METHODOLOGY OF SCIENCE

AN INTRODUCTION

Lukáš Bielik

COMENIUS UNIVERSITY IN BRATISLAVA · 2019

METHODOLOGY OF SCIENCE

AN INTRODUCTION

LUKÁŠ BIELIK

COMENIUS UNIVERSITY IN BRATISLAVA · 2019

This work was supported by the VEGA grant no. 1/0036/17.

Scientific review: prof. PhDr. František Gahér, CSc.
doc. PhDr. Igor Hanzel, CSc.

Typeset in EB Garamond using LaTeX

1st edition, 232 pages, 10.75 publisher's sheets
Published in electronic form by
Comenius University in Bratislava, 2019
ISBN 978-80-223-4782-2

CONTENTS

LIST OF ABBREVIATIONS

D-N	deductive-nomological
H-D	hypothetico-deductive
I-S	inductive-statistical
INUS	insufficient but non-redundant part of an unnecessary but sufficient (condition)
SR	statistical relevance

LIST OF SYMBOLS

$(\exists x)$	existential quantifier; "there exists (at least one) x such that…"
$(\forall x)$	universal quantifier; "for all x's it holds that…"
$\neg A$	logical not (negation); "it is not the case that A" or "non-A"
$A \wedge B$	logical and (conjunction); "A and B"
$A \vee B$	logical or (disjunction); "A or B"
$A \rightarrow B$	material conditional (implication); "if A, then B"

$A \leftrightarrow B$	material biconditional (equivalence); "A if and only if B"
$A \vDash B$	logical entailment; "A entails B" or "B follows from A"
$A \nvDash B$	negative logical entailment; "A does not entail B" or "B does not follow from A"
$F(a)$	"a has property F" or "a is an F"
ϕ, ψ	the Greek letters "phi", "psi" (typically denoting properties)
$\{X, Y\}$	a set of elements X, Y
$\langle Y, X \rangle$	a tuple of elements X, Y
$x \in S$	"x is an element of set S"
$S \cap T$	"the intersection of sets S and T"
$S \cup T$	"the union of sets S and T"
$S \setminus T$	"the difference of sets S and T"
$S \times T$	(on sets) "the Cartesian product of sets S and T"
xRy	"x is in relation R to y"; equivalently, $R(x, y)$
$P(X \mid Y)$	"the probability of X assuming that Y is true"
$[0, 1]$	"a closed interval from zero to one"
$\sum_i x_i$	"the sum of x_i over all values of i"
$\lvert x \rvert$	"the absolute value of x"
$\emptyset$	empty set
$\mathbb{N}$	"the set of all natural numbers"
$\mathbb{R}$	"the set of all real numbers"

PREFACE

The capacity to think critically about one's scientific discipline, the ability to identify and analyze the general assumptions underpinning it, and the capacity to recognize the methodological purposes of a method are skills that are prized not only by those wishing to thoroughly understand the discipline they study, but also by those interested in the (general) possibilities and frontiers of knowledge provided by science. Those who study social science and the humanities (among others) typically have to learn, understand and develop the ability to critically appraise the *system of knowledge* the discipline engages with, whilst also learning about the *principles* and *methods* that generated it. Although disciplines vary in terms of what is studied (the object of investigation), the goals pursued, and their research methods, this variety does not preclude us from studying the *methodological features* they share.

This book arose out of a desire to contribute to explanations of the methodological features that are sufficiently *general* as to apply to (almost) all the empirical sciences – and, therefore, to the natural and social sciences, just as much as the humanities. The topics we will cover relate to the *general philosophy* (or *methodology*) *of science*. Although the material is suitable for use as a textbook for introductory courses in the methodology or philosophy of science, the selection and treatment of some of the topics are reflective of the author's views. I would encourage readers to approach the chapters critically, patiently, and with an open mind.

The book is divided into six chapters. The first, "Science and its methodological characteristics", tackles the question of how we can distinguish science from other fields, such as "common sense", pseudoscientific systems and religion. In the second chapter, "A toolbox of scientific methods", we zoom in on some of

the methods of science used at the linguistic (conceptual) and empirical levels of scientific research. Chapter 3, "Types of scientific research: the H-D model" introduces the hypothetico-deductive model, a model of scientific research which is illustrated using (certain parts of) Durkheim's study of *Suicide*. It offers a basic characterization of the main stages of (empirical) research and looks at the roles played by the various methods. The fourth chapter, "Hypotheses and empirical evidence", concentrates on the relationship between empirical data (or evidence) on the one hand and (testable) hypotheses on the other. We introduce some of the main approaches and basic concepts used to test hypotheses. In some disciplines, "causal language" is used to describe or explain certain phenomena. Therefore, the fifth chapter, "Causation and its role in science" provides an overview of some of the philosophical approaches to questions such as "What are causes?" and "Under what conditions can an event, phenomenon, or fact, be identified as the cause of another event, phenomenon or fact?". Finally, the last chapter, "Scientific explanation", offers a bird's-eye view and critical analysis of the main models of scientific explanation that dominated the methodological debates in the latter half of the 20th century.

When writing this textbook, I drew on various parts of my more extensive book, *Methodological Aspects of Science* (in Slovak). In particular, Chapters 1, 2 and 4 of this textbook use material from Chapters 1, 3, and 5 (respectively) of my Slovak monograph. Chapter 3 of this textbook shares a common framework with Chapter 4 of *Methodological Aspects of Science*, but the case study and its analysis presented here are completely new. Chapters 5 and 6 of this textbook represent reduced and modified versions of Chapters 6 and 7 of the Slovak monograph. All chapters in this textbook are accompanied by a series of questions for review which highlight the main points of the material discussed. I hope the textbook succeeds in bringing together substantial material for all those eager to think critically about their own discipline.

I want to thank prof. PhDr. Darina Malová, PhD. (Department of Political Science) and my colleague, Mgr. Juraj Halas, PhD. (Department of Logic and the Methodology of Sciences), for the idea of preparing an English-language textbook for students at the Comenius University's Faculty of Arts. Moreover, with-

out Juraj's willingness to produce an English translation of the original Slovak manuscript, I would not have embarked on this project. I also wish to thank Catriona Menzies for editing and proofreading the final version.

Special thanks is due to my reviewers, prof. PhDr. František Gahér, CSc., and doc. PhDr. Igor Hanzel, CSc., whose perspicacious comments have been the source of many improvements to the original version of the text.

Lukáš Bielik
Bratislava, April 26, 2019

1 SCIENCE AND ITS METHODOLOGICAL CHARACTERISTICS

1.1 Science and society

Scientific knowledge has a huge impact on our everyday life. Scientific theories underpin our beliefs about the universe, society and our place in the matrix of natural and social relations. Moreover, our substantial reliance on a range of technologies – smartphones, navigation devices, credit cards – that enable us to perform our day-to-day activities is indicative of the practical applicability of scientific knowledge.

Science, and the theoretical and practical knowledge associated with it, has its origins in European civilization, but its image as an effective method of investigating both natural and social phenomena has spread far beyond the "Old Continent".[1] Even today, products or activities are often described as "scientific" (or "scientifically proven" etc.) in marketing as a means of highlighting their reliability and effectiveness. There is a widespread social awareness of the importance, significance and brilliance of science, although there are exceptions. Science is commonly believed to produce knowledge, which has traditionally been associated with justified true belief. Nonetheless, one can always find individuals, or even organized groups, that contest or reject the results of scientific inquiry, all

[1] See e.g. Losee (2001) for a comprehensive history of the main theories of scientific method from antiquity to the 20th century.

the more so if they don't fit in with their personal (religious or ideological) beliefs.[2]

Our task is to critically examine the scope and limits of scientific inquiry generally and to look at how scientific methods are used to solve different kinds of epistemic problems.[3] It is not enough to explore science merely in terms of its end result – scientific knowledge. Methodologically, the *most important* aspect of science is the specific *way* in which it *investigates* its object. To better understand science's historical successes, but also its failures, we need to look at the main components of scientific research.

In this chapter, we shall focus on those characteristic aspects of science that distinguish science from other ways in which we form our beliefs about the world and our place in it. Specifically, we will look at how science differs from "common sense" knowledge and from religious belief systems, as well as from ideologies and pseudo-scientific approaches and theories.

1.2 The problem of demarcation

One effective way of avoiding ambiguity in communication is to define the term or concept being used. Definitions are a means of identifying, establishing or explaining the meaning of terms that are crucial to a particular language or a context of communication. However, before presenting a definition of science, we shall look at some of the *methodologically relevant characteristics* of science. By *methodologically relevant* we mean those characteristics of science that capture the *basic presuppositions* underlying all scientific activity. Having examined them, we will propose a working definition of *science* that will underpin our methodological inquiries in the chapters to come.

Whether our goal is to define or simply characterize science (from a methodological point of view), in the course of doing so we will encounter what is known

[2] Here I am referring to e.g. the "scientific creationists" who believe, among other things, that the Earth is no older than approximately ten thousand years. For a critique of these views, see e.g. Kitcher (1982).

[3] An "epistemic problem" is a problem related to knowing some fact.

in the philosophy of science as **the problem of demarcation**. This is the problem of *identifying* and explicitly *distinguishing the boundary* between science (scientific knowledge) and non-scientific or unscientific (e.g. pseudoscientific) beliefs. We shall refer to all these areas as "cognitive fields". A cognitive field can be defined as (i) a *belief system* (or systems) that (ii) relies on a particular *source of belief* (such as sensory experience, inference, intuition, revelation or a book such as the Bible or the Quran etc.). Or, to put it another way, a cognitive field has a particular *means of justifying* such beliefs. In this sense, science, religion (both as a particular system of faith and as a system of religions), ideology, common sense knowledge, philosophy, but also the various pseudoscientific or unscientific approaches to knowledge (astrology, creationism, homeopathy, fortune telling etc.), are all cognitive fields.

The problem of demarcation has a long history stretching back to the thinking of the Ancient Greeks. It is a philosophical problem, since the question, "What makes science different from other approaches to knowing the world?" is closely related to what we identify as the goals and methods, ontological or epistemological presuppositions of science and the competences we ascribe to science.[4] The fact that certain elements of this theoretical view of science have changed over the long history of science may lead us to suspect that attempts to demarcate the boundaries of science are a pointless pursuit. But we should be wary of drawing hasty conclusions.[5] Instead, we shall limit ourselves to describing the methodological characteristics of modern science – science as it began to be shaped from the 16th and 17th centuries on – thus making our task easier. Nonetheless, even in modern science, there are several competing philosophical theories that differ

[4] By ontological presuppositions, we understand a sort of very general philosophical assumptions as to the kinds of "things" there are in the world or the sorts of entities (objects, properties, relations, etc.) assumed to be real in a given scientific discipline or theory. Epistemological presuppositions, then, express those conditions of *knowing* the world that we take for granted in a given discipline or theory.

[5] However, the philosopher Larry Laudan has suggested, in his influential article on "The Demise of the Demarcation Problem", that given the failure of our previous attempts to demarcate science, we should give up on this problem. See Laudan (1983).

in the way they identify the essential characteristics of science. These competing theories are also known as "theories of science" (and include approaches such as hypothetico-deductivism, falsificationism, the methodology of scientific research programs etc.).

For example, the best-known attempt at solving the demarcation problem in the 20th century was that of the Austrian-born British philosopher, Karl Raimund Popper (1902–1994), in *The Logic of Scientific Discovery* (published in English in 1959 as an expanded version of the original *Logik der Forschung*, 1934). Popper described the problem of demarcation as the "problem of finding a criterion which would enable us to distinguish between the empirical sciences on the one hand, and mathematics and logic as well as 'metaphysical' systems on the other" (Popper 2002, 11). Popper thought a metaphysical system could be a philosophical theory (such as the Ancient theory of atomism), but also any other theory that was *not empirically testable* (such as astrology, but also mathematics because mathematics does not investigate the empirical world). Popper subsequently proposed a solution to the demarcation problem: we should only consider as scientific those systems of statements (theories) which are (in principle) *falsifiable* – capable of being *refuted* by experience.

However, it later became clear that it was not possible to satisfactorily solve the demarcation problem based on criteria such as the property of statements, a specific method (Popper: falsifiability and the method of falsification) or as a characteristic approach to scientific inquiry (T. S. Kuhn: the solving of problems or puzzles; I. Lakatos: progressive research programs).[6]

Some of today's philosophers of science have therefore opted for a different approach to solving the demarcation problem (see esp. Bunge 1996; Mahner 2007; or Tuomela 1987). Rather than limiting themselves to one or two basic criteria, they suggest a much richer concept should be used: an epistemic field comprising ten to twelve characteristic traits. These traits can then be used to distinguish and

[6] Contemporary approaches to the problem of distinguishing science from non-scientific and unscientific systems of belief are discussed in Pigliucci – Boudry (2013). Laudan (1983) adopts a skeptical position on the problem of demarcation, mentioned above.

identify the cognitive fields.[7] An alternative to attempting to identify the boundaries of science using the concept of the epistemic field is the cluster approach proposed in Mahner (2013). It entails two main steps:

1. To specify 20 (or more) methodologically relevant traits that are typical of a scientific discipline. However, since the various disciplines differ in their characteristics one cannot assume that each trait (criterion) will be satisfied (to the same degree) in each discipline. A second step is therefore necessary.

2. This entails selecting a minimal *number* of traits, perhaps 15 out of 20, that any given discipline (activity) must satisfy in order to be scientific. Note that the point is not to select a set of 15 traits out of 20, but to select a *minimal number* of traits – for example 15. Then, the possible combinations of (at least) 15 of the 20 methodological traits will represent all the possible ways in which a discipline can satisfy the criteria and be considered scientific.

Although these two approaches – demarcation at the level of epistemic fields and the cluster approach – are interesting and show promise, we will opt for a simpler solution.

Our approach will be as follows: we will focus on those methodologically relevant aspects that enable us to outline a minimal *model of science*.[8] It is a model that will prove useful in later chapters. It combines, modifies and adds to the proposals put forward by Viceník (2000a, 81–84), Nola – Irzik (2005, 201–204), and Tuomela (1987, 82–88).

[7] Mahner (2007) draws on Bunge's approach and defines an epistemic field as comprising the following elements: a community of knowledge seekers; the society hosting this community; the philosophical background; the formal background; the specific background; the collection of problems; the fund of knowledge; the aims or goals; and the collection of general and specific methods.

[8] For now, a "model of science" can be viewed as a theoretical construct that represents only some aspects of science, relevant to the goals we want to pursue in this book.

1.3 Methodological traits of science

We will characterize science (primarily empirical scientific disciplines) through its *object*, *methods*, *epistemic values*, *methodological rules*, *system of knowledge*, *language*, *goals* and the *attitudes and activities of scientists*. (The characterizations proposed here are based on *descriptions* and *thinking about* scientific activity in practice, but they can also be seen as the *idealized and optimal features* exhibited, to greater or lesser degrees, in scientific disciplines and scientific activities. The resulting model comprising these characterizations leaves many other aspects of science out of the picture.)

1. The *object of inquiry* is always a non-empty set of objects that are empirical (concrete) or abstract in nature, or are a segment of the empirical reality the members of the scientific community set out to discover and to investigate the properties, relations, and attributes of (see Viceník 2000a, 81; Šefránek 1969, 11–12). The elements of the area of inquiry, be they minerals (in geology), atoms, molecules, or subatomic particles (in physics, chemistry or molecular biology), plants (botany), animals (zoology), linguistic entities (linguistics), or historical documents and events (history) and so forth, are (supposed to be) *objective* in the sense that *they exist independently of the subject* investigating them.[9] They are also objective in that any scientist who acquires the necessary theoretical and practical knowledge, and skills can investigate them (see Tuomela 1987, 85). Therefore, the object of scientific inquiry is also *intersubjectively accessible*.

We will leave it to readers to resolve the question of whether (and to what extent) the object of cognitive areas such as religion, common sense knowledge or pseudoscience is objective and intersubjectively accessible.

2. The *language of science* (i.e. the language of a discipline or a scientific theory) is a sort of hybrid language which, besides containing *special terms* (the specific terminology of that discipline), and possibly some artificial language terms (such as the language of a mathematical theory, e.g. algebra and set theory or the lan-

[9] In philosophy, a range of arguments is frequently used to question the claim scientists make in their theories about the objective existence of objects. We shall come back to these later. However, we believe that most scientists subscribe to this belief (or variants thereof).

guage of a particular logical system), also contains the ordinary natural language used in daily communication (Viceník 2000a, 83–84). Language is essential for the subsequent dissemination of scientific knowledge. It is used to formulate research problems and questions, and to express what is available to our sensory experience and thinking. Language is the objective medium of knowledge.

Of course, the language of a scientific discipline will differ in the degree to which it employs its own terminology, but also in how precise the meanings of its key terms are. The more theoretically advanced disciplines tend to insist on greater precision or semantic accuracy in the language of the discipline (Viceník 2000a, 84; Šefránek 1969, 11–30). However, even in disciplines with a less developed theoretical apparatus (such as history and archaeology), specific terms (such as "war", "historical fact", "historical event", "evidence") are used differently from the way they are used in everyday language.

There are two reasons for attempts to achieve greater precision in the use of terms in a discipline: firstly, there are many terms in natural language that have more than one meaning but are not distinguished through different spellings. Examples of this are the words "fact", "evidence" or "crown". Secondly, some terms do not have a precise and unambiguous meaning. "Young", "bald", "a lot" and many of the names of colors are typical examples. It is through defining and explication (which we discuss in Chapter 2) that we can reduce *polysemy or ambiguity* on the one hand and *vagueness* on the other hand.

3. The results of scientific inquiry are represented by a *system of knowledge*, another methodologically relevant aspect of science. Scientific knowledge is systematic and represented in language: we can identify various relations between its elements – statements or propositions that are tested on the basis of evidence (Viceník 2000a, 84).[10] For example, we can say that two or more statements that represent knowledge are logically consistent (non-contradictory) or that they relate to the same topic. We can say that a set of beliefs (statements or propositions) inductively supports another belief (statement or proposition) or we can say that

[10] A proposition is usually understood as either to refer to the meaning of a given indicative sentence (or statement), or as the truth-conditions the sentence denotes.

a certain piece of knowledge follows logically from another set of knowledge base and so on. Statements comprising scientific theories and those representing the results of empirical examination of the object always form a particular structure, which includes a plethora of relations.

Logically, the systematization of scientific knowledge minimally requires *consistency* in its beliefs (parts of knowledge), which are expressed in a theory. (Note that this applies only to *intratheoretical* consistency, or within-theory consistency; *intertheoretical* consistency, or between-theory consistency is not a general requirement.) We will leave to one side the questions of whether belief systems in other cognitive fields exhibit within-theory consistency and the extent to which non-scientific and unscientific knowledge systems can be characterized in terms of deductive and inductive relations.

4. *Methods – values – rules*. At the heart of our methodological characterization of science lie the scientific methods, epistemic or cognitive values, and methodological rules that prescribe the methods used to pursue a particular goal (obtain or realize a particular value) (see Nola – Sankey 2000; 2007). In this section, we shall briefly define our use of the word method in this book, noting in the process the link between the use of scientific methods and the values which we may (or ought to) pursue in scientific work. We will return to some of the important scientific methods in the next chapter.

Let's begin with the concept of method. Methods can be defined descriptively as well as normatively. Our discussion of the descriptive approach will draw on a simplified and modified concept of method found in the work of Vojtech Filkorn (see esp. Filkorn 1960 and 1972; but also Riška 1968 and Viceník 2000a). The normative approach is explained in a series of papers (Bielik et al. 2014a,b,c,d) and in a book by Zouhar et al. (2017).

We understand a *method* to be a *general* and *repeatable sequence* of conceptual (theoretical) or empirical *operations* (in a few cases a single operation) that, when *applied* to an *appropriate starting point* (the area the method can be applied to), will *lead us* through a finite number of steps to a particular *outcome* (the goal, or product of the method). These operations are the "steps" that make up the method. But we can also define *method normatively* as a *system of instructions*

specifying the process of getting (at least in principle) from the *problem* or *task* to the *solution* (see Bielik et al. 2014a; Zouhar et al. 2017).

It is useful to further divide scientific methods into: (a) conceptual or theoretical methods, such as definition, explication, analysis (conceptual, linguistic, logical etc.), synthesis, classification, abstraction and idealization, various modes of deductive and inductive inference (argumentation), and so forth; and (b) empirical or practical methods, such as observation, data-collection surveys, structured interview, measurement (e.g. of physical magnitudes) and experiments. When both theoretical and empirical methods are used a range of additional methods also have to be used and we call these (c) complex methods. Examples of complex methods are the testing and evaluation of hypotheses, descriptive and inferential statistics, (the application of models of) scientific explanation, causal analysis and the design and testing of scientific models. We will look more closely at some complex methods in later chapters.

The selection and use of a method is largely dependent on the *epistemic values* relating to the process of obtaining knowledge. These values include such things as the *testability* of an empirical theory and *selecting the theory that best approximates truth*. Scientific activity (at either the theoretical or practical level) therefore involves the (implicit) use of methodological rules and principles that specify the method or methods to be used to obtain a particular epistemic (cognitive) value.

When coming across the use of the term "the scientific method" (or "scientific methods") in a methodological discussion, we may discover philosophical debates on whether in fact the term "the scientific method" can be used to describe a single thing. Note that the debate is not about whether the methods mentioned above exist (analysis, classification, observation etc.), but about whether they can be placed *in a unique characteristic sequence*. The question, then, is whether there exists a single and generally accepted sequence of methods that is valid for every single discipline (or for science in general).

We shall come back to the question of whether there is a unique and generally valid *method of science* later on when we turn to the structure of scientific research. Here, we will just note that historically the various methodological views of science – views on how science should proceed – have often adopted various

positions on this. For example, according to falsificationism, science begins by proposing bold theories that give rise to predictions that can (and are supposed to be) be subjected to strict testing. If the theory is disproved through testing, then it is shown to be false; we can say it has been falsified. If the attempt at falsification fails, we can consider the theory to have been corroborated.[11]

5. *The fundamental goal of science* is to provide knowledge of the world, knowledge of truths (or facts). Basic scientific research is driven, not by the practical use of scientific knowledge, but by *curiosity* and consequent attempts to arrive at a truthful description and adequate *explanation* of the phenomena of the world we live in. Therefore, scientific activity strives to *explain and understand facts*, but also to *predict* and *reconstruct* them (retrodiction). The fact that in the history of science the search for answers to theoretical questions has led to considerable practical and technological applications is a most fortunate by-product. Nonetheless, as motivations for scientific research, they are often quite secondary or marginal. Explanation, prediction and retrodiction lie at the heart of science's overarching goal, which is to obtain reliable, truthful knowledge of the world and society we live in.

6. *Attitudes and activities*. The scientific endeavor is also associated with some characteristic attitudes rarely found in other cognitive fields. Of the many attitudes, we can highlight those that scientists hold in relation to their own or competing theories, or to scientific activity in general. For instance, *being open to criticism* on both the theories and methods we use. Another is the *willingness to revise our belief where there is good reason to doing so* (*self-correction*). Both these attitudes are closely tied to yet another: the *commitment to produce empirically testable theories*. Other methodologically relevant attitudes include striving to *use appropriate methods*, and *selecting and accepting theories regardless of our religious or political views* (see Tuomela 1987, 86–88 and Nola – Irzik 2005, 202–203).

[11] "Corroboration" is the technical term used by Popper. Popper wanted to avoid using the term "confirmed" to describe a theory that despite all best efforts could not be falsified. He therefore introduced the term "corroboration" to indicate that a theory had withstood our attempts at falsification. However, corroboration tells us nothing about the future successes or failures of the given theory. See Popper (2002).

Scientific activity can therefore be characterized as the *actualization of these six fundamental methodological traits*. It is an activity scientists pursue when investigating a certain segment of reality using certain methodological rules. The results of this activity are expressed in language and systematized in theories, models, laws or hypotheses in such a way that the object of investigation is satisfactorily described and explained, while keeping to the principles of openness and self-correction.

In addition to the aspects described above, scientific activity is also characterized by a considerable degree of autonomy. As Tuomela put it, science does not allow any "external checks of validity" (Tuomela 1987, 87). If science itself cannot correct its results (i.e. knowledge), or the methods of obtaining them, then no one else can. When formulating or justifying our hypotheses, scientists can refer only to those sources and instruments of knowledge that are also available to other members of the research community.

Not having strayed too far, in our idealized characterization, from what science is like in reality, we can now suggest a working definition of *science* based on these features:

> *Science* is *the systematic, goal-oriented activity of a research community focused on the production of reliable knowledge of the world, in which scientists investigate the objective and intersubjectively accessible objects in the world, their properties and relations, using adequate methods and rules, while expressing the results in the language of the given discipline and systematizing these results, with regard to the testability of theories, allowing for their revision or self-correction and accepting the autonomy of scientific methods and systems.*

This definition fits the goals we will pursue in later chapters. We shall come back to some of the methodological aspects of science in more detail in later chapters.

Finally, scientific disciplines are sometimes classified according to various criteria. One criterion is the distinction between *factual* and *normative* disciplines: where the goal of the former is to *describe* (as well as to explain, predict or reconstruct) part of reality, and the goal of the latter is to *prescribe* how things should be.

Physics, geology, psychology and history are all factual disciplines. Examples of normative disciplines are the various specializations in jurisprudence. (This classification does not include disciplines that neither describe the world nor prescribe how it should be, such as mathematics or logic.) Scientific disciplines can also be classified based on the type of object investigated in the given discipline. In this sense, we can distinguish between *natural science*, *social science* and *the humanities*, based on whether they investigate natural phenomena, social phenomena or man-made (social) artifacts, respectively. To make this classification exhaustive, we could add a fourth category: the *formal sciences* of mathematics and logic, which are concerned with formal methods and abstract objects or structures.

Other classifications can be found in the relevant literature (see e.g. Mahner 2007). However, here we are concerned with the empirical disciplines (i.e. the factual disciplines of natural science, social science, and the humanities).

1.4 Methodology of science: a science of science

The various scientific disciplines usually take a segment of natural or social reality as their object of investigation. However, there are also disciplines that investigate science itself. We sometimes call them "meta-sciences". *History of science*, *sociology of science, scientific policy*, *psychology of science*, *economics of science* and *ethics of science* are all examples of meta-sciences. (For a more detailed explanation of the objects of investigation and competences of these disciplines, see e.g. Viceník 2000b, 197–201). In this book, though, our focus is on the methodological features of science. These are the object of investigation in two complementary disciplines: the *methodology of science* and the *philosophy of science*.

Methodology of science is a *meta-scientific discipline* that *describes* or *prescribes the methods used in the construction, testing and justification of scientific hypotheses and theories*, as well as *the* epistemic and cognitive *values* that (should) guide *scientists* in their research. It also *analyzes* the *logical structure of the methods used* and reconstructs the *ontological and epistemological assumptions and consequences of using the given scientific methods* to pursue the goals of science.

If a methodological approach merely describes the use of methods in a particular scientific context or discipline, we call it a *descriptive methodology*. However, if it prescribes the activities and methods a scientist should use when pursuing certain goals, we call it a *normative methodology*. For more on the different ways of classifying methodologies, see Viceník (2000b) and Černík – Viceník (2011, Chapter 1).

Philosophy of science can be viewed as a complementary meta-scientific discipline that studies the *logical*, *epistemological* and *ontological* assumptions and consequences of the various scientific disciplines and of science in general. For example, while scientists produce explanations and predictions, philosophers of science ask about the conditions under which a scientific explanation can be accepted as adequate, or in which a prediction can be viewed as reliable. Similarly, while scientists put forward laws, philosophers of science are interested in questions such as "What makes something a (scientific, natural, social) law?", or "What are the consequences of conceiving of laws in this way?".

We will end this chapter with a note about terminology. The term "science" comes from the Latin "*scientia*", which is a translation of the Greek "*epistémé*", used since Ancient times to denote rigorously obtained systematic knowledge, as distinct from conjectures, guesses and unjustified beliefs. In English-speaking countries, the term "science" is usually used in a narrower sense than the Slovak "*veda*" or German "*Wissenschaft*" to refer to natural science only. In the following chapters, our inquiries on science will relate to social science and the humanities as well, so our use of the term "science" will be rather close to the Slovak or German meaning.

Study questions

1. What is a "cognitive field"?
2. What is the problem of demarcation?
3. What was Popper's solution to the demarcation problem?

4. Which of the attitudes held by scientists are methodologically relevant to their scientific work? (Specify at least two.)

5. Which methodologically relevant properties are typical of (any) of the objects of science?

6. Which general (theoretical or empirical) or specific methods are typical of the research in your discipline? (Please list at least four.)

7. What is the minimal requirement of any system of scientific knowledge (theory)?

8. What is the methodology of science concerned with and what does philosophy of science typically investigate?

2 A TOOLBOX OF SCIENTIFIC METHODS

Science's success, represented by scientific knowledge and technological advances, boils down to the application of scientific methods. These can be viewed as the abstract tools that have proved over human history to be especially reliable in expanding our knowledge. In this chapter, we will look at the basic approaches involved in scientific research. These relate to the linguistic representation of our knowledge or represent certain basic forms of reasoning. Others are aimed at obtaining and processing empirical data. Scientific methods can be viewed as the extremely useful tools scientists have at their disposal when addressing a particular scientific problem.

2.1 What makes a method scientific?

In common parlance, the term "method" refers to a *guideline or prescribed means of doing or achieving something* but also a *recipe* or *procedure* that leads to the intended result. These guidelines can be expressed as *instructions,* given in the form of imperatives such as: "Do *X*!", "Do *X* or *Y*!", "Do *X* and *Y*!", "If your aim is *Z*, do *X*!". A method can also be expressed by a *description of the steps* that lead (under standard conditions) to the intended result (or type of result).

