predominance of public utility as the actual social standard. He explains that 'it appears to be a matter of fact, that the circumstances of *utility*, in all subjects, is a source of praise and approbation: That it is constantly appealed to in all moral decisions concerning merit and demerit of actions: That it is the *sole* source of that high regard paid to justice, fidelity, honour, allegiance, and chastity: That it is inseparable from all other social virtues, humanity, generosity, charity, affability, lenity, mercy, and moderation. And, in a word, that it is a foundation of the chief part of morals'.[414] He unifies the different types of utilitarian behaviour he observed into a general principle of utility:

> Usefulness is only a tendency to a certain end; and if it is a contradiction in terms, that any thing pleases as means to an end, where the end itself nowise affects us. If usefulness, therefore, be a source of moral sentiment, and if this usefulness be not always considered with reference to the self; it follows, that everything, which contributes to the happiness of society, recommends itself to our approbation and good-will. Here is a principle, which accounts, in great part, for the origin of morality.[415]

Hume was very pleased with this empirical finding as a key element of his science of man, as he was aware that no 'moral writer' in the past explained morality using utility as a fundamental universal criterion.[416] By analogy with the explanatory and unifying power of gravity in Newtonian physics, he attributes the same explanatory and unifying power to utility, which, he claims, binds all individuals with each other just like gravity keeps the planets orbiting around the sun.[417] Utility therefore becomes a universal principle, which is used in the moral

[414] D. Hume (1772), SBN231.

[415] *Ibid.,* SBN219.

[416]*Ibid.,* SNB212. See Francis Hutcheson (1726, p. 125) for an earlier formulation of a similar principle of utility.

[417] D. Hume (1772), SBN204; Hume makes an explicit reference to Newton's rule number two on the unification power of causal generalisations: '*The causes assigned to natural effects of the same kind must be, so far as possible, the same.* Examples are the cause of respiration in man and beast, or the falling of some stones

judgements and choices made on different kinds of behaviour. Hume did not produce a moral argument justifying the choice on utility as the new moral value and principle for distinguishing virtue from vice, he only stated the principle and used it, as a scientist does, for classifying particular cases of behaviour as vice or virtue. Therefore, the origin and justification of such a principle remain unexplained in his work.

There are two basic kinds of utilitarian virtuous action, namely self-interested and humanitarian or benevolent. The first one is caused by the passion of self-interest; the second one is caused by the ability of humans to sympathise with the pain of others. As evidence of the first kind of action, Hume quotes the cases of the statesman who serves the public interest and the patriot soldier who risks his life for others.[418] In contrast, the behaviour of a tyrant is subject to social disapproval because of the harm it inflicts onto society as well as the behaviour of the industrious person who withholds a number of social benefits by being also a miser. The tyrant and the miser are self-interested but their actions are not virtuous, while the actions of the statesman and the patriot are both self-interested and virtuous. Hume explains that the 'open demand for praise and admiration' and 'an impatient desire for applause'[419] from the statesman and the patriot do not undermine the virtuous character of their actions because of public benefits they create. In contrast, the self-interested actions of monks and friars are not virtuous because 'celibacy, fasting, penance, mortification self-denial, humility, silence, solitude, and the whole train of monkish virtues' are socially useless, they create no social benefit. More specifically, Hume quotes the cases the Roman Catholic Saints Dominic and Ignatius of Loyola as examples of such moral vice.[420]

in Europe and America, or of the light of a kitchen fire and the sun, or of the reflection of the light on our earth and planets.' Newton (1726), p. 795.

[418] *Ibid.,* SBN227, SBN265-266.

[419] *Ibid.,* SBN265-6.

[420] *Ibid.,* SBN270, SBN342.

As evidence of the second kind of utilitarian behaviour, Hume quotes the case of the generous industrious person who gives charity to the poor and needy[421] and the three cases discussed in the previous section, namely the rich patron grieving the death of a poor man, benevolence to political enemies and motherly love. Hume observed that actions of benevolence and humanity are also socially praised as virtuous because of their utilitarian properties, he explains that 'nothing can bestow more merit on any human creature than the sentiment of benevolence in an eminent degree; and *that* a *part*, at least, of its merit arises from its tendency to promote the interest of our species, and bestow happiness on human society.'[422] This set of actions form a separate source of virtue and utilitarian value, which Hume defines as follows:

> The notion of morals implies some sentiment common to all mankind, which recommends the same object to general approbation, and makes every man, or most men, agree in the same opinion or decision concerning it. It also implies some sentiment, so universal and comprehensive as to extend to all mankind, and render the actions and conduct, even of the persons the most remote, an object of applause or censure, according as they agree or disagree with that rule of right which is established. These two requisite circumstances belong alone to the sentiment of humanity here insisted on.[423]

With those examples, Hume wants to demonstrate that this universal sentiment of humanity meets those two 'require circumstances', namely approval from public opinion and the rule of right, which is the utilitarian principle quoted above. Recall that this universal sentiment exists because of the ability humans have for sympathising with the pain and suffering of others.[424] Both self-interest and sympathy constitute the psychological foundations of utilitarian morality,

[421] *Ibid.*, SBN234-235.
[422] *Ibid.,* SBN181; see also SBN230-231.
[423] *Ibid.,* SBN272.
[424] D. Hume (1739-40), SB316-319

which regulates behaviour by promoting disinterested acts of benevolence and self-interested actions in pursuit of wealth, fame and admiration.

Hume explains how the utilitarian morality is maintained through the judgements the society as a whole makes upon each individual in combination with an inner psychological mechanism. This mechanism consists of the 'constant habit of surveying ourselves' examining how our actions 'appear in the eyes of those, who approach and regard us' in our 'continual and earnest pursuit of a character, a name, a reputation in the world', which 'is the surest guardian of every virtue.'[425] As a mechanism it is purely social and psychological or positive without being attached to any specific set of moral norms. It is positive because it also regulates criminal behaviour within organised crime such that of the robbers and pirates quoted by Hume.[426] Only when a definition of virtue is added does a moral sense emerge, which ensures the social reproduction of a specific set of moral norms.

Hume argues that such moral sense is natural because of the existence of those natural 'generous sentiments' of humanity, where even the weakest one can produce a basic inner sense of right and wrong.[427] An internal moral sense, which 'nature has made universal in the whole species'[428] makes moral judgements possible. The simplest natural foundation of this sense are the sensations of pleasure and pain, so that 'virtue is distinguish'd by the pleasure, and vice by the pain'.[429] No moral judgment can be made without these sensory grounds. Reason alone cannot be the source of moral judgements. Nonetheless, it holds an important function performing the calculation of the amounts of pain and pleasure expected from the different choices available. This is important in any social choice because having a concern for 'justice' and the 'happiness of mankind' in the design of 'municipal laws' and the 'debates of civilians'

[425] D. Hume (1772), SBN276.
[426] *Ibid.,* SBN209.
[427] *Ibid.,* SBN271.
[428] *Ibid.,* SBN173.
[429] D. Hume (1739-40), SB475.

and the 'reflections of politicians', '*reason* instructs us in the several tendencies of actions, and *humanity* makes a distinction in favour of those, which are useful and beneficial.'[430]

In this way, as an early modern scientist Hume produces a basic moral sociology which consists of two major empirical claims, namely the existence of a new moral principle and a psychosocial mechanism supporting its implementation. Judgements from society and laws and policies from the government shape individual behaviour giving it a new utilitarian face, such a social engineering is accomplished by relying on a psychosocial mechanism, which adjusts individual behaviour to the demands placed by the environment. This moral sociology and design are supported on a moral psychology with three main claims, namely the existence of natural self-interest, natural sentiments of humanity and a universal moral sense. A natural psychological ability to sympathise enables those sentiments of humanity by association with sensations of pain and pleasure, and those sentiments constitute a natural moral sense, which self-interest alone cannot produce.

As part of this moral psychology, Hume makes a further distinction between natural and artificial virtues, which is also relevant for the utilitarian moral sociology he is advancing. For instance, he identifies humanity and industry in work as natural, and justice and chastity as artificial. By artificial he means that behaviour which can be elicited by changing the environment, in particular the structure of incentives and penalties. Against Hume, I argue that acts of humanity and generosity he quotes such as charity and beneficence, and self-interested ones such as industry in work, and in general the highly diversified expressions of self-interest he observed in Britain and other parts of Europe in the early eighteenth century, were not natural but artificial. My argument relies on the moral psychology of universal natural self-interest advanced by Bernard Mandeville. The argument and analysis on 'artificial virtue' in this chapter is an extension of my argument and analysis on 'artefactual behaviour' in chapter two. I use terms 'artefactual' and 'artificial' as synonyms with each other. Both terms refer to

[430] D. Hume (1772), SBN286.

behaviour which does not exist naturally, however 'artefactual' carries a stronger technological connotation.

In the *Treatise* Hume asks whether virtue in general is natural or artificial. On the one hand, he explains that 'if ever there was any thing, which cou'd be call'd natural in this sense, the sentiments of morality certainly may, because there never was any nation of the world, not any single person in any nation, who was utterly depriv'd of them'. On the other hand, he recognises that ''tis absurd to imagine, that […] these sentiments are produc'd by an *original* quality and *primary* constitution. For as the number of our duties is, in a manner, infinite, 'tis impossible that our original instincts shou'd extend to each of them'.[431] Therefore, a considerable number of virtues are artificial. This conclusion has far-reaching consequences for the kind of folk social design and engineering performed in early modern times.

The manipulation of the environment via the provision of rewards and punishment, praise and blame makes the difference between artificial and natural virtues. Hume explains that justice, allegiance to government, obligation of promises, chastity and sexual fidelity are artificial because they 'may be chang'd by motives of reward and punishment, praise and blame. Hence, legislators and divines, and moralists, have principally applied themselves to the regulating these, and have endeavour'd to produce additional motives for being virtuous.'[432] In contrast, virtues such as 'meekness, beneficence, charity, generosity, clemency'[433] are natural because they are based on the natural sentiment of humanity and benevolence. Other virtues such as 'industry, perseverance, patience, activity, vigilance, application, constancy'[434] are also natural because they are based on natural 'qualities of the mind', which are not voluntary and cannot be elicited with the provision of rewards and punishment either. We know, he says, 'that to punish a man for folly, or exhort him to be prudent and sagacious, wou'd have but little effect; tho' the same punishments and exhortations, with regard to justice and injustice, might have a considerable

[431] D. Hume (1739-40), SB473-474.
[432] *Ibid.*, SB609.
[433] *Ibid.*, SB578.
[434] *Ibid.*, SB610.

influence.'[435] In contrast, I argue that charity, beneficence, clemency, industry and perseverance are all artificial, that is to say, they are brought about by incentives and penalties supplied by legislators, the clergy, moralists, parents and the society as a whole.

The psychosocial mechanism described above, where individuals 'survey themselves' by examining the opinions others have of them in order to calculate the overall balance of their own pleasure and pain is a positive psychological description with no moral significance. That is to say, those opinions and the pain and pleasure attach to them are mere 'impressions'. Therefore, Hume recognises that 'the next question is, of what nature are these impressions, and after what manner do they operate upon us? Here we cannot remain long in suspense, but must pronounce the impression arising from virtue, to be agreeable, and that proceeding from vice to be uneasy.'[436] Hence, impressions become morally significant only after a notion of virtue and vice is attached to them. Mandeville describes a similar mechanism for the origin of morality, where shame and pride are the grounds upon which virtues are socially constructed through the work of politicians and the clergy.[437] Like Mandeville, Hume also accepts the effect education and the work of politicians have on the creation of artificial virtues but he makes a warning against the extension of this claim to all virtues, that is to say, against the claim that all virtues can be constructed without being 'founded on the original constitution of the mind', so that 'Tho' the rules of justice be *artificial*, they are not *arbitrary*.'[438]

The previous chapter showed how Mandeville saw the transition from a largely agrarian society to a precapitalist commercial one as the product of large-scale reforms on property rights, division of land and crop production, diversification of manufacturing and the expansion of domestic and foreign trade. All these reforms lead to the transformation of self-interest from a 'dormant' state to an intensely active state. The concept of 'dormant passions' in agrarian societies complements Mandeville's ideas on design and artificial virtues. He explains that

[435] *Ibid.*, SB609.
[436] *Ibid.,* SB470.
[437] B. Mandeville (1732), vol. I, p. 36-37
[438] D. Hume (1772), SBN214; D. Hume (1739-40), SB484.

'dormant passions' keep every individual as a 'lumpish Machine, without the Influence of his Passions, may be justly compar'd to a huge Wind-mill without a breath of Air', so 'Man never exerts himself but when he is rous'd by his Desires: While they lie dormant, and there is nothing to raise them, his excellence and Abilities will be for ever undiscover'd'.[439]

He observed that sloth, ease and ignorance in an agrarian society are also cases of self-interest, which are opposed to the standards of utilitarian morality. In this way, self-interest became far more active and diversified because of the new incentives provided with the new reforms. Envy, greed, industry in work and politeness were constructed and reshaped under the new economic laws. Therefore, self-interest remains active in both an agrarian and a precapitalist commercial society, and during the transition between them. The different expressions of an intensively active and diversified self-interest such as industry, perseverance, activity, application, constancy, vanity, love of fame and others discussed by Hume were the artificial product of the political and economic reforms implemented after the Glorious Revolution, which led to a rapid economic growth; by 1750 England was the largest economy in Europe.[440] A recent case with similar artefactual behavioural changes and an economic boom took place in the city of Shenzhen in China, where the first capitalist market was established in 1979 as a special economic zone; Shenzhen was rapidly transformed from a coastal city with a weak local economy based on fishing and agriculture to a thriving industrial commercial port.

The same analysis and conclusions on the artificial character of self-interest in England during the eighteenth century can also be applied to the disinterested actions of benevolence and charity, which Hume argues are natural because they based on a natural sentiment of humanity. Mandeville criticised Lord Shaftesbury for mistakenly trying to build a moral philosophy upon 'the ruins of Christianity'[441], which had already experienced radical changes through the protestant movements and the economic and political reforms of the period. I

[439] B. Mandeville (1732) vol. I, p. 199; see also pp. 145, 381, and vol. II, pp. 92, 119.
[440] See D. Ormrod (2003), pp. 307-313; and D. C. Coleman (1977).
[441] B. Mandeville (1732) vol. II, p. 432.

argue that Hume's defence of benevolence and humanity as natural was a secularised response to the decline of Christian morality as it had been constructed and maintained by the Roman Catholic Church for more than ten centuries.

The Scottish and the English Reformation movements of the sixteenth century gradually implemented less demanding Christian moral standards, which Hume witnessed in the eighteenth century. In fact, Hume's rejection of the monkish virtues and his defence of the sentiment of humanity and benevolence reflect the Protestant choice, which retained charity and beneficence from the Roman Catholic morality. Hume's choice is remarkable because he rejected those virtues from Roman Catholicism, which were more demanding in terms of self-denial, and therefore less compatible with the new economic order such as fasting, penance, celibacy, silence, solitude and mortification. He retained only humanity and benevolence which still require self-denial, although comparatively less, and also because these virtues were consistent with the new utilitarian morality.

The previous chapter showed how numerous actions of self-denial were not perceived as such by multitudes thanks to the skills of the clergy, who devised a large number of rules, rituals and other reinforcement mechanisms. In his book *An Enquiry into the Origin of Honour and the Usefulness of Christianity in War*, Mandeville recognises the great skills the Roman Catholic clergy had, which successfully restrained the self-interested passions of large multitudes with actions such as fasting, chastity and no consumption of meat on certain days. He also observed that Roman Catholic leaders were more skilful at doing all this than any Protestant or Muslim religious leaders at the time. He explained that the quality of those skills could be appreciated when individuals 'contract a Habit of Virtue, so as to practice it, without being sensible of self-denial and even that they may take Pleasure in Actions that would be impracticable to the vicious. But then it is manifest that this Habit is the Work of Art, Education, and Custom; and it never was acquired, where the conquest over the Passions had not been already made.'[442] He described this as 'taught nature' instead of 'artificial behaviour'.

Therefore, what Hume saw as disinterested benevolence was the product of moral standards, which had been designed and successfully implemented by Roman Catholicism and retained by the Church of Scotland.[443] They were the product of the design, education and continuous reinforcement performed across many generations making the restraint of self-interest and self-denial almost insensible.

The utilitarian taste, the intense and diversified self-interest and the acts of charity and beneficence Hume observed in eighteenth century Britain, were the artefactual product of economic reforms implemented after the Glorious Revolution and the retention of moral virtues originally implemented by the Roman Catholic Church. The new utilitarian morality based on self-interest and supplemented with acts of benevolence and humanity reflected the new Protestant morality and the emergence of a precapitalist commercial economy. Unlike Hume, Mandeville shows a sharper awareness of the effects economic reforms and religion have on moral behaviour. This explanation tries to answer the question of the origin of the utilitarian flavour and the related social behaviour observed by Hume in Britain during the early eighteenth century. A historical explanation of virtue built upon Mandeville's work complements the descriptive sociology and psychology from Hume. A historical explanation of virtuous behaviour reduces the scope of naturalistic explanations such as that of Hume, and it opens new possibilities for constructivist explanations, which place a significant challenge on the idea of human nature and natural behaviour with important implications for design and engineering in the social sciences.

[442] B. Mandeville (1732b), pp. x-xi, 110-113.

[443] David Hume was baptised and raised as a Presbyterian, at least three generations back his family belonged to the Presbyterian Church; later in his life he became an atheist; see E. C. Mossner (1980), pp. 12-13, 31-34, 51, 64.

5.3. The engineering of utilitarian behaviour

The science of human nature David Hume wanted to erect was a science of matters of fact. Nevertheless, he became aware of the gap between those matters of fact, or actual states of affairs, and those states of affairs projected in some social and political blueprints available at the time. Such an ontological gap implied a new scientific challenge and a shift from a science concerned with establishing matters of fact to a science concerned with how to bring about states of affairs projected in social blueprints; that is to say, a shift from a factual science to a science of design and engineering. In his work, Hume devoted less space to the discussion of aspects related to design and engineering, nonetheless they are crucial for the social and political matters he was concerned with. In this section, I present the ideas on design and engineering which can be extracted from his work.

A moral psychology of self-interest, a sentiment of humanity and the ability to sympathise does not naturally lead to a utilitarian morality. There is a gap between what Hume called human nature and the realisation of any moral standard. There are two important challenges related to this gap. The first one consists of accomplishing a successful first implementation of new moral norms; the second consists of preventing defection, deviation and different forms of corruption of such moral norms. Another important challenge is the moral conversion or cooperation from those who remain opposed to the new norms. These challenges belong to a science of design and engineering, which has to devise the rules, policies and further conditions needed for constructing and maintaining the new behaviour. Furthermore, such rules, policies and conditions have to be consistent with the moral psychology adopted.

Hume identifies two important problems for the realisation of the utilitarian morality. The first one consists of the narrow scope of generosity, which is limited to a small circle of family and friends:

> When experience has once given us a competent knowledge of human affairs, and has taught us the proportion they bear to human passion, we perceive, that the generosity of men is very limited, and that it seldom extends beyond their friends and family, or, at most, beyond their native country. Being thus acquainted with the nature of man, we expect not any impossibilities from him; but confine our view to that narrow circle.[444]

The second one consists of the dominant role of self-interest as a cause for behaviour, which needs to be constantly redirected in order to prevent any harm to society:

> no affection of the human mind has both a sufficient force, and a proper direction to counterbalance to love of gain, and render men fit members of society, by making them abstain from the possessions of others. Benevolence to strangers is too weak for this purpose […] There is no passion, therefore, capable of controuling the interested affection itself, but the very affection itself, by an alteration of its direction.[445]

Therefore, self-interest is dominant while generosity is weak. Hume claims that nature provided us only with the generosity and benevolence necessary for keeping together families and a small group around them. Hence, any extension of generosity beyond the close circle and any restraint of self-interest are artificial. This is important because it presents the policy-maker and the social scientist with the task of fixing and completing the unfinished job of nature by trying to make women and men fit for any large or small society, such as nation-states and the eventual union of some of them in even larger polities such as the European Union. Any large society can function and remain together only through artificial means. From the daily work of making new policies and laws to the constitutional changes and the creation of new institutions, the designer and engineer face the challenge of devising the means for constructing and

[444] D. Hume (1939-40), SB602

[445] *Ibid.,* SB492.

maintaining public spirit in society, extended generosity, patriotism, honesty and promise-keeping among many other types of artificial behaviour.

Hume explains that self-interest can be redirected from socially harmful greed and theft to a socially beneficial greed, which is also beneficial to the individual. This can be done by showing the larger benefits to be gained by restraining greed and by cooperating with strangers in society, in contrast to 'running into the solitary and forlorn condition' if harm is done to society. Unrestrained self-interest is narrow and myopic, and it is therefore less rational or 'folly'.[446] By redirecting it, it becomes 'sagacious'[447], that is, it becomes rational or enlightened self-interest. Therefore, the psychological lesson to be learnt by the designer is not to rely on other means except self-interest itself. Hume criticised John Locke's ideas on a social contract and the Whig doctrine of consent because they ignored this important lesson by relying instead on promises made among the parties making no provision for self-interest. He insisted on claiming that 'interest is the *first* obligation to the performance of promises.'[448] The case shows again the crucial role moral psychology plays in design and engineering.

How can generosity be extended outside of the close circle of kin relations and friends to the wider national society and any foreign one? This can be done by exciting the imagination with relevant images of pain and suffering, so that sympathy is artificially produced exciting a sentiment of humanity for any person outside the close circle. Sympathy is a fundamental psychological mechanism that plays a crucial function in Hume's moral psychology. It is a 'communication of sentiments' which takes place in our imagination from the sensory input we receive from observing other individuals.[449] It operates through the capacity the human mind has for representing the pain felt by others by associating it with our own. Because of the actual use of *empathy* in psychology and the social sciences as a means for *understanding* others, it is important to explain that for David Hume and Adam Smith sympathy was fundamentally and

[446] *Ibid.*, SB492.

[447] *Ibid.*, SB492.

[448] *Ibid.,* SB523, see also D. Hume (1772) SBN199-201.

[449] *Ibid.,* SB316, SB324, SB363, SB576 and SB579.

distinctively about the communication of the sensations of pain and pleasure and the production of the related emotions. It was not an intellectual ability allowing us to *understand* others remaining emotionally unaffected.[450]

'Limited generosity', Hume explains, is the psychological foundation of 'justice and property', while 'extended sympathy' is the psychological foundation of virtue.[451] Thus, generosity to strangers both local and distant can only be extended through the psychological ability of the human species to sympathise with the pain of others. We can sympathise with the hardship and pain of the poor feeling compassion for them, and we can also sympathise with the pleasure the rich enjoy feeling envy and admiration. But sympathy can also have negative social effects since it is also the source of 'popular sedition, party zeal, [and] a devoted obedience to factious leaders'[452], and it can also encourage 'idleness and debauchery' through charity and beneficence.[453] Therefore, extended sympathy has to be both excited and curbed, which are important tasks for the designer and the engineer.

Hume mentions some of the means available for promoting and reinforcing the norms of a utilitarian morality, such means include conversation, church services, school education and theatre. More generally, he argues that reason and reflection can also be used as means for correcting self-interest and extended sympathy, when they are misled by narrow or irrational motives, or when they become biased by proximity with those living around us. Thus, 'the intercourse of sentiments, therefore, in society and conversation, makes us from some general unalterable standard, by which we may approve or disapprove of characters and manners.'[454]

[450] Adam Smith (1790, p. 10, see also p.30) explains that 'Pity and compassion are words appropriated to signify our fellow-felling with the sorrow of others. Sympathy, though its meaning was, perhaps originally the same, may now, however, without much impropriety, be made use of to denote our fellow-feelings with any passion whatever.' In sociology Max Weber ([1922 pp. 4-22) distinguishes empathetic understanding from rational understanding. More recently the economist Ken Binmore (2005, pp. 101, 114) differentiates empathy from sympathy; in psychology Lauren Wispé (1991, pp. 67-82) also makes a similar distinction between sympathy and empathy.

[451] D. Hume (1739-40), SB586.

[452] D. Hume (1772), SBN224, and SBN221-223.

[453] *Ibid.,* SBN180.

[454] *Ibid.,* SBN229; see also n. 1, SBN274-275; and D. Hume (1739-40), SB602-SB603.

Earlier, reason only had a *calculative* function, working out the utilities of different actions. Now, as a *discursive* means, it plays a major role in correcting both self-interest and sympathy.