Any *method* can be generally defined either (i) *normatively* as a *series of prescriptions* or *instructions to be performed in order to achieve the intended aim* (see Bielik et al. 2014a,b,c,d and Zouhar et al. 2017) or (ii) *descriptively* as a *description of a series of steps* or *procedures we perform to obtain a particular (type of) result* (see Filkorn 1998, Gahér 2016 and Gahér – Marko 2017).

For a system of instructions or steps to be considered a method, it must meet the following requirements:

1. the instructions or steps (operations), and the system they comprise, must be *realizable* under standard conditions – they must not contradict the laws of logic or the (assumed) laws of nature;

2. the instructions or steps (operations), and sequence thereof, must be *general* – in principle, anyone who satisfies certain basic requirements should be able to use the method;

3. the instructions or steps (operations), and sequence thereof, must be *replicable* – the same method can be used at different times and in different places.

We thus think of methods in relation to the *activities performed in pursuit of a goal* (or *type of goal*), *ideally* the goal should also be *achieved* using the method (see Riška 1968, 27; Filkorn 1972, 225). In this sense, *driving a car*, *tying one's shoelaces*, *changing a light bulb*, *baking bread* or *brewing beer* are all goal-oriented activities – as such, they are all activities directly constituted by or at least regulated by certain *methods*.

In light of the above, we can propose the following definition:

> **DEFINITION OF A METHOD**
> *A method* is *a system of instructions* or *operations* (*procedures*) that we can perform to get from our (kind of) *starting-point* (such as a problem or task) to the *goal* (solution, product).

The various methods (conceptual, empirical and complex) differ in the *type* of *instructions* (*operations*) used, as well as in the *relationship* between these instructions (operations). For example, when we follow a certain instruction as part of a method we presume that other, mutually independent instructions have already been followed. In some methods, the relationships between the instructions are much "looser", so the user can change the order in which they are performed or

even omit some of the instructions and still achieve the goal the method is intended to achieve.

How do scientific methods differ from other, non-scientific or unscientific methods? The differences can be seen at various levels:

First, scientific methods are characterized by the *specific order* of the *steps* (to which the type of instruction or operation corresponds) comprising the method. We will better understand this abstractly formulated difference once we learn about the scientific methods themselves, found further on in this chapter. Second, we use scientific methods to achieve goals that relate (directly or indirectly) to our knowledge of the world. We can therefore say that scientific methods are aimed at *cognitively relevant goals* (or types of goals).[12]

Furthermore, scientific methods are – *insofar as the history* of our knowledge of the world is concerned – the best tools we have at our disposal. We have good reason to think that scientific methods are a *reliable* way of achieving their intended goals. In other words, a reliable scientific method achieves its intended goal – *most* of the time it is used and under standard conditions.

Using these characteristic features we can now propose a working definition of *a scientific method*, which we shall then use in the remainder of the book:

> **DEFINITION OF A SCIENTIFIC METHOD**
> *A scientific method* is a *specific system of instructions* (*operations, procedures*) performed to get from one initial *cognitive state* (a problem or a task) to a final *cognitive state* (the solution to the problem or task) and which *reliably* achieves the final state.

In this chapter, we will introduce a number of scientific methods. Some of these will be defined normatively – as systems of instructions aimed at a certain goal; while others will be defined descriptively – as a series of operations which, when performed, may lead to the intended cognitive goal.

[12] A cognitively relevant goal (or kind of goal) is a goal that, for some epistemic agent or community of epistemic agents, has the potential of modifying their knowledge (or beliefs supported by evidence) or of adequately representing it.

By looking at a method as a system of instructions or operations that, when applied to a given starting-point, lead (under standard or ideal conditions) to a given result (goal), we may be giving the impression that scientific methods are *algorithms* (or that their use in tackling scientific tasks and problems is algorithmic).[13] This requires a brief comment. First, only some methods can be considered algorithms – those where the starting-point (i.e. domain of operations/instructions) and all operations/instructions are defined recursively. Our definition of method may resemble an algorithm in certain respects, but that does not mean methods generally need to be specified in such an exact way. Second, leaving aside the formal (or analytic) disciplines, such as mathematics and logic, the *use* of scientific methods (even those exhibiting the characteristics of algorithms) is rarely algorithmic. We can certainly identify algorithmic elements in the *concept* of method introduced here, as well as in certain *uses* of methods, but in general, a method is not an algorithm.

In the sections of this chapter, we will also introduce some methods by characterizing the *results obtained using the method*. For example, we will use the products (or somewhat idealized schemes) of these methods, that is definitions and explications, to express the nature of the methods of definition and explication. We will also take a similar approach to the methods of deductive and non-deductive reasoning. We will note the criteria applying to the resulting products of these methods. (The instructions or operations arising from our definition of the concept of method will meet the criteria of an adequate definition, explication etc.)

The scientific methods introduced in this chapter can be divided into two groups: (a) theoretical (conceptual) methods; and (b) empirical (practical) methods. There are historical reasons for the distinction, but we use it here for instrumental purposes. Theoretical methods mainly involve instructions that operate on concepts (entities of various kinds) of a theory. The outcome is usually another concept (or the meaning of statements etc.). They focus on the relationships between concepts or the linguistic entities that those concepts represent. By

[13] An algorithm is an exact prescription of the finite number of steps (operations) leading from the input state(s) to the intended output state(s).

contrast, empirical (practical) methods, besides being reliant on certain concepts (theories), are primarily concerned with what these concepts represent in empirical reality. Theoretical methods include definition, explication, classification and idealization, as well as deductive and non-deductive reasoning, etc., while the basic empirical methods are observation, data-collection by survey, or measurement and experiment.

There are also other complex methods, comprising both theoretical and empirical methods, such as the methods of explanation and prediction, those of hypothesis testing etc., which we will encounter in later chapters.

2.2 Theoretical methods

In this section, we will deal with the scientific methods that are primarily concerned with meanings (i.e. concepts) or language generally. Language is crucial to knowledge – it is the means by which we express, represent and communicate our knowledge and beliefs. However, our linguistic (conceptual) tools are not always suited to this function. Fortunately, there are methods for improving these tools. It is impossible to represent knowledge without formulating the procedures that enable us to express the various relationships between the parts of our knowledge. Therefore, we will also focus on the methods of reasoning that form part of some of the more complex scientific methods, such as testing and evaluating empirical hypotheses or explaining and predicting phenomena.

However, before turning to the methods themselves, we will introduce some of the linguistic categories we shall use later on.

2.2.1 Language and meaning

We will use the term "language" to refer to *a system of signs determined by a set of syntactic* and *semantic rules*. A language's syntactic rules determine the basic sequences of signs that constitute its linguistic expressions. In turn, the semantic rules determine the kinds of objects the words and phrases refer to or, in other

words, the things denoted by that language's linguistic expressions (see Cmorej 2001a, 14).

We will use the terms "concept" and "meaning" synonymously to refer to *the thing meant (signified) by a linguistic expression*. Thus, *meaning* amounts to the objective (ideal) *simple or complex abstract procedures* that determine various kinds of objects: individuals, their characteristics and the relationships between individuals; numbers; relations; the truth-conditions of sentences; and so on. *Denotation* stands for the object that we grasp in our mind through the use of meaning and that is denoted by the linguistic expressions of the given language. For example, the term "planet of the Solar system" expresses the meaning <<planet of the Solar system>> and denotes the property of *being a planet of the Solar system*.[14] Currently, eight celestial bodies have this property, but if any ceased to exist (such as Mercury, which will be consumed by the expanding Sun in a few billion years), the number of celestial bodies possessing this property would change. The actual value of this property, and of the denoted object (property, relation, etc.) in general, is called *extension*. In our case, the extension of the property of being a planet of the Solar system is the *set of objects* that are classified as the planets of the Solar system.

Let us summarize the distinctions we have noted above using the following generalizations:

> **An expression** *conveys or codifies* **meaning**
>
> **Meaning** *determines* (*identifies*) **the denotation** (individual, relation, property etc.)
>
> **An expression** *denotes* **the denotation**
>
> *The value* of **the denotation** under current circumstances (in the current state of affairs) is its **extension**

[14] It may look as if we haven't really explained the meaning of the phrase "planet of the Solar system", as we just repeated "planet of the Solar system" but enclosed it in the symbols "<<" and ">>". However, the first occurrence of the expression in the sentence (in quotation marks) is simply an *utterance*, while the second occurrence (not in quotation marks) is used to identify the meaning of the English phrase.

We have stated that the terms "meaning" and "concept", or "meaning of an expression", and "the concept conveyed by an expression" will be used as synonyms. Thus, meanings or concepts are a sort of objective, extra-linguistic instrument (procedure) that enables us to make statements about physical and abstract objects and their properties. Meanings differ from the expressions conveying them (some expressions in a language, or even in different languages, express the same meaning – hence, two different expressions can share a meaning), but they also differ from the objects (entities) they identify. The term "metal" does not conduct electricity, but it makes sense to say that a piece of metal conducts electricity. A metal may be malleable and expand in the heat, but we cannot say that about the term "metal" or its meaning. Moreover, the terms "metal", "*Metall*", and "*kov*", are all words in different languages that can be said to have the same meaning.

Some of the semantic categories introduced above help us further define important concepts. For example, statements in a language (e.g. in the specific language of a scientific theory) can be classified according to the relation between their *meaning* and *truth-value*. The two classes are:

(a) the class of *analytic statements*

(b) the class of *empirical (synthetic) statements*

(a) Any statement in a given language (or in the language of a scientific theory) is analytic if a competent speaker of that language can determine its truth-value (i.e. whether the statement is true or false) on the basis of the meaning of that statement alone. For example, any competent user of English can determine that the statement "A cardiologist is a medical doctor who specializes in heart diseases" is true, based on the meaning of the statement. Users simply need to be able to understand the meaning of the statement. Similarly, any competent user of English will be able to say that the statement "The United Kingdom is not a monarchy" is false – simply on the basis of the meaning of the statement and without having to do resort to empirical investigation.

(b) An empirical (or synthetic) statement is one that is not analytic in the given language (theory). That means that a statement is only empirical if and only if its truth-value cannot be determined on the basis of meaning alone, but requires some form of empirical investigation or reliance on empirical evidence. Consider, for example, the statement "The modern Slovak Republic was founded on January 1, 1993" or "There is only one university in Bratislava". To find out whether these statements are true or false, we need not only to understand the meaning, but to require some knowledge of Slovak history and the capital city of Slovakia.

To be able to classify statements in this way requires us to have a certain level of competence in the given language (theory). Although we are unable to provide the criteria that would determine what constitutes linguistic competence, we can give a basic characterization:

> A competent user of language L understands the meanings of the linguistic expressions of that language (or at least the relevant subset) and, on that basis, can determine which objects, properties or relations the statement concerns.

This characterization is all that is required for us to understand the way in which statements are sorted as analytic or empirical (synthetic). If we find examples of statements that we have difficulty classifying as analytic or synthetic, we can eliminate some of the doubt by identifying the relevant natural language or scientific theory. For example, we might be unsure whether "Gothic is a Germanic language" is an analytic or synthetic statement. However, we could turn to a linguistic theory (see e.g. Čermák 2001, 65) that classifies languages into groups, and we would find "Germanic languages", and the subgroup of "extinct Germanic languages" to which Gothic belongs, and any competent user of this theory will be capable of establishing that the statement is analytic. That is, the statement is true based on the semantic relations established by the classification. (It is possible, as is the case here, that our success in classifying the statement depends on us being able to empirically investigate the origins of the languages and their resemblances.)

In philosophy, distinctions are commonly made not only between analytic and empirical (synthetic) statements, but also between *a priori* and *a posteriori* statements (or the knowledge expressed by the statements). An *a priori* statement is one we can ascertain the truth of, without having to resort to empirical matters or empirical experience. Conversely, an *a posteriori* statement is one we can only ascertain the truth or falsity of through empirical investigation.

For example, "$25 \times 1 = 25$" is an *a priori* statement, whereas "The sun rose at 06:27 today" is an *a posteriori* statement.

This classification is based on the relation between the *truth-value* of a statement and *how we ascertain* the truth-value. Most analytic statements are *a priori* and, conversely, most empirical statements are *a posteriori*. In the philosophy literature and related discussions there is much debate as to whether there are *analytic statements* that are *a posteriori* and whether there are *synthetic statements* that are *a priori*. However we shall not be covering these issues here.

The American philosopher Willard van O. Quine (1953) was critical of the distinction between analytic and synthetic statements, arguing it was vague and theoretically pointless. Although there is not the space here to examine Quine's arguments nor the criticism of them, we believe that the distinction we make can be defended against Quine's objections. (For arguments in favor of the *analytic/synthetic* (empirical) distinction, see e.g. Materna 2007.)

2.2.2 The method of definition

Some linguistic expressions have one of the following two semantic properties: either their meaning is imprecise – they are *semantically vague*, or one expression (syntactic component) has at least two meanings – it is *polysemic*. Both are undesirable if we need to express the information precisely, and should therefore be eliminated or reduced.

Semantically vague expressions are those that competent language users are unable to say what precisely they *denote* or which particular *extension* applies under the circumstances. For example, words such as "rich, "young", "old" or "bald" (and their associated grammatical forms) have no precise meaning in ordinary

language. We can of course find people who are (not) young or bald, but we can also find borderline cases where we cannot precisely determine whether they are young (bald, etc.). If scientists convey their knowledge and findings using vague terms, the value of that information is equally uncertain. The aim therefore is to eliminate vagueness in the language of science (or a theory).

Then there are *semantically ambiguous (polysemic)* expressions such as "V", which, in language L, has at least two non-equivalent meanings M_i and M_j (and the associated denotations D_i and D_j). For example, the word "crown" meaning <<a headdress worn by a monarch as a symbol of authority>> is graphically (syntactically) identical with the word "crown" meaning <<the top or highest part of something>>. We can syntactically and semantically distinguish between these polysemic meanings as follows: "crown$_1$", where the subscript $_1$ indicates the first meaning (headdress); and "crown$_2$" where the subscript $_2$ indicates the second meaning (the highest part). By so doing, we have *defined* these expressions.

But polysemy does not occur only in ordinary language. The language of a scientific discipline, such as physics, may contain expressions – say, "energy" – that have different meanings in different theories (for example, in classical mechanics and in quantum mechanics). Polysemy is not usually a problem in everyday communication, since the language user can determine the speaker's intended meaning based on the context. The same does not apply, however, to scientific language and communication. The semantic differences of polysemic theoretical terms are often less obvious, and on first reading, they are more difficult to discern than in ordinary contexts. The writer of a scientific text can, if desirable, eliminate polysemy by defining the term. This method enables the writer to specify the exact meaning of the *key terms* used in the text.

To eliminate polysemy and vagueness, we may use various *methods of definition*. As we shall see later, explication is another method that can be used. However, let us first look at the methods of definition.

Proper definitions

The result of using the method of definition is *a definition*, and it usually takes one of the following forms:

(DEq) *definiendum* $=_{df}$ *definiens*

(DEe) *definiendum* $\Leftrightarrow_{df}$ *definiens*

(SD) *definiendum* $\Leftarrow_{df}$ *definiens*

All proper definitions contain: (1) the expression whose meaning we wish to define (the *definiendum*) and (2) the expression(s) used to define the definiendum (the *definiens*); all definitions either *express* (DEq and DEe) or *constitute* (SD) the *relation between* the definiendum and the definiens. The relation in question can either be *the relation of definitional equality* or *the relation of definitional equivalence*. The difference between the two is as follows:

> **DEFINITIONAL EQUALITY**
> The definiendum is in *a relation of definitional equality* with the definiens if and only if the definiendum and the definiens express *one and the same meaning*.
>
> **DEFINITIONAL EQUIVALENCE**
> The definiendum is in *a relation of definitional equivalence* with the definiens if and only if the definiendum and the definiens have *equivalent meanings* – i.e. the *meaning* of the definiendum *identifies the same denotation* as the *meaning* of the definiens does.

For the sake of simplicity, we shall sometimes ignore this difference. The distinction between expressions that are semantically identical and semantically equivalent will be explained using the examples below.

Here we refer to definitions as "proper definitions" because they specify, describe or codify the *meaning of the definiendum* using *the meaning of the definiens*,

and the relationship between the definiendum and definiens is either that of semantic identity or of semantic equivalence. (Later in this section, we will also present some improper methods of definitions.)

In general, the various *methods of definition* can be described as the set of these two schematic instructions:

(I1) Select (identify) the expression you wish to define (i.e. the *definiendum*)!

(I2) Assign to the *definiendum* a *definiens* that meets the criteria of an adequate definition (CAD)!

To follow the second of the two instructions, we need to know the criteria of adequacy. Proper definitions usually conform to these criteria (CAD):

1. The definiendum and the definiens must denote the same object (property, relation, proposition); consequently, the definiendum and the definiens must have the same extension (in all cases).

2. The definiens should only contain expressions that are precise, unambiguous and clear in meaning.

3. The definiens should only contain positive expressions (no negative expressions).

4. The definiens should only contain expressions whose meaning can be assumed to be known to the target audience.

5. If the definiendum and the definiens are not identical in meaning but equivalent, then the definiens should contain terms that are semantically simpler, i.e. it should not be possible to define the definiens using the definiendum. (This rules out circular definitions.)

These criteria fulfill a heuristic role – they indicate the kind of definiens we should be looking for if we want to arrive at an adequate definition. They apply to *both the basic kinds* of a proper definition: *analytic* and *synthetic* definitions.[15]

Analytic definitions In analytic definitions, the meaning of the definiendum is already established and fixed in the given language (of a scientific theory). The definiens *expresses* or *describes* the same meaning as the definiendum. Alternatively, the definiens *expresses* a meaning that is equivalent to that of the definiendum. In other words, the definiens denotes the same object as the definiendum.

Let us look at a few examples:

(a) a medical doctor $=_{df}$ a person who is qualified to treat people who are ill

(b) cardiology $=_{df}$ the branch of medicine that deals with heart disease

(c) a house $=_{df}$ a building for human habitation

(d) a morpheme $=_{df}$ the smallest unit of a language that has a distinct meaning

We can see in example (a) that the definiendum and the definiens do not have the same meaning; however, they point to the same denotation – the same property (that of *being a medical doctor*). The definiens is "a person who is qualified to treat people who are ill" which in English is equivalent in meaning to the definiendum, "medical doctor". In example (b) the definiendum and definiens express the same meaning. We can say that "cardiology" is the short form for expressing the meaning of the terms used in the definiens. In example (c), "house" and "a building for human habitation" do not mean exactly the same thing, but they are equivalents, i.e. the definiendum and definiens denote *the same property* (the

[15] Here, we draw on Salmon (1995), Štěpán (2001), Gahér (2003, addendum VII) and Hurley (2006, Chapter 2). For a subtler classification of definitions and definitional relations in proper definitions, based on hyperintensional semantic theory (see e.g. Materna 2004), that distinguishes between semantically identical and semantically equivalent terms, see Bielik et al. (2010). For alternative approaches to definitions and the process of defining, see Glavaničová (2017) and Zouhar (2015a,b).

denotation is the same). And so we won't find a house that is not also a building for human habitation; and we won't find a building for human habitation that is not also a house. What about example (d) though? We have noted that any competent user of linguistic theory (in English) would say that the definiendum and the definiens mean exactly the same thing, and that the definiens is simply another way of expressing what the definiens expresses. Someone who is unfamiliar with the language of the particular general linguistic theory may not be able to determine whether the definiendum and definiens in example (d) express the same meaning. However, the defining must always be done in relation to the language in question – in a scientific context that language will be the language of the relevant scientific theory.

In all these examples, the definition conveys the established meaning of the expression on the left-hand side of the definitional equation. In an analytic definition, the definiens should always express the established meaning of the definiendum in the relevant language (the language of the scientific theory), and so *it makes sense to ask whether the analytic definition is true or not*. The examples we have given are all analytic definitions that are true (in English).

Synthetic definitions These are used to *propose*, *introduce* or *codify* the meaning of the definiendum in the relevant language. In science, we sometimes also introduce the meaning of a definiendum into the language of a *theory*. Here, the definiendum is either (i) a new linguistic expression with no established meaning; or (ii) an existing linguistic expression, and the definiens *determines*, *selects* or *specifies* the meaning of the definiendum in that context. Thus, synthetic definitions are terminological conventions used to introduce or further specify the meaning of the definiendum. In an analytic definition, the definiens contains expressions that have a fixed meaning in the language. Synthetic definitions can be expressed thus:

(SD) $\quad$ *definiendum* $\Leftarrow_{\mathrm{df}}$ *definiens*!

and read as: "Let the definiendum have the meaning expressed by the definiens!".

We can illustrate synthetic definitions using the words "astronaut" and "streetcar". Although both are now part of English vocabulary, this was not always the case. Before these expressions were introduced into English, they were simple strings of letters with no meaning. These were then codified (expressed by a synthetic definition) through a process of acceptance until the terms acquired the meanings they have today. We can represent this process thus:

(e) astronaut $\Leftarrow_{df}$ a person trained to travel in a spacecraft!

(f) streetcar $\Leftarrow_{df}$ a mass transport vehicle that runs on rails and is powered by electricity!

The meanings of these terms are now explained using an analytic definition (and so "$\Leftarrow_{df}$" becomes "$=_{df}$").

Since synthetic definitions are basically proposals or conventions on how to use the expression in the definiendum, there is no point in asking whether the definition is true or false.

We cannot overstate the importance of synthetic definitions in eliminating the ambiguity and vagueness of a term (or the use of a term). Methodologically their function is to help us convey more precisely the meaning of a term (the definiendum), as used in a scientific text for example. In other words, synthetic definitions enable us to select one of the several possible meanings of a term.

Another example of a synthetic definition is the use of *abbreviations*. An abbreviation could be defined thus: "Let 'DF' mean the 'definiendum'!" "DF" would then represent the meaning of the term "definiendum" in the language of the relevant theory (or in a text).

The use of analytic and synthetic definitions helps eliminate ambiguity and vagueness in language (in science). These types of definitions are perhaps generally the ones used most, but we shall also look at some of the other kinds, which may differ from (DEq), (DEe) and (SD).

Further types of (improper) definitions

Operational definitions In operational definitions, the aim is not so much to define the meaning of the definiendum as to specify the conditions under which an object (entity, set of objects) has a certain property that is not directly observable, or whether there is a certain relationship between objects (entities) that is not directly observable. The meaning of the definiendum of some theoretical terms does not necessarily identify a property that can be directly observed as its denotation. However, we can use an operational definition so that the definiens expresses a certain operation (a test) that specifies the conditions under which the property (or relation) can be attributed to the entity (entities). In general, operational definitions take the form:

(OD) $(\forall x)\,[T\,(x) \Leftrightarrow_{\text{df}} \text{If } O\,(x)\,, \text{ then } E\,(x)]$

where $T\,(x)$ is the definiendum containing at least one theoretical term T whose conditions of applicability to object x are expressed in the definiens. The statement in the definiens expresses the fact that if x is subjected to observable operation O, then x will display certain observable features E.[16] Consider the following examples:

(g) Liquid A is an *acid* $\Leftrightarrow_{\text{df}}$ If litmus paper is dipped in the liquid, it turns pink.

(h) X is *of above-average intelligence* $\Leftrightarrow_{\text{df}}$ If X completes a standardized IQ test, then X's score will be in the interval [111, 120].

Example g) is an operational definition of the term "acid", while example h) is an operational definition of the term "of above-average intelligence". In both cases,

[16] The fact that the definiens of an operational definition has the form of a (material) implication (an "If α, then β"statement) has an unfortunate consequence. For that statement to be true, we simply need the antecedent (i.e. α) to be false. In our case, simply by not subjecting x to operation O the complex statement "If $O\,(x)$, then $E\,(x)$" will be true, and so will be $T\,(x)$. This problem can be solved pragmatically by limiting operational definitions to objects that we test under the observable conditions described by the antecedent of the definiens.

the terms express (theoretical) properties that are not directly observable. Simply by looking at the liquid, we cannot determine whether it is an acid or an alkali. This is also true of the attribute of being *of above-average intelligence*. (In both of these definitions, universal quantification is assumed. In other words, the definitions hold for any liquid and any person X, respectively.)

In these definitions the definiens specifies operations (in the form of a complex conditional statement) that enable us – if the observable or measurable effect of the operations occurs – to ascribe the property denoted by the definiendum to the object (i.e. to state that the liquid is an acid, or that X is of above-average intelligence).

Operational definitions are an important instrument for formulating the conditions under which indirectly observable properties (relations, etc.), denoted by the terms of the theory, can be ascribed to objects (phenomena, events). On the other hand, these definitions cannot be considered the *definitions of the meaning* of these terms. They merely indicate the conditions under which one can indirectly verify whether the objects have the given theoretical properties or not. If we were to equate the meaning of the operationally defined terms with the operations described in the definiens of the definitions, any change in the operation would also lead to a change in the meaning of the term. It is evident, however, that the meaning of the term "acid" in the language of chemical theory will not change simply because we replace the litmus operation with another procedure.

To summarize, if we use terms such as "content", "aggressive", "rare", "conservative" in our research, it is advisable to provide operational definitions for these (and similar) terms in addition to the standard definitions of their meaning.

Inductive (recursive) definitions Another type of definition is the inductive definition. This type is often used in mathematics, computer science and law, but also to define the rules of some games. We can use these definitions to specify *the objects* we wish to denote using a particular term.

All inductive definitions define the object denoted by the definiendum in three steps: (1) the definiens must contain an inductive base – a statement or formula that specifies the basic objects denoted by the definiendum; (2) the definiens must

also contain an inductive step that specifies the means of getting to all the other objects denoted by the definiendum; this step can be formulated using several rules (or steps); and finally (3) an inductive closure that states that no other objects are denoted by the defined term than those determined by steps 1) and 2).

We can illustrate this schematically using the definition of the term "descendant". It comprises the following steps:

1. Every child of its biological parents is their descendant.
2. If y is the descendant of x, and z is the descendant of y, then z is the descendant of x.
3. No objects apart from those satisfying conditions 1 or 2 are descendants.

The definiendum in this inductive definition is the term "descendant". It denotes the relation between two or more individuals, the second (third, fourth, etc.) of whom is the descendant of the first. This relation is then specified in the definiens via three steps. The first two steps specify the tuple (or n-tuple) of individuals, while the third states that *only* the tuples (n-tuples) of individuals that satisfy the conditions listed in the first two steps are or can be in this relation.

Inductive definitions can thus help us define a property, relation, set, etc., (denoted by the expression in the definiendum) by means of rules that specify exactly which objects have (or can have) the property in question, or are (can be) elements of the given set, etc.

Ostensive definitions Although ostensive definitions are not truly definitions, they are used to teach the (empirical) terms used in a given language.

An ostensive definition is one where we say the expression and gesture towards one or more objects found within the extension of that expression. Consequently, we can only use ostensive "definitions" when the extension is an object that can be seen. For example, we can ostensively "define" the meaning of the term "book" by pointing to a particular book (an element within the extension of the expression) that the target audience can see or observe. On the other hand, we cannot

ostensively define the meaning of terms such as "state", "idea" or "the number ten", since the entities currently found in the extensions of those terms cannot be seen, are not observable.

Moreover, even when the defined expression has an observable extension, its meaning cannot be properly defined using an ostensive definition. The meaning and denotation of terms are not reducible to an extensional object. Therefore, the methodological uses of this type of definition are extremely limited.

Verbal extensional definitions As in the previous type, the definiens of a verbal extensional definitions should contain one or more of the elements comprising the extension of the expression being defined. However, these definitions differ from ostensive ones in the way that these elements are selected. Their definiens usually contains the (proper) names of some or all of the elements belonging to the extension of the term being defined. For example, the "meaning" or the (incomplete) extension of the expression "president of the United States" can be (partially) defined using the following verbal extensional definition:

(VED) the president of the U. S. $=_{\text{ext}}$ {Barack Obama; Donald Trump}

As with ostensive definitions, their methodological potential for eliminating ambiguity and vagueness is rather limited.

2.2.3 The method of explication

As was the case with the method of definition, explication can be approached in two ways. It can be viewed both as a method, i.e. the process of explication or a methodological activity, and as the product or result of that activity.

In an intensional definition, the definiendum and the definiens should express the relation of *semantic identity* or the relation of *semantic equivalence*. In contrast, in explication, the *meaning* (concept) of the *expression* (or expressions) *being explicated* is *replaced* by another, *non-equivalent* (and therefore non-identical) *meaning* (concept) of the *expression* (or expressions) explicating it.

The basic form of an explication (product of the method of explication) can be schematically expressed as follows:

(Exp) explicandum $\Rightarrow_e$ explicatum!

The explicandum is therefore the starting point in the method of explication. It is an expression whose meaning we intuitively understand and use in certain communicative situations, but which cannot be used more systematically to solve theoretical problems. The goal of explication is to find an explicatum that is as precise and unambiguous as possible, and can then be used to deal more systematically with theoretical problems. Generally, explication can be delineated using the following instructions:

(I1) Select (identify) the *explicandum*, an expression with an imprecise or theoretically unfruitful meaning!

(I2) Replace the *explicandum* with an *explicatum*, an expression that meets the criteria of adequate explication (CAE)!

Before turning to the general criteria of adequate explication (CAE), we should note the fundamental distinction between an explication and a definition. When introducing the concept of analytic definitions, we noted that analytic definitions express a preexisting (already constituted) relation between the meaning of the definiendum and the meaning of the definiens. Analytic definitions therefore respect (implicit or explicit) linguistic conventions. By contrast, in our discussion of synthetic definitions we saw that synthetic definitions are a means of introducing these conventions into language (or a theory). Therefore both these kinds of definitions – analytic and synthetic – are in some way related to linguistic conventions. The former express them, while the latter can be used to introduce them.[17] However, the relations between explications and conventions are much *looser*. When formulating explications, our goal is not to express a meaning that

[17] We do not go into the conditions that must be met for synthetic definitions to be successfully integrated into language.

has already been established. At most the fixed meaning of the term being explicated (i.e. the explicandum) serves as the starting point of explication, because we want to replace that fixed meaning with another meaning: the meaning of the term (or terms) explicating it. Moreover, when explicating a term, our aim is not necessarily to create and codify a new linguistic convention, as is typically the case with synthetic definitions. An explicatum may represent a larger conceptual whole (such as a theory) that goes beyond the process of definition.