Education, religion, conversation and public entertainment become the means for such correction ensuring the realisation and adequate functioning of utilitarian morality. Hume acknowledges the power of these means in the creation of artificial virtues, which in some cases can surpass nature. He explains that 'precept and education must so far be owned to have a powerful influence, that it may frequently increase or diminish, beyond the natural standard, the sentiments of approbation and dislike; and may even, in particular instances, create, without any natural principle, a new sentiment of this kind'.[455]

Socially beneficial sympathy becomes an important feature of human psychology because it makes wider socialisation possible by extending emotional affectation and bonds beyond the close circle of family and friends. Any extension of sympathetic feelings beyond our close circle is artificial, so extended generosity crucially depends on it. The redirection of self-interest is also artificial and it crucially depends on education. Recall that before being restrained, self-interest is intensely activated with the economic reforms that transform any agrarian society into a commercial precapitalist one. Therefore, the success of a utilitarian morality crucially depends on the balance between self-interest and extended sympathy; the first one has to be redirected while the second one has to be both excited and curbed. Hence, utilitarian behaviour is artificial; it is the product of design and social engineering.

Recently, economists and game theorists in social choice theory and welfare economics have been using sympathy as a psychological mechanism, which can help solve problems of interpersonal utility comparisons as well as for extending generosity beyond close kinship and friendship. Kenneth Arrow uses extended sympathy to justify social choices.[456] John Harsanyi

[455] *Ibid.,* SBN214.

[456] K. Arrow (1977).

uses sympathy as a justification for utilitarian norms of distribution,[457] while Ken Binmore distinguishes sympathy from empathy, using the latter as the psychological foundation for extending our concerns on fairness to the whole society.[458]

To sum up, the implementation of utilitarian morality requires of the following five tasks:

- Narrow self-interest must be redirected.
- Limited sympathy must be extended.
- Socially harmful cases of sympathy must be prevented.
- Moral norms must be promoted in schools, churches, public entertainment centres and in conversation.
- Deviations from the utilitarian norms must be corrected using rational discussion and reflexion.

These five tasks can be summarised into the following four principles, which can be described as *principles of design and engineering* :

i) Redirection of self-interest.
ii) Excitement and restraint of sympathy and related sentiments of humanity.
iii) Propaganda and reinforcement.
iv) Rational discussion and reflexion.

The first two principles are closely related to the utilitarian morality described by Hume, while the last two have a more general scope. As it can be appreciated, the implementation of morals norms is a difficult scientific and social task. Again, self-interest, sympathy and the sentiment of

[457] J. Harsanyi (1977).
[458] K Binmore (2005), pp. 101, 113.

humanity do not naturally lead to a utilitarian morality, and there is actually no necessary connection between the two. Such a moral psychology could support different sets of moral norms. No psychology supports one unique set of moral norms. By analogy with theories in the natural sciences, we can say that the same moral psychology can support more than one moral philosophy. That is to say, moral philosophies are underdetermined by psychological facts and theories. This claim can be appreciated further in the next section where two different cases of distributive justice are discussed.

The first two principles listed above are concerned with a utilitarian morality, which holds implicit a criterion of distributive justice challenged by alternative criteria existing at the time; Hume discusses and resists some of them. They are groups deviating from utilitarian morality, 'common sense' and the 'common life' such as the Covenanters in Scotland, the Levellers in England, the Camisards in France and the Anabaptists in Germany, which are described as 'superstitious' or 'enthusiasts'.[459] Ironically, Hume himself was an enthusiast of the new economic order and the new morality it brought with it. The case of the Levellers and the Commonwealth of Oceana from James Harrington are discussed in the next section.

5.4. Artificial justice

Justice, allegiance to government and obligation of promises are all artificial. This is because there is no element in the psychological constitution of the human species supporting the behaviour expected from these rules. Self-interest, sympathy and the sentiment of humanity do not imply a natural inclination to them. Distributive justice deserves attention because it was a subject discussed by Hume in the context of his own utilitarian view and rival views on justice at the time, and because the concerns and views of that period are still relevant today.

[459] See essay 'Of Superstition and Enthusiasm' in D. Hume (1777).

Hume explains that property is 'the object of justice' which comprises three basic general rules, namely 'the stability of possession, of its transference by consent and of performance of promises.'[460] With these rules 'justice evidently tends to promote public utility and to support civil society' but the related sentiment of justice attached to any idea of justice is not natural but artificial, because it is 'derived from our reflecting on that tendency'.[461] More specifically, justice is the product of convention and education:

> Unless, therefore, we will allow, that nature has establish'd a sophistry, and render'd it necessary and unavoidable, we must allow, that the sense of justice and injustice is not deriv'd from nature, but arises artificially, tho' necessarily from education, and human conventions.[462]

Hence, there is no instinct, no passion, no affection or any other part of human nature that supports a commitment to the rules of justice. First, he explains that 'property' is not a sensible quality of any material object, such as an acre of land, so it cannot be considered to be part of the land itself or any other material object. Second, he compares justice with virtue and vice, which can be defined by degrees whereas the dominion over land or any other material object is complete. Third, he describes how proximity and self-interest lead to biased judgements on the distribution of property, which are contrary to the abstract character and generality of the rules of justice.[463] Two further reasons for the artificial character of property are the high variation of local 'municipal laws', and the 'finer turns and connexions of imagination, and from the subtleties and abstractions of law-topics and reasonings.'[464]

[460] D. Hume (1739-40), SB526.
[461] D. Hume (1772), SB201.
[462] D. Hume (1739-40), SB483.
[463] *Ibid.,* SB526-533.
[464] D. Hume (1772), SBN202-203; see also SBN201, SBN209-210

Therefore, a commitment to the rules on property can only come from 'reason', 'reflexion', and 'forethought' on the 'whole plan or system.'[465] Rational discussion and reflexion must appeal to considerations of common interest and public utility in order to redirect narrow self-interest: 'The same self-love, therefore, which renders men so incommodious to each other, talking a new and more convenient direction, produces the rules of justice, and is the *first* motive of their observance.'[466] Once it has been redirected by reflection, self-interest is best satisfied by following the rules of justice for "tis evident, that the passion is much better satisfy'd by its restraint, than by its liberty, and that in preserving society, we make much greater advances in the acquiring possessions, than in the solitary and forlorn condition, which must follow upon violence and an universal licence.'[467] Therefore, a good design on the distribution of property must on the one hand be consistent with self-interest, and on the other it must rely on education and continuous reinforcement for the prevention of theft and violence.

Besides self-interest, sympathy and the related sentiment of humanity are also fundamental in Humean moral psychology, while in his descriptive sociology moral utilitarianism is the dominant theme. Given that justice and the rules of property are artificial, there is the important political question closely related to the distribution of property: what kind of distributive justice is consistent with a population of self-interested sympathetic individuals who have a preference for utilitarian morality?

As was discussed earlier, within utilitarianism one of the most important artificial tasks of design and engineering is to extend sympathy in order to create overall a more egalitarian distribution of wealth and property for the benefit of the poor unemployed and low income working classes. Thus, equality becomes relevant because it seems to be consistent and the next step within the utilitarian rationale, which relies on the extension of sympathy and the related sentiments of humanity and benevolence. On the one hand, there is no precise answer on how

[465] *Ibid.,* SBN306-SBN309.
[466] D. Hume (1739-40), SB543.
[467] *Ibid.,* SB492.

much wealth should be redistributed without spoiling the motivation and output from the industrious self-interested individuals, who can also be sympathetic. On the other hand, there is no precise answer on how much inequality should be allowed without creating dangerous tensions that lead to conflict and violence from the poor unemployed and low income working classes. The amount of inequality can vary significantly. Moreover, any constitutional reform must also consider other aspects such as the current form of government, manners, climate, religion and commerce.[468]

In his essay 'The Idea of a Perfect Commonwealth', Hume considers *Oceana* from James Harrington as the 'only valuable model of a commonwealth' in comparison with Plato's *Republic* and Thomas More's *Utopia*.[469] Harrington criticised absolute monarchy, pure aristocracy and regulated monarchy because all of them excluded the majority of the population from owning any share of land. In absolute monarchy 'one man has the whole, or two parts in three of the whole land or territory'; in a pure aristocracy 'a few men have the whole, or two parts in three of the whole land or territory' having no monarch ruling over them; and in a regulated monarchy those few men having 'the whole, or two parts in three of the whole land or territory' are ruled by a monarch. In contrast, 'if the many, or the people, have the whole, or two parts in three of the whole land or territory, the interest of the many or of the people is the predominant interest, and causes democracy.'[470]

Harrington devised an agrarian law, which was the fundamental component of the blueprint for a republic based on an egalitarian distribution of land. This agrarian law prescribed £2000 as the maximum monetary value of any share of land an individual could own in this new republic. He assumed that £10,000,000 was the total rental value of the land in England, so the minimum number of land owners would be 5000, which represented a large redistribution.

[468] D. Hume (1772), SBN196.

[469] D. Hume (1777), p. 514.

[470] J. Harrington (1656), p. 593.

Moreover, the number of owners would grow equalising property further because this law also prescribed inheritance of land in equal parts to all children in a family.[471]

Harrington claims that political power and representation must be supported on wealth and property. He explains that 'empire is of two kinds, domestic and national, or foren and provincial. Domestic empire is founded upon dominion' and 'Dominion is property real or personal, that is to say, in lands, or in mony and goods.' [472] Political representation with no wealth and property means little or no power, so widespread land ownership would be the 'foundation' of the new English polity, which would be reflected in the 'superstructures' of political representation, namely a Senate with 300 members and Representatives with 1,050 members. Individuals with a yearly income lower than £1000 would constitute the majority of Representatives by filling four-sevenths of the places in order to constitute a popular government.[473]

Hume dismisses the whole blueprint of the commonwealth of Oceana. He rejects the economic foundation because it is 'impracticable. Men will soon learn the art, which was practised in ancient Rome, of concealing their possessions under other people's names', while rotation of the political superstructure 'is inconvenient, by throwing men, of whatever abilities, by intervals, out of public employments', and it 'provides not a sufficient security for liberty, or the redress of grievances' between the Senate and the Representatives.[474] He then presents his own blueprint on the superstructure of a republic describing the different rules and composition of the Senate and the Representatives. However, he does not advance any economic reform supporting such a political representation. Epistemically and methodologically this is an important step back, which ignores the important political discovery Harrington made relating the distribution of wealth and property to any design in politics, arguing that the latter one must

[471] *Ibid.*, see pp. 106, 108, 155 and 159; see C. Webster (ed) (1974), pp. 23-61 C. B. for further details on Harrington's Agrarian Law.

[472] *Ibid.*, p. 91, see also pp. 287-193.

[473] *Ibid.*, pp. 306 and 448.

[474] D. Hume (1777), p. 515.

be adequately supported by the former. Therefore, political blueprints must be consistent with the distribution of wealth and property. A blueprint in politics supported on an inadequate distribution of wealth and property is a bad blueprint; it constitutes an error of design.

Note that Hume has rejected equality because it is impracticable but not because it is inconsistent with a utilitarian morality. In the *Enquiry*, he discusses the egalitarian reforms proposed by the Levellers; there he recognises that 'it must, indeed, be confessed, that nature is so liberal to mankind, that, were her presents equally divided among the species, and improved by art and industry, every individual would enjoy all the necessaries. It must also be confessed, that, whenever we depart from this equality, we rob the poor of more satisfaction than we add to the rich.'[475]

Hume accepts that the 'the rule of equality' 'would be highly *useful'* and quotes the cases of Sparta, Rome, and 'many Greek cities' as successful examples of such equality. However, he still rejects such a rule again because it is 'impracticable', and also because it is 'pernicious'. It can be pernicious because it can produce opposite effects by increasing poverty instead of reducing it. This can happen because any accomplished equality would almost immediately break apart due to individual differences in 'art, care and industry'. First, because if individual art, care and industry are supervised this can affect productivity causing 'most extreme indigence', so that 'instead of preventing want and beggary in a few, render it unavoidable to the whole community.' Second, because in order to maintain strict equality the government needs extensive powers, which can easily lead to tyranny. Third, because perfect equality destroys the grounds for authority and subordination by reducing 'all power to a level, as well as property.'[476]

On the whole, Hume's argument against equality is unbalanced, and therefore his conclusions are not entirely justified. He quotes at least three cases of success implementing equality and makes four objections, namely the possibility of concealing the value and size of

[475] D. Hume (1772), SBN193-194.

[476] All quotations in this paragraph are taken from D. Hume (1772), SBN193-195.

property, widespread indigence and poverty, tyranny and the extinction of authority and subordination. Although, these objections are plausible they remain hypothetical because no example or evidence is quoted, and they may also clash with the cases of success he quotes. He does not explain why equality was successfully accomplished in those cases. A comparison of laws and policies between successful and unsuccessful cases of equality would have made a balanced argument with stronger conclusions. For instance, by comparing whatever similarities may be found with the Commonwealth of England, which seems to be a case Hume had in mind. Surely, there would be important design and engineering lessons to be learnt from both successful and unsuccessful cases, which would be in the interest of utilitarian morality.

I reply to Hume's objections by referring to some of the solutions discussed in previous chapters taken from current design. The four principles of design and engineering listed in the previous section are used in the next section, where the problem of self-interested knaves is discussed. Harrington does not anticipate the possibility of people concealing the value and size of their land, so this objection from Hume is a fair one; it poses a challenge for the designer who has to devise the mechanisms which can prevent such behaviour. In chapter two this problem is discussed as a problem of 'cheating', where cheating can be made look legal. Currently, it is the job of mechanism design theorists to try to solve this problem by devising mechanisms for the revelation of true preferences.[477]

The prediction Hume makes on the creation of widespread indigence and poverty deserves two replies. First, in a general unspecific sense, constant control and the requisition of part of the productive output from any individual could indeed cause a decrease in productivity and the overall output from society. However, just as the problem of concealing the value and size of land can be solved with the design of appropriate mechanisms, the bad effects of equality may have in reducing productivity could also be solved. Decentralisation, small

[477] See L Hurwicz (1960), (1973) and (2006).

government and direct local distribution of social outputs are some of the recent solutions devised for this problem and for the prevention of tyranny and oppression.

Second, in a specific concrete sense widespread indigence and poverty existed already before and during the eighteenth century when the new English commercial economy was booming, and the new utilitarian morality had been adopted. Charity and beneficence were clearly insufficient to tackle the problem. This prompted and justified the debate for an egalitarian distribution, which would not only reduce indigence and poverty preventing the outbreak of war and violence, but it would also increase the social output by stimulating self-interest and productivity through widespread ownership of land and new opportunities for income. Both James Harrington and the Levellers did not argue for the abolition of private property, where self-interest may be severely reduced. Instead, they argue for an egalitarian distribution of property, which essentially relied on self-interest.[478]

The objection Hume makes on the extinction of authority and subordination caused by a more egalitarian distribution of property stands against the long struggle for liberty, which started with the English Civil War and continued through the Glorious Revolution, and which the Levellers and James Harrington were a part of. Such an extinction seems to be consistent with utilitarian morality because of the public benefits it can create by reducing the size of government, particularly the part tackling the bad effects of unemployment, low income and lack of property or land, such as poverty, crime, illiteracy, illness and food deprivation. Furthermore, Hume does not provide a utilitarian argument which justifies the preservation of authority and subordination. His defence of authority seems to represent the values inherited from the Restoration, which materialised with the formation of the Parliament of Great Britain and the bureaucracy of ministries and offices attached to it.

[478] See C. B. Macpherson (1962), P. Baker and E. Vernon, Elliot (ed) (2012); and M. Levy (1983) for different accounts on self-interest in the work of James Harrington and the reforms championed by the Levellers.

If the replies to Hume's objections are accepted, equality becomes practicable and its possible pernicious effects can be prevented, while its consistency with a utilitarian morality and a psychology of self-interest, sympathy and sentiments of humanity is retained. The progress Hume made on design is comparatively less than the progress he made on moral psychology. Both Hume and Mandeville laid the foundations of modern moral psychology, making a lasting contribution. One of the most important of these contributions has been the distinction between natural virtues and artificial virtues such as justice. This distinction opened up a vast new domain for the design and engineering of artificial behaviour in politics and economics, which can meet the requirements of distributive justice.

5.5. Self-interested knaves

In section 5.3., the redirection of socially harmful greedy self-interest became crucial for human sociality and for the realisation of utilitarian morality. Hume explains that only 'the degrees of men's sagacity or folly' should be considered for turning harmful narrow self-interest into an enlightened socially useful one. So, the question 'concerning the wickedness or goodness of human nature, enters not in the least into that other question concerning the origin of society.'[479]

Most likely, the rejection of natural wickedness was a response to Mandeville, who argued that humans are already wicked in the state of nature, so any explanation of human sociality must include this psychological feature. 'Sagacity' and 'folly' are semantically equivalent to the current terms 'rationality' and 'irrationality', so according to Hume a fool or irrational greedy self-interested individual would break the rules by seizing wealth and property she is not entitled to. She would then be expelled from society, and as a consequence her own expected utility would be drastically reduced. In contrast, a rational greedy self-interested individual keeps the same expected utility by following the rules.

[479] D. Hume (1739-40), SB492.

A third possibility opens when wickedness is added to rationality, greed and self-interest giving rise to the knave. The first case of a knave to be discussed belongs to party politics, the second to civil society. In the first case, self-interested knaves act openly and insensibly; in the second case they act secretly and sensibly. The aim of the analysis and discussion of these two cases is to evaluate Hume's psychological theory on the means it offers to the designer for preventing knavish behaviour, which harms society and therefore subverts utilitarian morality.

Knaves in politics. In 1741, twenty-one years after Mandeville introduced the knaves principle, Hume reintroduced the principle to politics in his essay 'Of the Independence of Parliament', where he discusses a failure in the design of the mixed government established after the Glorious Revolution. Mandeville was mainly concerned with the abuses and absolute power exerted by the Crown, so he argued for a constitutional reform reducing the power of kings and queens. Hume had similar concerns for the Commons but unlike Mandeville he did not argue for a constitutional reform to solve the problem.

In his essay, Hume does not acknowledge Mandeville as the author of the knaves principle; he makes only a generic reference to those 'political writers' who have established this maxim. It is important to point this out because currently economists, political scientists and philosophers refer to Hume as the proponent of the knaves principle and some ideas related to it. Hume presents the knaves principle as follows:

> Political writers have established as a maxim, that, in contriving any system of government, and fixing the several checks and controuls of the constitution, every man ought to be supposed a *knave*, and to have no other end, in all his actions, than private interest. By this interest we must govern him, and, by means of it, make him, notwithstanding his insatiable avarice and ambition, co-operate to public good. Without this, say they, we shall in vain boast of the advantages of any constitution, and shall find, in the end, that we have no security for our liberties or possessions except the good will of our rulers; that is, we shall have no security at all. It is, therefore, a just *political* maxim, *that every man must be supposed to a knave.*[480]

[480] D. Hume (1777), p. 42.

Like Mandeville, Hume regards self-interest as a dominant psychological motivation in politics, that is to say, public interest is not the dominant motivation. Because greed from self-interested politicians can actually harm society, Hume argues for placing constraints on them, so that public interest can be met. This can be done by preventing the concentration of wealth and power in the hands of a few politicians by redistributing wealth and power among the different branches of government. He explains that 'where the power is distributed among several courts, and several orders of men, we should always consider the separate interest of each court, and each order; and, if we find that, by the skilful distribution of power, this interest must necessarily, in its operation, concur with the public, we may pronounce that government to be wise and happy.'[481]

Hume was concerned with the lack of effective controls the Crown and the Lords had over the Commons. The Commons had become too independent since they had full power to decide many laws and other important decisions such as the allocation of a budget to the Crown. New bills voted in both houses were simply approved by the Crown, which made no use of the veto power it had to stop laws that were harmful to their interest and that of the Lords. The king did not have any real power since he needed the approval from the Commons on many decisions and actions, and the Commons were not expected to give away any share of their power. To think otherwise would be foolish and inconsistent with the moral psychology of self-interest.[482]

Hume explains that 'Honour, is a great check upon mankind'[483] which can help prevent abuse and corruption mostly when women and men act individually, but not when they act collectively organised in groups such as political parties. Abuse, knavery and other forms of corruption in politics become worse when there is no enforcement and no punishment considered in the law. He argues that after the Glorious Revolution the distribution of power

[481] *Ibid.*, p. 43.
[482] *Ibid.*, pp. 44-45.
[483] *Ibid.*, p. 43.

became corrupted because there was no control over the actions of the Commons. The mixed form of government, which was devised as the means for curbing the Crown by giving more power to Commons, relied on a mistaken moral psychology applied to political behaviour. It did not consider the universality of self-interest and its bad effects; it ignored the knaves principle advanced by Mandeville twenty years earlier.

Consistently with considerations of self-interest and knavery, it would be equally naïve to grant more power to the Crown, simply because this would create a harmful imbalance of power, placing at risk the democratic progress made with the Glorious Revolution. The allocation of more power to the Lords was not an option either because they were allies of the Crown, so it would create a greater imbalance of power. Hume correctly describes these imbalances as the 'paradox of limited monarchy', he regarded these imbalances as a form of political 'corruption and dependence',[484] which he ultimately accepts making no suggestion on a new constitutional design. Although, he believes that a republican form of government is a better alternative because 'checks and controls are more regular in their operation' and 'authority is distributed among several assemblies or senates', he makes no commitment and no argument for it, he just accepts that 'a limited monarchy admits not any such stability.'[485] Politically and economically Hume was largely conservative; he maintained a generally moderate position and supported only a few progressive reforms. In France he was actually praised as a conservative intellectual.[486]

A shift in the behaviour of the Commons from self-interest to public interest could provide a solution to the paradox. As an example, he points out the inefficiencies created in the delivery of public services by the imbalance of power, which occurs when the Commons places

[484] *Ibid.*, p. 46.

[485] *Ibid.*, p. 46.

[486] See L. Bongie (1965, 2000) for an account on the political influence Hume had among conservative intellectuals and politicians in France before and after the revolution. D. Livingston (1984) claims that Hume instead of Edmund Burke should be considered as the founder of modern conservatism. N. McArthur (2006) describes Hume's political position as a precautionary conservatism. See E. Miller (1962) on the debate over the Whig or Tory position of Hume.

further conditions for approving the budget or by slowing down its release. He argues that this problem, and in general the paradox of a limited monarchy, could be solved if the 'honest and disinterested'[487] part of Commons cooperated with the Crown, acting with moderation by voluntarily restraining their own self-interest. He believes that the king could reciprocate the move by also restraining his own self-interest seeking the preservation of the monarchy; the Crown 'will always command the resolutions of the whole so far, at least, as to preserve the ancient constitution from danger.'[488] However, these hypotheses and arguments are inconsistent with the knaves principle itself, namely the claim that self-interest and knavery dominate in party politics and government. The solution Hume offers calling for patriotism and honesty enters into conflict with such a principle. This conflict follows from a moral psychology with two fundamental features, namely self-interest and sentiments of humanity.

The conflict may find a solution in the four principles of design and engineering discussed in section 5.3. Such principles must be tested against the prospects they have for turning self-interested knaves into patriots with a public interest. Principle number (i) dictates redirection of self-interest. In this case it cannot be applied because there is no better alternative offered to the Commons, and there is no threat either to dissolve government or have a pro-monarchy uprising, which would place the interest of the Commons at risk. Principle number (iii) had already failed, because no education, conversation or religious indoctrination had the expected effects on the self-interested knaves because they were actually harming the public interest.

Principle number (ii) requires the prevention of cases of sympathy which harm the society; party zeal is one of them quoted by Hume himself. Because no redirection of self-interest is available, the designer can only rely on sentiments of humanity and benevolence, which could be excited through sympathy. This actually is what Hume is trying to do when he calls for patriotism and honesty. Principle (iv) requires the holding of a rational discussion,

[487] *Ibid.*, p. 45.
[488] *Ibid.*, p. 45.

which appeals to the moral principles regulating the whole society and the government, that is, to utilitarian principles. Hume does not discuss this option in his essay but we can imagine the debates held in the Parliament at the time making calls to the Commons to curb their zealous partisan attitude, so that the public interest could be properly met by releasing the budget in a timely manner to the Crown and by sharing power with it.

Hume argues that the patriot and the statesman are motivated by 'generous humanity'.[489] However, the Commons are not patriots because they care more for their own interest than for the public interest. In contrast, Hume seems to believe that kings, queens and their appointed ministers are patriots who have just been stopped from fully meeting the public interest because of the constraints imposed on them by the Commons. Such ambivalent claims and the null effects of principle (ii) are the product of a psychology which includes sentiments of humanity as one of its foundations, in contrast to a psychology of universal self-interest which is simpler and more likely to be effective in preventing knavery in politics.

It is a major problem not to have the means for preventing knavery and restraining self-interest which are harmful to the public interest. The moral psychology and the principles of design and engineering offer no effective means for solving the paradox, and because Hume dismisses the possibility of a republican constitutional reform, no further design solution is available. Therefore, the British society of the eighteenth century has no option but to accept the paradox of a limited monarchy and the bad effects it has on the public interest. Furthermore, the problem is not exclusive of politics but it extends to civil society, where there are also self-interested knaves.

Knaves in civil society. Hume recognises the existence of self-interested knaves in civil society as a problem for justice and for the utilitarian morality. An honest greedy self-interested individual follows the rules of justice because she is aware of the comparative advantages of doing so, when she considers the negative economic and social effects of violating those rules.