Yet we should also note that a synthetic definition might become part of the explication, if we wish to associate the new meaning of the explicatum with the expression originally occupying the position of explicandum. We will give examples of this distinction, but before doing so, let us turn to the criteria of adequate explication (CAE).

Despite the fact that there are no strict rules governing explication and there is considerable room for creativity, there are some general criteria that determine whether an explication (the product of the method) should be accepted as adequate.[18] The following criteria (see Carnap 1962, 5–8; Kuipers 2007, vii–ix) must be followed when replacing the explicandum with the explicatum:

1. the *meaning* of the explicatum should be *precise*; it must be given in an *exact* form;

2. the meaning of the explicatum should be theoretically *fruitful*; i.e. it should be useful for formulating (new) theoretical statements;

3. the explicatum should be as *simple* as possible; if several explicata (of the same explicandum) are considered, the simplest one is preferred;

[18] Rudolf Carnap states, in his discussion of whether an explication is adequate or right, that: "[…] if a solution for a problem of explication is proposed, we cannot decide in an exact way whether it is right or wrong. Strictly speaking, the question whether the solution is right or wrong makes no good sense because there is no clear-cut answer. The question should rather be whether the proposed solution is satisfactory, whether it is more satisfactory than another one, and the like." (Carnap 1962, 4).

4. the explicatum should be *similar*, in some respects, to the explicandum; including in being applicable to all the unproblematic cases to which the original explicandum can be applied. Moreover, the meaning of the explicatum should preserve all the conditions of adequacy that apply to the explicandum.

Condition 3 is the least specific. What exactly do we mean by "simplicity" here? Two clarifications can be made. Firstly, each explicatum must be assessed on its own merits (without comparing it to the other explicata). Simplicity can be understood to mean that the explicatum should be economical in postulating (assuming) the existence of certain entities; and that its use in formulating and solving (conceptual) scientific problems is straightforward. Secondly, when comparing the explicatum to other explicata, we can determine its simplicity based on the entity and type of entity it postulates and how straightforward it is to use. (For a view on understanding the criteria of simplicity, see Bielik 2018).

The remaining criteria should be clear: we covered precision of meaning when discussing definitions, so it should be clear why precision is required in the current context. The theoretical fruitfulness of an explication hinges on the explicatum allowing a more precise and more comprehensible formulation of a theoretical problem or its solution. Lastly, the reasons as to why the explicandum should be similar to the explicatum are explained in condition number 4.

We have already suggested that we can compare explications in terms of how well they satisfy the CAE. There is little sense therefore in thinking of an explication as being "right" or "wrong" in an absolute sense. Instead we should assess whether the various explications are adequate in terms of the shared theoretical goal. (Explication E_1 may be more satisfactory than explication E_2 with respect to theoretical goal G.)

To illustrate the method of explication, we shall give several examples from different areas:

(a) truth $\Rightarrow_e$ correspondence between the statement and the fact (state of affairs) described by the statement

(b) being close to $x \Rightarrow_e$ being within 15 meters of x

(c) evidence e confirms hypothesis $h \Rightarrow_e$ the probability of h, assuming that e is true, is higher than the prior probability of h before considering e

(d) many human lives $\Rightarrow_e$ more than 50 million human lives

(a)–(d) represent explications from different areas. While a) could be a philosophical explication of the pre-theoretical term "truth", (b) could be an attempt at precision or at finding an alternative to "being close to x" – for instance during the questioning of the witnesses of a crime. Of course, depending on the theoretical goals, the way in which we specify distance from x may change. For example, in a discussion between astronomers on the distance of two cosmic objects (such as stars or galaxies), the term "being close to x" could be replaced with "being within 136 light years of x" (or with a different specification) etc. Hence, how the explicatum is specified will depend on the particular area of inquiry and the problem the explicatum relates to. Example (c) is one possible explication of the expression "confirming hypothesis h using evidence e", while (d) is a possible explicatory substitution for "many human lives", e.g. in the context of summarizing the consequences of World War II.

Carnap (1962) considered paradigmatic cases of explication to be cases where *qualitative concepts* are replaced with *comparative concepts* and the latter with *quantitative concepts*, or *qualitative ones* directly with *quantitative concepts*. Carnap's "qualitative concept" can be equated with the meaning of a term denoting a property (of an individual), such as "being tall", while "comparative concept" can be understood as denoting a relation comparing the extent to which a property is present, such as "being taller than". Finally, a quantitative concept is the quantitative expression of a certain magnitude, such as "being 178 cm tall".

2.2.4 Methods of analysis

The term "analysis", as used in everyday and specialist language, has several meanings. However, the procedures denoted by the term all share something in common (see Beaney 2015; Kosterec 2016).

Any *analysis* can be identified as a series of general instructions specifying how to get from one theoretical starting point (a whole) to the final state of analysis – the identification of the structure of the whole.

(I1) Identify (delineate conceptually) the object or starting point of the analysis – a whole of some type!

(I2) Using concepts of an appropriate theory T, identify some or all of the elements comprising the whole!

(I3) Check whether these are the basic elements!

(I4) If the elements are not basic elements and should be basic elements, repeat step (I2)! If the elements are basic elements, proceed to the next instruction.

(I5) Using concepts or theoretical instruments of theory T, determine the properties of the elements and the relations between them that are relevant to the present inquiry!

(I6) Using the previous steps, formulate the structure of the whole!

When analyzing a theoretical or empirical whole we always have to rely on certain pre-theoretical or theoretical concepts and categories in order to intellectually grasp the whole and its integral parts. Any analysis will therefore be based on the conceptual system containing the concepts within which the object is grasped and analyzed.

We can illustrate the method of analysis using a simple example from modern theories of syntax:

Suppose we have the sentence "Jane greeted John". Our background theory of analysis is modern syntax theory. Our goal is to express the structure of the sentence. We can explicitly set out the steps in this kind of sentence analysis thus:

(a1) Identify the object of analysis!
(Result: "Jane greeted John".)

(a2) Using concepts from modern syntax, identify some or all of the elements comprising the whole!
(Result: "Jane", "greeted", "John".)

(a3) Verify whether these are basic elements!
(Result: Yes, these are the basic elements. The sentence cannot be broken down to any simpler elements.)

(a4) If these are not basic elements, and if necessary, repeat step (I2)! If these are basic elements, proceed to the next instruction.
(Result: We move onto the next instruction.)

(a5) Assign the respective category of constituents to each element of the sentence!
(Result: "Jane" ⇒ subject; "greeted" ⇒ predicate; "John" ⇒ object)

(a6) Express the syntactical parts!
(Result: *Subject – predicate – object*)

This example has been simplified, but nonetheless is an illustration of how the method of analysis can be used in relation to the background syntactical theory.

There are a variety of different analytical approaches because the theories that can be selected as the background theory of analysis have different conceptual systems. We can therefore talk of *linguistic analysis*, *conceptual analysis*, *textual analysis*, *frequency analysis*, *causal analysis*, *statistical analysis*, *chemical analysis* and many others.

2.2.5 Classification

Classification is another conceptual method. One of the basic ways of systematizing the field of interest in which the objects (entities) whose properties and relations are of interest to us belong is to classify the objects of that universe of discourse into classes or orders of classes. The basic principle for classifying objects (of a certain kind) is to select the appropriate *classificatory property* (*properties*) used to sort the objects into classes. Successful classification depends on

whether the *classificatory property* (*properties*) are clearly specified and whether they are (easily) recognizable. The process of selecting the *appropriate* classificatory properties is dependent on the *theoretical role* the classification is supposed to play (in relation to the theoretical goal).

We can illustrate this by looking at an example of an ambiguously specified property. Suppose we are in an apple orchard and our task is to only select species of trees that have red apples from among the many varieties. When distinguishing between the class of apple trees with red apples and all the other trees, we may encounter cases where no such distinction can be made. Imagine, for example, that some of the apples were partly red and partly yellow or green. We would not know how to classify them. Some trees would remain unclassified. The problem is that the expression "red apple" is not sufficiently precise for us to unambiguously classify the apples on that basis. Since the meaning of the term "red" is not specified exactly, it makes it difficult to determine which property (of redness) the expression identifies and which objects (apples) in fact have that property.

If we need classifications that are based on precisely defined properties (or relations), then we have to select terms with meanings that allow us to clearly identify the classificatory properties (or relations). Moreover, if the property is an empirical one, we must also be able to identify whether the objects exhibit it, or at least be able to indirectly test whether it is present.

Depending on the nature of the classification process, we can distinguish the following types of classifications:[19]

1. *analytic classification*
2. *synthetic classification*
3. *order-based classification*

Analytic classification This is a type of classification in which the initial set of objects (of a certain type) is divided into subclasses based on a classificatory

[19] Our explanation of the method of classification is based on Bunge (2005a, 82–89) and Filkorn (1960).

property (or properties). This method can also be represented as a set of two instructions:

(I1) Identify the initial set of objects – universe of classification U – that will be the object of classification!

(I2) In relation to theoretical goal C, select the classificatory properties $F_1, \ldots, F_n$ of the initial set of objects U that satisfy the following two requirements:

(i) $U = F_1 \cup \ldots \cup F_n$, where $n \geq 2$;

(ii) $F_i \cap F_j = \emptyset$ for any two properties $F_i, F_j \in \{F_1, \ldots, F_n\}$.

The starting point of analytic classification is always a collection (set) of objects, such as a set of words, plants, real numbers, elementary particles, an author's literary works etc. The aim is to divide initial set U into mutually exclusive classes represented by the properties $F_1, \ldots, F_n$.

Condition (i) of instruction (I2) states that based on the properties $F_1, \ldots, F_n$ belonging to the objects in initial set U, all the objects in the set will be classified into *at least one* of classes $F_1, \ldots, F_n$. In other words, no object from the initial set should remain unclassified. Condition (ii) states that each classified object belongs to *a single class at most*. When the two conditions are combined, each object in U will belong to *one class only* in the resulting classification. If both conditions are met, we can say that the analytic classification is adequate and complete.

The simplest example of an analytic classification is dividing the initial set into two classes. All we need to do is select any property F exhibited by at least one object in the initial set. In most cases, however, more than one object will exhibit the property. Property F enables us to distinguish a (non-empty) class of objects with that property. The remaining objects in the set will belong to the second class – the class of objects *not exhibiting* property F, i.e. exhibiting property non-F. A set of elements that has been classified into two mutually exclusive classes is also known as a "dichotomy". Richer analytic classifications are based on the elements in a set containing more than two properties.

We can illustrate analytic classifications using two simple examples:

Example 1

Suppose that our initial set is a set of integers {…, −1050, −2, −1, 0, …, 5, 6, …, 478 985, …}. Examples of the many properties exhibited by (some) elements in this set are the properties of *being an even integer*, *being an integer divisible by three*, *being a number greater than 1000* etc. For our classification, we will pick two mutually exclusive properties: *being a positive integer* and *being a negative integer*. The properties enable us to divide the universe of discourse (the set of integers) into two classes: a class of *positive integers* and a class of *negative integers*. But is this classification adequate and complete?

Clearly, no integer can be both positive and negative. Condition (ii) is thus met. What about condition (i)? Have all the elements in the initial set (the set of all the integers) been classified? No. The number zero remains unclassified, since it is neither positive nor negative (the set of positive integers including zero is usually called "the set of non-negative integers"). Therefore, this classification is not complete. To complete it we have to include another classificatory property, such as that of *being an integer that is neither positive nor negative*. Having adjusted the classification, the initial set has now been exhaustively (completely) divided into three mutually exclusive classes.

Example 2

This time our initial set is the set of Slovak consonants. We will select three classificatory properties: *being a soft consonant*, *being a hard consonant* and *being an unmarked consonant*. In Slovak, all consonants belong to one of three classes:

A: class of soft consonants = {c, dz, j, ď, ť, ň, ľ, č, dž, š, ž}

B: class of hard consonants = {g, h, ch, k, d, l, n, t}

C: class of unmarked consonants = {b, f, m, p, r, s, v, z, x}

We should again ask whether this classification is both adequate and complete. We can see that condition (i) has been met. Has condition (ii) been satisfied as

well? It seems to have been. We can therefore state that the classification is both adequate and complete. However, the classification could suffer from another defect that cannot be eliminated by meeting conditions (i) and (ii). In Slovak, consonants are classified as soft, hard and unmarked based on grammatical rules, but words that have been adopted from other languages do not necessarily follow these rules. In Slovak words whether a consonant is followed by "i" or "y" is determined by the class of the consonant ("i" follows a hard consonant) but words of foreign origin don't necessarily follow this classification (in "cylinder" the "y" follows a soft rather than a hard consonant).[20] In addition to the conditions listed above, the classification can be considered adequate if the elements exhibit the classificatory properties without exception and if they exhibit them under the circumstances in which the classification was assessed.

When checking to see if an analytic classification is both adequate and complete, we thus need to consider whether conditions (i) and (ii) have been met and whether the objects being classified in fact exhibit the classificatory properties ascribed to them in the classification.

Synthetic classification Synthetic classification is another means of sorting objects. Similarly to the previous type of classification, it involves an initial set of objects or, more precisely, a set of classes of objects (expressions, numbers, empirical data, etc.). The objects or the classes of objects are classified into larger classes – superclasses.

Synthetic classification can be schematically represented as follows:

(I1) Identify the initial set of objects belonging to the different classes (in other words, identify the different classes of the objects)!

(I2) Take the properties of the objects in the initial set and select one property (or more properties) that can be used to categorize the elements of the different classes into one common superclass or multiple superclasses!

[20] On the classification of consonants in Slovak, see Mistrík (2002, 195–197).

Synthetic classification therefore aims at finding an unambiguous common property that can be used to group elements in the original different classes into a superclass. This procedure is the inverse of analytic classification.

Let us briefly look at an example of this type of classification. Suppose we want to find a common property that would enable us to group together states such as Bulgaria, Croatia, Czech Republic, Denmark, France, Germany, Lithuania, Slovak Republic and Turkey into a common class. Some of the states belong to the class of Visegrad countries {Czech Republic, Slovak Republic}, some are Balkan countries {Bulgaria, Croatia}. However, (as of January 2019) all the countries have the property of being members of the North Atlantic Treaty Organization. These countries (and, of course, others) are therefore elements of the class of NATO member states. It holds that NATO = {..., Bulgaria, Croatia, Czech Republic, Denmark, France, Germany, Lithuania, Slovak Republic, Turkey, ...}.

Order-based classification The third type of classification we shall introduce is known as *order-based (relation-based) classification*. This classification is used to order objects or classes of objects based on a relation or sequence. The starting principle is finding the appropriate *relation* or *order* that enables us to place the objects (or classes of objects) in a characteristic sequence. The classificatory relation could be the relation of *being greater than or equal to* ("$\geq$"), the relation of *being a subset* ("$\subset$") etc.

Order-based (relation-based) classification can be specified using the following instructions:

(I1) Identify the classes of objects $F_1, \ldots, F_n$ in universe of discourse U!

(I2) Order the classes of objects $F_1, \ldots, F_n$ into an n-tuple $\langle A_1, \ldots, A_m \rangle$ or a relation $R(A_1, \ldots, A_m)$ such that each class A_j, where $1 \leq j \leq m \leq n$, is identical to *exactly* one class F_i, where $1 \leq i \leq n$.

Instruction (I2) states that the original classes in the universe of discourse should be placed in a sequence or a characteristic relation, where each class in the sequence ($\langle A_1, \ldots, A_m \rangle$ or relation $R(A_1, \ldots, A_m)$) is identical to exactly one of

the original classes $F_1, \ldots, F_n$. This assumes that no class A_j in the set of classes $A_1, \ldots, A_m$ is identical to any of the other classes in the set.

An example will help us understand this general characterization. Suppose the universe of discourse comprises all the chemical elements (or classes of chemical elements) from hydrogen (H) to Ununoctium (Uuo). Our classificatory relation is *being an element with a lower proton number than*, denoted by the symbol "$<_p$". The goal is to place the chemical elements into a characteristic sequence based on this relation. The resulting sequence (or part thereof) can be expressed thus:

$$\mathrm{H} <_p \mathrm{He} <_p \mathrm{Li} <_p \ldots <_p \mathrm{Uus} <_p \mathrm{Uuo}$$

Another example of an order-based classification is the well-known biological taxonomy or Linnaean System named after the Swedish botanist Carl Linnaeus (1707–1778). It is a hierarchy of living organisms (animals and plants) from the most general category – taxa – to the most specific, based on certain biological features. The taxonomy takes the following form:[21]

$$\text{Domain} \supset \text{Kingdom} \supset \text{Phylum} \supset \text{Class}^{22} \supset \text{Order} \supset \text{Family} \supset \text{Genus} \supset \text{Species}$$

The symbol "$\supset$" represents the relation of *being a proper superset*, where the first set (class) is a superset of the second set (class). However, the taxa also contain other elements that can be classified in other ways within the taxon – typically using analytic classifications.

Classifications are the basic methodological tool for systematizing knowledge about objects in the universe of discourse being investigated. By classifying objects, we can divide the original universe into smaller wholes and then examine them, or conversely, we can express the most general features of the original whole, or represent the hierarchy of the features revealed by the objects of our inquiry.

[21] See e.g. the entry on Carolus Linnaeus by Staffan Müller-Wille in the online version of the Encyclopaedia Brittanica: `britannica.com/biography/Carolus-Linnaeus/`.

[22] The term "class" has a specific meaning in Linnaeus' classification, which differs from the "set-theoretical" meaning used in this book.

We need not emphasize that the prior and adequate classification of objects is crucial to applying other conceptual and empirical methods. It also lays the foundations for adequate scientific explanations, interpretations and reliable predictions.

2.2.6 Methods of abstraction and idealization

Methods of abstraction and idealization are crucial tools for constructing and formulating scientific theories, (theoretical) models, scientific laws, computations, explanations and so on.

Jones (2005, 175) describes the basic difference between abstraction and idealization (as both a method and its product) as follows: abstraction involves omitting a truth about an object, while idealization amounts to the deliberate misrepresentation of an object. Due to space constraints, we will limit ourselves to describing the basic structure and function of these methods. We refer the reader to Jones' paper (2005), which forms the basis of this subsection; see also McMullin (1985), Weisberg (2007), Halas (2015a,b,c, 2016a,b) and Hanzel (2008; 2015).

Abstraction

In applying the method of abstraction, we begin from the fact that object o, which is the object of our investigation, displays a number of properties $\phi_1, \ldots, \phi_n$. (In fact, all objects have an infinite number of properties, but only a few of these are accessible to us. These are the ones we have identified as $\phi_1, \ldots, \phi_n$.) Thus, we can generally truthfully state that object o is (or has the property of being) ϕ_1, or ϕ_2, …, or ϕ_n. However, in some situations – in our case in the context of scientific research and in relation to some theoretical goals G – we may be interested in only *some* of the properties $\phi_1, \ldots, \phi_n$, while the remainder are of no theoretical significance. We therefore disregard or *abstract* from the remaining properties. For example, if property ϕ_i, where $1 < i < n$, is not relevant to our theoretical goals G, we *abstract* from it. Suppose we are investigating the behavior of a group of adults in a situation where they are subjected to short-term stress (for example, performing a complex mathematical task). We want to find out which factors

affect their ability to perform the task effectively. We thus abstract from a variety of the properties the adults exhibit: for example, their clothes, eye color, month of birth etc. These properties (and many others) are irrelevant to goals G and we therefore pay no attention to them.

The *method of abstraction* can generally be given thus:

(I1) Identify object o that is known to have the properties $\phi_1, \ldots, \phi_n$!

(I2) Select each property ϕ_i from the set of properties $\{\phi_1, \ldots, \phi_n\}$ known to be unimportant to object o given the theoretical goals G!

(I3) Identify the set of remaining properties $\{\phi_1, \ldots, \phi_n\}$ that are relevant, given G, with the set $\{\phi_j, \ldots, \phi_m\}$, where $1 \leq j \leq m < n$.

(I4) Represent the object o using only properties $\phi_j, \ldots, \phi_m$.

The method of abstraction leads us away from the realization that object o exhibits certain (different) properties and towards a situation in which we focus only on the properties that are relevant to the goal. It is indispensable to the effective application of other methods – those that use a conceptual/theoretical apparatus, and those in which we use our senses and various measuring, observational or experimental instruments.

Idealization

Similarly to the method of abstraction, the method of idealization also assumes that object o, which is the object of our investigation, exhibits a variety of properties $\phi_1, \ldots, \phi_n$. Unlike in abstraction, an additional step is required in idealization in which it is claimed or assumed, that for theoretical purposes G, o has at least one other property ψ that differs from the properties $\phi_1, \ldots, \phi_n$. But in reality object o does not have, and cannot have, property ψ. Moreover, from the statement that object o has property ψ it follows that object o does not have at least one of the properties $\phi_1, \ldots, \phi_n$, whereas in actual fact it has.

To illustrate this abstract characterization of the method of idealization, let us return to our example of research into the behavioral responses of individuals subjected to short-term stress. Suppose we make the following assumption: "Persons $o_1, \ldots, o_n$ who took part in the test had *approximately the same* sleep pattern the preceding night." This assumption is almost certainly false. It is highly probable that the *length* and *quality* of sleep was different for each participant. We know this assumption is (probably) false, but by accepting it, we can focus on other aspects of the problem. Accepting idealizing assumptions of this kind is an important part of the process of solving some research problems, and is a fundamental part of the process of constructing scientific models.

The *method of idealization* can be represented using the following sequence of instructions:

(I1) Identify object o that is known to have the properties $\phi_1, \ldots, \phi_n$!

(I2) Given theoretical goals G, select at least one property ψ where (i) object o does not in fact have property ψ; but (ii) if object o had property ψ, this would enable us to solve problem P; and (iii) it follows from the assumption that o has property ψ that object o does not have at least one property ϕ_i from the set of properties $\{\phi_1, \ldots, \phi_n\}$!

(I3) We accept the assumption that o has property ψ (as well as properties ϕ_1, $\ldots, \phi_n$ but not property ϕ_i)!

The method of idealization typically fulfills a constitutive function in the construction of theoretical models, where the model system is a simplified and deliberately "distorted" representation of another – usually empirical – system.

Abstraction and idealization play an important role in simplifying the investigation of certain phenomena, reducing the computational complexity involved, isolating causal factors, identifying the general properties of certain systems etc. Typically, they are part of a host of conceptual and empirical methods and contribute to their effectiveness.

2.2.7 Reasoning as a method

Our discussion of reasoning, the types and products (i.e. inferences or arguments) of reasoning, and their basic logical and methodological properties is based chiefly on Cmorej (2000; 2001b), Salmon (1995), Skyrms (2000) and Viceník (2001a,b).

Reasoning is one of the intellectual activities we use throughout our lives (excluding the first months). In reasoning, the goal is always to express a relation between at least two thoughts that are expressed (or may be expressed) in statements. In what follows, we will treat reasoning as the derivation of statement(s) (propositional expressions) from (the set of) other statement(s). A statement is a sentence that has (or may have) one of two truth-values, True or False. A broader definition of reasoning would include not just statements, but statement forms as well. A statement can be transformed into a statement form by replacing at least one expression in the statement with a suitable type of variable. Consider the statements

(S1) Tom is the brother of Jane.

(S2) It is not true that Rupert is sick.

One of the possible statement forms (SF1) can be obtained by replacing the expression "Tom" with a so-called individual variable. We could obtain another form by additionally replacing the expression "Jane" with (another) variable:

(SF1) x is the brother of Jane.

(SF1*) x is the brother of y.

From (S1), we may also obtain the statement form

(SF1′) xRy or, equivalently, $R\,(x, y)$.

Here, R is a variable replacing the expression "is the brother of". Hence, it is a (binary) predicate variable. We read (SF1′) as "x is in relation R to y".

On the other hand, we can also transform statements into statement forms by replacing the entire statement with a variable. For example, if in (S2) we replace

the statement "Rupert is sick" with the statement (propositional) variable "*p*", we obtain:

(SF2) It is not true that *p*.

Statement forms are neither true nor false. By making appropriate substitutions (for example, substituting proper names, the names of properties and relations, statements etc.), we can transform statement forms back into statements.

If we call both statements and statement forms *propositional expressions*, we can also define reasoning more precisely. *Reasoning* is a thought-process or act in which we derive a propositional expression from other propositional expressions (i.e. a non-empty set of propositional expressions) (see Cmorej 2000, 329). This derivation always has a certain logical structure that can be represented as an inference. By inference, we do *not* primarily mean the *thought-process* whereby we derive the *conclusion* – a statement (statement form) – from certain statements (statement forms) also known as *premises*. What we are interested in is the result of this activity.

An *inference* (also *argument*) is a tuple $\langle P, C \rangle$ where P is a (non-empty) set of statements called the *premises*, and C is a statement called the *conclusion* of the inference. P may contain one or more statements which may also be the conclusions of other inferences.

An *inference form* can be obtained from an *inference* by replacing each statement in the inference with a statement form. For example, if we have the inference:

(I1) It is not true that Rupert is sick.

Rupert is healthy.

we can construct one of the possible inference forms by replacing the expression "is sick" with the predicate variable "*F*", the expression "is healthy" with the variable "*G*", and the name "Rupert" with the individual constant *a*:

(IF1) It is not true that $F(a)$.
$G(a)$

or:

(IF1′) $\neg F(a)$.
$G(a)$

Inference forms (or inference schemes) can be expressed using inference rules that in general take the following form:

(IR) If the premises $P_1, \ldots, P_n$ have the form $A_1, \ldots, A_n$, derive conclusion C whose statement form is B.

Inference rules (which are themselves inference forms) can express a *logical form* or *structure* that is common to several inferences that differ in content and grammar. The logical form of inferences is an inference form whose statement forms (appearing in the premises and conclusion) contain *logical constants* and *variables* only, or *variables* only. Logical constants are expressions denoting *logical operators*, such as *sentential connectives* (negation, conjunction, disjunction, implication, equivalence), *quantifiers* (universal, existential etc.) or *logical predicates* (=).

Let us illustrate the concept of logical form or structure of inference using these examples:

(I2) Any metal heated to temperature T expands.
This is a metal which we heat to temperature T.
This metal expands.

(I3) All Slavic languages are Indo-European
(i.e. they belong to the family of Indo-European languages).
Slovak is a Slavic language.
Slovak is an Indo-European language.

Clearly, inference (I2) differs grammatically and content-wise from inference (I3). These inferences concern two completely different areas of knowledge: the first is about the physical properties of metals, the second is about the classification system of languages. Despite these differences, inferences (I2) and (I3) have the same logical form (if we simplify some of their inferential features):

(LF1) $$\frac{(\forall x)\,(F\,(x) \rightarrow G\,(x))\qquad F\,(a)}{G\,(a)}$$

We read the inference form (LF1) as

> "For all x, if x is F (e.g. x is a metal heated to temperature T, or: x is a Slavic language), then x is G (e.g. x expands or x is an Indo-European language). Further, a is F (e.g. This is a metal heated to temperature T; or: Slovak is a Slavic language); and, finally, the conclusion states: a is G (e.g. This metal expands; or: Slovak is an Indo-European language)."

What is the point, though, of revealing this abstract structure? After all, in scientific research, we are interested in the content, i.e. in what the statements are about, and the information that can be derived using inferences. It is true that the logical or inference form provides us with no information about the intellectual content of statements or inferences. However, it enables us to explicitly express what several inferences – different in terms of grammar and content – have in common. By recognizing the logical form of inferences, we are (in principle) capable of distinguishing those inferences (or their forms) that are logically (deductively) valid from those inferences (or their forms) that are not, and hence are not deductive. Moreover, from among those inferences that are not deductive, we can identify, based on their logical form, some that can be used to formulate hypotheses and scientific theories, or in testing and evaluation in light of the empirical data.

In the previous paragraph, we used the terms "logically valid" and "deductive" when speaking about inferences, but did not define these terms. We shall do so

shortly. However, let us first turn to the role of reasoning or inference as a method (or, more precisely, a collection of methods) in scientific research.

Argumentation is a process whereby we put forward statements as reasons for accepting (or rejecting) other statements. The result of argumentation is none other than an inference (argument), in which the statement (or statement form) that is justified or supported (the conclusion) can be distinguished from the statements (statement forms) used to justify the given statement (the premises). The premises of an inference are put forward as reasons for accepting the conclusion of an inference.

When assessing the reliability and adequacy of the argumentation and its results (i.e. arguments or inferences), we are faced with the following two basic questions:

1. Are all the premises, put forward as reasons for accepting the conclusion, true – or are only some of them true?

2. Assuming that the premises are true, do they guarantee the conclusion is true? Alternatively, how do the premises (assuming they are true) support the conclusion?

The first question concerns the factual truth of the premises – i.e. whether the statements put forward as reasons for accepting a thesis are indeed true. Answering the second question requires us to determine whether the (assumed) truth of the premises suffices to demonstrate the truth of the conclusion or provides some support for the conclusion.

To answer the first question, we need to compare the statements appearing as premises with the state of affairs. (Of course, this process for testing the truth of statements applies only to (some) empirical statements, not analytic statements. When determining the truth-value of an analytic statement, we simply understand what they mean, and there is no need to compare them with the world out there.) But, where (empirical) statements are concerned, it is only through this process of empirical testing that we can decide whether they are true or not.

Determining whether the truth of the premises guarantees the truth of the conclusion, or if it provides sufficient support for the conclusion, depends primarily on the *logical form of the inference*. The logical form of some types of arguments is such that their logical constants and variables guarantee their validity.