[489] D. Hume (1772), SBN227, see also SBN256.

The insensible greedy self-interested individual violates the rules of justice exposing herself to easy punishment and losing all economic and social benefits. The sensible greedy self-interested individual violates the rules of justice without exposing herself to punishment, so she keeps all economic and social benefits for herself.

The rules of justice comprise civil laws, laws of war, laws of nations and rules of property,[490] which as we saw in the previous section consist of three rules, namely stability of possession, of its transference by consent and of the performance of promises.[491] Hume explains that all rules of justice have a utilitarian justification, that is to say, 'public utility is the sole origin of justice; and that reflections on the beneficial consequences of this virtue are the sole foundation of its merit.'[492] Therefore, sensible and insensible greedy individuals who violate the rules of justice affect public utility. The insensible knave is punished by society also for utilitarian reasons[493] but the sensible knave escapes punishment, so her case represents a difficult challenge.

Within Humean moral psychology and the related principles of design, the prospects of turning the sensible knave into an honest self-interested member of society are no better than those of turning self-interested knaves in politics into patriots with a public interest. Hume explains that:

> A sensible knave, in particular incidents, may think, that an act of iniquity or infidelity will make a considerable addition to his fortune, without causing any considerable breach in the social union and confederacy. That honesty is the best policy may be a good general rule; but it is liable to many exceptions.[494]

[490] *Ibid.*, SBN187, 196-197, 205.
[491] D. Hume (1739-40), SB526.
[492] D. Hume (1772), SBN183
[493] *Ibid.* SBN187.
[494] *Ibid.*, SBN283; see also D. Hume (1739-40), SB620.

The four principles of design and engineering from section 5.3. also fail to deal with the sensible knave. Knaves cannot be publicly exposed and punished because they 'cheat with moderation and secrecy'.[495] For the same reasons, no rational discussion can be held with them, so they cannot be called into reflection and change, and there is no possibility either of directly producing in them sympathy and sentiments of humanity. As long as they continue breaking the rules of justice it becomes clear that no personal reflection has been powerful enough, and that no general attempt at producing sentiments of humanity and benevolence has succeeded either. Education, religion and conversation with others have also failed.

Hume makes a last attempt at minimising the value of the 'worthless toys and gewgaws' and 'the feverish empty amusements of luxury and expense' enjoyed by the knaves, who could instead enjoy the 'conversation, society, study, even health and the common beauties of nature, but above all the peaceful reflection of one's own conduct.'[496] But individual sensible knaves and those acting in organised crime also enjoy the benefits just listed by Hume and, contrary what he claims, they have no remorse. They enjoy peace of mind and they often feel proud of their actions and the material benefits they get just like robbers and pirates do, which are the groups Hume quotes as also needing rules of justice for their survival and adequate functioning.[497] Knaves know that allegiance to the rules of justice is artificial, that is to say, a 'noble lie' and an 'artificial duty' as Marcia Baron points out. Because of this awareness and their greedy self-interest, they 'attach less importance to acting justly and will be guided in their actions by personal interest, attachments to friends, or the variable, moral standards associated with the natural virtues rather than the inflexible rules governing the artificial virtues.'[498] David Gauthier also believes that Hume has lost the battle against the knave, and agrees with Baron

[495] *Ibid.*, SBN283.
[496] *Ibid.*, SBN283.
[497] *Ibid.,* SBN209.
[498] M. Baron (1982), p. 555.

that allegiance to the rules of justice is a lie. However, unlike her, he claims that 'people need not to be lied to, because they lie to themselves' creating 'an imaginary motive'.[499]

Indeed, in one of his last essays Hume accepts defeat; he explains that the knave has no 'humanity, no sympathy with his fellow-creatures, no desire of esteem and applause […] no remorse […] I must repeat it, my philosophy affords no remedy in such a case'.[500] From this conclusion, it follows that any honest self-interested individual in society becomes 'the cully' of his own 'integrity'.[501] Generalised knavery severely weakens society and it can easily lead to chaos and violence. In contrast, moderate knavery can keep the society functioning, particularly when the probability of spotting it is low and when the benefits the knaves get are large, which can be used to pay for protection from the police, judges and other parts of government. The sensible knave challenges the allegiance to rules of justice and the rationality from honest greedy self-interested individuals, who now may become fools unless further reasons are provided justifying their choices and behaviour.

Both the sensible knave in society and the knave in politics remain an unresolved anomaly in Hume's moral psychology and for the related principles of design and engineering. In contrast, a moral psychology of universal self-interest like that from Mandeville seems to provide better means for preventing knavish behaviour. This is because it shows the need for a smarter design of incentives, penalties and controls in society and adequate checks and balances in politics, which could have been implemented two decades before Hume published his own psychological theory of self-interest and sympathy and his descriptive sociology of utilitarian morality.

Currently, the Humean psychological theory of sympathy and the related sentiments of humanity and generosity prevail in social choice theory and welfare economics. However, this theory has lost the debate in public choice theory and neoclassical economics, where universal

[499] D. Gauthier (1992), pp. 421, 427.
[500] D. Hume (1777), pp. 169-170.
[501] D. Hume (1939-40), SB535.

self-interest has been adopted. Mistakenly, some neoclassical economists and public choice theorists quote Adam Smith, and occasionally David Hume, as the original source of the psychology of self-interest, when Mandeville should be quoted instead. The moral psychology of universal self-interest from Mandeville is quoted by Right-Libertarian economists such as Friedrich Hayek and James M. Buchanan in support of their ideas. However, as I have argued such a moral psychology and the related ideas of design and engineering should also be considered by Left-Libertarians. Similarly, they could retrospectively be applied to the republican egalitarian blueprint from James Harrington and the egalitarian reforms championed by the Levellers.

—O—

REFERENCES

Ahmed, Amel (2013) *Democracy and the Politics of Electoral System Choice: Engineering Electoral Dominance*. Cambridge: Cambridge University Press.

Alexandrova, Anna (2008) 'What Experimental Economics Teaches Us About Models', *Journal of Economic Methodology* 15 (2): 197-204

Arrow, Kenneth (1977) 'Extended Sympathy and the Possibility of Social Choice', *The American Economic Review* 67 (1): 219-225

________ and Hurwicz, Leonid (ed.) (1977) *Studies in Resource Allocation Processes*. Cambridge: Cambridge University Press.

Atkinson, Arthur B. (1998) *Poverty in Europe*. Oxford: Blackwell.

Bacon, Francis (1620) *Novum Organon*, *The Works of Francis Bacon, vol. 4*, edition by J. Spedding, R. L. Ellis, and D.D. Heath (1857-1874). London: Longman & Co.

Baier, Annette C. (1992) 'Artificial Virtues and the Equally Sensible Non-Knaves: A Response to Gauthier', *Hume Studies* XVIII (2): 429-440.

Bailer-Jones, Daniela (2009) *Scientific Models in Philosophy of Science*. Pittsburgh: Pittsburgh University Press.

Baker, Philip; Vernon, Elliot (ed) (2012) *The Agreements of the People, the Levellers, and the Constitutional Crisis of the English Revolution*. Hampshire, UK: Palgrave Macmillan

Baron, Marcia (1982) 'Hume's Noble Lie: An Account of His Artificial Virtue', *Canadian Journal of Philosophy* 12 (3): 539-555.

Bartha, Paul (2010) *By Parallel Reasoning: The Construction and Evaluation of Analogical Arguments*. Oxford: Oxford University Press.

Barry, Norman (1982) 'The Tradition of Spontaneous Order', *Literature of Liberty* 5 (2): 7-58.

Becker, Gary (1976) *The Economic Approach to Human Behaviour*. Chicago: The University of Chicago Press.

Bengson, John and Moffett, Marc (ed.) (2011) *Knowing How: Essays on Knowledge, Mind, and Action*. New York: Oxford University Press.

Bentham, Jeremy (1833) *The Rationale of Reward, in The Works of Jeremy Bentham, Vol. 2*, John Bowring (ed.), Edinburg: William Tait Publisher.

Binmore, Ken (1994) *Game Theory and the Social Contract, vol. 1, Playing Fair*. Cambridge, USA: The MIT Press.

________ (1998) *Game Theory and the Social Contract vol. 2, Just Playing*. Cambridge, USA: The MIT Press.

________ (2005) *Natural Justice*. Oxford: Oxford University Press.

________ (2009) *Rational Decisions*. Princeton: Princeton University Press.

________ and Klemperer, Peter (2002) ' The Biggest Auction Ever: The Sale of the British 3G Telecom Licences', *The Economic Journal* 112 (478): C74-C96.

Bjørnskov, Christian and Potrafke, Niklas (2011) 'Politics and privatization in Central and Eastern Europe', *Economics of Transition* 19(2): 201–230.

Black, Max (1962) *Models and Metaphors*. New York: Cornell University Press.

________ (1993) 'More about metaphor', in *Metaphor and Thought*, A. Ortony (ed.), New York: Cambridge University Press.

Bongie, Laurence (2000) *David Hume: Prophet of the Counter-revolution*. Indianapolis: Liberty Fund.

Boonin-Vail, David (1994) *Thomas Hobbes and the Science of Moral Virtue*. Cambridge: Cambridge University Press.

Borhnstedt, G. W., and Stecher, B. M. (eds.) (2002). 'What we have learned about class size reduction in California'. CA, USA: California Department of Education.

Boudon, Raymond (1974) *Education, Opportunity, and Social Inequality*. New York: John Wiley & Sons.

________ (1981) *The Logic of Social Action*. London: Routledge & Kegan Paul.

Boycko, Maxim et al. (1995) *Privatizing Russia*. Cambridge: The MIT Press.

Brennan, Geoffrey and Buchanan, J. M. (1981) 'The Normative Purpose of Economic 'Science': Rediscovery of an Eighteenth Century Method', *International Review of Law and Economics* I (2): 155-166.

Brennan, Geoffrey and Hamlin, Alan (2000) *Democratic Devices and Desires*. Cambridge: Cambridge University Press.

Brown, Peter M. (2002) 'Einstein's Gravitational Field'. CERN: European Council for Nuclear Research.

Buchanan, James M. (1979) 'Politics without Romance: A Sketch of Positive Public Choice Theory and its Normative Implications', *IHS-Journal* 3: B1-B11.

________ and Geoffrey Brennan (1983) 'Predictive Power and Choice Among Regimes', *Economic Journal* 93 (369): 89-105.

________ and Geoffrey Brennan (1985) *The Reason of Rules: Constitutional Political Economy*. With a Foreword by Robert D. Tollison and a Note by J. M. Buchanan, 2000, Indianapolis: Liberty Fund.

________ et al. (1990) *Europe's Constitutional Future*. London: Institute of Economic Affairs.

Bunge, Mario (1999) *The Sociology-Philosophy Connection*. New Brunswick: Transaction Publishers.

________ (2004) 'How Does It Work? The Search for Explanatory Mechanisms', *Philosophy of the Social Sciences* 34 (2): 182-210.

Campbell, Norman R. (1920) *Physics: The Elements*. London: Cambridge University Press.

Catania, Charles (ed.) (1988) *The Selection of Behaviour: The Operant Behaviorism of B. F. Skinner*. Cambridge: Cambridge University Press.

Carnap, Rudolf. (1928) *Logical Construction of the World and Pseudoproblems in Philosophy*. Edition with minor correction published in 2003 by Open Court Publishing Company, Illinois.

________ (1934) *The Logical Syntax of Language*. Edition from 1959 published by Littlefield, Adams and Co, New Jersey.

________ (1939) 'Foundations of Logic and Mathematics', *International Encyclopedia of Unified Science, vol. I*. Chicago: The University of Chicago Press.

________ (1945) 'On Inductive Logic', *Philosophy of Science* 12 (2): 72-97

________ (1962) *Logical Foundations of Probability*. Chicago: The University of Chicago Press.

________ (1980) "A Basic System of Inductive Logic Part II", in *Studies in Inductive Logic and Probability, vol. 2*, R.C. Jeffrey (ed.). Berkeley: University of California Press.

Carr, Edward H. (1945) *The Twenty Years' Crisis*. With a new introduction by Michael Cox, 2001, New York: Palgrave Publishers Ltd.

Cartwright, Nancy (1983) *How the Laws of Physics Lie*. Oxford: Claredon Press.

________ (1989) *Nature's Capacities and Their Measurement*. Oxford: Claredon Press.

________ (1992) 'Aristotelian Natures and the Modern Experimental Method', *Inference, Explanation, and Other Frustrations*, J. Earman (ed.), Berkeley: University of California Press.

________ (1999) *Dappled World*. Cambridge: Cambridge University Press.

________ (2007) *Hunting Causes and Using Them*. Cambridge: Cambridge University Press.

________ (2009) 'Evidence-Based Policy: What's To Be Done About Relevance?, *Philosophical Studies* 143 (1): 127-136.

________ and Stegenga, Jacob (2011) 'A Theory of Evidence for Evidence-Based Policy', *Proceedings of the British Academy* 171: 289–319.

________ and Hardie, Jeremy (2012) *Evidence-Based Policy: A Practical Guide To Doing It Better*. New York: Oxford University Press.

________ and John Pemberton (2013) 'Aristotelian Powers: Without them, What Would Modern Science do?', in *Powers and Capacities in Philosophy: The New Aristotelianism*, J. Greco, and R. Groff (eds.), 2013, New York: Routledge.

Cesarano, Filippo (2006) *Monetary Theory and Bretton Woods*. Cambridge: Cambridge University Press.

Chang, Hasok (2004) *Inventing Temperature: Measurement and Scientific Progress*. Oxford: Oxford University Press.

Chalesworth, James (ed.) (1967) *Contemporary Political Analysis*. New York: The Free Press.

Cherkaoui, Mohamed (2005) *Invisible Codes: Essays on Generative Mechanisms*. Oxford: The Bardwell Press.

Chiesa, Mecca (1992) 'Radical Behaviorism and Scientific Frameworks: From Mechanisms to Relational Accounts', *American Psychologist* 47 (11): 1287-1299.

Chipman, John et al. (ed.) (1990) *Preferences, Uncertainty, and Optimality: Essays in Honor of Leonid Hurwicz*. Boulder: Westview Press.

Chomsky, Noam (1971) 'Review of B.F. Skinner's Beyond Freedom and Dignity', *The New York Review of Books*, Dec 30: 18–24.

Clarke, Edward, H. (1971). "Multipart Pricing of Public Goods", *Public Choice* 11 (1): 17–33.

Coleman, D. C. (1977) *The economy of England 1450-1750*. Oxford: Oxford University Press.

Collingwood, Robin G. (1946) *The Idea of History*. Oxford: Clarendon Press.

Corbett, Krystilyn (1996) 'The Rise of Private Property Rights in the Broadcast Spectrum', *Duke Law Journal* 46: 611-650.

Cramton, Peter (1997) 'The FCC Spectrum Auctions: An Early Assessment', *Journal of Economics & Management Strategy* 6 (3): 431–495.

Da Costa, Newton and Newton, Steven (2000) "Models, Theories, and Structures: Thirty Years On', *Philosophy of Science* (67): S116-S127.

Davidson, Donald (2001) Essays on Actions and Events. New York: Oxford University Press.

De Finetti (1937) 'Foresight: Its Logical Laws, Its Subjective Sources' in *Studies in Subjective Probability* (1964), H. E. Smokler (eds), New York: John Wiley and Sons.

De Vries, Marc; Hansson, Sven; Meijers, Anthonie (eds.) (2012) *Norms in Technology*. Dordrecht: Springer.

Demeulenaere, Pierre (ed.) (2011) *Analytical Sociology and Social Mechanism*. Cambridge: Cambridge University Press.

Doran, Barbara G. (1975) 'Origins and Consolidation of Field Theory in Nineteenth-Century Britain: From the Mechanical to the Electromagnetic View of Nature', *Historical Studies in the Physical Sciences* 6: 133-260.

Doris, John M. (ed), (2010) *The Moral Psychology Handbook*. Oxford: Oxford University Press.

Duhem, Pierre (1906) *The Aim and Structure of Physical Theory*. Princeton: Princeton University Press.

Dupré, John (1993) *Disorder of Things: Metaphysical Foundations of the Disunity of Science*. Cambridge, USA: Harvard University Press.

________ (1994) 'Could There Be a Science of Economics', *Midwest Studies in Philosophy* 18: 363-378.

________ (1996) 'The Solution to the Problem of the Freedom of the Will', *Noûs* 30 (10): 385-402.

_________ (2001) 'Economics without Mechanism', in *The Economics World View*, U. Maki (ed), Cambridge: Cambridge University Press.

Durkheim, Émile (1887) 'La Science positive de la morale en Allemagne', *Revue Philosophique* 24: 33-58, 113-42, 275-84; translated to English as *Ethics and the Sociology of Morals* (1993) by Robert T. Hall, New York: Prometheus Books.

Dusheiko, Mark et al. (2007) 'The Impact of Budgets for Gatekeeping Physicians on Patient Satisfaction: Evidence from Fundholding', *Journal of Health Economics* 26: 742–762.

Einstein, Albert(1905) 'On the Electrodynamics of Moving Bodies', in M. N. Saha and S. N. Bose (1920) *The Principle of Relativity: Original Papers by A. Einstein and H. Minkowski*, University of Calcutta. Original publication in German: Einstein, A. (1905), "Zur Elektrodynamik bewegter Körper", Annalen der Physik 322 (10): 891–921

________ (1920) 'Ether and the Theory of Relativity', in *Sidelights of Relativity* (1922) Trans. by G. B. Jeffery and W. Perrett. London: Methuen & co. Ltd.

Elster, Jon (1989) *Nuts and Bolts for the Social Sciences.* Cambridge: Cambridge University Press.

________ (1995) 'Forces and Mechanisms in the Constitution-Making Process', *Duke Law Journal* 45 (2): 364-396.

________ (1999) *Alchemies of the Mind: Rationality and Emotions*. Cambridge: Cambridge University Press.

________ (2007) *Explaining Social Behaviour: More Nuts and Bolts for the Social Sciences*. Cambridge: Cambridge University Press

________ Offe, Claus; and U. Preuss (1998) *Institutional Design in Post-communist Societies*. Cambridge: Cambridge University Press.

Enthoven, Alain (1985) 'Reflections on the Management of National Health Service', Research Report. London: Nuffield Trust.

________ (1999) *In Pursuit of an Improving National Health Service*. London: The Nuffield Trust.

________ (2002) 'Introducing Market Forces into Health Care: A Tale of Two Countries', Research Report. London: The Nuffield Trust.

Eulau, Heinz (ed.) (1969) *Behavioralism in Political Science*. New York: Atherton Press.

Fahlman, Scott E. (1979) *NETL: A System for Representing and Using Real-world Knowledge*. Cambridge, USA: The MIT Press

Faraday, Michael (1846) 'Thoughts on Ray-Vibrations', in *Experimental Researches in Electricity vol. III* (1855), London: Taylor and Francis.

________ (1851) 'Twenty-Fifth series', in *Experimental Researches in Electricity vol. III* (1855), London: Taylor and Francis.

________ (1851) 'Twenty-Sixth Series', in *Experimental Researches in Electricity vol. III* (1855), London: Taylor and Francis.

________ (1855) *Experimental Researches in Electricity vol. III*. London: Taylor and Francis.

________ (1852) 'On the Physical Character of the Lines of Magnetic Force', *in Experimental Researches in Electricity vol. III* (1855), London: Taylor and Francis.

________ (1852) 'On Lines of Magnetic Force; Their Definite Character, Their Distribution within a Magnet and Through Space', in *Experimental Researches in Electricity vol. III* (1855), London: Taylor and Francis.

________ (1855) 'On Some Points of Magnetic Philosophy', in *Experimental Researches in Electricity vol. III* (1855) London: Taylor and Francis.

________ (1858) Addendum to 'On the Conservation of Energy', in M. Faraday (1859) *Experimental Researches in Chemistry and Physics*. London: Taylor and Francis.

Frey, Bruno (1997) 'A Constitution for Knaves Crowds Out Civic Virtues', *The Economic Journal* 107 (443):1043-1053

________ and Stutzer, M; Benz, M. (2001) 'Trusting Constitutions', *Économie Publique* 7 (1): 25-41.

Foster, Michael (1901). *Lectures in the History of Physiology during the Sixteenth, Seventeenth, and Eighteenth Centuries*. Cambridge: Cambridge University Press.

Fox, Robert (1971) *The Caloric Theory of Gases: from Lavoisier to Regnault*. Oxford: Claredon Press.

Frege. Gottlob (1879) *Begriffsschrift,* in *From Frege to Gödel: A source Book in Mathematical Logic* 1879–1931 (1971), Jean van Heijenoort (ed), Cambridge, USA: Harvard University Press.

________ (1979) *Posthumous Writings*, H. Hermes, F. Kambartel, and F. Kaulbach (eds.), Chicago: The University of Chicago Press.

Frigg, Roman (2006a) 'Models', in *The Philosophy of Science: An Encyclopedia, vol. 2*, S. Sarkar, J. Pfeifer, J. Garson (eds.). New York: Routledge.

Frigg, R. (2006b) 'Scientific Representation and the Semantic View of Theories', *Theoria* 55: 49-65.

Fulton, John F. (1966) *Selected Readings in the History of Physiology*. Illinois: Charles C. Thomas Publisher Ltd.
Galison, Peter (1987) *How Experiments End*. Chicago and London: The University of Chicago Press.

Gauthier, David (1982) 'Three against Justice: The Foole, the Sensible Knave, and the Lydian Shepard', *Midwest Studies in Philosophy* 7 (1): 11-29.

________ (1992) 'Artificial Virtues and the Sensible Knave', *Hume Studies*, XVIII (2): 401-428.

Geertz, Clifford (1973) *The Interpretation of Cultures*. New York: Basic Books

Gert, Bernard (1967) 'Hobbes and Psychological Egoism', *Journal of the History of Ideas* 28 (4): 503-520.

Giddens, Anthony (1984) *The Constitution of Society*. Cambridge: Polity Press.

Giere, Ronald (1988) *Explaining Science: A Cognitive Approach*. Chicago and London: The University of Chicago Press.

Gigerenzer, Gerd & P. Todd (1999) *Simple Heuristics that Make Us Smart*. New York: Oxford University Press.

Gilles, Donald (2000) *Philosophical Theories of Probability*. London and New York: Routledge.

Goldsmith, Maurice M. (2001) *Private Vices, Public Benefits: Bernard Mandeville's Social and Political Thought*. Christchurch, New Zealand: Cybereditions Corporation.

Gooding, David (1980) 'Faraday, Thomson, and the Concept of the Magnetic Field', *The British Journal for the History of Science* 13 (02): 91–120.

Gooding, Robert (ed.) (1996) The Theory of Institutional Design. Cambridge: Cambridge University Press.

Goodman, Nelson (1977). *The Structure of Appearance*. Dordrecht: Reidel Publishing Company.

Granovetter, Mark (1978) 'Threshold Models of Collective Behaviour', *The American Journal of Sociology* 83 (6): 1420-1443.

Greig, J. Y. T. (1932) *The Letters of David Hume, vol. I*. Oxford: Oxford University Press.

Groves, Theodore (1973). "Incentives in Teams". *Econometrica* 41 (4): 617–631

________ et al. (1987) *Information, Incentives, and Economic Mechanisms: Essays in Honor of Leonid Hurwicz*. Oxford: Basil Blackwell.

Guala, Francesco (2005) *The Methodology of Experimental Economics*. Cambridge: Cambridge University Press.

________ (2008) 'The Experimental Philosophy of Experimental Economics: Replies to Alexandrova, Hargreaves Heap, Hausman, and Hindriks', *Journal of Economic Methodology* 15 (2): 224-231

Hallyn, Fernand (ed.) (2000) Metaphor and Analogy of the Sciences. Dordrecht: Kluwer Academic Publishers.

Harman, Gilbert, (1965) 'The Best Explanation: Criteria for Choice Theory', *Philosophical Review*, 74: 88–95.

Harman, Peter M. (1990) The Scientific Letters and Papers of James Clerk Maxwell: 1846-1862, Vol. I. Cambridge: Cambridge University Press.

Harsanyi, John (1977) 'Morality and the Theory of Rational Behaviour', *Social Research* 4 (1): 623-656.

Harré, Rom (1960) 'Metaphor, Model and Mechanism', *Proceedings of the Aristotelian Society* 60: 101–122.

________ (1961) *Theories and Things: A Brief Study in Prescriptive Metaphysics*. London & New York: Sheed and Ward.

________ (1970) *The Principles of Scientific Thinking*. London: The Macmillan Press

________ (1972) *The Philosophies of Science*. Oxford: Oxford University Press.

________ (1986) *Varieties of Realism*: *A Rationale for the Natural Sciences*. Oxford: Basil Blackwell.

________ (1988) 'Where Models and Analogies Really Count', *International Studies in the Philosophy of Science* 2 (2): 118–133.