Deductive inference

What, however, do we mean if we say that an inference (argument) is logically (deductively) valid? *An inference is logically valid* if and only if its *conclusion is (logically) entailed by the premises*.[23] The concept of logical validity of an inference therefore depends on the concept of logical entailment. In the literature on logic, logical entailment is usually defined and specified in relation to a particular formal logical system or language. Our two definitions of logical entailment below will be somewhat simplistic but sufficient for our purposes. Although these definitions refer to the relation of logical entailment between the set of statements $\{S_1, \ldots, S_n\}$ and the statement S, where $n \geq 0$[24] we can extrapolate the concept of logical entailment to inferences or inference forms. Logical entailment can therefore be defined using the following two equivalent definitions:

(LE) Statement S is logically entailed by set of statements $\{S_1, \ldots, S_n\}$ if and only if, assuming that all statements in the set $\{S_1, \ldots, S_n\}$ are true, S must also be true.

(LE′) Statement S is logically entailed by set of statements $\{S_1, \ldots, S_n\}$ if and only if *it is not possible (logically conceivable)* that all statements in set $\{S_1, \ldots, S_n\}$ are true and statement S is false.

If we substitute conclusion C for statement S and premises $P_1, \ldots, P_n$ for the statements in set $\{S_1, \ldots, S_n\}$, where the number of premises is $n \geq 1$, we ob-

[23] We disregard the distinction between entailment and logical entailment often made in the literature on logic (see, e.g., Cmorej 2000, 335) since it has no bearing on our interpretation.

[24] A statement (or formula) can be entailed by an empty set of statements (formulae). In that case, the statement is a tautology, i.e. one that is logically true in every interpretation. Inferences require at least one premise, i.e. $n \geq 1$.

tain an alternative definition of the logical entailment as a relation between the premises and the conclusion of the inference.

In logic, it is standard to assess the logical validity of inferences based on the logical validity of their inference forms (or logical forms). An inference is valid if its logical form is valid. We then require a definition of the relation of logical inference between the set of statement forms $\{\mathcal{A}_1, \ldots, \mathcal{A}_n\}$ and statement form B. A working definition of the entailment between statement forms can be expressed thus (see, e.g. Cmorej 2000, 333):

(LEF) Statement form B is logically entailed by the set of statement forms $\{\mathcal{A}_1, \ldots, \mathcal{A}_n\}$ if and only if under no substitution for the variables of forms $\mathcal{A}_1, \ldots, \mathcal{A}_n$, and B is it possible to obtain true statements from $\mathcal{A}_1, \ldots, \mathcal{A}_n$ and a false statement from B.

Logical entailment therefore expresses a *specific relation of truth-dependence* between a certain set of statements (statement forms) and another statement (statement form). There is a big distinction between whether an inference or inference form is valid (i.e. its conclusion is entailed by the premises) and whether the premises or conclusion in the inference is true. In logic, there are effective methods for proving whether a set of statements logically entails another statement, but we shall not discuss those here. (The reader is referred to the work of Cmorej 2001b; Gahér 2003; Hammack 2009; and Zouhar 2008 on the various types of methods of proof).

The definition of logical entailment or argument validity will enable us to distinguish a number of situations that may occur when evaluating the premises and conclusion where the inference is valid. An inference may be valid and one of the following situations may occur:

(a) all the premises are true and the conclusion is true

(b) some premises are true, some are false, and the conclusion is true

(c) some premises are true, some are false, and the conclusion is false

(d) all the premises are false and the conclusion is true

(e) all the premises are false and the conclusion is false

The only case which is ruled out by the definition of logical entailment between the premises and the conclusion is one where all the premises are true and the conclusion is false. An inference in which all the premises are true and the conclusion is false is therefore invalid.

However, we must distinguish this situation from one where we have an inference where the premises and conclusion are known to be true. The fact that all the premises are true, and the conclusion is true, does not necessarily mean that the inference is valid. Even where invalid inferences are concerned all of the situations a)–e) described above may occur. The fact that the premises and conclusion are true does not tell us whether the inference is valid or invalid. Its validity or otherwise is determined by its logical form (or inference form). Inferences (I2) and (I3) listed above are logically valid because their logical form (LF1) is valid. Whatever (appropriate) natural language expressions we substitute for individual constant "*a*" and predicate variables "*F*" and "*G*", the conclusion of an inference that has this logical form will be true whenever both the premises are true.

To illustrate the logical form of an inference that is invalid, let us look at the following inference:

(I4) Some humans are mortal.
Socrates is human.

Socrates is mortal.

In inference (I4), both the premises are true, and the conclusion is true. (The formulation of the first premise may seem unnatural, since we believe all people are mortal, not just some. However, this does not rule out the premise being true. For it is true that there exists at least one human who is mortal, and that is exactly what the statement says.) The logical form of inference (I4) can be expressed in the language of predicate logic as follows:

(LF2) $(\exists x)(F(x) \wedge G(x))$
$F(a)$

$G(a)$

The logical constants in the first premise (i.e. the existential quantifier and conjunction) do not take a form (meaning) that would guarantee that in any inference of this logical form in which the premises are true, the conclusion must be true as well. Therefore, the logical form of this inference is invalid. We can construct another inference with a logical form identical to (LF2), but while its premises will be true, the conclusion will be false. Consider the inference:

(I5) Some singers are Chinese.
Michael Jackson is a singer.

Michael Jackson is Chinese.

Both the premises in (I5) are true (assuming we disregard the tense or relate the premises to when Michael Jackson was alive), but the conclusion is evidently false. But the logical form of (I4) and (I5) is identical. Because it is not logically valid, that means we can find examples of inferences that correspond to situation a), as well as inferences where the premises are true but the conclusions are false.

Below we list some of the infinite number of inference rules that are logically valid that we may encounter when reconstructing the logical forms of arguments that often appear in scientific research. Some are formulated using propositional logic, others in predicate logic, and some in both forms. We also include examples of inferences where the logical form is identical to the logical form of these rules:

$p \rightarrow q$	If we examine language diachronically, we account for its development.
p	We examine language diachronically.
q	We account for its development.

$p \rightarrow q$	If we examine language diachronically, we account for its development.
$\neg q$	It is not true that we account for the development of language.
$\neg p$	It is not true that we examine language diachronically.

$(\forall x)(F(x) \rightarrow G(x))$	All Slavic languages are Indo-European.
$F(a)$	Slovak is a Slavic language.
$G(a)$	Slovak is an Indo-European language.

$(\forall x)(F(x) \rightarrow G(x))$	All Slavic languages are Indo-European.
$\neg G(b)$	Hebrew is not an Indo-European language.
$\neg F(b)$	Hebrew is not a Slavic language.

$(\forall x)(F(x) \rightarrow G(x))$	All Slavic languages are Indo-European.
$(\forall x)(G(x) \rightarrow H(x))$	All Indo-European languages are natural languages.
$(\forall x)(F(x) \rightarrow H(x))$	All Slavic languages are natural languages.

Some of these logically valid rules will be used in the next sections; some we will look at later on.

We have not yet mentioned one specific type of inference. It is a logically valid inference (and hence satisfies the definition of logical entailment) but it has no cognitive value in argumentation and justification. It is an inference where the conclusion is also one of the premises. The following form represents the simplest variant of this type of inference:

(LF3) $\dfrac{p}{p}$

Here, "p" can be substituted with any statement. Inferences in which the statement in the conclusion is one of the premises are *trivially valid inferences*.

In the class of all valid inferences, we can distinguish between inferences that are trivially valid and inferences that are valid, but not trivially valid. We can call the latter *non-trivial inferences*. In the class of non-trivial inferences, we can identify another subclass: arguments where the premises are all true. We call these *sound inferences* (see Figure 1).[25]

[25] This terminology has been adapted from Cmorej (2001b). Cmorej also distinguishes rigorous inferences (or arguments) that are the (proper) subset of sound inferences. Rigorous inferences

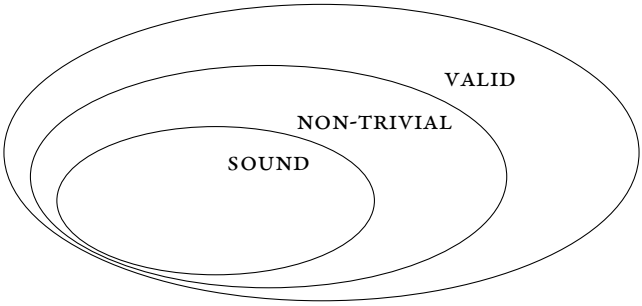

Figure 1. The classification of valid inferences.

To summarize, we have distinguished a class of inferences that are deductively (logically) valid. Within this class, we have identified inferences that are non-trivially valid, as well as inferences that are non-trivially valid *and* all their premises are true. The latter are called *sound inferences*. Their cognitive advantage is that they always lead from true assumptions to true statements. However, this property means their potential to extend our knowledge of the world is rather limited. The conclusion of a valid argument (including sound ones) does not contain *any new empirical information* that has not already been expressed (or was at least implicitly present) in the premises. Thus, if in non-trivially valid arguments we begin from true premises then we always arrive at true conclusions, but the truth of the conclusion is always "contained" in the truth of the premise. Nevertheless, valid arguments represent a precious tool for uncovering (new) analytic (i.e., non-empirical) information contained (implicitly) in the premises.

Many forms of deductive inference or reasoning are used in ordinary communication as well as in scientific discourse. Their use is one of the fundamental (essential) principles of rationality. Logically valid inferences (or, equivalently, deductive inferences) are used much more systematically in the fields of mathematics, logic, computer science and generally in the formal (analytic) disciplines than

are sound inferences where the premises are *known* to be true.

in most empirical science disciplines. The implication therefore is that empirical disciplines also rely on the many types of non-deductive reasoning, especially the various kinds of inductive and abductive inference.[26]

In the remaining part of this section, we will learn about several types of non-deductive reasoning. We shall again focus on two questions in our discussion of non-deductive inference as a conceptual method: 1. Are all the premises of the non-deductive argument true? 2. Does the fact that the premises are true provide (sufficient) support for the conclusion to be true? To answer the first question in relation to empirical statements, we need to apply an empirical method (such as observation, measurement etc.). The second question again depends on the logical form of non-deductive inferences, and on other factors that are key to the representation and evaluation of the inferences.

Non-deductive inference

What are non-deductive inferences? The basic definition is quite simple: *Non-deductive inferences* are inferences that *are not logically valid*. In other words, the conclusion of these inferences is not entailed by the premises. Non-deductive inferences are therefore all inferences where the logical form does not guarantee that whenever all the premises are true, the conclusion is also true. Inference (I5) above is a type of non-deductive inference, since its conclusion is not entailed by the premises.

However, we will not look at all types of non-deductive inference, just the many forms and varieties used in science. Although they do not provide us with a tool for obtaining infallible beliefs or knowledge, they are important for generating and testing our beliefs about the world.

We have noted that non-deductive inferences are those that are not logically valid. But how do we identify the non-deductive inferences that are better at generating knowledge? There is still no entirely satisfactory answer to this question. However, there are several theoretical approaches (for example, theories of non-

[26] Reasoning by analogy is sometimes distinguished from inductive inference. However, we shall treat reasoning by analogy as a kind of inductive reasoning.

monotonic inference, attempts at constructing an inductive logic, many theories in formal epistemology or argumentation theory) that provide a set of criteria for representing and evaluating many cognitively relevant non-deductive inferences. These approaches differ in terms of their goals, as well as in the criteria for non-deductive inferences. We will limit ourselves to giving a neutral overview of the *logical form* and some of the *methodological criteria* for certain types of *cognitively relevant* non-deductive inferences. Cognitively relevant inferences are those that are used in scientific (expert) discourse (and sometimes also in non-scientific discourse) to generate and test predictive and explanatory hypotheses (or theories).

The *reliability* or *cogency* of non-deductive inferences depends on various factors: (a) on the mutual *relevance* of the premises and conclusion; (b) on the (kind of) mutual *support* between the set of premises and the conclusion; and (c) on the *truth-value* of the premises of these inferences.

The relations between *relevance* and *support* are defined variously in the different philosophical approaches. In general, we may say that the former relation can be defined in two basic ways (see also Schurz 1991):

(a1) Premise P_i that is an element in set of premises $\{P_1, \ldots, P_m\}$ is relevant to conclusion C if set $\{P_1, \ldots, P_m\}$ is consistent and $\{P_1, \ldots, P_m\}$ entails C, but C is not entailed by set $\{P_1, \ldots, P_m\} \setminus \{P_i\}$ from which premise P_i has been eliminated (i.e. C is not entailed by the difference of sets $\{P_1, \ldots, P_m\}$ and $\{P_i\}$).

(a2) Premise P_i is relevant to conclusion C if the probability of C, assuming that P_i is true, is different from the probability of C (without taking the premise into account), i.e. $Pr(C \mid P_n) \neq Pr(C)$.

In general, premise P_i is relevant to conclusion C if and only if either (a1) or (a2) is true.

However, it is much more difficult to provide a theoretically neutral definition of the relation of *support* between the premises and conclusions of cognitively relevant non-deductive inferences. The existing approaches model this relation, and other related factors, in different ways. We shall simply point out three simple

principles that underlie several approaches to defining the criteria for evaluating non-deductive inferences. These criteria establish the relation of support between the conclusion and the premises of an inference:

(b1) The relation of support is modeled using a conditional probability function $Pr(X \mid Y)$ defined on the elements of some language L, where some of the elements appear as premises and another element(s) as a conclusion: $Pr(C \mid P_1 \wedge \ldots \wedge P_n)$.[27] More precisely, the relation of support is defined as the degree of probability the premises confer on the conclusion.

(b2) The relation of support is modeled using schemes of *defeasible* arguments. Here, the premises of cognitively relevant non-deductive inferences are viewed as the *prima facie* reasons for accepting the conclusion, unless there is a known *counterargument* that would challenge the truth of the premises, conclusion or the relation between the premises and the conclusion in the original argument. The conclusion of a defeasible argument can be accepted temporarily, relative to the premises and the available knowledge base, but it can later be rejected in the light of new information.[28]

(b3) The relation of support is modeled by an explanatory relation between the premises (or their components) and the conclusion. If the conclusion (or more precisely, the truth of the conclusion) of a non-deductive argument can explain the truth of the particular premise(s) of that argument (better than other, alternative conclusions), then this explanatory function is considered to be a reason to accept the conclusion (or the belief about the truth, plausibility or probability of the conclusion).[29]

[27] We read this as "The probability that C is true assuming that all the premises $P_1, \ldots, P_n$ are true". We will provide a more detailed explanation in the subsection on inductive arguments.

[28] On modeling non-deductive inferences as defeasible arguments, see e.g. Pollock (1987).

[29] This approach is typical of so-called abductive reasoning or inference to the best explanation. See e.g. Harman (1965) or Lipton (2004).

These three principles underlie a variety of approaches where the goal is to analyze, reconstruct, represent and evaluate non-deductive arguments. We shall characterize them in greater detail in the next two subsections.

Finally, even though the *actual truth of the premises* is not a necessary condition for evaluating the probabilistic modeling of the relation of support, it remains an important factor in evaluating non-deductive arguments.

The three factors discussed above (relevance, relation of support and the truth-value of the premises) do not rule out the possibility (or need) to include other criteria that fulfill a methodological or logical role in the representation and evaluation of inferences or inference forms.

The non-deductive inferences we will be most interested in can be divided into two groups: *inductive* inferences and *abductive* inferences.

Inductive inferences One approach to delineating a (non-empty) class of cognitively relevant inductive inferences relies on the concept of probability. More precisely, it makes use of the conditional probability function that expresses the probability of a statement (proposition) representing the conclusion of an inference, assuming that the premises of that inference are true. This use of the concept of probability (probability function) enables us to define two different, non-equivalent concepts of inductive support.

The first option is to define the concept of an inductively strong argument:

(ISA) An argument is inductively strong if and only if (i) it is not deductively valid; and (ii) the probability of its conclusion, assuming that all the premises are true, is greater than the probability of the negation of the conclusion, assuming that all the premises are true:

$$Pr(C \mid \{P_1, \ldots, P_n\}) > Pr(\neg C \mid \{P_1, \ldots, P_n\})$$

This definition is partially based on Carnap's concept of absolute confirmation (see Carnap 1962; but also Hájek – Joyce 2008 and Crupi 2015). However, Carnap thought the project of constructing an inductive logic as a generalization of the relation of deductive entailment, and so did not accept condition (i) of the (ISA)

definition. The (ISA) definition is close to the way inductive arguments are typically defined in the literature (see e.g. Skyrms 2000, 17). To express the idea behind (ISA) less formally, we could say that an inductively strong argument is one whose premises make the conclusion (highly) probable. Just how probable? The (ISA) definition establishes minimal probability only: the premises must make the conclusion more probable than its negation. Moreover, since according to probability theory, $Pr(C \mid \{P_1, \ldots, P_n\}) + Pr(\neg C \mid \{P_1, \ldots, P_n\}) = 1$, it follows that $Pr(C \mid \{P_1, \ldots, P_n\}) > 0.5$.[30]

The second option is to compare the prior probability of conclusion $Pr(C)$ with conditional probability $Pr(C \mid \{P_1, \ldots, P_n\})$, as expressed in the definition of incremental inductive support:

(IIS) Premises $P_1, \ldots, P_n$ inductively support the conclusion C of an (inductive) inference if and only if the probability of the conclusion, assuming that all the premises are true, is greater than the prior probability of the conclusion (i.e. the probability of the conclusion before taking the premises into consideration):

$$Pr(C \mid \{P_1, \ldots, P_n\}) > Pr(C).$$

The (IIS) definition does not require the premises to make the conclusion more probable than its negation. It suffices if they increase the original probability assigned to it before the premises are taken into account. (IIS) also originated in Carnap's work (1962) and is now one of the standard definitions of the Bayesian theory of confirmation (see Crupi 2015, Hájek – Joyce 2008, Hawthorne 2017, Howson – Urbach 2006).

We will come back to the theory of confirmation that makes use of the concept of incremental inductive support (IIS) in Chapter 4. It has to be noted, though, that neither of these definitions of inductive support tells us anything about the meaning of the concept of probability *Pr* that appears in it. For a detailed introduction to several interpretations of the concept of probability, see e.g. Gillies

[30] This is a logical consequence of the axioms of probability theory.

(2000), Hájek (2012) or Childers (2013). Here, we will simply briefly characterize two (epistemic) interpretations of the concept of probability associated with the (ISA) and (IIS) definitions: the concept of *probability as the degree of belief* of a rational agent and that of *logical probability*. According to the first, *subjectivist interpretation*, the concept of probability expresses the degree of belief of an (ideal) agent who represents the differing "strength" of their beliefs using probabilities in the interval $[0, 1]$ of real numbers $\mathbb{R}$. The *system* of the degrees of the agent's belief should be *coherent*. Coherence is understood as compatibility between the system of degrees of the agent's belief and the axioms of probability theory. On the other hand, according to the *logical interpretation*, the concept of probability expresses the degree to which the conclusion is *partially entailed* by the premises (as formulated in a – formal – language).

Next, we shall abstract from the details of the particular approaches to modeling the relation of inductive support. We will focus on characterizing the logical form and the basic methodological criteria associated with the given type of inductive inference.

Before doing that, let us briefly look at the following two inductive inferences:

(I6)
I met stranger *a*.
Person *a* borrowed money from me,
saying they would pay me back within the week.
———
Person *a* will pay me back within the week.

(I7)
I met friend *b*.
Friend *b* borrowed money from me,
saying she would pay me back within the week.
Friend *b* has always kept her word.
———
Friend *b* will pay me back within the week.

Both inferences share something in common, but also differ substantially. They share a similar, although not completely identical, logical form. They differ in the degree to which the premises support the conclusion. Their logical form is

not sufficient to determine which premises are inductively strong and which are inductively weak, or, in other words, to determine which inference contains the premises that confer greater probability on the conclusion. In both inferences, the conclusion may be false despite their premises being true. However, it seems reasonable to say that the probability of the conclusion being false seems quite high in the first inference, and that we would probably expect the conclusion to be true, rather than false, in the second inference. If we were to explicitly formulate our reasons for believing that, we could say that our implicit assumption leads us to doubt the conclusion of (I6): "Strangers who ask others for money, promising they will pay them back, usually do not keep their word." And since we have met someone to whom this generalization applied, we expect this person won't keep their promise. Conversely in (I7) our willingness to accept that the conclusion is true or highly probable is related to our belief that "Friends who have kept their word in the past and could be relied on will also be reliable in the future." Even though we do not consider this belief infallible, the fact that it is representative of our experiences so far leads us to accept that conclusion (I7) is highly probable or more probable than its negation. In assessing the reliability or strength of the argument, we have (implicitly) relied on the concept of inductive strength as defined in (ISA).

Although we are not always capable of distinguishing precisely between reliable and unreliable inductive inferences – since it may be difficult to specify the numeric probabilities appearing in the (ISA) and (IIS) definitions – we are usually able, at least in principle, to specify the reasons for deeming a certain inference reliable or unreliable.

Let us now look at and briefly characterize the types of inductive arguments commonly listed in the literature as schemes of inductive inference (see e.g. Carnap 1962, Chapter 4, §44; Gustason 1994; Salmon 1995, Chapters 4–5; Skyrms 2000, Chapter 5; Viceník 2001a,b).

Enumerative induction The basic form of enumerative induction consists of one or more premises stating that in number n of cases, object a had (characteristic) property F, while the conclusion states that in the next $(n+1)$ case, object

a will also have property *F*. A stronger variant might state that object *a* always (in any situation) has property *F*.

The premises and conclusion of an enumerative induction may be more complex in structure. Consider the following inference:

(I8) Jozef is a Comenius University ("CU") graduate and had found a job within 6 months of graduating.
Petra is a CU graduate and had found a job within 6 months of graduating.
Júlia is a CU graduate and had found a job within 6 months of graduating.
…
…
Kamil is a CU graduate and had found a job within 6 months of graduating.

The next CU graduate will find a job within 6 months (…).

We can express the logical form of this inference using predicate logic (simplified somewhat) thus:

(EI) $F(a) \wedge G(a)$
$F(b) \wedge G(b)$
$F(c) \wedge G(c)$
…
…
$F(n) \wedge G(n)$

$F(n+1) \rightarrow G(n+1)$

or in its stronger form:

(EI′)
$$\begin{array}{l} F(a) \wedge G(a) \\ F(b) \wedge G(b) \\ F(c) \wedge G(c) \\ \ldots \\ \ldots \\ F(n) \wedge G(n) \\ \hline\hline (\forall x)(F(x) \rightarrow G(x)) \end{array}$$

The premises in (I8) of the logical form (EI) or (EI′) state that n objects have properties F and G. In our case, the objects are individuals who have the properties of *being a CU graduate* and *having found a job within 6 months of graduating*. The conclusion states that the next (or every other) object (person) to have property F will also have property G.

The double horizontal line marking off the premises from the conclusion indicates that the inference is not a deductively valid one, but one whose premises – depending on a number of factors – confer a certain probability on the conclusion, ideally a high one.

Sometimes expressions like "F" or the properties denoted by them are called *reference predicates* or *reference properties*, while expressions like "G" or the properties they denote are called *attribution predicates* or *attribution properties*.

Typically, all the premises in an enumerative induction express a certain state of affairs that has been observed or otherwise empirically identified. Therefore, these statements are sometimes called *observational statements*. They express what has been observed or empirically tested with observable results. While the statement appearing in the conclusion of an enumerative induction of the form (EI) is sometimes called a *predictive statement* or simply a *prediction*. Stronger conclusions of the form (EI′) are usually called *inductive generalizations* or simply *generalizations*.

We have noted that the logical form of an inductive argument does not in itself conclusively point to which inductive inferences (if any) are reliable or strong (or more reliable, stronger). In the methodological literature, a range of requirements can be placed on reliable inferences of the form (EI) or (EI′):

1. All the premises of an enumerative induction must be true.

2. The number n of observed cases in which objects exhibiting property F also exhibited property G must be sufficiently large relatively to population size.

3. No case contradicting the inductive generalization in the conclusion has been observed; i.e. no case has been observed of object i exhibiting property F but not property G (i.e., $F(i) \wedge \neg G(i)$, where $i \leq n$).

4. Property G was examined (or identified) in the most diverse number of cases in which objects had property F.

These four conditions do not guarantee the reliability of any examples of the enumerative kind of inference satisfying them. However, they do hint at candidates that may potentially be reliable enumerative inductions. It is generally the case that we can characterize the relation of support between the premises and the conclusion in inferences of form (EI) or (EI′) using probabilistic approaches that rely on the (IIS) definition or theories of defeasible reasoning (see e.g. Pollock 1987, 489).

Let us briefly outline the interpretation of inductive support provided by the (IIS) definition. Let the probability of the conclusion in enumerative induction (EI) or (EI′) be $Pr(C) = r_1$. Let the probability of the conclusion of (EI) or (EI′), assuming that the premises are true, be $P(C \mid \{P_1, \ldots, P_n\}) = r_2$. If $r_2 > r_1$, the premises inductively support the conclusion. Alternatively, we could interpret a situation in which the concept of probability from the (ISA) definition is used to evaluate an argument of the form (EI) or (EI′).

We might similarly use one of the approaches representing schemes of defeasible arguments to interpret inductive support. The plausibility of an inductive inference of the form (EI) or (EI′) is based on its premises being the *prima facie* reason for accepting the conclusion. If we had an object with reference property F but not attribution property G, we would reject the conclusion (see condition (iii) above).

Nelson Goodman (1983, 72–83) established that enumerative induction only works with *certain suitable* properties or predicates. He showed that a purely syntactical definition of (enumerative) induction (i.e. one based on a logical form only) can lead to cases in which the same empirical evidence can be used to formulate (at least) two different inferences in which the premises are true, but the conclusions are mutually inconsistent. Several attempts have been made in the literature to deal with this problem. On Goodman's "new riddle" of induction and its solution, see Fitelson (2008) and Schwarz (2011).

Despite the problems associated with it, enumerative induction is often used in scientific discourse. Whenever partially observed results are extrapolated to the entire, untested population of objects, we make an inference based on the scheme (or rule) of enumerative induction.

In characterizing the remaining forms of induction we will limit ourselves to introducing their logical form. The methodological conditions under which the relevant forms of inference can be characterized as reliable or strong are similar to those applying to enumerative induction. We will leave it up to the reader to specify them.

Statistical generalization This is another type of inference that is similar in some respects to enumerative induction. Typically, the premises of an inference of the statistical generalization type state that *m*% of the elements of a sample of objects with reference property *F* also had property *G*. In the conclusion the percentage of objects exhibiting property *F* that also exhibit property *G* is extrapolated to the whole population. This type of inference can be expressed thus:

(SG) *m*% of the objects tested (observed) that have property *F* have property *G*.

---

m% of all objects that have property *F* have property *G*.

This type of argument is representative of the type of inference found in the statistical theory of point estimation. In point estimation, we take a certain feature of the sample, such as the mean or proportion, as the base of our point estimate of

a parameter (e.g. the mean or proportion) of the entire population of objects. We shall not discuss statistical methods further here. But, we will add that there are standard statistical methods for estimating the parameters of a population based on the statistical indicators of a (random) sample. Making inferences in the form of a statistical generalization is one of the main forms used in such methods.

Probabilistic and statistical inferences Having briefly introduced statistical generalization as a type of statistical inference, we shall now turn to another type of statistical inference. In this type the premises presuppose a statistical generalization of the kind represented in the conclusion of the preceding type of inference (SG). Consider the following inference:

(I9) 92% of Slovaks are religious.
Person *a* is Slovak.
═══════ [92%]
Person *a* is religious.

The general form of statistical inference can be represented thus:

(SI) *m*% of objects that have property *F* also have property *G*.
$F(a)$
═══════ [*r*]
$G(a)$

Instead of using percentages, we can refer to probability *r* from the interval [0, 1] of real numbers and express the equivalent logical form of inferences of this type thus:

(PI) $Pr(G \mid F) = r$
$F(a)$
═══════ [*r*]
$G(a)$

The premises of statistical and probabilistic inferences of this kind characteristically include at least one statement that expresses a statistical generalization or

a probabilistic hypothesis. In inferences (SI) and (PI), the first premise does so (in (PI), we read the first premise as "The probability of any object having property G assuming that it has property F is equal to the real number r"). However, reliable statistical (or probabilistic) arguments must also satisfy the condition that the percentage (or probability) value is greater than 50% (or 0.5). Moreover, the higher the value, i.e. the nearer it is to 100% (or $r = 1$), the greater the probability an object that has property F as indicated in the first premise will also have property G (which is indicated in the conclusion of this type of inference).

In addition, the reference class of objects established by property F must be homogeneous. A class of objects with property F is homogeneous if further dividing it into subclasses has no effect on the probability of the object having property G.

If probability m is very high (approaching 100%) and the given statistical hypothesis has not been rejected in tests and the reference class satisfies the homogeneity requirement and we have the information that a particular object belongs to that reference class, then we may relatively reliably infer that there is m% probability that the object will also have attribution property G.

Reasoning by analogy Analogical inference is sometimes considered to be a specific type of non-deductive reasoning within inductive inference. However, we shall treat it is a kind of inductive inference.

Analogical inference is a type of inference in which we can infer, based on premises that state that two or more cases share certain characteristic features, that the cases in question also share another similar feature. We can illustrate reasoning by analogy using the following two examples:

(I10) Peter is a student and he uses email, Skype and Facebook to communicate with his peers.
Jane is a student and she uses email, Skype and Facebook to communicate with her peers.
Tanya is a student and she uses email and Skype to communicate with her peers.

Tanya also uses Facebook to communicate with her peers.

(I11) Mice and humans have similar physiological mechanisms. to communicate with his peers.
Mice suffering from illness X and treated with drug D exhibited symptoms Y.
Humans also suffer from illness X.

If we treat humans suffering from illness X with drug D, they will exhibit symptoms Y.

Typically, analogical inferences can be reconstructed using one of the following two schemes:

(AI1) Object a has the characteristic properties $F_1, \ldots, F_n, G$.
Object b has the characteristic properties $F_1, \ldots, F_n, G$.
…
Object n has the characteristic properties $F_1, \ldots, F_n$.
[There is an evidential connection between properties $F_1, \ldots, F_n, G$.]

[It is probable that:] Object n also has characteristic property G.

(AI2) System s_1 has the properties $F_1, \ldots, F_n$ in common with system s_2.
System s_1 also has property G.

[It is probable that:] System s_2 also has property G.

We shall use the now standard terminology (see Hesse 1966; Bartha 2010) and refer to properties $F_1, \ldots, F_n$ (with respect to objects $a, b, \ldots, n$ and systems s_1 and s_2) as

positively analogous properties and property G (with respect to the same objects) as a *hypothetically analogous property*. If there is evidence showing that properties $F_1, \ldots, F_n$ (of objects in a system) are somehow related to property G, and the set of objects (or system s_1) shares properties $F_1, \ldots, F_n$ in common with another object (or system s_2), we have *prima facie* reason to believe that the second system also has property G.

What do we mean, though, when we state that there is some evidential connection between properties $F_1, \ldots, F_n$ and property G? There are many explanations in the literature. For example, properties $F_1, \ldots, F_n$ may be *causes* or *causal factors* related to effect G. Alternatively, G may be the *cause* or *causal factor* (partial cause) of properties $F_1, \ldots, F_n$. Or there may be a positive correlation between properties $F_1, \ldots, F_n$ and property G; all of the properties may be the effects of a common cause. In other contexts, properties $F_1, \ldots, F_n$, and G may be elements of a certain *structure* etc. (see Bartha 2010).

Whether this type of reasoning is plausible depends, among other things, on whether we know of any relevant differences between the systems (objects) that would cast doubt on the conclusion of a hypothesis stating that a system (object) exhibits property G.

We can also assess analogical reasoning based on the probability model (IIS) or the defeasible argument model. According to the probability model, the probability of the conclusion of an analogical inference, assuming its premises are true, is greater than the initial (prior) probability of the conclusion. According to the defeasible approach, the conclusion of an analogical inference (whose premises are true) can be accepted until such time as there is relevant and reliable evidence refuting it.

Eliminative induction The last type of inductive inference we deal with here is *eliminative induction*, also known as *Mill's canons (methods) of induction*. They are a set of inferential rules explicitly and systematically presented by the British philosopher John Stuart Mill (1806–1873) in his *System of Logic Ratiocinative and Inductive* (see Mill 1886, Book III, Chapter VIII). The system comprises five rules – the methods of *agreement*, *difference*, *agreement and difference*, *con-*

comitant variations and *residue*. Mill believed that by using these rules and a set of phenomena or factors A, B, C, D, …, that we consider to be potential causes of another phenomenon (or event) E, we can establish which of the phenomena or factors is the probable cause (see also Viceník 2001b, 203–209). The methods of eliminative induction are used on the assumption that during empirical testing we focused on or directly modified certain initial parameters (A, B, C, D, …), and then observed whether their presence or absence leads (led) to phenomenon (effect) E.

Simplifying somewhat, and using mathematical concepts to describe the way in which the cause of the phenomenon is identified, we may state that parameters A, B, C, D, …, represent the *independent variables* and phenomenon (event) E the *dependent variable*. The aim in using Mill's canons is to establish a relation of functional dependence between any of the independent variables A, B, C, D on the one hand and dependent variable E on the other. The (ideal) type of dependence sought is *causal dependence*.

For the first three rules, and the fifth rule, the *value* of the independent and dependent variables is the *presence* or *absence* of factor A, B, C, D, …, or event E. The respective lowercase letter, i.e. a, b, c, d, …, or e, is used to indicate the presence of the parameters, while their absence is noted as "–". Each column in the rules of eliminative induction represents a single parameter of the test conditions that is a potential cause of E. The results of the test (observation, experiment) are typically recorded in each row as the value of the respective variable. The number of tests performed is indicated by the number of rows under each parameter.

We will discuss the first three methods of eliminative induction in schematic form:

(i) Method of agreement

	A	*B*	*C*	*D*	*E*
1.	*a*	*b*	*c*	*d*	*e*
2.	*a*	–	*c*	*d*	*e*
3.	*a*	*b*	–	*d*	*e*
4.	*a*	*b*	*c*	–	*e*
5.	*a*	–	–	–	*e*

A is the probable cause of *E*.

The table shows that in the tests (in this case, five) involving parameters *A*, *B*, *C*, and *D*, factor *A* occurred whenever event *E* occurred. The method of agreement states that based on the assumption (implicit and not included in the premises of this rule of inference) that parameters *A*, *B*, *C*, and *D* represent the potential causes of phenomenon *E*, and based on the assumption that *A* was the only parameter present during the tests when *E* occurred, we can infer that parameter *A* is the probable cause of event *E*.

In principle, the maximum number of factors appearing as potential causes of *E* is not limited. However, there must be at least two parameters (potential causes) of a finite number.

Let us examine how this rule would work in a simplified situation. Suppose we want to determine which of the various motivational factors makes participants in a random sample of citizens willing to participate in a survey on, for example, interest in environmental issues. Phenomenon *E* represents the event of the participant completing and returning the survey offered to them. The participant completing and returning the survey will be represented by value "*e*" of the variable "*E*". The first motivational factor (*A*) is the information that participants will be paid €20 for completing the survey. Factor *B* is the information that participants will receive brochures on environmental activism. The third factor (*C*) is the information that participants will receive a bottle of mineral water. Let us assume that all the promises are kept and the participants are in no doubt they

will receive their reward after completing the survey. Imagine we approach 50 participants, all of whom we offer reward *A*. Only some of them (say, 20) are also offered reward *B* or *C* or both *B* and *C*. Let us assume that all the 50 participants return their completed survey (phenomenon *E*), regardless of whether they receive either reward *B* or *C* in addition to reward *A*. The method of agreement would lead us to infer that the financial award was the probable cause of (or reason for) the participants completing the survey.

But this conclusion is not certain. The assumption is that in all cases parameter *A* was the only reason for completing the survey. However, we can hardly rule out the possibility that participants who were offered reward *B* and/or *C* in addition to *A* were motivated by factor *B* or *C* (or both), or by the combination of all three factors.

The method of agreement relies on several philosophical assumptions (principles) that are not necessarily met in situations of this type. They are: 1. The method of agreement assumes that (in the given situation) we have considered *all the potential causes* that lead or may lead to phenomenon (event) *E*; 2. The method also assumes that only one factor is the active cause (the other factors do not have an effect); 3. The use of this method requires the potential causes to be observable or otherwise empirically identifiable.

Therefore, the (assumed) truth of the conclusion inferred by this rule depends not only on whether one of the potential factors leading to *E* was always present (unlike the other factors), but also depends on whether the three assumptions above are true in the given situation (when using this rule).

A weaker – conditional – version of the rule appears more plausible, schematically represented thus:

(MA*) One of factors A, B, C, D is the cause of E.
Factor B was not present in test t_i and E occurred.
Factor C was not present in test t_j and E occurred.
Factor D was not present in test t_k and E occurred.
None of the factors B, C, D were present in test t_m and E occurred.
Factor A was present in all of the tests $t_1, \ldots, t_m$ and E occurred.

Factor A is the cause of E.

This type of inference is, in fact, deductively valid. In this sense, it may be plausible. However, the plausibility of the conclusion in (MA*) depends on whether all of the premises are true. Whether premises 2–6 are true depends on what was observed in the tests. However, we cannot test whether the first premise is true through observation. Indeed, there may be reasons to assume we haven't fully thought through the causes of event E. Therefore, the conclusion of this (deductively reconstructed) argument may be false.

Let us turn to the next canon, which also serves to identify which of the potential causes is the active one:

(ii) Method of difference

	A	B	C	D	E
1.	a	b	c	d	e
2.	–	b	c	d	–

A is the probable cause of E.

If in one test we observe the presence of parameters A, B, C, and D, followed by phenomenon E, while in another test, we find that E did not occur even though all the parameters (potential causes) except A were absent, we may infer that parameter A, absent in the second test, is the probable cause of E. In other words, the second test revealed that the presence of parameters B, C, and D is *not sufficient* for phenomenon E to occur.

However, this rule does not eliminate the possibility that the actual cause of E is factor A combined with another factor (say, B), so the cause of phenomenon E would not be A, but A combined with B.

The third rule combines the preceding two in a single eliminative method:

(iii) Method of agreement and difference

	A	B	C	D	E
1.	a	b	c	d	e
2.	a	–	–	–	e
3.	–	b	c	d	–

A is the probable cause of E.

This rule, by itself, cannot eliminate the possibility that what appears to be potential cause A under the test conditions may in fact just be the symptom of another cause. Nonetheless, the method of agreement and difference is a more complex and effective tool for testing how a change in the value of an independent variable translates into a change in the value of the dependent variables.

To illustrate this rule, let us return to our example of the factors motivating survey participation. Assume that 20 of the 50 participants were offered rewards A, B, and C, and all of them completed the survey. Assume that another 10 participants completed the survey after receiving reward A only. Finally, the remaining 20 participants refused to return the completed survey despite being offered rewards B and C (but not A). In this case, we may infer that if A, B, and C are the potential causes of E, then it is probably factor A that was the probable cause of (reason for) the participants completing and returning our environmental survey.

For the sake of completeness, we list Mill's remaining two canons in schematic form:

(iv) Method of concomitant variations

	A	B	C	D	E
1.	a	b	c	d	e
2.	$a*$	–	–	–	$e*$

A is the probable cause of E.

(v) Method of residue

	A	B	C	D	E_1	E_2
1.	a	b	c	d	e_1	e_2
2.	B & C & D is the cause of E_1.					

A is the probable cause of E_2.

Applying eliminative induction involves identifying, under test conditions, those factors we can call the causes of phenomenon (or of combined phenomena) E. These methods can be used to identify or eliminate what we call the *causally necessary* or *causally sufficient condition* of a phenomenon. The two concepts can be defined thus (for now we will leave aside the problems associated with the use of these concepts in some contexts):

(CNC)
A is a *causally necessary condition* for B if and only if it holds that if A does not occur, B does not occur.

(CSC)
A is a *causally sufficient condition* for B if and only if it holds that if A occurs, B *occurs*.

Note that the *method of agreement* enables us to eliminate those parameters that *are not causally necessary conditions* – i.e. parameters B, C, and D. Why? Because

if a causally necessary condition does not occur, phenomenon E will not occur either. In the 2nd to 5th row of the corresponding table, we saw that factors B, C, and D were absent but phenomenon E still occurred.

The *method of difference* shows (in the 2nd row) that factors B, C, and D (either by themselves, or in combination) *are not causally sufficient conditions*. Why? Because whenever a causally sufficient condition occurs, it always leads to phenomenon E occurring, as it is the cause. As the 2nd row of the corresponding table shows, the presence of parameters B, C, and D was not enough for E to occur.

Finally, the *method of agreement and difference* shows that parameter A is both a necessary and a sufficient causal condition for phenomenon E to occur.

Eliminative induction is the last of the inductive forms of reasoning we look at, so we shall now turn to another significant class of non-deductive inferences that plays an important role in scientific and non-scientific discourse.

Abductive inferences *Abductive inferences* are a specific type of non-deductive inference. Their premises typically include (i) statement(s) about the state of affairs, situation or fact to be explained, and (ii) statements that involve potential explanations (causes or reasons) for the state of affairs described in (i). The conclusion of an abductive inference is the potential explanation that appears to be the best one given the knowledge base. There is much debate on which factors determine which of the hypotheses is the "best" (available) explanation.

We can trace the term "abduction", and the explicit definition of abductive inferences, back to the work of the American philosopher Charles Sanders Peirce (1839–1914).[31] However, before turning to the logical form of abduction, let's look at an informal example of the use of abductive inference:

Imagine that, walking through the woods, you come upon an old, abandoned shack. You notice that the lock has been pried open and the door is ajar. You go in and see a room covered in dust, with cobwebs on the windows and walls. The

[31] See e.g. Peirce (1992, 186–199). On defining abductive arguments and the problems associated with their use, see Harman (1965), Lipton (2004), Niiniluoto (2004) or Aliseda (2006).

glass in the window is broken. You conclude that the shack is uninhabited, long abandoned. The hypothesis that the shack is abandoned is the best available explanation for its current state. Observing that the shack is desolate, you assume that the hypothesis – that the shack has been abandoned is the best available explanation – is true. You could be wrong. It may be that a tramp uses the shack as shelter and doesn't look after because it provides enough shelter as it is. Nevertheless, since there are no further clues to indicate the shack is inhabited, you assume that in the circumstances the best (most natural) explanation is provided by the hypothesis that the shack is abandoned.

Expressed in the language of propositional logic, the form of this abductive inference can be formulated thus:

(AB) $p \rightarrow q$
q
$\overline{\overline{p}}$

The premises and conclusion of the inference are:

$p \rightarrow q$: "If the shack is abandoned, then it is ramshackle."

q: "The shack is ramshackle."

p: "The shack is abandoned."

Alternatively, the first premise can be expressed "If the shack *were* abandoned, then it *would* be ramshackle." This formulation linking the explanatory hypothesis with the evidence (fact) is closer to how the typical user of abduction would initially think about the hypothesis. It can be semi-formally expressed using the following scheme:

(AB′) If hypothesis H were true, then E would be true.
E is true.
$\overline{\overline{\text{It is plausible that } H \text{ is true.}}}$

For our purposes, the difference between schemes (AB) and (AB′) can be disregarded, although it should be noted that both schemes only express simple cases of abduction in which the premise is a single explanatory hypothesis "p" . However, we frequently encounter cases where the fact can be explained by several alternative hypotheses. In such situations, our goal is to eliminate the hypotheses that appear to be false or less probable given the empirical information available. If we are able to select one hypothesis that provides a *good enough* explanation of the phenomenon and offers the *best* (most probable) explanation (relative to the context), we infer that the hypothesis is probably true. (For a critique and defense of abductive arguments, see Douven 2017.) This more complex abductive inference is represented by scheme (AB*):

(AB*) $p_1 \rightarrow q$
$p_2 \rightarrow q$
$\ldots$
$p_i \rightarrow q$
q
$\underline{\left(\neg p_2 \wedge \ldots \wedge \neg p_i\right)}$
p_1

Statements $p_1, p_2, \ldots, p_n$ are components of the conditionals that constitute the premises of form (AB*) and represent the alternative hypothesis that, in these circumstances, potentially explain the phenomenon described by statement q. One of the premises states that the phenomenon described by q occurred. The last premise – $\left(\neg p_2 \wedge \ldots \wedge \neg p_i\right)$ – expresses the ideal case in which the available information allows us to eliminate all the alternative hypotheses except for one, hypothesis p_1. The latter is therefore the best explanation, in the circumstances, of why q is true. And that is precisely what the conclusion states. Hypothesis p_1 is therefore the best explanation (given the evidential base) of phenomenon q. Obviously, this does not mean that the hypothesis we abductively infer is actually true.

As our example of the use of abductive inference in science, we will compare two traditional theories in physics: the corpuscular theory of light and the wave theory of light (see Thagard 1978, 78). Phenomena such as the *reflection*, *refraction*, *diffraction* or *polarization* of light have long been known in physics. These phenomena represent the facts an adequate physics theory of light should be able to explain. Corpuscular theory could explain reflection and refraction, but diffraction and polarization presented more of a challenge. In other words, if corpuscular theory were true, it would explain why light reflection and refraction occur. But other light phenomena also required explaining, including diffraction and polarization. The alternative theory, wave theory of light, was capable of explaining not just reflection and refraction, but also diffraction and polarization. Therefore, at the beginning of the 19th century physicists preferred wave theory, which they thought better explained optical phenomena (than the competing theory). We can reconstruct this as follows:

> The phenomena of light reflection, refraction, diffraction and polarization exist.
> If corpuscular theory were true, it would explain the existence of light reflection and refraction (but not diffraction and polarization).
> If the wave theory of light were true, it would explain the existence of light reflection, refraction, diffraction and polarization.
> ---
> It is plausible that wave theory is true.

Quantum mechanics later showed that neither theory provided a complete and true description of the nature of light and optical phenomena. Nonetheless, given the knowledge available in the 18th and 19th centuries, wave theory provided a better explanation than corpuscular theory at that time.

Abductive inference plays a key role whenever we are considering which of the available hypotheses is true and underpins the process of finding a plausible hypotheses. Moreover, abductive inference also has a role to play in the application of a suitable theory (or hypothesis) to explain a state of affairs.

2.3 Empirical methods

The theoretical methods discussed above are indispensable to scientific research. But research is also empirical and we use a range of empirical methods to help us obtain and verify data through sensory experience. We have already noted that empirical methods cannot conceivably be used without concepts and theoretical methods. Furthermore, empirical methods require us to use our sensory organs and various observational, measuring or experimental tools, as well as the many techniques for obtaining, recording, evaluating and analyzing information about the world we live in.

In this section, we will introduce the three fundamental methods of empirical research: observation, measurement and experimentation. The forms of these three general empirical methods differ according to the scientific discipline in question. However, we shall focus on some of the common features shared by many of the methods (or techniques) of observation, measurement and experimentation. Although all three methods are interlinked in scientific research, we will try to characterize them in isolation. (Our discussion draws especially on Čížek et al. 1969, Chapter 8; Berka 1983; Bunge 2005b, Chapters 12–14; and Cohen – Nagel 1934.)

2.3.1 Observation

As is the case with the names of other methods, the term "observation" can be used to refer to (a) a whole set of specific types of observation; or (b) denote a simplified, generally stated procedure that captures several of the shared aspects of the different methods of observation. Unless otherwise specified, we refer to observation in the second sense.

Observation is one of the most *fundamental* methods in empirical science. Although it is not the only empirical method, it is part of several other, empirical and complex methods, such as measurement, experimentation, the testing of empirical hypotheses and scientific explanation.

Observation can be characterized as the *intentional, planned, focused and systematic perception of the outside world* using *sight* and *optical instruments as aids* (see Čížek et al. 1969, 402). Unlike ordinary observation, scientific observation is the *intentional* and *systematic* investigation of the world from within the wider context of scientific research. It has two basic methodological functions: (a) it is the basic *tool for obtaining empirical information* about the world; (b) it is the main *tool for testing empirical hypotheses* about the world. It enables us to systematically collect the data used to formulate hypotheses or that forms the empirical basis on which the hypotheses are tested.

Although observation procedures differ according to the type of research, we can identify some of the basic steps they share in common. Schematically, these steps can be represented using the following instructions:

(I1) Determine the starting point, object and goal of the observation!

(I2) Isolate, either in the mind or practically, the object of observation from other phenomena associated with the object being observed!

(I3) Intentionally and consciously observe the main and secondary aspects of the object being observed!

(I4) Using concepts of initial theory T, describe and record the observed aspects of the object!

(I5) If you have achieved your goal, terminate the observation and analyze the data collected! If not, return to one of the previous instructions and repeat the process!

These instructions have been greatly simplified; nonetheless, they capture the basic steps of a range of observation methods. We will now briefly discuss what lies behind these instructions.

The first instruction requires us to establish the circumstances under which we want to carry out the observation. Naturally, we are only interested in the circumstances we assume relevant to our object – in other words, we need some

prior knowledge about the object to be observed. Observation is an intentional and planned activity so we need to establish what the intended goal is from the start. Do we want to obtain data needed for further analysis? Do we want to collect data to test an existing hypothesis?

Our *existing knowledge* will influence our determination of the context, object and goal of observation as set out in instruction (I1). This includes our *pre-theoretical beliefs* or intuitions relating to the object of observation, as well as the *theoretical knowledge and assumptions* we rely on throughout the observation. We typically use the *concepts* of scientific theories to conceptually grasp the object of observation. Moreover, as we have noted, scientific observation is always carried out with a particular goal in mind. This goal determines the role the process of observation will play in the wider structure of scientific research.

The second instruction, (I2), is closely related to the first. Our pre-existing knowledge as observer also affects the intended focus and what we abstract. The ability to isolate, in the mind or in practice, the object of observation from its wider environment can be crucial to the effectiveness of the collection, recording, analysis and interpretation of the data.

Instructions (I3) and (I4) characterize the core of the method of observation – i.e. the process of concentrating on the object and recording what is seen. Unlike in ordinary observation, in scientific observation the linguistic (symbolic) record of the features observed is typically *expressed using the concepts of the initial theory* used to determine the object and goal of observation (in step (I1)).

Finally, the last instruction determines when the process can be terminated and what to do if the observation fails to achieve its goal.

We shouldn't be too cautious in our understanding of the process outlined in the instructions. Some of the observation activities can be performed in parallel and we may return repeatedly to some of the previous instructions. Nonetheless, the scheme provided above sets out some of the features the otherwise distinct methods of observation share in common.

Similarly to many other methods, observation forms part of other, more complex methods, such as those for testing hypotheses (confirmation, falsification etc.), discovering the causes of phenomena or for measuring or experimenting.

But equally, at certain stages, the method of observation involves the use of other (e.g. conceptual) methods such as abstraction and idealization, and several methods of reasoning and analysis.

The various methods of scientific observation can be classified using a range of criteria. In the methodological literature (such as Čížek et al. 1969, 403), a distinction is made between *direct observation* (using only sensory organs) and *indirect observation* (using optical instruments and aids as well). A distinction is also made between *qualitative observation* (which results in qualitative data) and *quantitative observation* (which produces quantitative data). Finally, observation can be classified as *simple/natural* (no modification of the environment or context of observation) or *experimental observation* (involving the manipulation of the conditions).

Looking at observation in terms of the components involved, we may distinguish the following elements:

(1) The observer Observers initiate the process of observation and select all the other components of the process. However, observers are also a common source of error. Their knowledge of the methods of observation (or techniques involving e.g. measurement), or lack thereof, affects the *quality of the data* collected. The observer's pre-existing knowledge can affect the choice of data or the selection of the recording technique. This, in turn, may negatively affect the final evaluation of data (e.g. with respect to the hypothesis tested).

(2) The object of observation The object of observation is an object in space-time whose observable properties we are interested in. Alternatively, we might focus on unobservable properties that can be accessed via operationalization. If the objects of observation are people, they should not include the observer.

(3) Medium of observation The medium or means of observation are the observer's organs (natural medium), as well as optical aids or measuring instruments (artificial medium). In both cases, they must be reliable mediators of the data being observed (or measured).

(4) Context and environment Each observation is embedded within a stage of scientific research. The research is therefore the main context within which the observation takes place. We have noted that all observations are "laden" with the concepts and categories of a theory. This theory forms the theoretical context of the observation. All observations take place within a natural, social or cultural environment. Therefore, the factors within this environment will also exert an influence on the process of observation.

These four elements lie at the heart of all processes of scientific observation.

2.3.2 Measurement

If we want to determine whether object *a* has property *F*, we can express our finding using either the statement "*a* has property *F*" or the statement "*a* does not have property *F*". For example, if we ask whether *Barack Obama* is the *President of the United States*, the answer will either be yes or no. By contrast, some objects exhibit properties *to lesser or greater degrees*. For example, we can ask whether *Barack Obama* is *taller* than *Donald Trump* or if *Hilary Clinton* has been *involved in politics for longer than Sarah Palin*. In the first case, we are interested in whether the object Donald Trump exhibits the property of *physical height* ("tallness") to a greater degree than the object Barack Obama does. In the second case, we are asking who has spent more time in active politics – Hilary Clinton or Sarah Palin. We call the properties that can be used to compare objects in terms of the degree to which the property is present, and to order the objects on that basis (for example, from lowest to greatest degree) *comparative properties* or *comparative magnitudes*.

With some properties, the difference in the degree to which they are present can be both ordered and numerically expressed on a scale. These are *quantitative properties*.

Although the literature includes several, non-equivalent definitions of the concept of measurement, for our purposes, it will suffice to define "measurement" (in the proper sense of the term) as *the method of assigning numerical values to ele-*

ments of an empirical system or as *the method of assigning numerical values* ***and*** *units to elements of an empirical system*.[32]

The birth of modern science at the turn of the 16th and 17th centuries was accompanied by attempts to use mathematical language to describe nature. This was evident in the tendency to use quantitative expressions in relation to various phenomena, as in "How fast will a body fall to Earth if dropped from a certain height?" or "How far away from Earth are the other planets?".

Today, a number of methods and techniques of measurement are used in the traditional disciplines of natural science, but also in research in the various social science disciplines. In this section, we shall introduce some of the relevant terms and presuppositions relating to various different methods of measurement.

Method of empirical counting

The first and most basic method of measurement, also an essential component of many others, is that of *empirical counting*. Note that this method is not identical to the arithmetic operation of summing ("+"), although it makes use of summing (and many other operations). In empirical counting, the goal is to find the number of elements in a set.

When counting the elements in a set we assume that each element in the set is assigned *one and only one natural number* (1, 2, 3, 4, . . .) and that *one and only one* element in the set corresponds to each natural number. (In the language of set theory, it is a one-to-one function from a set of objects O to the set of natural numbers $\mathbb{N}$.) Counting thus enables us to establish the number of elements in a set of (empirical) objects or to compare sets of objects in terms of the number of elements.

[32] Formally, measurement can be defined as a function $f : D \rightarrow \text{Num} \times \{u_f\}$ that projects objects of an empirical system D onto the Cartesian product of a numerical system Num with a given unit of measurement (see Schurz 2014, 102). Similarly, Berka (1983, Chapter 2) draws on the definition of measurement as a projection homomorphism of an empirical system onto a numerical relational system.

For example, when working with a sample (or population) of people, we might be interested in the number of men and women in the sample. Alternatively, we could be interested in whether the percentage of unemployed people in a country has increased, decreased or remained constant from one year to another. To answer this question, we need to count the number of unemployed people as well as all the people in the labor force, and compare that ratio to the corresponding ratio for the following year. Thus, counting forms part of many other methods and measurements, especially those used in descriptive statistics.

It should be noted that counting the number of elements in a set presupposes that this set (and hence its elements) has been clearly delineated. For example, if we wanted to extract "smart" students from a sample of students, and count the number of them, we would run up against the problem of determining what exactly a "smart student" is. We therefore have to clearly identify the elements in a set before we can count them. In its most general form, the method of empirical counting involves the following instructions:

(I1) Clearly identify set S of empirical objects that you want to count!

(I2) Progressing from 1 to N, assign one and only one natural number to each of the objects in the set such that each number is assigned to no more than one object!

(I3) If every element in S has been assigned a number from 1 to N, then N is the number of elements in the set (the cardinality of the set).

Although the third instruction is an explicit declaration, and not an imperative, it can also be read as a recommendation for the researcher: Identify number N with the number expressing the total count of elements in a given set.

Before turning to the other characteristics of methods of measurement, we shall introduce some of the basic concepts associated with measurement generally.

Variables

For our purposes, the term "variable" will denote (a) *a property* that can *acquire at least two different values*; or (b) a term denoting such a property. In either case, we are talking about variables in relation to their values and these may change in relation to changes in the conditions.

Variables can thus denote the varying conditions that we focus on in research. The properties expressed by variables acquire at least two (usually more) different values. In observational and experimental research, we are usually interested in whether any change in the value of at least one variable is associated with a change in the value of another variable. For example, one variable could be the *time spent studying for a test*. Its values will be, for example, *60 minutes*, …, *90 min.*, …, *240 min.*, …, *960 min*…,. Another variable could be the *test score*, whose value would be the number of points acquired on a test by the subject, ranging from *0* to *100 points*. In our research, we might be interested in whether the subject's test score increases according to the amount of time spent studying for the test (in minutes or hours).

The variables we assume could lead to changes in other variables are called *independent variables*. Variables that are affected by the values of independent variables are called *dependent variables*, because we view any change in their value as resulting from a change in the value of the independent variables. In our example in the previous paragraph, *time spent studying for the test* is an independent variable, while *test score* is a dependent variable.

In observational research using only simple observation and involving no manipulation of the environment, our goal is to track and record whether any change in the value of the independent variable is associated with any change in the value of the dependent variable. In experimental research, we manipulate the value of the independent variables; in other words, we are able to modify these values. In this sense, variables may represent *magnitudes*, i.e. properties that determine size.

Measurement scales

The value of a variable (dependent or independent) can vary in nature. When observing and measuring certain properties (magnitudes, parameters), different measurement scales can be used to determine the type of value possessed by the variable. Therefore, measurement involves both the *measurement process* itself and a *measurement scale*. Generally, a measurement scale is a *measure* or a *scale* (conceptual or material) characterized by an "ordered interval of numerical values, the so-called *scale values*, which can be theoretically assigned to the measured magnitudes…" (Berka 1983, 85). In the methodological literature, we can find the following classification of variables based on the types of scales that characterize the properties of their values:

(1) Nominal scales Nominal scales are the basic instrument for differentiating between the elements in a set. The number, letter or name of the value of a variable (or other symbol or expression) assigned to a particular object does *not* in this case express the degree to which the property is present. It is merely the *code* or *label* assigned to the object to clearly differentiate it from the other elements in the set. An example of a variable with a nominal scale would be the *month of the year*. The names of the months of the year, i.e. *January*, *February*, …, *November*, *December*, then become the values of the property of *being a month of the year*. Nominal scales represent qualitative categories that enable us to classify, compare or analyze objects. They do not provide quantitative data that can be compared in terms of size (i.e. the degree to which the property is present), but they do provide a starting point for the use of other methods, such as empirical counting. To express the possible relations between the values $x_1, x_2, \ldots, x_n$ of variable X using a nominal scale, we could either state that $x_1 = x_2$ or $x_1 \neq x_2$.