________ (1999) 'The Rediscovery of the Human Mind: The Discursive Approach', *Asian Journal of Social Psychology* 2 (1): 43–62.

________ and Secord, Paul F. (1972) *The Explanation of Social Behaviour*. Oxford: Basil Blackwell

________ and Madden, Edward (1975) *Causal Powers: A Theory of Natural Necessity*. Oxford: Basil Blackwell.

________ and Arronson, J. L., and Way, Eileen C. (1995) *Realism Recued: How Scientific Progress is Possible*. Illinois: Open Court.

________ and Arronson, J. L., and Way, Eileen C. (2000) 'Apparatus as Models of Nature' in *Metaphor and Analogy in the Sciences* (2000) F. Hallyn (ed), Dordrecht: Kluwer Academic publishers.

Hart, Oliver and Moore, John (1994) 'A Theory of Debt Based on the Inalienability of Human Capital', *The Quarterly Journal of Economics* 109 (4): 841-879.

Harrington, James (1656) *The Oceana and Other Works*. Edited by John Toland, London: Becket and Cadell, 1771. Quotations are taken from the digitised edition published by Liberty Fund, 2004, Indianapolis.

Hausman, Alan (1979) 'Goodman's Perfect Communities', *Synthese* 41: 185-237.

Hausman, Daniel (1998) 'Rationality and Knavery' in W. Leinfellner, and E. Köhler (eds.) *Game Theory, Experience, Rationality*, 1998, Dordrecht: Kluwer Academic Publishers.

Hawley, Katherine (2003) 'Success and Knowledge-How', *American Philosophical Quarterly* 40 (1): 19-31.

Hayek, Friedrich (1943) 'A Commodity Reserve Currency', *The Economic Journal* 53 (210/211): 176-184.

________ (1945) 'The Use of Knowledge in Society', *The American Economic Review* 35 (4): 519-530.

________ (1967) *Studies in Philosophy Politics and Economics*. London: Routledge & Kegan Paul.

________ (1978) *New Studies in Philosophy, Politics, Economics and the History of Ideas*. London: Routledge & Kegan Paul.

________ (1982) *Law, Legislation and Liberty: A New Statement of the Liberal Principles of Justice and Political Economy*. London: Routledge.

Hedström, Peter and Swedberg, Richard (1996) 'Social Mechanisms', *Acta Sociologica* 39 (3): 281-308.

________ (ed.) (1998) *Social Mechanisms: An Analytical Approach to Social Theory*. Cambridge: Cambridge University Press.

Hedström, Peter (2005) *Dissecting the Social: On the Principles of Analytical Sociology*. Cambridge: Cambridge University Press

Helman, David. (ed.) (1988) *Analogical Reasoning: Perspectives of Artificial Intelligence, Cognitive Science and Philosophy*. Dordrecht: Kluwer.

Hempel, Carl. (1965) *Aspects of Scientific Explanation and Other Essays*. New York: The Free Press.

Hesse, Mary (1953) 'Models in Physics', *British Journal for the Philosophy of Science* 4 (15):198-214.

________ (1961) *Forces and Fields: The Concept of Actions at a Distance in the History of Physics*. London: Thomas Nelson and Sons Ltd.

________ (1964) 'Analogy and Confirmation Theory', *Philosophy of Science* 31 (4):319-327.

________ (1965), 'The explanatory function of metaphor', in *Logic, Methodology and Philosophy of Science*; Y Bar-Hillel (ed), Amsterdam: North-Holland

________ (1966) *Models and Analogies*. Indiana: University of Notre Dame Press.

________ (1970) 'An Inductive Logic of Theories', *Minnesota Studies in the Philosophy of Science IV*, M. Radner and S. Winocur (eds.), Minneapolis: University of Minnesota Press.

________ (1975) 'Bayesianism and Initial Probabilities' *Minnesota Studies in the Philosophy of Science VI*, G. Maxwell and R. Anderson (eds.), Minneapolis: University of Minnesota Press.

________ (1988) 'Theories, Family Resemblances and Analogy' in D. Helman (ed) *Analogical Reasoning*. Dordrecht: Kluwer.

________ (1993) 'Models, Metaphors and Truth', in *Knowledge and Language Vol. III: Metaphor and Knowledge*, F. R. Ankersmit and J.J.A. Moij (eds), Dordrecht: Kluwer.

Herschel, John (1831) *A Preliminary Discourse on the Study of Natural Philosophy*. Original edition reprinted in 2009 by Cambridge University Press.

Hitlin, Steven and Vaisey, Stephen (eds.) (2010) *Handbook of the Sociology of Morality*. New York: Springer.

Hindriks, Frank (2008) 'The Scope of Experimental Economics', *Journal of Economic Methodology* 15 (2): 216-222.

Hobbes, Thomas (1642) *De Cive*. Critical edition by Howard Warrender, 1983, New York Oxford University Press.

_________ (1651) *Leviathan, or the Matter, Form and Power of Commonwealth, Ecclesiastical and Civil.* Critical edition by Noel Malcom, 2012, Oxford University Press.

Hume, David (1739-40) *A Treatise of Human Nature*. Critical edition in two volumes by David F. Norton and Mary J. Norton for The Clarendon Edition of the Works of David Hume, 2007. Oxford: Oxford University Press.

________ (1748) *An Enquiry Concerning Human Understanding*. Critical edition by Tom L. Beauchamp for The Clarendon Edition of the Works of David Hume, 2000. New York: Oxford University Press.

_________ (1772) *An Enquiry Concerning the Principles of Morals*. Critical edition by Tom L. Beauchamp for The Clarendon Edition of the Works of David Hume, 1998. New York: Oxford University Press.

________ (1777) *Essays: Moral, Political and Literary*. Revised edition by E.F. Miller, 1987. Indianapolis: Liberty Fund.

Hurwicz, Leonid (1960) 'Optimality and Informational Efficiency in Resource Allocation Processes', in *Mathematical Methods in The Social Sciences*, K. J. Arrow, S Karlin, and P. Suppes (eds.) Stanford: Stanford University Press. Quotations taken from a reprint in K. J. Arrow and Hurwicz (1977) Studies in Allocation Resource Processes. Cambridge: Cambridge University Press.

________ (1972) 'On Informationally Decentralized Systems', in C. B. McGuire and R. Radner (eds.) *Decision and Organization: a Volume in Honor of Jacob Marshak* (1972) Amsterdam and London: North-Holland. Quotations taken from a reprint in K. J. Arrow and Hurwicz (1977) Studies in Allocation Resource Processes. Cambridge: Cambridge University Press.

________ (1973) 'The Design of Mechanisms for Resource Allocation', *The American Economic Review* 63 (2): 1-30.

________ and Reiter, Stanley (2006) *Designing Economic Mechanisms*. Cambridge: Cambridge University Press.

Hutcheson, Francis (1726) *An Inquiry into the Original of Our Ideas of Beauty and Virtue in Two Treatises*. Edition by Wolfgang Leidhold, 2004, Indianapolis: Liberty Fund.

Jackson, Frank and Smith, Michael (eds.), (2005) *The Oxford Handbook of Contemporary Philosophy*. Oxford: Oxford University Press.

Jacobson, Marcus (1993) *Foundations of Neuroscience*. New York: Plenum Press.

Jones, Bence (1870) *The Life and Letters of Faraday vol. II*. London: Longmans, Green & Co.

Kahneman, Daniel; Slovic, Paul, and Tversky, Amos (1982) *Judgement Under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.

Kavka, Gregory (1986) *Hobbesian Moral and Political Theory*. Princeton: Princeton university Press.

Keynes, John Maynard (1921) *A Treatise on Probability*. London: MacMillan and Co. Limited.

________ (1923) *A Tract on Monetary Reform*. London: Macmillan and Co. Limited.

________ (1943) 'The Objective of International Price Stability', *The Economic Journal* 53 (210/211): 185-187.

Klemperer, Paul (2004) *Auctions: Theory and Practice*. Princeton: Princeton University Press.

Kripke, Saul (1980). *Naming and Necessity*. Cambridge, USA: Harvard University Press.

La Rochefoucauld, François de (1678) *Collected Maxims and Other Reflections*. Critical edition by E.H. Blackmore, A.M. Blackmore, and F. Giguère, 2007, New York: Oxford University Press.

Laudan, Rachel (1982) The Role of Methodology in Lyell's Science', *Studies in the History of Philosophy of Science* 13 (3):215-249.

Layton, Edwin (1987) 'Through the Looking Glass, or News from Lake Mirror Image', *Technology and Culture* 28 (3): 594-607.

Ledyard, John et. al. (1997) 'Experiments Testing Multiobject Allocation Mechanisms', *Journal of Economics and Management Strategy* 6 (3): 639-675.

Le Grand, Julian; Mays, N; and Mulligan J. (eds) (1998) *Learning from the NHS Internal Market*. London: King's Fund.

________ (2003) *Motivation, Agency, and Public Policy: Of Knights & Knaves, Paws & Queens*. Oxford: Oxford University Press.

Levy, Michael (1983) 'Freedom, Property and the Levellers: The Case of John Lilburne', *The Western Political Quarterly* 36 (1): 116-133.

Lewis, David (1986) *On the Plurality of Worlds*. Oxford: Blackwell Publishing.

Little, Daniel (1991) *Varieties of Social Explanation*. Colorado: Westview Press

________ (1998) *Microfoundations, Method and Causation*. New Brunswick and London: Transaction Publishers:

________ (2013) 'Disaggregating Historical Explanation: The Move to Social Mechanisms in the Philosophy of History', *Social Epistemology Review and Reply Collective* 2 (8): 1-7.

Lipton, Peter (2004) *Inference to the Best Explanation*. London: Routledge.

Livingston, Donal (1984) *Hume's Philosophy of Common Life*. Chicago: University of Chicago Press.

Losee, John (2001) *A Historical Introduction to the Philosophy of Science*. Oxford: Oxford University Press.

Macpherson, Crawford B. (1962) *The Political Theory of Possessive Individualism: Hobbes to Locke*. With a new introduction by F. Cunningham, 2011, Oxford: Oxford University Press.

McAfee, Preston and McMillan, John (1996) 'Analyzing the Airwaves Auction', *The Journal of Economic Perspectives* 10 (1): 159-175.

Mandeville, Bernard (1720) *Free Thoughts on Religion, the Church and National Happiness*. Edited with an introduction and textual notes on the first (1720) and second (1729) editions, bibliography and supplementary index by Irwin Primer. 2001. New Jersey: Transactions Publishers.

________ (1732) *The Fable of the Bees or Private Vices, Publick Benefits, 2 vols.*, with a Commentary, Critical, Historical, and Explanatory by F.B. Kaye, 1988, Indianapolis: Liberty Fund.

________ (1732a) *Letter to Dion*, with an Introduction by Jacob Viner, 1953, Los Angeles: University of California Press.

________ (1732b) *An Enquiry into the Origin of Honour and the Usefulness of Christianity in War*. Facsimile edition with an Introduction by M. M. Goldsmith, 1971, London: Frank Cass & Co. Ltd.

Martin, Thomas (1932-1936) *Faraday's Diary: Being the Various Philosophical Notes of Experimental Investigation Made by Michael Faraday during the Years 1820-1862, vol. V*. London: G. Bell and Sons Ltd.

Maskin , Eric and J. J. Laffont (1979) 'A Differential Approach to Expected Utility Maximizing Mechanisms', in *Aggregation and Revelation of Preferences* (1979) J. J. Laffont (ed.), Amsterdam: North Holand.

________ & S. Baliga (2003) 'Mechanism Design for the Environment', in *Handbook of Environmental Economics*, Vol. 1, K.G. Mäler and J. Vincent (eds.), Amsterdam: Elsevier Science.

Maxwell, James Clerk (1855-1856) 'On Faradays Lines of Force', Transactions of the Cambridge *Philosophical Society*, Vol. X, Part I.

________ (1861-62) 'On Physical Lines of Force', *The London, Edinburg and Philosophical Magazine and Journal of Science*; in four parts: Vol. XXI, pp. 161-175, 281-291, 338-348 parts 1-III, Vol. XXIII, pp. 12-24, 85-95 parts III-IV. Also, in *The Scientific Papers of Clerk Maxwell, vol. I* (1890), W. D. Niven (ed.), Cambridge University Press.

Maynard, Alan (1986) 'Performance Incentives in General Practice', in *Health Education and General Practice*, Teeling Smith G. (ed.), London: Office of Health Economics.