(2) Ordinal scales Ordinal scales show the type of value possessed by variables, and these can be put in order. For example, the subjects of a test can be ranked from least successful (i.e. the subject with the lowest test score) to most successful. Similarly, minerals can be ordered in relation to *being harder than*. Cohen

– Nagel (1934, 295) give an example definition of this type of relation: An object (such as a diamond) can be said to *be harder than* another object (such as a piece of glass) if and only if the first object can scratch the second object, but the second object cannot scratch the first object. Moreover, an object *is as hard as* another object if neither of the two objects can scratch the other. Based on this definition, we can order the minerals in progression. Ordinal data can always be put in order based on the intensity of the property exhibited. However, ordinal scales tell us nothing about the difference between the first and the second value, or the second and the third, the third and the fourth etc. Ordinal scales can express the following relations between values $x_1, x_2, \ldots, x_n$ of variable X: $x_i < x_j$, $x_i > x_j$, $x_i = x_j$. However, if we put 50 objects, such as minerals, in order from softest to hardest, we cannot say that the difference in hardness between the 50th and 40th object is equal to the difference in hardness between the 35th and 25th object. This is because the difference between the values is not defined on a scale like this.

(3) Interval scales Interval scales are like ordinal scales in that they represent the type of values (data) that can be ordered from smallest to greatest. Interval scales have a conventionally established zero point and a unit of measurement (such as °C) which can be used to compare the extent to which one value is greater (or smaller) than another (see Schurz 2014, 102–105). If, for example, we measure the temperature of a substance in degrees Celsius, the difference between the values 34 °C and 35 °C is equal to the difference between the values 100 °C and 101 °C. In both cases, the difference is 1 °C. In addition to all the relations available on ordinal scales, the difference between the values of the variables ($|x_i - x_j|$) is defined on interval scales. The conventionally established zero point is another important feature of an interval scale. It is important because when measuring the temperature in degrees Celsius, for example, point 0 °C (conventionally) indicates the temperature at which water changes from a liquid state to a solid state (ice). As temperatures can also have a negative value in degrees Celsius, the zero point does not indicate that the object exhibits the minimal possible value of that magnitude.

(4) Ratio scales Ratio scales are similar to interval scales in that they indicate the difference between (any) two values of magnitude. In both cases, there is a conventionally established unit of measurement. In contrast to interval scales, however, ratio scales have an *absolute zero point,* which is not based on convention. It is assumed to exist objectively. For example, the mass of 0.000… kg represents the absolute zero point for masses measured in kilograms. To say a mass has a negative value (such as −16.79 kg) is meaningless. Similarly, no object can have a temperature lower than the (theoretical) temperature of 0 K (Kelvin).

Sometimes the methodological literature distinguishes between *intensive* and *extensive magnitudes* (*properties*); see e.g. Cohen – Nagel (1934, 293–297). This distinction is based on whether the values of these magnitudes (variables) can be meaningfully added together. Consider temperature or intelligence (IQ): It makes no sense to say that a person with an IQ of 160 has *double* the intelligence of a person with an IQ of 80. Magnitudes whose values cannot be added together are called *intensive magnitudes*. Magnitudes whose values can be added together (such as weight or length) are called *extensive magnitudes*.

Where methods of measurement are concerned, an object is assigned a number expressing the degree to which it exhibits a property (magnitude). Although we don't provide a general characterization of measurement in the form of a sequence of instructions here, we can state that when performing measurements we have to select the variable and the values we want to observe. In order to characterize the values of a variable, we must first establish the type of scale to be used. Finally, the method of observation and the relevant empirical methods for determining the values of the magnitudes being measured form an essential part of the methods and techniques of measurement.

In conclusion, let's just point out the difference between a *measured value* and an *actual value*. Whenever we measure a magnitude, it is possible that random factors and systematic errors may occur. We have no control over the first, but we can avoid systematic errors.

2.3.3 Experimentation

In experimentation, like simple observation, we select the observable properties of the phenomenon under investigation and use our sensory organs, along with any appropriate optical aids or instruments. Unlike in simple observation, however, in experiments we also have to modify or manipulate the conditions of the initial observational situation, while controlling or fixing others at a constant value. We then observe whether this change in circumstances is manifested in the results of our observation.

In an experiment we intervene in the situation being observed by influencing the values of at least one independent variable and tracking whether this change leads to a change in the values of the dependent variable (or variables). An *experiment* is thus *the observation of previously selected properties in controlled and intentionally manipulated conditions*.

Like observation and measurement, experimentation can be viewed as comprising a range of different, albeit related, methods. The common features of these methods can be expressed using the following instructions:

(I1) Select the phenomenon you wish to study experimentally!

(I2) Isolate the phenomenon you wish to study from the factors that could affect it!

(I3) Determine the aspect of the phenomenon that will be represented by the *independent variable* and the aspect that will be represented by the *dependent variable*! Similarly, determine the *extraneous variables* whose values could affect the values of the dependent variables!

(I4) Manipulate the values of the independent variables, while controlling for extraneous variables, and track whether the change in the values of the independent variables leads to any change in the values of the dependent variables!

(I5) Record the results of the experiment!

(I6) Analyze and systematically express the results of the experiment!

It is clear from these instructions that experiments are carried out in modified or consciously produced conditions. Although not expressed explicitly in these instructions, experiments are typically repeated if the circumstances allow.

Another condition required for an experiment to be adequate is the random, or as random as possible, sampling of the subjects or objects of testing. This is because the adequacy of an experiment depends on whether the selected properties are tested on a representative sample of the population (or entities), or whether the choice of experimental group is influenced by factors that prevent representative sampling.

Instruction (I3) refers to, among other things, *extraneous variables*. These are variables which are secondary to the goals of the testing, but which may affect the result of the experiment if not controlled for. There are two main strategies for ensuring an experiment is reliable despite the presence of such secondary factors.

The first strategy is to control for the values of the extraneous variables when manipulating the values of the independent variables. The extraneous variables are fixed at constant values throughout the various tests. This strategy is not always available to us, though. We may not be able to control some of the extraneous variables. The second strategy is to experimentally test two groups of the test object – an *experimental group* and a *control group*.

This strategy is applied when we are unable to identify all the relevant extraneous variables or fix them at a constant value. The idea behind this approach is simple: first, we need a random sample in which we can unambiguously identify the features (of a larger population of objects) that correspond to the independent and dependent variables. We then use random sampling again to divide this set of objects (the sample) into two groups: the *experimental group;* and the *control group*. Since the first sample (from the population) and the second sample (from the first sample) are both random, we may assume that the values of the extraneous variables are equally represented in both groups. Therefore, even though we cannot assume that the values of the extraneous variables in our sample are

constant, we can assume that their various values are present in both groups to a comparable degree.

In both the experimental and the control group, we then identify features X and Y that represent the independent and the dependent variables of the phenomenon. The crucial difference between the experimental group and the control group is that we intervene in the values of independent variable X in the experimental group so they differ from the values of variable X in our control group. Sometimes variable X is not explicitly present in the control group, but this can be represented as the zero value of this variable. For example, when testing the effects of drug D, independent variable X could be "administering D to the patient". In the control group, this variable has the value of 0, representing the fact that the drug was not administered to the patients in this group. In the experimental group, the value of X is (for example) 1, representing the fact that the drug was administered to the patients in this group. We now want to see whether this difference between the values of the independent variables in the experimental group and in the control group will be reflected in any change in the value of dependent variable Y. For example, when testing drug D, we could track the differences in the symptoms of the disease, represented by value Y, between the experimental group and the control group. If we spot any differences in the symptoms, we can ascribe these to the difference in the values of X, since the impact of all the other, extraneous variables was approximately equal in both groups.

Naturally, as was the case with measurement, when experimenting we have to use other methods as part of our theoretical and material preparations of the test conditions. In addition, as in observation and measurement, experiment is part of other, more complex methods used in scientific research.

Study questions

1. Characterize the concept of *method*.
2. Do methods amount to algorithms?

3. Describe the difference between analytic and empirical (synthetic) statements and provide an example of both from your (or a similar) discipline.
4. Characterize the difference between a priori true or false statements and a posteriori true or false statements.
5. Definitions and explications help us eliminate two semantic phenomena from the language of science. Name and briefly characterize them.
6. Choose any two expressions from your discipline. Provide a definition of them and characterize the type of definition.
7. Briefly characterize the difference between analytic and synthetic definitions. (When do we use each of them?)
8. State the basic difference between (proper) definitions and explications!
9. List and briefly characterize the four (Carnap's) criteria of adequate explication!
10. Which two basic conditions must analytic classifications meet for them to be adequate? Briefly characterize them. Provide an example of an analytic classification from your discipline.
11. Briefly characterize the concepts of *reasoning* and *inference* (*argument*)!
12. What is the purpose of identifying the logical form of arguments?
13. Which two basic questions must be taken into consideration when evaluating the reliability of an argument (in the context of scientific argumentation)?
14. When is an argument logically valid, when is it non-trivial and when is it sound?
15. Three basic factors determine the reliability (strength) of non-deductive inferences. What are they?

16. Pick any type of non-deductive inference (such as enumerative induction) and briefly characterize it (with reference to logical form and other properties).

17. Explain: (a) when A is a causally necessary condition for B; (b) when A is a causally sufficient condition for B.

18. Briefly characterize the method of observation and list its basic functions in scientific research.

19. Briefly characterize the four basic components of observation: the observer, the object of observation, the medium and the context (environment).

20. Explain the difference between independent variables and dependent variables.

21. What are the different scales of measurement?

22. Briefly characterize what an experiment is and what experimentation involves.

23. What two strategies can we use to limit the influence of extraneous variables in an experiment? Briefly explain both strategies.

3 TYPES OF SCIENTIFIC RESEARCH: THE H-D MODEL

Now that we are already familiar with the basic theoretical and empirical methods, we can turn to the question of how they take part in the process of scientific inquiry or *scientific research*. We shall see that scientific research comes in various forms, depending on the goals and kinds of data used. Despite these differences, *all* types of scientific research share certain elements in *common*. In this chapter, we will examine these elements by looking at an intentionally simplified *research model* – the model of *hypothetico-deductive research*. Its name reflects the fact that the *hypothetico-deductive method* is used to test and evaluate hypotheses. It is a model that provides a basic scheme with a number of variations, specifically for testing and evaluating scientific hypotheses.

In section 3.1 of this chapter, we deal with the basic typology of scientific research. Then, in section 3.2, we look at research into the social causes of suicide by the French sociologist Émile Durkheim to illustrate the structure of hypothetico-deductive research. Finally, in the last section of the chapter, we shall return to our basic characterization of the hypothetico-deductive model and explain the components that make up the structure of research.

3.1 Typology of research

Scientific research (in empirical disciplines) can be characterized as the systematic and (in principle) replicable application of (theoretical and empirical) scientific methods for collecting, analyzing and evaluating empirical data (given a certain

theoretical background) where the goal is to find a solution to a scientific problem.

Scientific research is therefore the use of scientific methods to solve a cognitively relevant problem (see Zouhar et al. (2017)). A cognitively relevant problem is one that, when resolved, will (potentially) enrich (modify, extend) our knowledge. Thus, scientific research can be seen as a system of methods connected in time and in substance. These methods form the structure of the research process. The process is structured such that it begins with a particular epistemic (theoretical) problem and ends with a set of information relevant to solving that problem. In between the first and last stages of research come the methods that are best suited to solving the given problem.

Although there are many different types of scientific research, and each method (or its particular use) has a characteristic sequence, this does not mean that the sequence is always *specified strictly and in full* by the given type of research. Rather, each stage in our model of research will represent a *basic* step towards solving the cognitive problem while also allowing the researcher to, for example, return to an earlier step. This will become clearer as we learn more about the hypothetico-deductive research model.

To some extent, the structure of the research will depend on the nature of the data and the research goals. These also underpin the various typologies of scientific research (see Kumar 2011, Chapter 1). For example, if we look at research from the perspective of the types of data that are collected, analyzed and interpreted, we can distinguish the following categories:

1. *qualitative research*
2. *quantitative research*
3. *mixed research*

(1) Qualitative research In *qualitative research*, we use *nominal* (*categorical*) or *ordinal* data, but not *interval* or *ratio* data. Generally, qualitative research is about obtaining a detailed description of a phenomenon, situation, person, event

or process. The data used are not quantifiable. A qualitative research project could be about identifying and presenting the views of a group of people on a particular phenomenon. It could also describe related events and order them in (chronological) sequence or it could describe the living conditions of a community or social stratum. Usually the results obtained by examining the subjects, objects, situation and so forth cannot be generalized to other examples, even those that are similar in certain respects.

(2) Quantitative research In *quantitative research*, we use *interval* or *ratio* data. Generally, research is quantitative if the data (and methods of collection, analysis and interpretation) contain information about the *degree* to which a property is exhibited or the *rate at which the phenomenon changes*. This information can be expressed in numbers and, usually, suitable units. Properties that are expressed in this way are sometimes called *magnitudes*. For example, suppose we want to identify the degree of support for various political parties during a certain time interval using a poll. If our sample from the population of potential voters is representative and the results show that x% of the respondents express support for party A, while y% support party B, and so on, then we could state that the values x%, y%, …, represent a (point) *estimate* (or the basis of an interval estimate) of political party support for the political parties within the population as a whole. Results obtained through quantitative research are usually *generalizable* to the entire population (of subjects or objects) represented by the sample provided the sample is representative.

(3) Mixed research As one might expect, *mixed research* uses both qualitative and quantitative data, depending on the research questions.

Other typologies we may come across include, for example, typologies based on the *general research goals*. We can distinguish between:

1. *exploratory research*
2. *descriptive research*

3. *correlational research*

4. *explanatory research*

(1) Exploratory research The goal of *exploratory research* is to gain information about a phenomenon or area on which we lack information. It is used to establish a knowledge base that can then be used when developing further research tasks, problems and hypotheses. The structure of this type of research is relatively loose and flexible. Usually, there are no explicit hypotheses. However, partial (working) hypotheses may be formulated in the course of the research. For example, using the participant observation method, the researcher could join a close-knit community (group) to learn about the habits, behavior and values of its members. At this stage, the findings are not used to test an existing hypothesis. Rather, they generate a *system of data* that can be used to formulate new problems or hypotheses for further research. However, in order to continue investigating the phenomenon, the researcher will require theoretical knowledge.

(2) Descriptive research In *descriptive research*, our efforts are aimed at providing a systematic description of the phenomenon, problem, situation or process. Although *universal* hypotheses are not commonly found in descriptive research, *singular* hypotheses are used (for more on types of hypotheses, see Chapter 4). These express the assumption that a particular object, subject, situation or phenomenon has certain properties. For example, the goal of a descriptive research project might be to describe the types of services provided by a particular institution or firm. Alternatively, we could be interested in analyzing the administrative structure of an organization (or political party), describing the environment in which a particular writer or thinker worked or describing the themes of their work etc. In this case, our hypotheses will relate to the particular institution (firm), organization, author or literary work etc. (Hence, they differ from the universal hypotheses that relate to all institutions of a certain type, or to any organization, writer or literary work etc.).

(3) Correlational research When doing correlational research, we are interested in whether there is a *positive* or *negative* correlation between (at least) two variables that represent certain factors of the phenomenon under investigation. In other words, we want to know whether the values of one variable are accompanied by *an increase or decrease* in the values of another variable, or whether no such change occurs. If the values of both variables increase, that is an example of positive correlation, while if the value of one variable increases and the other decreases, that is an example of negative correlation. If there is no dependency between the values of the two variables, the correlation is zero. In correlational research, we therefore formulate hypotheses in which it is assumed there is a relation of positive correlation or of negative correlation. These are *statistical* hypotheses.

(4) Explanatory research The final type of research is the most complex. It not only involves elements of descriptive research (in the collection of data) and correlational research (in assuming a relation of dependence between at least two variables), but also attempts to answer questions such as "Why did this event occur?" or "Why does phenomenon x have such and such a property?". Put differently, in explanatory research, the desire is usually to *explain* a phenomenon. Correlational research might lead us to the conclusion that stress levels positively correlate with the incidence of heart attacks; or that the domestic environment correlates (positively or negatively) with children's grades. These results then form the basis of explanatory research. We might be interested in finding out whether stress does in fact increase the risk of a heart attack and, if so, why. Similarly, we might ask in what way the domestic environment (positively or negatively) influences children's results. In this type of research, we seek to *explain* why certain phenomena occur, why a particular event has occurred or what mechanisms underpin the processes.

Finally, when thinking about research in terms of the *uses* of the results, we may distinguish two types:

1. *basic research*

2. *applied research*

(1) Basic research In basic research, although we formulate and test hypotheses (theories) relating to a cognitive problem, these may in fact have no practical future use.

(2) Applied research On the other hand, in *applied research* we are seeking solutions to problems (cognitive or technological) that directly affect society, the individual, politics, a country's economy etc.

We have noted that the structure of scientific research depends, to some extent, on the nature of the data we are working with, and on our research goals. In what follows, we shall draw on this distinction, while concentrating on the basic structure of the hypothetico-deductive model of research. Before doing that, we will take a closer look at an example of a research problem (or, more precisely, part of a research problem). Later on, we will use this example to illustrate the structure of our model of research.

3.2 Émile Durkheim on suicide

Émile Durkheim (1858–1917), a French sociologist, investigated the social factors of suicide in the late 19th century. The results were first published in his 1897 book *Suicide*. It will serve as our example of the hypothetico-deductive research model. However, a word of caution to the reader: we are not concerned here with providing a complete and historically accurate description of the procedures Durkheim used. Nonetheless, our account and reconstruction is based on his work. It will be used in the next section to characterize the basic structure of hypothetico-deductive research in more detail.[33]

[33] The idea of using Durkheim's *Suicide* to illustrate the hypothetico-deductive research model came from Johannson (2016, 48–49). The hypothesis discussed below is also borrowed from his work. However, our analysis also includes some broader aspects on which Johansson draws. For a thorough discussion of Durkheim's *Suicide*, see Pickering – Walford (2000).

In *Suicide* (2005) Durkheim analyzes suicide as a specific social phenomenon related to distinctive social factors that cannot simply be reduced to the psychological (pre)dispositions of the individuals involved. More specifically, Durkheim was not interested in suicide as an *individual act* in which the mental state of an *individual* leads them to "voluntarily" end their life. On the contrary, Durkheim considered *suicide* to be a *social phenomenon* that can be investigated indirectly within sociological research by investigating the *suicide rate* in different societies. Therefore, his approach was to identify the *social factors* that could explain the differences in suicide rates in societies. Before analyzing the available data and formulating hypotheses that might explain it, Durkheim defined some of the basic concepts he used in his work. For example, he defined "suicide" as follows:

> "[T]he term suicide is applied to all cases of death resulting directly or indirectly from a positive or negative act of the victim himself, which he knows will produce this result." (Durkheim 2005, xiii)

"Suicide-rate" is then defined as

> "the rate of mortality through suicide, characteristic of the society under consideration." (Durkheim 2005, xiv)

It must be noted that Durkheim's analysis of (the differing rates of) suicide is based primarily (albeit not exclusively) on European countries and societies. His conclusions should therefore be interpreted (and perhaps viewed critically) as applying to the "Western civilization" in the main.

In addition to the critical, negative part of his work in which he objected to the causes of suicide being equated with "mental alienation" or other extra-social factors, Durkheim also offered a positive theory that distinguished three (main) categories of suicide: (i) egoistic suicide; (ii) altruistic suicide; and (iii) anomic suicide. He also formulated a theory of the social causes of each.

In what follows, we will not deal with Durkheim's "crusade" against psychological explanations of (the causes of) suicide. Nor will we criticize Durkheim's anti-psychologism. Moreover, we will disregard the categories of altruistic and

anomic suicide. Instead, we will concentrate on the part of Durkheim's research that deals with the causes of the first type of suicide – egoistic suicide.

Durkheim characterized egoistic suicide as resulting from an individual becoming less and less integrated in society (see Durkheim 2005, 165ff). If we were to ask which social factors were (causally) responsible for the differences in suicide rates between societies, Durkheim would have answered that one cause would be the *different degree of social cohesion*, or, more simply, the different *degree of integration* between societies. Durkheim thought *the degree of integration in a society was inversely proportional to the suicide rate in that society*. In other words, the more integrated the society, the lower its suicide rate.

Durkheim tested this (main) hypothesis by testing three partial hypotheses formulated as follows (see Durkheim 2005, 167):

(H_1) Suicide varies inversely with the degree of integration of religious society.

(H_2) Suicide varies inversely with the degree of integration of domestic society.

(H_3) Suicide varies inversely with the degree of integration of political society.

Let us examine the first in more detail. It states that *the higher the suicide rate* in a *religious* society, *the lower the degree of integration* typically found in that society. Conversely, *the lower the suicide rate* in a religious society, *the more integrated* that society. Note that this hypothesis does not state which is the *cause* and which is the *effect*. It is obvious, though, that Durkheim's goal was to identify the causes of the different suicide rates. Therefore, the hypothesis must be interpreted as stating that the *degree of integration of a religious society* is the *cause*, while the *suicide rate* is the *effect*. A better and more explicit way of expressing the causal nature of the hypothesis would be

(H_{1*}) The degree of integration of a religious society varies inversely with its suicide rate and the former is the cause of the latter.

The question we now face is this: If hypothesis H_1 (or H_1*) were true, how would it manifest itself? One of the possibilities – call it E – can roughly be described

thus: If we encountered (any) two (or more) religious societies that differed in degree of social integration, their suicide rates should also differ (in inverse proportion). More precisely, if religious society S_1 had a higher degree of integration than religious society S_2, then suicide rate $\mathcal{M}_1$ in S_1 should be lower than suicide rate $\mathcal{M}_2$ in S_2. Indeed, Durkheim believed that the statistical data on suicide, when viewed in the context of the religious profile of a (European) country, would reveal something similar.

His analysis (see Durkheim 2005, Book II, Chapter 2) of the first type of suicide was based on the available statistics on suicide rates in European countries at that time. Durkheim noted that predominantly Catholic countries such as Portugal, Spain and Italy had a noticeable lower suicide rate (usually expressed as the number of suicides per million inhabitants) than typically Protestant countries such as Prussia, Saxony and Denmark (see Durkheim 2005, 105).

It could be objected, though, that this difference was due to other cultural differences between these countries. To control for the influence of other cultural differences between these countries, Durkheim thought that the effect of Catholicism and Protestantism should be compared on the suicide rates *within* a society (country). In his view, Switzerland provided suitable empirical "material" for this as it contained Catholic cantons and Protestant cantons and these were inhabited by both French and German national communities.

The data available provided the following information (see Durkheim 2005, 108):

FRENCH CANTONS		GERMAN CANTONS		TOTAL OF CANTONS (ALL NATIONALITIES)	
Catholics	83/mil.	Catholics	87/mil.	Catholics	86.7/mil.
Protestants	453/mil.	Protestants	293/mil.	Protestants	326.3/mil.
				Mixed	212/mil.

Table 1. Data on suicide-rates used by Durkheim.

The numerical data in the table refer to the number of suicides per million inhabitants. It is clear from the table that the suicide-rate was lower in the Catholic cantons and higher in the Protestant ones, irrespective of whether the population was French or German. In principle, we can just look at the "All cantons" column and the figures for the Catholic and the Protestant cantons. These show that the suicide rate in the Protestant cantons is almost four times that in the Catholic cantons. Using other statistical data, Durkheim also identified similar differences in the suicide rates for other areas, such as Bavaria which had a predominantly Catholic population and a lower suicide rate than the minority Protestant population.

Turning back to hypothesis H_1* and one of its implications, denoted above as E, we can see that apart from a variation in the suicide rate, H_1* and E also refer to *varying degrees of integration* of religious society. But how can we tell if one religious society is more integrated than another? In relation to the data analyzed by Durkheim, we can be more specific: When comparing the Catholic and the Protestant cantons of Switzerland, on what basis could we state that the former show a higher rate of integration than the latter?

A fairly extensive part of the second chapter in Book II of Durkheim's work is dedicated to answering this question. In brief, we may state that "[t]he only essential difference between Catholicism and Protestantism is that the second permits free inquiry to a far greater degree than the first" (Durkheim 2005, 112). The difference between the two religions does not concern suicide *per se*. The latter is morally inadmissible in both. Where Catholicism differs from Protestantism, though, is in the roles ascribed to the individual and to the religious community in interpreting and living according to the religious doctrine. Ultimately, this is also reflected in the relationship between the individual and society:

> "…the Catholic accepts his faith readymade, without scrutiny. He may not even submit it to historical examination… The Protestant is far more the author of his faith. The Bible is put in his hands and no interpretation is imposed upon him." (Durkheim 2005, 112)

Catholic manifestations of spiritual life are far more closely connected to the Church, which regulates many areas of its members' private lives. Protestant societies, in contrast, typically exhibit a greater degree of individualism. Consequently, its members' private lives are less regulated. As Durkheim put it, "…the greater concessions a confessional group makes to individual judgment, the less it dominates lives, the less its cohesion and vitality" (Durkheim 2005, 114). His answer to the question posed above is that "the superiority of Protestantism with respect to suicide results from its being a less strongly integrated church than the Catholic church" (Durkheim 2005, 114).

Let us summarize the results we have obtained. Durkheim's goal was to investigate the social causes of suicide as a social phenomenon. One of the hypotheses he proposed was that there exists a class of suicides that are related to the degree to which the individual is integrated in society. More precisely, Durkheim argued that there is a relation of inverse proportionality between the degree of integration in society and the suicide rate of that society. The difference in the degree of integration is the cause of the difference in the suicide rate. This hypothesis was tested by means of three more specific hypotheses that concentrate on three kinds of groups: religious societies, households and political societies. We have looked at the first of these three hypotheses. It states that the degree of integration in a society is inversely proportional to the suicide rate. If we take two societies, i.e. a society represented by the Swiss Catholic cantons and a society represented by the Swiss Protestant cantons, and assume that the former is more integrated than the latter, we would expect the suicide rate to be lower in the former than in the latter. The statistical data available to Durkheim attested to that.

This brief digression into Durkheim's complex research project is sufficient for us to illustrate the structure of scientific research in the following section.

3.3 The structure of scientific research: The hypothetico-deductive model

In section 3.1, we noted that there are various types of scientific research. In this section, we will introduce the basic structure of a specific type of research – research based on the hypothetico-deductive (H-D) method. First, we will identify the six basic stages of this kind of research. We shall then attempt to reconstruct the part of Durkheim's investigations discussed above so we can show the different components. Finally, we will generalize our description of the basic steps of the H-D research model and provide a more detailed characterization.

In the philosophical and methodological literature the *hypothetico-deductive* research model (H-D research) is considered to exhibit various degrees of complexity. Our approach will be based mainly on Hempel's work (1966, Chapters 2–3).[34] H-D research usually comprises the following six basic stages:

1. Formulating the research problem
2. Proposing a hypothesis (hypotheses) as the potential solution to the problem (an explanation)
3. Deriving test implications (predictions) from the hypothesis (hypotheses)
4. Designing the methods of data collection and analysis
5. Testing the derived implications of hypotheses
6. Evaluating (interpreting) the results of the testing

These steps form the (ideal, simplified) core of H-D research, which can be further developed and amended.[35] There is also the possibility of returning to a previous

[34] See also Bunge (2005a), Kitchener (1999, Chapter 12) or Giere et al. (2006, Chapter 2).

[35] Some add further steps to this sequence. For example, Bunge (2005a, 10) includes an intermediate step between 4 and 5, in which the methods proposed for testing the hypotheses are themselves tested for relevance and reliability. Kitchener (1999, Chapter 12) introduces eight stages instead of six. Then there are those who limit themselves to presenting the H-D method of hypothesis testing and evaluation (i.e. stages 2, 3, and 6).

step and then proceeding through the remaining steps. In this section we will demonstrate that Durkheim's suicide research, when suitably reconstructed, can be divided into the six basic stages.

(1) Formulating the research problem (Durkheim)

In Durkheim's research into suicide the assumption is that the latter can be investigated at the psychological level, where we can attempt to identify both an individual (and often unreliable, according to Durkheim) motive that led the person to take their life, and a social phenomenon related to the specific character and structure of the society in which it occurs. If, Durkheim says, we look at particular societies (e.g. countries, religious communities), we can see that they have varying suicide rates. Durkheim's interest in this was such that it prompted him to formulate several research questions. One of these was:

(Q) What social factors operate as causes of the (significantly) different suicide rates in the European societies being compared?

Durkheim also noted that suicide rates vary in Catholic countries such as Portugal, Spain or Italy, and in typically Protestant countries such as Prussia, Lower Saxony or Denmark. He therefore based his research on the following question as well:

(Q') What social factors are responsible for the lower suicide rates in European Catholic societies (countries) relative to the suicide rates in Protestant societies?

(2) Proposing a hypothesis as the potential solution to the problem

The part of Durkheim's research we introduced above *partly answers* research question (Q) and *partly answers* variant (Q'). Both answers can be formulated as hypotheses – that is, as statements that, if true, would (partly) solve the research problem (expressed by questions Q and Q').

Durkheim's hypothesis (H) is thus a partial answer to research question (Q):

(H) *One of the causes* of the different suicide rates in any two (European) societies S_i and S_j is the *different degree* of *social integration* (cohesion) such that the *degree of integration of society* S_i (S_j) *is inversely proportional to the suicide rate in* S_i (S_j).

In response to question Q', which assumes that suicide rates differ in societies that have a different religion, Durkheim offers the following hypothesis as a partial answer:

(H_1*) The degree of integration of a religious society varies inversely with the suicide rate and the former is the cause of the latter.

However, if we apply hypothesis H_1* to a comparison of two (or more) religious societies, the answer to question Q' can be formulated more precisely:

(H') For any two *religious societies* S_i and S_j, it holds that if the degree of social integration of S_i is greater than the degree of social integration of S_j, then the suicide-rate in S_i is lower than the suicide rate in S_j.

Hypothesis H' is a special variant of hypothesis H_1* that contrasts two random religious societies.

(3) Deriving testable implications from the hypothesis

Note that none of hypotheses H, H_1* and H' can be immediately evaluated as true or false. Therefore, we have to derive test implications from them – statements that tell us what should be observed or what empirical data should be collected if the hypotheses were true. In other words, test implications should indicate how we will know the hypothesis is true. In standard cases, test implications are logical consequences of the hypotheses. However, further assumptions (also known as "auxiliary" assumptions or hypotheses) are usually required to derive the implications. These are not tested, but we rely on them when testing the implications of the given hypothesis.

In what follows, we shall focus only on hypothesis H'. What would we expect to observe if it were true? Assume, for example, that there are two religious societies: S_1 represented by the Catholic Swiss cantons and S_2 represented by the Protestant Swiss cantons. Usually, an assumption of this kind expresses what we call "antecedent conditions". For the sake of simplicity, we can include it among the auxiliary assumptions. Let us further assume that it is true that the Catholic Swiss cantons exhibited (at the time of Durkheim's research) a greater degree of social cohesion than the Protestant ones. (We will leave aside the question of how the degree of integration of a society can be objectively assessed.) This information will be included as an auxiliary assumption. Finally, from hypothesis H' and the two auxiliary assumptions (denoted as "A"), we may derive statement E, which represents the test implication of hypothesis H':

(E) The suicide rate in society S_1 is lower than in society S_2.

In other words, the suicide rate in the Catholic cantons is lower than in the Protestant ones.

(4) Designing the methods of data collection and analysis

In Durkheim's research, the available statistical data on the number of suicides in the (mostly) European countries in the 1840–1870 period played a key role. Durkheim relied on existing data. Nonetheless, at this research stage, the goal was to acquire relevant data that would help establish the suicide rates in the Catholic and Protestant Swiss cantons. If no such data had been available, Durkheim would probably have had to use the survey method and contact the church and state institutions which typically keep such information (usually in the form of parish records).

(5) Testing

The statistical data on deaths in the Catholic and Protestant cantons available to Durkheim enabled him to establish the suicide rates in these two religious societies and compare them with test implication E. In the Catholic cantons, the

suicide-rate was around 86.7 persons per million inhabitants, while in the Protestant cantons, the figure was 326.7 persons per million inhabitants. The data was thus compatible with test implication E – that the suicide rate in the Catholic cantons should be lower than in the Protestant ones.

(6) Evaluating (interpreting) the results of the testing

The fact that the data were compatible with the test implication of the hypothesis H' can, under certain circumstances, be interpreted as evidence indicating the hypothesis is true. We could therefore say that the data *confirm* hypothesis H', because they are compatible with one of its implications (the prediction) and, at that moment, there are no available data that would disconfirm or refute the hypothesis.

Having briefly reconstructed Durkheim's research, we can look at the six stages once more, this time from a general perspective covering lots of different examples of research.

Before returning to these stages, it will be useful to set out a logical scheme underpinning the hypothetico-deductive testing of hypotheses. Although the next chapter will be devoted to the issue of hypothesis testing, we will also briefly look at two possible situations in which the results of testing (i.e. empirical evidence) are used to evaluate a hypothesis – *confirmation* and *disconfirmation* (*falsification*) of a hypothesis:

CONFIRMATION	**DISCONFIRMATION**
If H and A, then E.	If H and A, then E.
E.	It is not true that E.
H.	It is not true that (H and A).

The argument representing the confirmation of a hypothesis is non-deductive. In contrast, the argument illustrating the disconfirmation (falsification) of a hypothesis is deductive (it conforms to the *modus tollens* rule).

In both cases, the schemes capture the following steps of the H-D method. The first premise represents stages 2 and 3: proposing the hypothesis and deriving the test implications from the hypothesis (and the auxiliary assumptions). The second premise records the evidence (positive or negative) and expresses the data obtained during hypothesis testing (stage 5). The conclusion of both schemes corresponds to stage 6 – the final assessment of whether the results confirm or disconfirm the hypothesis.

Let us now look again at the stages of H-D research and characterize them at a more general level:

(1) Formulating the research problem

All research is motivated by a certain (cognitive) problem that we wish to solve, eliminate, understand better etc. Here the term "problem" has two closely related meanings: (i) *problem* $=_{df}$ *a difficulty that cannot be overcome immediately*; and ii) *problem* $=_{df}$ *a question for which no answer is immediately available*. In science, we assume that at least some kinds of scientific (research) problems can (at least partially) be solved. The problems motivate the research and, ideally, the research leads to them being solved.

In the process of scientific inquiry, we have to be as exact as possible when specifying the problem we are interested in. Exactness and unambiguity can be achieved – at least to some extent – by using suitable conceptual methods such as those of definition or explication. It is crucial that we clearly define the key terms relating to the area we are working in. To succeed in the later stages of research, we must be sure to modify the language so as to eliminate or at least reduce polysemy and vagueness.

Moreover, it may sometimes be useful to formulate secondary (partial) problems whose resolving can help us find the solution to the main problem.

However, before we can state the problem we have to select the data and facts we consider relevant. The research problem may have been motivated by the fact that the information we have is incomplete, in which case, our goal will be to obtain a complete solution. Alternatively, the problem could have been motivated

by the existence of inconsistent information. Then, the goal would be to identify which parts of that information were true etc.

Finally, when formulating problems, existing theories and previous research findings are of unquestionable importance. Therefore, the preparatory stage of research is always about gaining a full picture, the "state of the art", on the problem or the partial (secondary) problems, which if solved, may shed light on the main problem.

(2) Proposing a hypothesis as the potential solution to the problem

As in the previous stage, the goal here is to be as precise as possible when creating the hypothesis. In scientific research, a hypothesis is a statement with an unknown truth-value which we wish to discover (using certain methods) or at least approximate. We therefore propose hypotheses in the form of conjectures that represent a (preliminary) answer to the problem. In explanatory research, hypotheses also have the potential to explain why a phenomenon occurred.

Usually, several alternative hypotheses $H_1, \ldots, H_n$ can be considered in relation to the research problem. These may be (i) mutually compatible (when confirmed or falsified); (ii) compatible in part; (iii) mutually incompatible (at most a single one may be true or confirmed).

(3) Deriving testable implications from the hypothesis

We have already seen that this is the stage where we describe the circumstances we would expect to observe if the hypothesis is true. If the hypothesis contains theoretical terms – terms referring to the properties or relations that are only indirectly observable – we can use operational definitions when formulating the implications.

Returning to the schemes of confirmation and disconfirmation (see above), we see that the first premise is of the form:

If H and $\mathcal{A}$, then E.

Variable "*H*" represents the main hypothesis (from stage 2), while variable "*A*" represents the auxiliary assumptions, and variable "*E*" the test implication of hypothesis *H*. In general, test implication *E* needn't always express a simple statement as it did in Durkheim's research. In place of *E*, we could substitute a conditional statement of the form "If observable (or empirically identifiable) conditions *C* occur, then observable conditions *O* occur".

Moreover, at this stage of the research, a single hypothesis that has multiple observable implications can prove useful. This is because the more different kinds of implications we consider, the greater the possibility of refuting the hypothesis and identifying the errors in our thinking.

(4) Designing the methods of data collection and analysis

At this stage, we need to obtain the kind of data that will help us determine whether the testable implication is true or false. The type of design used to collect and analyze the data depends on the nature of the testable implications. In other words, in the preceding stage we partially establish the kind of data we require if we are to probe the test implication of the hypothesis. In this step, we have to select the methods that will best enable us to collect suitable data. As we have seen, Durkheim used data that were already available. In general, at this stage, we employ all the relevant empirical practical methods that will enable us to obtain suitable data for testing: (simple) observation, participant observation, various measurement methods and techniques, experimentation, collection of data via surveys or unstructured and semi-structured interviews, empirical data comparison, analysis of historical documents, methods of descriptive and inferential statistics etc.

(5) Testing

Having established which methods of data collection and analysis we are going to use, we may turn to the data collection, coding (recording) and analysis. We therefore organize the data into a predetermined structure so we can assess whether they are compatible with the implication of the hypothesis being tested.

Naturally, at this stage we may use some of the conceptual designed to eliminate ambiguity (defining, explication) or to systematize (classify) data.

When testing hypotheses, we also have to be careful to control the test conditions, making sure it goes to plan. Ideally, the tests are repeated in varying conditions to eliminate or reduce the influence of unexpected factors.

(6) Evaluating (interpreting) the results of the testing

The final phase gives us an opportunity to compare the results obtained during testing with what we derived from the hypothesis (its implications). We want to determine whether the results *confirm* or *disconfirm* the hypothesis. Again, at this stage, we use some of the conceptual methods – especially those related to inference – to do this. If we state that the hypothesis has been confirmed or disconfirmed, this signals that we have used either non-deductive inference (where it is confirmed) or deductive inference (where it is disconfirmed).

Finally, at this stage, we return to the original research problem (question) and state whether it has been solved – completely or partially. In turn, this may lead us to consider new questions that could serve as the basis for further research.

The six steps discussed above represent the core of the basic H-D research model. Although this is only one of several types of research, it contains all the essential components typically found in the other types of research. This is because all research deals with research questions, hypotheses, auxiliary assumptions and methods of data collection or analysis. Moreover, most research presupposes the use of some model for testing and evaluating the hypotheses. An approach that involved methods that did not contain some of these components could hardly be called scientific research.

Study questions

1. Briefly characterize empirical scientific research.

2. Briefly explain the differences between basic and applied research.

3. Briefly explain the differences between qualitative and quantitative research.
4. List the basic stages of hypothetico-deductive research.
5. Provide an example of a research problem and propose a hypothesis that could represent the solution to it. The example should be from your area of study (or a similar one).
6. Characterize the concept of a *problem*.
7. What (empirical) methods could be suggested in stage 4, "Designing the methods of data collection and analysis" (in H-D research)?

4 HYPOTHESES AND EMPIRICAL EVIDENCE

In the previous chapter, we have noted that scientific research is generally about solving a particular theoretical or practical cognitive problem. The solution is usually a hypothesis that provides a potential answer to the research (i.e., epistemic) question. Under what circumstances can this potential answer be considered adequate?

In this chapter, we will look at some of the fundamental aspects of testing and evaluating hypotheses. We shall introduce the concepts commonly used to describe this process. The most common are those of *hypothesis*, *data* and *evidence*, but also other notions concerning the possible relations between hypotheses and evidence: *verification*, *falsification*, *confirmation* and *disconfirmation*. We will explore the three main approaches to using evidence to confirm or disconfirm a hypothesis: the *instantial model*, the *hypothetico-deductive model* and the *Bayesian model*.

4.1 Data, evidence and the logical form of hypotheses

An inherent part of the research process is the phase in which the proposed hypothesis (or multiple hypotheses) is submitted to a serious test. The goal in testing is to ascertain whether the hypothesis corresponds to the data. Therefore, it is always assumed that there is data of some kind to test. Moreover, the data must be relevant to the hypothesis being tested. For example, the data a political scientist uses when formulating or testing hypotheses could take the form of statements by politicians and public officials. Similarly, a microbiologist uses a microscope

to obtain data on the various kinds of cells and their structure. The main data sources used by a historian are various types of historical documents or other material artifacts produced by individuals or human societies. In psychological research, the relevant data may be the records (descriptions, photographs, audio and video recordings) of the behavior of research participants or their statements. In all the scientific disciplines, we work with empirical data resulting from a particular empirical (natural or social) process.

By "data" we shall understand the *information* or *information indicators* that can be used in the research process to formulate or test hypotheses. That data could be *numbers* (or numerals representing numbers or the values of certain magnitudes), *sentences*, *responses (from respondents)*, the *values of a certain variable (as captured in our measurements)*, *records produced during an observation or an experiment*, *documents*, *images*, *charts*, *statements*, *films*, *photographs* etc.

Usually, the researcher has a large quantity of diverse data at their disposal that has to be sorted using certain criteria – for example, based on the *validity* of the data (i.e. relevance to the research problem) or the *reliability* of the data (i.e. whether they were reliably obtained).[36] The data obtained in an empirical process that have not yet been analyzed are called *raw* data, while data that have been sorted through analysis are called *analyzed* data (see e.g. Kitchener 1999, 214).

The information provided by the analyzed data can be expressed in the form of statements. If the information in a statement relates to a particular hypothesis or to phenomenon of interest, that statement is called an *evidential statement* or more simply *evidence*. If the data *support* (verify, confirm) hypothesis H "in some sense" (to be specified below), we say that they constitute *positive evidence*

[36] On validity and reliability, see e.g. Gott – Duggan (2003, 7ff). In a nutshell, validity concerns the question whether data (or their analysis and representation) correspond (i.e., are relevant) to a research question or hypothesis. Reliability of data derives from the reliability of methods or instruments that produced them. A method is reliable if and only if it produces either the same data or data within a specified interval of values. For instance, when measuring distance (e.g., in cm) and using a measuring device (such as a rod) properly, the data produced by are both valid and reliable.

E for H. Conversely, if the data in some sense *refute* hypothesis *H* (i.e. falsify, disconfirm it), we say that they constitute *negative evidence E for H*.

If the (analyzed) data are *irrelevant* (to hypothesis *H*), insufficient in quantity (due to the nature of the hypothesis) or quality (degree of validity or reliability), we say that the data *constitute neither positive nor negative evidence* for *H*.

Empirical data can also underpin our thinking about the existence of or characteristic features of natural or social phenomena. By phenomenon, we mean a *theoretically postulated entity*, whose relation to data and scientific theories can be characterized thus:

> "Data, which play the role of evidence for the existence of phenomena, for the most part can be straightforwardly observed. However, data typically cannot be predicted or systematically explained by theory. Phenomena are detected through the use of data, but in most cases are not observable in any interesting sense of that term. Examples of data include bubble chamber photographs, patterns of discharge in electronic particle detectors and records of reaction times and error rates in various psychological experiments. Examples of phenomena, for which the above data might provide evidence, include weak natural currents, the decay of the proton, and chunking and recency effects in human memory." (Bogen – Woodward 1988, 305–306)

Thus, in contemporary methodology, a *phenomenon* does *not* mean something that can be directly observed using our senses. A phenomenon is an *entity* whose *existence and properties* we only consider *on the basis of data* and other available theoretical assumptions. For example, political scientists might examine the phenomenon of the credibility of politician *X* or political party *P*. The credibility of *X* or *P* is not something that can be directly observed. However, if the political scientist acquires (e.g. from a sociologist) poll data that show high support for *X* or *P* among all those polled, that data can then be used as the *indicators* of the degree of credibility of *X* or *P*. The credibility of *X* or *P* is thus a *phenomenon* that could be the subject of further theoretical research.

Typically, the objects of scientific research are phenomena, states of affairs, events or processes and their properties. We express our beliefs about the nature of these entities using hypotheses. A *hypothesis* is a *statement* (or a proposition, i.e. the meaning of a statement) with an *unknown truth-value* that we are trying to *establish* or at least *approximate* using certain *evidence*. Scientific hypotheses are always part of the broader research process: they represent a potential solution – complete or partial – to the research problem or are designed to explain a certain phenomenon. A *prediction*, i.e. a statement saying what should occur if some antecedent conditions are satisfied, can also be a hypothesis.

There are three contexts in which hypotheses are used: the *discovery* (or *formulation*) of a hypothesis, the *justification* (i.e. of testing and evaluation) of a hypothesis and the *application* of a hypothesis:

- The *context of discovery* includes all the psychological, social and theoretical circumstances that lead a scientist (or research team) to discover a potential solution to the problem that is subsequently expressed in the form of a scientific hypothesis.

- The *context of justification* is the systematically prepared conditions under which this hypothesis is tested and positive or negative evidence obtained.

- Lastly, the *context of application* refers to the conditions in which the *successful* (confirmed) hypotheses are used in processes that lead to related theoretical or practical problems being solved. Generally, this could be the use of theoretical knowledge expressed as a hypothesis in designing technologies etc.

In this chapter, we do not go any further into the contexts of discovery, justification and application; suffice it so say that all hypotheses, as formulated in the context of discovery, must conform to certain standard requirements. Some of these are (cf. Cohen – Nagel 1934, 207–215):

- logical consistency

- testability (in principle)
- simplicity

Hypotheses that satisfy these criteria can be subjected to empirical tests. Scientific hypotheses may express various kinds of potential facts. For example, some hypotheses suppose that certain objects have such and such a property. Others state that there is a correlation between two (or more) variables (properties) with such and such a coefficient. Hypotheses may also postulate a causal relation between events of a certain kind etc. Let us look at some examples of the various types of hypotheses:

1. There exists a planet, Vulcan, whose orbit lies between the Sun and the orbit of the planet Mercury.
2. No two democratic states are at war with each other.
3. Napoleon died of arsenic poisoning.[37]
4. Regular smoking increases the risk of lung cancer.
5. There is water on the surface of some planets outside the Solar System.
6. For each light-ray incident on the interface of two optical environments (with different optical properties), it holds that the angle of refraction β (subtended between the incident ray and the normal to the interface) is equal to the angle of incidence α.

These hypotheses differ not only in content, but also in logical form.

Since any given evidential relation between a hypothesis and empirical evidence depends not only (i) on the state of affairs in the world but also (ii) on the logical form of the hypothesis and the evidential statements, we will distinguish between four basic categories of hypotheses in terms of their logical form.

[37] See e.g. Broad's article (2008) in the *New York Times*.

(1) Singular hypotheses

Singular hypotheses attribute a certain property or relation to one or more objects identified by name or description. For example, statement 3 is a singular hypothesis. The logical form of this type of hypothesis can be generally expressed using one of the following schemes:

- $F(a)$
- $F(a) \wedge F(b)$
- $R(a, b)$
- …

The first scheme states that object a has property F. The second scheme states that object a has two properties, F and G. The last scheme states that object a is in relation R to another object b. Naturally, singular hypotheses can get more complicated. What they all have in common, though, is that they attribute at least one property to at least one object (named or described) or they ascribe at least one relation to at least one tuple (triple, …, n-tuple) of particular objects. Moreover, singular hypotheses may contain also a negation; in general, they may assume that it is *not* the case that such-and-such an object has such-and-such a property.

(2) Existential hypotheses

Existential hypotheses attribute at least one property to an object that is not further specified (or named) or express a relation between at least one tuple (triple, …, n-tuple) of objects that are not further specified. Statement 5 above is an example of an existential hypothesis, since it states that there exists at least one planet x outside our Solar System that has water on its surface. We can capture the logical form of existential hypotheses using the following schemes:

- $(\exists x)\, F(x)$
- $(\exists x)\, (F(x) \wedge G(x))$

- $(\exists x \exists y)\, R\,(x, y)$
- …

The first scheme states that there exists an object x which has property F. The second scheme states that there exists an object x which has the properties F and G. Finally, the third scheme can be read as stating that there exist objects x and y such that x is in relation R to y. The difference between singular and existential hypotheses is that while singular hypotheses name objects to which certain properties are ascribed (or denied), existential hypotheses do not directly name such objects – they simply assume such objects exist.

Similarly to singular hypotheses, existential hypotheses can also include negation. However, it is important to note to which part of the statement the negation relates. If, in a hypothesis with existential quantifier ($(\exists x)$), negation ($\neg$) appears after the quantifier, the hypothesis is an existential hypothesis. This is the case in the following schemes:

- $(\exists x)\, \neg F\,(x)$
- $(\exists x)\,(F\,(x) \wedge \neg G\,(x))$
- $(\exists x \exists y)\, \neg R\,(x, y)$
- …

If, however, the negation appears before the existential quantifier, the hypothesis is a universal, not an existential, hypothesis. Consider the following schemes:

- $\neg\,(\exists x)\, F\,(x)$
- $\neg\,(\exists x)\,(F\,(x) \wedge G\,(x))$
- …

The first scheme expresses the fact that there is no such object x that has property F. This is the equivalent of saying that for every object x it holds that it *is not* F. The second scheme expresses the fact that there is no such object x that has both properties F and G. In other words, for each object x it holds that *it is either not F or not G*. Hence, these hypotheses are not existential but universal, despite containing the existential quantifier.

(3) Universal hypotheses

The structure of universal hypotheses can get fairly complicated, but what they all have in common is that they ascribe a certain property to *all* objects (that satisfy a certain other condition or have a certain characteristic property). For example, statement 6 is a universal hypothesis. To every light-ray satisfying a certain condition (i.e. that it is an incident on the interface of two optical environments) it ascribes a certain property (i.e. that the angle of the ray's refraction equals the angle of incidence). The following logical form represents the simplest case of a universal hypothesis:

$$(\forall x)\,[F\,(x) \rightarrow G\,(x)]$$

We read this scheme as "For all (objects) x it holds that if x has property F, then x has property G", or simply "All Fs *are* Gs".

Universal hypotheses can also contain negation. As with existential hypotheses, it is important to note the part of the statement in which the negation appears. Cases where the negation occurs after the universal quantifier are universal hypotheses. For example, the schemes

- $(\forall x)\,[\neg F\,(x) \rightarrow G\,(x)]$
- $(\forall x)\,[F\,(x) \rightarrow \neg G\,(x)]$
- $(\forall x)\,[\neg F\,(x) \rightarrow \neg G\,(x)]$

are all cases of universal hypotheses. The first scheme expresses the fact that for each object x it holds that if x *is not* F, then x *is* G. The second scheme states that

for each object x it holds that if *x is F*, then *x is not G*. The third scheme expresses the fact that every object x that *is not F is also not G*.

If, however, the negation is inserted before the universal quantifier, we get an existential statement. For example, the scheme

$$\neg\,(\forall x)\,[F\,(x) \rightarrow G\,(x)]$$

can be read as "It is not true that each object x that is F is also G", which can be equivalently expressed as "There exists at least one object x that is F and is not G". Thus, the negation of an entire universal statement expresses an existential statement. (Conversely, as we have seen above, the negation of an entire existential statement expresses a universal statement.)

(4) Statistical/probabilistic hypotheses

Generally, statistical or probabilistic hypotheses express the probability of a certain phenomenon occurring. Alternatively, they may express the probability of a phenomenon occurring assuming that some other phenomenon has already occurred. For example, "The probability that it will rain in Bratislava tomorrow is 75%" or "The probability that it will rain in Bratislava tomorrow, assuming that the temperature today was −10 °C, is low (e.g. less than 5%)". In the first case, the probability concerns a single phenomenon. This is known as *unconditional probability*. The second case is an example of *conditional probability*: the fact that it was −10 °C on one day may affect the probability of it raining the next day. Since a probability expressed as a percentage can be replaced with real number r from the interval $[0, 1]$, where 0 represents 0%, 1 represents 100%, and the numbers between these extremes represent percentages between 0 and 100%, statistical hypotheses can be expressed in the following basic form. A statistical hypothesis states that the probability of a certain object having a certain property G, assuming the object belongs to reference class F, is equal to real number $r \in [0, 1]$. Schemes of statistical hypotheses therefore have the following form:

n% of objects which are F are G

or

$$p(G \mid F) = r, \text{ where } r \in [0, 1].$$

If we look at the examples of hypotheses listed above, we can note that statement 4 is a statistical hypothesis, despite it not explicitly mentioning percentages or probability. It could be rephrased thus: "The probability that a person will contract lung cancer, assuming they smoke, is high." We have not explicitly expressed how high that probability is, but we can assume, for example, that it is higher than 0.75 (or 75%). A limiting case would be a hypothesis stating that the probability of a phenomenon (or property) occurring, assuming that some other phenomenon (or property) occurred, is equal to 1 (or 100%). In this case, the statistical hypothesis would be equivalent to a universal hypothesis.

We may also encounter hypotheses where it is not clear whether they are existential, universal or statistical (probabilistic) hypotheses. In some cases, the *logical* interpretation of a particular hypothesis may be problematic – for example, if the quantification in the statement is *implicit*, i.e. the sentence does not contain a quantifier ("some", "all", "none" etc.). In most cases, however, we are able to decide (given the research context and the research problem) which (logical) type of hypothesis it is. To avoid misunderstandings, it is always advisable to be as precise as possible when expressing the hypothesis, and to include explicit quantification. Where singular hypotheses are concerned, the object to which the property is ascribed should be given an unambiguous name or description.

In practice, hypotheses are often complex combinations of the basic types of hypothesis. Below, we shall limit ourselves to the testing and evaluation of singular, existential and universal hypotheses.

Before moving on, let us first go back to the concept of *empirical evidence*. We have noted that certain kinds of *data* can *support* or *confirm* a hypothesis. It would be more precise to say that the data relevant to the testing of the hypothesis are typically expressed using a singular (or, in some cases, existential or statistical) statement (or combination thereof). A statement of this kind is called an *evidential statement*.

Thus, in testing hypotheses, we compare *hypotheses* and *evidential statements*, assuming that *evidential statements* are *true*, that they have been verified as true or that we have *accepted* them as true in the empirical situation. Evidential statements therefore express what has been observed or what has been measured in a certain environment using the necessary instruments. As such, evidential statements or the truth-value of these statements can be viewed as the result of using the given empirical method.

Of course, all observation or empirical identification relies on the system of concepts and theoretical or pre-theoretical knowledge used in observations (and sensual perceptions in general). However, our observations are not always reliable. The instruments used to obtain the data do not always work the way we expect them to. The data obtained need not always be the result of a reliable process. Sometimes an *evidential statement* that we assume is true turns out not to be true. In the philosophy of science, admitting that *errors* can occur even in relation to basic *evidential statements* is known as *fallibilism*. Fallibilism involves recognizing that we may be wrong even when we have no reason to doubt a particular statement. A new evidential statement that is accepted as true may cast doubt on another evidential statement that had previously been accepted as true.

However, all scientific reasoning and all hypothesis testing relies on the basis of evidential statements which we either have no reason to doubt at that moment or their uncertainty is within an accepted interval. In other words, unless there is specific evidence indicating that we are wrong to consider an evidential statement true, we treat that statement as if it were true.

In what follows, we will therefore assume that we have access to evidential statements. We will be concerned with the kinds of possible relations between evidential statements and hypotheses, assuming that the evidential statements are true and the hypotheses have one or other logical form.

4.2 Verification, falsification and models of confirmation

Our discussion of the evidential relations between hypotheses and evidential statements is based on some of the key literature on *confirmation theory*, especially Carnap (1962) and Hempel (1945), but also Crupi (2015), Earman (1992), Earman – Salmon (1999), Hájek – Joyce (2008) and Schurz (2014).

We have noted that in *hypothesis testing* the aim is to compare the hypothesis with evidential statements describing the relevant data obtained by means of observation, measurement or experimentation, or another specific empirical method (survey, interview, content analysis etc.). We have also stated that the *evidential statement* used to test the hypothesis is *taken to be true* unless there is independent conflicting evidence that casts doubt on its truth. If that assumption is not satisfied, there is no point comparing the statement with the hypothesis.

When testing hypotheses, we normally end up with one of the following relations. Hypothesis H may be

1. *verified* as true
2. *falsified* as untrue
3. *confirmed*
4. *disconfirmed*

on the basis of the *empirical evidence* expressed by evidential statement E.

Whether the evidence verifies/confirms or falsifies/disconfirms the hypothesis depends on two factors: (i) the state of affairs, i.e. what the reality is and, therefore, whether the evidential statement is true or false; and (ii) the logical form of the hypothesis and the evidential statement. In the following subsections, we will look at the situations in which hypothesis H is verified, confirmed, falsified or disconfirmed by evidential statement E.

4.2.1 Verifiable and verified hypotheses

Let us begin by making an important distinction. There is a fundamental distinction between cases in which the hypothesis is *verifiable* and cases in which it is also *verified*.

Hypothesis H is said to be *verifiable* if it is logically conceivable (admissible) that there exists an evidential statement E which – if true – would verify that H was *true*. To be more precise, hypothesis H is verifiable by evidential statement E if and only if H is logically entailed by E (symbolically, $E \vDash H$), where E expresses an observable state of affairs.[38] Hypothesis H is therefore verifiable if and only if it can (in principle) be verified (as true) by evidential statement E.

However, not all *verifiable* hypotheses are in fact *verified*. For hypothesis H to be verified, another condition must be satisfied. Hypothesis H is *verified* by evidential statement E (as true) if and only if H is logically entailed by E and E *is true*. In other words, hypothesis H is verified by evidential statement E as true if and only if E being true guarantees that H is also true and we know (or have good reason to believe) that E is a true statement.

To illustrate this, let us suppose that we want to test hypothesis H_1:

(H_1) Some people have lived to be over 120 years old.

We could transform our hypothesis into the following logical form:

(H_1*) $(\exists x)\,(H\,(x) \wedge L_{\geq}\,(x, k))$

Predicate symbol "H" represents the property "…*is human*", symbol "$L_{\geq}$" represents the relation "… *lived to be over*…" and individual constant "k" represents 120 years. We read the formal notation H_1* as "There is at least one x such that x is human and x lived to be 120 years old".

Using data from Wikipedia's "Oldest people" entry, which relies on data from reliable sources such as the Gerontology Research Group and the Guinness World Records, we can state the following evidential statement, E_1:

[38] Recall the notion of logical entailment defined in Chapter 2: Statement B is logically entailed by statement A if and only if B is true whenever A is true. In other words, A logically entails B if and only if whenever A is true, it is not possible for B to be false.

(E_1) Jeanne Calment (from France) lived to be 122 (and 164 days) years old.[39]

If we replace the name Jeanne with the individual constant "a", the logical form of statement E_1 can be expressed as follows (some modifications notwithstanding):

(E_1*) $H(a) \wedge L_{\geq}(a, k)$

This is equivalent to "Human a lived to be over 120 years old".

Assuming that E_1 is true, we can see that its being true also guarantees the truth of the hypothesis H_1. In other words, whenever E_1 is true, H_1 must also be true. We can express this in the statement that E_1 logically entails H_1.