________ Marinker, Marshall and Pereira, Denis (1986) 'The Doctor, The Patient, And Their Contract: I. The General Practitioner's Contract: Why Change It?, II. A Good Practice Allowance: Is It Feasible?, III. Alternative Contracts: Are They Viable? ', *British Medical Journal* 292(6531): 1313-1315, (6532): 1374-1376, (6533): 1438-1440.

McMillan, John (1994) 'Selling Spectrum Rights', *The Journal of Economic Perspectives* 8 (3): 145-162.

Menger, Carl (1871) *Principles of Economics*. Alabama: Ludwig von Mises Institute, 2007, with a foreword by Peter G. Klein and introduction by F. A. Hayek.

Merton, Robert (1968) *Social Theory and Social Structure*. New York: Free Press.

________ (1936) 'The Unanticipated Consequences of Purposive Social Action', *American Sociological Review* 1 (6): 894-904.

Milgrom, Paul (1994) 'Access to Airwaves: Going, Going, Gone', *Stanford Business School Magazine*, June 1994.

________ (2004) *Putting Auction Theory to Work*. Cambridge: Cambridge University Press.

Miller, Eugene (1962) 'David Hume: Whig or Tory?', *New Individualist Review* 1 (4): 19-27.

Monro, Hector (1975) *The Ambivalence of Bernard Mandeville*. Oxford: Claredon Press.

Morgan, Mary and Morrison, Margaret (eds) (1999) *Models as Mediators: Perspectives in the Social and the Natural Science*. Cambridge: Cambridge University Press.

Mossner, E. C. (1980) *The Life of David Hume*. Edinburg: Nelson.

Myerson, Roger B. (1979) 'Incentive Compatibility and the Bargaining Problem', Econometrica 47: 61–74.

________ (1981) 'Optimal Auction Design', *Mathematics of Operation Research* 6: 58-73.

________ & Satterthwaite, M. (1983) 'Efficient Mechanisms for Bilateral Trading', *Journal of Economic Theory* 29: 265-81.

Nersessian, N. (1984) *Faraday to Einstein: Constructing the Meaning in Scientific Theories*. Dordrecht: Marinus Hijhoff Publishers.

Newton, Isaac (1717) *Opticks*. London: William Innys Printer.

________ (1726) *The Principia: Mathematical Principles of Natural Philosophy*. New translation by J. B. Cohen and A. Whitman with a Guide by J. B. Cohen, 1999, California: University of California Press.

Norris, Pippa (2004) *Electoral Engineering: Voting Rules and Political Behaviour*. Cambridge: Cambridge University Press.

North, Douglas (1990) *Institutions, Institutional Change, and Economic Performance*. Cambridge: Cambridge University Press.

________ L. Aston, and T. Eggertsson (1996) *Empirical Studies in Institutional Change*. Cambridge: Cambridge University Press.

Nurmi, Hannu (1998) *Rational Behaviour and the Design of Institutions*. Cheltenham: Edward Elgar Publishing Inc.

Oakeshott, Michael (1947) 'Rationalism in Politics', *Cambridge Journal* 1 (2): 81-98, 145-157. Quotations taken from M. Oakeshott (1991) *Rationalism in Politics and Other Essays*. New and Expanded Edition with a Foreword by Timothy Fuller. Indianapolis: Liberty Fund. A new edition of this paper published with minor revisions in 1962, which was reprinted in 1991.

Ormrod, David (2003) *The Rise of Commercial Empires: England and the Netherlands in the Age of Mercantilism, 1650-1770*. Cambridge: Cambridge University Press.

Ortony, Andrew (1993) *Metaphor and Thought*. New York: Cambridge University Press.

Ostrom, Elinor (2005) *Understanding Institutional Diversity*. Princeton: Princeton University Press.

Parsons, Talcott (1939) 'The Professions and Social Structure', *Social Forces*. 17 (4): 457-467.

________ (1947) 'Certain Primary Sources and Patterns of Aggression in the Social Structure of the Western World', *Psychiatry* 10 (2): 167-181.

________ (1951) *The Social System*. London: Routledge & Kegan Paul Ltd. Quotations taken from the 1991 edition with a New Preface by Bryan S. Turner also published by Routledge & Kegan Paul.

Persaud, T. V. N. (1997) *A History of Anatomy: The Post-Vesalian Era*. Illinois: Charles C. Thomas Publisher Ltd.

Peters, B. Guy (2012*) Institutional Theory in Political Science*. New York: Continuum Publishing Group.

Pettit, Philip (1997) *Republicanism: A Theory of Freedom and Government*. Oxford: Oxford University Press.

________ (2002) *Rules, Reasons and Norms*. Oxford: Oxford University Press.

Pitt, Joseph (2001) 'What Engineers Know', *Tecné* 5 (3): 17-30.

Plott, Charles R. (1997) 'Laboratory Experimental Testbeds: Application to the PCS Auction', *Journal of Economics & Management Strategy* 6 (3): 605–638.

Polanyi, Michael (1940) *The Contempt of Freedom*. London: Watts & Co.

Popper, Karl (1961) *The Poverty of Historicism*. London: Routledge.

________ (1966) *The Open Society and Its Enemies*. London: Routledge. Quotations taken from the 2002 one-volume hardback edition with a 'Preface' by Václav Havel, and 'Personal Reflections' by E. H. Gombrich, also published by Routledge.

Quine, Willard van O. (1969) *Ontological Relativity and Other Essays*. New York: Columbia University Press.

Rawls, John (1999) *A Theory of Justice*. Oxford: Oxford University Press.

________ (2001*) Justice as Fairness*. Cambridge, USA: The Belknap Press of Harvard University Press.

Reid, Thomas (1785) *Essays on the Intellectual Powers of Man*. Critical edition by D. Brookes, 2002, Edinburg: Edinburg University Press.

Reilly, Benjamin (2001) Democracy in Divided Societies: Electoral Engineering for Conflict Management. Cambridge: Cambridge University Press.

Rodriguez-Pereyra, Gonzalo (1999) 'Resemblance Nominalism and the Imperfect Community', *Philosophy and Phenomenological Research* 59: 965-982.

Rothschuh, Karl E. (1973) *History of Physiology*. Florida: Krieger Publishing Co.

Ruse, Michael (1976) 'Charles Lyell and the Philosophers of Science', *British Journal for the History of Science* 9 (2): 121-131.

Ryle, Gilbert (1946) 'Knowing How and Knowing That', Proceedings of the Aristotelian Society 46: 1-16.

Sartori, Giovanni (1968) 'Political Development and Political Engineering', *Public Policy* 17: 261-298.

________ (1997) *Comparative Constitutional Engineering*. New York: New York University Press.

Schelling, Thomas (1978) *Micromotives and Macrobehavior*. New York: W. W. Norton & Company. Quotations taken from the 2006 edition also published by W. W. Norton & Company.

Sen, Amartya Sen (1977) 'Rational Fools: A Critique of the Behavioural Foundations of Economic Theory', *Philosophy and Public Affairs* 6 (4): 317-344.

________ (1982). *Poverty and Famines*. Oxford: Oxford University Press,.

Shaftesbury, A. A. Copper, Earl of (1732) *Characteristicks of Men, Manners, Opinions, Times. 3 vols.*, Douglas den Uyl (ed.), 2001, Indianapolis: Liberty Fund.

Skinner, B. F. (1938) *The Behaviour of Organisms*. New York: Appleton-Century-Crofts.

________ (1948) "Superstition' in the Pigeon', *Journal of Experimental Psychology* 38 (2): 168-172.

________ (1958) 'Reinforcement Today', *American Psychologist* 13(3): 94-99.

________ (1963) 'Operant Behaviour', *American Psychologist* 18 (8):503-515.
________ (1971) *Beyond Freedom and Dignity*. Indianapolis: Hackett Publishing Company.

________ (1981) 'Selection by Consequences', *Science* 213 (4507): 501-504.

________ and C. B. Ferster (1957) *Schedules of Reinforcement.* New York: Appleton-Century-Crofts.

Schilpp, Arthur (ed) (1963) *The Philosophy of Rudolf Carnap*. Illinois: Open Court.

Scott-Taggart, J. M. (1966) 'Mandeville: Cynic or Fool?', *The Philosophical Quarterly*, 16 (64): 221-232.
Sidgwick, Henry (1907) *Methods of Ethics*. Indianapolis: Hackett Publishing Company.

Simon, Herbert (1996) *The Sciences of the Artificial.* Cambridge: The MIT Press.

Smith. Adam (1790) *The Theory of Moral Sentiments.* Critical edition by D. D. Raphael and A. L. Macfie. Glasgow Edition of the Works and Correspondence of Adam Smith, 1976. Photographic reproduction published by Liberty Fund, Indianapolis, 1982.

Sober, Elliot and Wilson, David. S. (1999) *Unto Others: The Evolution and Psychology of Unselfish Behaviour.* Cambridge, USA: Harvard University Press.

Spoor, Christopher and Munro, James (2003) 'Do Budget-holding Physicians Respond to Price? The Case of Fundholding in the UK', *Health Services Management Research* 16 (4): 261-267.

Stanley, John (2011) 'Knowing (How)'*, Noûs* 45 (2): 207–238

________ and Williamson, Timothy (2001) 'Knowing How', *The Journal of Philosophy* 98 (8): 411-444.

Steil, Benn (2013) *The Battle of Breton Woods*. Princeton: Princeton University Press.

Tao, Y. and Zhiguo, L. (2012) *China's Economic Zones: Design, Implementation and Impact. Reading*, UK: Paths International Limited.

Taylor, David W. (1924) 'Comparison of Model Propeller Experimentation in Three Nations', *Transactions of the Society of Naval Architects and Marine Engineers* 32: 61-83.

Thagard, Paul R. (1978) 'The Best Explanation: Criteria for Theory Choice', *The Journal of Philosophy* 75 (2): 76-92.

Than, Mya and Tan, Joseph L. H. (ed.) (1993) *Vietnam's Dilemma and Options*. Singapore: Institute of Southeast Asian Studies.

Thompson, Silvanus P. (1910) *Life of William Thomson*. London: Macmillan & Co. Limited.

Thomson, William (1847) 'On a Mechanical Representation of Electric, Magnetic and Galvanic Forces', *Cambridge and Dublin Mathematical Journal* (2) 61-64.

________ (1851) 'A Mathematical Theory of Magnetism', *Philosophical Transactions of the Royal Society of London* 141 (1851): pp. 243-268.

________ (1872). *Reprint of Papers on Electrostatics and Magnetism*. London: MacMillan & Co.

Van Fraassen, Bas (1980) *The Scientific Image*. Oxford: Claredon Press.

Valdés, Juan Gabriel (1995*) Pinochet's Economists: The Chicago School in Chile*. Cambridge: Cambridge University Press.

Vallentyne, Peter and Steiner, Hillel (ed.) (2000) *Left-Libertarianism and Its Critics*. New York: Palgrave.

Vickrey, William (1961) 'Counterspeculation, Auctions, and Competitive Sealed Tenders', *The Journal of Finance* 16 (1): 8-37.

Vincenti, Walter (1990) *What Engineers Know and How They Know It.* Baltimore: The John Hopkins University.

Voigt, Stefan (ed.) (2002) *Constitutions, Markets and Law*. Cheltenham, UK: Edward Elgar Publishing Limited.

________ (ed.) (2013) *Design of Constitutions*. Cheltenham, UK: Edward Elgar Publishing Limited

Weber, Max (1922) *Economy and Society*, edited by G. Roth and C. Wittich, 1978, Berkeley: University of California Press.

Weggel, Oskar (2007) 'Vietnam's Policy of Economic Zoning', *Journal of Current Southeast Asian Affairs* 26 (6): 79-97.

Weimer, David (ed.) (1995) *Institutional Design*. Boston: Kluwer Academic Publishers.

Weisbrot, Mark and Johnston, Jake (2012) 'Venezuela's Economic Recovery: Is it Sustainable?', Research Paper, Washington: Center for Economic and Policy Research

Webster, Charles (ed) (1974) *The Intellectual Revolution of the Seventeenth Century*. Oxford: Routledge.

Williamson, Oliver (2000) 'The New Institutional Economics: Taking Stock, Looking Ahead', *Journal of Economics Literature* 38 (3): 595-613.

Wispé, Lauren (1991) *The Psychology of Sympathy*. New York: Plenum Press.

Zeng, Douglas Z. (ed) (2010) *Building Engines for Growth and Competitiveness In China*. Washington: The World Bank.

Zielonka, Jan (2001) *Democratic Consolidation in Eastern Europe, vol. 1: Institutional Engineering*. New York: Oxford University Press.

Zwolinski, Matt (2007) 'Sweatshops, Choice, and Exploitation', *Business Ethics Quarterly* 17 (4): 689-727.

—O—