Not all hypotheses are *verifiable*, i.e. it is not true that for every hypothesis there is a conceivable evidential statement that would entail the hypothesis. For example, the hypothesis

(H_2) All Comenius University graduates are able to find a job within six months of graduating.

is not verifiable by any finite conjunction of singular evidential statements. Even assuming that we have an evidential statement that states that up to now number n of Comenius University graduates have all been able to find a job within 6 months of graduation, the truth of this statement *cannot guarantee* that hypothesis H_2 is true. We can generalize this and state that only three kinds of hypotheses are verifiable: *singular* hypotheses, *existential* hypotheses and those *universal* hypotheses that concern a finite number of objects (of some kind) which can all (in principle) be identified. Neither *universal hypotheses with an unlimited universe of discourse* nor *statistical (probabilistic) hypotheses* can be verified.

4.2.2 The falsification of hypotheses

Another possible relation between a hypothesis and an evidential statement is that of falsification. Similarly to the previous case, it is important to distinguish

[39] See `https://en.wikipedia.org/wiki/Oldest_people`.

between the concept of a *falsifiable* hypothesis and the concept of a *falsified* hypothesis.

When can a hypothesis be said to be *falsifiable*? Hypothesis H is falsifiable if evidential statement E is logically conceivable and would *logically contradict* the hypothesis. Again, there are hypotheses that are *falsifiable* but that are not *falsified*. Hypothesis H is only *falsified* if and only if it is *falsifiable* (i.e. an evidential statement E that logically contradicts it is conceivable) and we know that statement E is true.

For example, given a universal hypothesis of the form

$$(H) \qquad (\forall x)[F(x) \rightarrow G(x)]$$

and an evidential statement

$$(E) \qquad F(a) \wedge \neg G(a)$$

that we know is true, we may state that E logically contradicts H and, therefore, E *falsifies* H as (conclusively) false. If a hypothesis states that all Fs are Gs and we discover an object which is F but is not G, that discovery falsifies the hypothesis.

We can be more precise in defining both falsifiability and falsification. H is *falsifiable* by evidential statement E if and only if E *can be true* and E *logically entails* $\neg H$. Hypothesis H is falsified by evidential statement E if and only if E is a *true statement* and E *logically entails* $\neg H$.

For example, if hypothesis H_2 ascribes the property of "*having found a job within 6 months of graduation*" to all Comenius University graduates, we just need to find one graduate who (despite all efforts) failed to find a job within 6 months to have evidence that falsifies our hypothesis.

However, there are certain reasons why we may not always be able to determine whether a hypothesis has been falsified even if we have negative evidence disputing the hypothesis. We therefore need to think about whether there is a weaker relation between the evidence and the hypothesis, and we look at this below as the relation of disconfirmation. Only three kinds of hypotheses are falsifiable: *singular* hypotheses, *universal* hypotheses and those *existential* hypotheses which concern a finite universe of objects. *Existential hypotheses with an unlimited universe*

of discourse and *statistical (probabilistic) hypotheses* cannot by their very nature be falsified.

4.2.3 Models of the confirmation and disconfirmation of hypotheses

Every evidential statement E that verifies hypothesis H (as true) also confirms it. However, the relation of *confirmation* between evidence E and hypothesis H is weaker than the relation of *verification*. In most cases where evidential statement E *confirms* hypothesis H, the fact that E is true *does not guarantee* that H is true; nonetheless, E provides *positive support* for H. The various theories of confirmation are attempts to establish what "positive support" means exactly.

For the sake of simplicity, we will not concern ourselves here with the difference between confirmable and confirmed hypotheses (nor disconfirmable and disconfirmed hypotheses), but readers can easily work out the difference based on our remarks above.

There has been much discussion on the theory of confirmation and on defining the relations of non-deductive support (or countersupport) between hypothesis H and evidence E (see Crupi 2015).

The main theories of confirmation are the *instantial model of confirmation*, the *hypothetico-deductive model of confirmation* and the *Bayesian model of confirmation* (see Crupi 2015; Earman – Salmon (1999); Norton 2005).

The instantial model of confirmation

The starting point of this model is a universal hypothesis, i.e. a hypothesis of the form

$$(H) \qquad (\forall x)[F(x) \rightarrow G(x)]$$

According to this theory, universal hypothesis H is confirmed by its positive instances, i.e. instances where we have identified objects that have both property F and property G.

Evidence of this type can be expressed using evidential statements of the form:

$$(E) \qquad F(a) \wedge G(a); F(b) \wedge G(b); \ldots; F(n) \wedge G(n)$$

This form of statement states that object *a*, object *b*, …, object *n* have both properties *F* and *G*.

Assuming that the evidential statements of the form *E* are true, this model of confirmation allows us to state that *E confirms H*. Consider hypothesis H_2 (see above) on the ability of Comenius University graduates to find employment. If our evidence consisted of a record of all the graduates who had found a job within 6 months of graduating, that evidence would confirm hypothesis H_2 in the instantial model.

Of course, it only makes sense to say universal hypothesis *H* has been confirmed if there is no evidence (expressed as an evidential statement) of object *k* having property *F* (e.g. being a Comenius University graduate) but not property *G* (e.g. not finding a job within 6 months). A case where we have observed (or otherwise empirically identified) that object *k* has property *F* and *does not have* property *G* (i.e. $F(k) \wedge \neg G(k)$) would *falsify* and *disconfirm* a universal hypothesis of form *H*.

There are a number of problems with this model of confirmation. One is the *Raven paradox* (see Hempel 1945), which arises if, in addition to the instantial model (that positive cases confirm a general hypothesis), we also accept the *equivalence condition*. The equivalence condition states that *if E confirms (disconfirms) one of two logically equivalent hypotheses, it also confirms (disconfirms) the other*. This means that if two hypotheses say the same thing using different words, they must be evaluated identically in relation to the evidence. Let's take the hypothesis "All ravens are black" as our example. We can formulate another hypothesis that is equivalent to the first hypothesis – "No non-black objects are ravens". Suppose we now formulate a true evidential statement describing my white socks. This statement is an instance that confirms the hypothesis that "No non-black objects are ravens". This hypothesis is equivalent to the hypothesis "All ravens are black", and so according to the condition of equivalence my white socks (more precisely, the statement describing them) are an instance confirming the hypothesis about the color of all ravens. We won't go into the various possible solutions to this

paradox, but instead refer readers to the rich literature on this problem (see also Hempel 1945; as well as Crupi 2015, Earman – Salmon 1999).

Another problem with this model is its *limited* applicability. Instantial confirmation can only be considered in relation to hypotheses that denote (directly) *observable properties*. If the property F that appears in universal hypotheses of the form H is *not directly observable*, then it is not possible to think about a situation in which we could state that we *had observed* a case in which object n had property F and property G. Put simply, we would not be able to observe n having property F.

Despite these difficulties, many scientists use the instantial model, and if they talk of having confirmed empirical hypotheses and refer to observable properties (relations), they may well be using this model of confirmation.

The hypothetico-deductive model of confirmation

The hypothetico-deductive model of confirmation is also used in relation to *universal hypotheses*. In contrast to the instantial model, the hypotheses tested using this model may also refer to theoretical properties (or relations), i.e. entities that are not directly observable or directly empirically identifiable.

The basic idea in this model is summed up in the thesis that a *hypothesis* is *confirmed* by its *successful* (i.e. *true*) predictions. In this case, *predictions* are none other than the test implications derived from the given hypothesis (and other, auxiliary statements). We encountered the hypothetico-deductive model of hypothesis testing in the previous chapter.

In its basic form, a hypothetico-deductive confirmation of a hypothesis can be formulated using the following scheme:

> If hypothesis H is true, then prediction E is also true.
> Prediction E is true.
> ———————————————
> Hypothesis H is (probably) true.

We have already noted that we can usually only derive test implications (E) from hypothesis (H) if we use other (so-called) auxiliary hypotheses (generally denoted

as "$\mathcal{A}$"). These hypotheses may concern (a) the test conditions or the collection of empirical data; or (b) theoretical knowledge that is not being tested directly but that is being indirectly relied on during testing. The resulting scheme has the following form:

$$\frac{(H \wedge \mathcal{A}) \rightarrow E \qquad E}{(H \wedge \mathcal{A})}$$

This inference scheme is non-deductive: the truth of the premises does not guarantee the truth of the conclusion. Therefore, even though we may be able to derive prediction E from a hypothesis (and auxiliary statements) and E is eventually shown to be *true*, E being true *does not exclude the possibility* that hypothesis H is not in fact true. Prediction E being true does, however, lend some kind of (non-deductive) support to hypothesis H.

We can illustrate this using a historical example (see Giere et al. 2006, 58–63). Galileo Galilei attempted to put forward an argument in support of Copernicus' *heliocentric theory* that – in short – assumed the Earth orbits around the Sun which remains stationary. From this theory and a set of auxiliary statements, one could derive the prediction that one of the planets in the Solar System, Venus, exhibits observable phases that are similar to those exhibited by the Moon. What was particularly interesting was the consequence that, during a certain interval of its orbit around the Sun, Venus could be observed from Earth fully lit up. By contrast, Ptolemy's geocentric theory holds that Venus cannot be observed from Earth fully lit up because (according to an assumption in geocentric theory) it orbits the Earth in epicycles, and the Sun, orbiting Earth, lies beyond Venus. In which case, if geocentric theory were true, the Sun would light Venus from behind making Venus appear mostly dark or only partly lit up when seen from Earth. Unfortunately, when looking at the night sky through the naked eye, we can only see Venus as a tiny sparkling dot and so cannot discern whether it indeed exhibits all of the phases that the Moon does. In other words, none of these predictions could be tested using our unaided eyes. Galileo, however, constructed

a telescope and was able to test whether Venus is ever fully lit up. Galileo's observations of the phases of Venus, including the full phase, were in keeping with heliocentric theory (and, conversely, disconfirmed the prediction of Ptolemaic theory), and so his observation data were positive evidence. In other words, they hypothetico-deductively confirmed the heliocentric hypothesis. His argument and testing can be schematically expressed using the following inference:

> *If* heliocentric theory is true and Venus' orbit lies between the Earth's orbit and the Sun, and *if* the data gathered from the telescope are reliable, *then* Venus' disk will be fully lit up by the Sun during one of the phases of the planet's orbit.
> (We have observed that) Venus' disk was fully lit up by the Sun during one of the phases of the planet's orbit.
>
> ---
>
> Heliocentric theory is (probably) true.

Although this argument is not deductively valid, Galileo accepted it as a case of a successful prediction derived from the theory he tested.

This model, too, has several problems. One of them is the *problem of the underdetermination of hypotheses (theories) by evidence*. It arises because a single prediction E can be logically derived not just from a single hypothesis H, but from an infinite number of other hypotheses that are incompatible with H. This leads theoretically to a situation in which no finite amount of evidence E is sufficient for us to be able to select one of the multiple (competing) hypotheses because the evidence can be deductively derived from all of them. In practice, this problem can be partially solved (i.e. the multiplicity of competing hypotheses can be reduced) by deriving multiple, different predictions from the hypotheses, and then attempting to eliminate the hypotheses whose predictions are shown to be false.

Nonetheless, the hypothetico-deductive model is suitable for testing universal hypotheses (and as an approximation of the testing of statistical hypotheses). It is not suitable for confirming singular or existential hypotheses.

We have seen that when testing universal hypotheses we frequently rely on auxiliary hypotheses denoted as "$\mathcal{A}$". Often, it is only because of these auxiliary as-

sumptions that we are able to derive certain predictions E from hypothesis H. In such cases, if prediction E is shown to be false, we cannot simply conclude that hypothesis H is also false. The correct inference, in this case, would have the form:

$$\begin{array}{l} (H \wedge \mathcal{A}) \rightarrow E \\ \neg E \\ \hline \neg\,(H \wedge \mathcal{A}) \end{array}$$

The first premise states that if hypothesis H and auxiliary statements $\mathcal{A}$ are all true, then prediction E should also be true. The second premise states that *E is not true* (i.e. $\neg E$ is true). Therefore, the conclusion of this (deductively valid) argument is that *it is not true that the conjunction of H and $\mathcal{A}$* is true. In other words, the conclusion tells us that either hypothesis H is false, or that auxiliary statements $\mathcal{A}$ are false, or that both *H and $\mathcal{A}$* are false. Neither this argument nor evidence $\neg E$ tells us which of these three possibilities obtains. This does not necessarily mean that we would not be able to obtain richer evidence through further testing that would enable us to determine if hypothesis H was wrong or if any of the auxiliary hypotheses $\mathcal{A}$ were wrong. However, merely obtaining a negative result during testing would not necessarily be enough to show that the hypothesis had been falsified. In such cases, we can say that the hypotheses have been *disconfirmed* but we cannot say that they have been conclusively falsified.

The Bayesian model of confirmation

These two models of confirmation – the instantial model and the hypothetico-deductive model – are *qualitative*. The aim is to determine whether an evidential statement confirms or disconfirms the hypothesis being tested (or is neutral relative to it). Bayesianism is a *quantitative* approach to confirmation. It begins with the assumption that hypothesis H, evidential statement E and the compound statements involving both H and E can all be ascribed (given background knowledge K) a real number r from the interval $[0, 1]$ that expresses their respective

probabilities. Bayesianism comes in various forms, but they all share certain core ideas.

Bayesian theory uses the mathematical theory of probability, and above all one of its theorems, *the Bayes' theorem*, to model the evidential relations between hypothesis H and evidence E with respect to background knowledge K. More precisely, the relations of confirmation, disconfirmation and irrelevance (independence) between H and E are modeled using a probability function P (or a set P of such functions). This probability function assigns a real number $r \in [0, 1]$ to each of the elements in set S consisting of statements or propositions A, B, C, …, in line with the axioms and theorems of probability theory. Set S is closed under the operations of *negation* and (countable) *disjunction*. That is, if A and B are statements or propositions of S, then $\neg A$ (and $\neg B$) and $A \vee B$ are also statements or propositions of S. The probability function P must satisfy the following four axioms (see Hájek – Joyce 2008, 118):

(A1) The probability of any statement (proposition) $A \in S$ is a real number r from the interval $[0, 1]$:

$$0 \leq P(A) \leq 1$$

(A2) If A is a tautology (a logically/necessarily true statement), then $P(A) = 1$.

(A3) If A and B are mutually incompatible statements (i.e. if it is true that $\neg (A \wedge B)$), then:

$$P(A \vee B) = P(A) + P(B).$$

(A4) If $P(B) > 0$ then:

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

While axioms A1) – A3) are concerned with unconditional probabilities, axiom A4) is concerned with conditional probability. It states that the probability of statement A, given that B is true, is equal to the probability of conjunction $A \wedge B$ divided by the probability of statement B (where $P(B) > 0$).

Bayesian agents who use probability function P in accordance with the axioms and theorems of the theory of probability can put all the information from their knowledge base K into the function and use it to model their *degree of rational belief* in the hypothesis (or hypotheses) being tested or in the evidence.

The Bayesian theory of (dis)confirmation is essentially about comparing two probabilities. On the one hand, it is concerned with the probability $P(H)$ we assign to hypothesis H based on our existing knowledge and beliefs (from set K), before we have even begun to consider the influence evidence E has on the hypothesis. On the other hand, it is concerned with probability $P(H \mid E)$, which expresses the probability of hypothesis H, assuming that evidential statement E is true. Probability $P(H)$ is also known as *prior* probability, while $P(H \mid E)$ is known as *posterior* or *conditional* probability. The Bayesian definition of the concept of *incremental* confirmation and its complementary concepts are based on comparing these two probabilities (see Howson – Urbach 2006, 91–93):

DEFINITION OF CONFIRMATION
Evidence E confirms hypothesis H if and only if

$$P(H \mid E) > P(H).$$

DEFINITION OF DISCONFIRMATION
Evidence E disconfirms hypothesis H if and only if

$$P(H \mid E) < P(H).$$

DEFINITION OF NEUTRAL EVIDENCE
Evidence E is neutral (irrelevant) to hypothesis H if and only if

$$P(H \mid E) = P(H).$$

In order to compare these two probabilities and thus decide whether the evidence confirms, disconfirms or is neutral to the hypothesis, we need to know the values of these three probabilities: $P(H)$, $P(E \mid H)$ and $P(E)$.

We have denoted the first as the *prior* probability of hypothesis H. It is the probability assigned to the hypothesis (by a Bayesian agent) before the latter is

confronted with evidence E. In this sense, $P(H)$ may represent the (ideal) rational agent's *subjective* degree of belief given their (current) background knowledge K, or the degree of belief of a group of rational agents (e.g. members of a research team).

$P(E \mid H)$ expresses the probability of evidence E assuming that hypothesis H is true. It is the *likelihood* of certain predictive consequences of the hypothesis given that the hypothesis is true. There are two cases that limit the likelihood of a hypothesis being true: If evidence E is the logical consequence of H, then $P(E \mid H) = 1$. If $\neg E$ is the logical consequence of hypothesis H, then $P(E \mid H) = 0$. Most cases, however, presuppose some kind of a non-deductive relation of probabilistic dependence between hypothesis H and evidence E. If statistical hypotheses are involved, then $P(E \mid H)$ expresses a certain probabilistic model; more precisely, it is the probability that certain evidence E will be observed assuming statistical hypothesis H is true.

Finally, $P(E)$, also known as the "expectability" of evidence, expresses the probability of E being true given all that we know (see Hájek – Joyce 2008, 119). This probability can be computed using the theorem of total probability:

THEOREM OF TOTAL PROBABILITY
If $P(H_1 \vee \ldots \vee H_n) = 1$, and for each $i \neq j$ it holds that $\neg (H_i \wedge H_j)$, then:

$$P(E) = P(E \wedge H_1) + \ldots + P(E \wedge H_n)$$
$$= P(E \mid H_1) \times P(H_1) + \ldots + P(E \mid H_n) \times P(H_n)$$

The relation between these three elements – $P(H)$, $P(E \mid H)$, and $P(E)$ – and posterior probability $P(H \mid E)$ is expressed by Bayes' theorem, named after the English reverend Thomas Bayes (1701–1761), who first formulated it:

BAYES' THEOREM – VERSION I

$$P(H \mid E) = \frac{P(E \mid H) \times P(H)}{P(E)}, \text{ where } P(E) > 0$$

If we express probability $P(E)$ using the theorem of total probability, while taking only hypothesis H and $\neg H$ into account, for which it holds that $P(H \vee \neg H)$

$= 1$ (and, trivially, $\neg (H \wedge \neg H)$), then Bayes' theorem can be formulated as follows:

BAYES' THEOREM – VERSION 2

$$P(H \mid E) = \frac{P(E \mid H) \times P(H)}{[P(E \mid H) \times P(H)] + [P(E \mid \neg H) \times P(\neg H)]}$$

If we now consider the set $\{H_1, \ldots, H_n\}$ of all mutually incompatible hypotheses (in pairs), then the denominator $P(E)$ can be expressed in full, in accordance with the theorem of total probability, as:

BAYES' THEOREM – VERSION 3

$$P(H_i \mid E) = \frac{P(E \mid H_i) \times P(H_i)}{\sum_j P\left(E \mid H_j\right) \times P\left(H_j\right)}$$

Hence, Bayes' theorem is a systematic way of calculating the conditional posterior probability that expresses the probabilistic effect of the evidence on the hypothesis, and of comparing it with the prior probability of the hypothesis. If prior probability $P(H)$ (or $P(H_i)$) is lower than posterior probability $P(H \mid E)$ (or $P(H_i \mid E)$), then evidence E confirms hypothesis H in the sense of the definition of (incremental) confirmation given above. If, on the other hand, probability $P(H)$ is higher than $P(H \mid E)$, then E disconfirms H. Finally, if the values of the prior and posterior probabilities are identical, the evidence is neutral with respect to the hypothesis.

However, Bayes' theorem does not tell us how we should update our prior probabilities if we discover that evidential statement E is true. Apart from the theory of probability and a specific interpretation of the concept of probability (as degrees of rational belief), Bayesianism therefore also makes use of the rule of conditionalization (although there are alternatives to it within Bayesianism):

RULE OF CONDITIONALIZATION

If agent I at time t did not know whether E was true, and if the only thing he discovers at a later time $t + 1$ is that E is true, then the new

> unconditional probability at $t + 1$, $P_{t+1}(H)$ equals the conditional probability at time t, $P_t(H \mid E)$:
>
> $$P_{t+1}(H) = P_t(H \mid E).$$[40]

The conditionalization rule instructs Bayesian agents on how to update (modify) their prior probabilities in cases where the only change in their knowledge is the fact that E is true.

Three basic elements – probability theory, the subjectivist interpretation of probability and the conditionalization rule – make up the core structure of the Bayesian model of confirmation. We shall not remain at the level of abstract principles, but turn to an example that illustrates the basic features of this approach.

Example

Imagine we have the information that 1% of the population suffers from illness Z. To determine whether a person is suffering from Z, we run tests. A positive test result generally means the patient has Z. More precisely, the probability that the test will be positive, assuming the patient has Z, is 97% (or 0.97). There is also a 5% probability that the test will be positive despite the patient not having Z. Let's say a patient was tested and the result came back positive. What is the probability that the patient has Z if the patient tested positive?

Let "H" denote the hypothesis that the individual has Z. Similarly, let "$\neg H$" represent the hypothesis that the patient does not have Z. The statement that the (patient's) test result was positive will be denoted as "E". Conversely, "$\neg E$" will denote the statement that the test result was negative. The information from the preceding paragraph can now be expressed the following way:

$$P(H) = 0.01$$
$$P(\neg H) = 1–0.01 = 0.99$$

[40] There is also a generalization of the Rule of conditionalization: the so-called Jeffrey conditionalization. It is also related to contexts where an agent does not come to know E, but she ascribes some probability P (other than 0 or 1) to E. See, e.g., Earman (1992, 34).

$$P(E \mid H) = 0.97$$
$$P(E \mid \neg H) = 0.05$$

We need to determine the posterior probability $P(H \mid E)$. Entering the above figures into Bayes' theorem (version 2), we get:

$$P(H \mid E) = \frac{(E \mid H) \times P(H)}{[P(E \mid H) \times P(H)] + [P(E \mid \neg H) \times P(\neg H)]}$$
$$P(H \mid E) = \frac{0.97 \times 0.01}{[0.97 \times 0.01] + [0.05 \times 0.99]} = 0.169$$

We can see that the posterior probability of the patient having Z is almost 17%, assuming the test result was positive. Comparing the prior probability of hypothesis H before the effect of the test was considered (i.e. $P(H) = 0.01$) with the posterior probability $P(H \mid E)$, we may note that the latter is greater than the former, $P(H \mid E) > P(H)$. Therefore, evidence E confirms hypothesis H (the difference being about 16 percentage points). Moreover, if the patient decides to undergo one more test, the conditionalization rule would require us to use the conditional probability $P(H \mid E)$ from the first test as our new prior probability $P(H)$. Hence, the value of $P(H)$ would now be 0.17 instead of 0.01 and the value of the alternative hypothesis $P(\neg H)$ would be 0.83 instead of 0.99. (We will leave the reader to calculate posterior probability $P(H \mid E)$ after the second test, using these updated values.)

The Bayesian model of confirmation can handle many of the problems encountered in the previous approaches (for example, the Raven paradox, the problem of underdetermination). However, there are other problems associated with it. Explaining these and the Bayesianist defense against them is beyond the scope of this book. Once again, we refer the reader to the rich literature on the Bayesian theory of confirmation: see e.g. Howson – Urbach (2006); Earman (1992); Crupi (2015); and Hájek – Joyce (2008).

There are a number of theoretical and practical problems that arise when testing and evaluating whether hypotheses have been *verified*, *confirmed*, *falsified* or

disconfirmed. Nonetheless, scientists often talk about *confirmed* or *falsified* hypotheses in their work. It is therefore always advisable to ask which model of confirmation they used.

Study questions

1. What are data? Think of an example of data used in your (or a related) discipline.
2. What is the difference between data and empirical evidence? When do data become empirical evidence?
3. Characterize the concept of *phenomenon* as used in the methodology of science.
4. Characterize the concept of *hypothesis*.
5. Explain the differences between the contexts of discovery, justification and application of hypotheses.
6. What are the methodological requirements that apply to hypotheses in the discovery and formulation phase?
7. Characterize singular hypotheses and give at least one example.
8. Characterize existential hypotheses and give at least one example.
9. Characterize universal hypotheses and give at least one example.
10. Characterize statistical hypotheses and give at least one example.
11. Describe the nature of evidential statements. Do they express infallible and reliable knowledge?
12. Explain when a hypothesis is verifiable, and when is it verified.

13. Explain what constitutes the confirmation of a hypothesis in the instantial model of confirmation. Provide an example of a hypothesis and an evidential statement that confirms that hypothesis.

14. Explain when a hypothesis is falsifiable, and when it is falsified.

15. Explain when a hypothesis is disconfirmed only, and what the methodological consequences of this are.

16. What are the three basic elements of the Bayesian theory of confirmation?

17. Define the concepts of confirmation, disconfirmation and neutral evidence with reference to the Bayesian theory of confirmation.

5 CAUSATION AND ITS ROLE IN SCIENCE

5.1 Introduction

Information about the causes of natural and social phenomena is an important part of scientific knowledge. If we can identify the cause of an event, this enables us (potentially) to explain why that event came about. Similarly, if we have knowledge about the kind of conditions leading to a given event, that enables us to derive a prediction – anticipate the kind of event that will occur if these conditions are satisfied. Knowledge about the causes of phenomena and events is thus an important tool that can be used to explain and predict (or historically reconstruct) other phenomena and events. In general, the conditions leading to a certain phenomenon (event, state of affairs, fact etc.) are known as the *causes* (of the given phenomenon etc.), while the phenomenon, event (state of affairs, fact etc.) they caused is called the *effect* of that cause.

Scientists formulate explanations just as the rest of us do. However, they are not particularly concerned with what saying that one phenomenon is the cause of another actually means. Conversely, philosophers (of science) do not spend much of their time producing causal statements (at least not relating to the philosophy of science), but pose questions instead, such as "*Under what circumstances* is something *the cause* of a certain phenomenon (event etc.)?" or "*What conditions must be satisfied* for us to *identify* something *as the cause* of something else?". In other words, philosophers (of science) inquire as to the *ontological* and *epistemological* assumptions of our causal statements.

In what follows, we will look at some of the chief concepts of causation. These play – or so it seems – an important role in science and underpin our efforts to understand and predict the course of the events that we encounter.

5.2 Causation: concepts and approaches

"Long-term stress damages the brain", "One of the causes of the extinction of the dinosaurs was an asteroid hitting the coastal region of Yucatan", "Long-term smoking leads to a range of cardiovascular diseases", "The introduction of teaching method *X* in subject *Y* increased the students' GPA". These and other statements show that in some situations, we consider certain events (phenomena, conditions) to be the causes of other events (phenomena, conditions). It is not only statements of fact like these that indicate causal thinking. It is often expressed in questions such as "Is climate change today (primarily) the consequence of human activity?", "Does a family's economic conditions affect their children's educational outcomes?", or "What led Hitler to attack the Soviet Union?".

This "causal language" is used more frequently in some disciplines than in others. For example, a scientist working in pharmacological research would probably pose causal questions more often than a linguist would. The former is often interested in the desirable and undesirable effects of a drug or a specific dosage of the drug. A linguist may be interested in the social or cultural factors that led to a linguistic norm or in finding out why a particular dialect was selected for the codification of a language. However, in the linguist's case, such causal questions would seem to be the exception rather than the norm.

Moreover, even within a single discipline (such as physics), there may be theories that contain more causal expressions than other theories in that discipline (for example, classical mechanics versus quantum mechanics).

Despite such differences, if we are to gain a better understanding of our causal thinking, it is crucial that we understand the meaning of causal expressions, and the role played by some of the ontological and epistemological assumptions underpinning them. In the rest of this chapter, we will introduce some of the clas-

sic approaches as well as some of the more recent approaches to specifying what causes (effects) actually are and what we mean when we say X is the cause of Y.

Philosophically, the most fundamental question here is whether causation is a feature of our world that can be reduced to some other (non-causal) features of the world. Approaches that consider this to be the case are known as *reductionist* approaches, while those that reject this possibility are *non-reductionist* approaches.

When we say that X is the cause of Y, we can think of X and Y in two distinct ways. We may view them as particular events or as aspects of a state of affairs that obtained at a particular point in time and space, such as in the statement "One of the causes of the extinction of the dinosaurs was an asteroid hitting the coastal region of Yucatan". Alternatively, we may think of causes and effects as certain types of events or circumstances that are not fixed at a particular point in time and space, such as in the statement "Long-term stress damages the brain".

Some philosophical theories view causes as particular events (or types of events) whose occurrence (or absence) makes a difference to some circumstances. Such approaches are known as "difference-making accounts". Other theories present causes as things that produce certain phenomena. These theories are known in the literature as "production accounts" (see Illari – Russo 2014).

Let us take a closer look at some of these theories, namely, the regularity account, the INUS account, the counterfactual and probabilistic approaches, as well as the manipulationist account.

As far as preferences regarding any of the approaches are concerned, we are open to methodological pluralism. This is because there are research contexts to which one particular approach is well suited while in others another approach would be better. We will limit our discussion to briefly introducing the basic approaches to causation and pointing out some of the problems with them.

5.2.1 Regularity theories of causation

The intellectual ancestor of modern regularity theories of causation was David Hume (1711 – 1776). In his *An Enquiry Concerning Human Understanding* (1777),

Hume formulated two non-equivalent characterizations (definitions) of the concept of cause. The first defines a cause as

> "[…] *an object, followed by another, and where all the objects similar to the first are followed by objects similar to the second*." (Hume 1975, Sec. VII, Part II, Paragraph 60)

The second informs another, counterfactual approach to causation:

> "Or in other words *where, if the first object had not been, the second never had existed*." (ibid.)

However, let us turn to the first definition. Using more up-to-date language, we can restate it thus:

Assume that *c* and *e* are particular events in space-time, and that *C* and *E* are types of events of which *c* and *e* are particular instances. Then, *c* is a cause of *e* if and only if:

1. *c* is contiguous in space-time with *e;*
2. *e* follows in time after *c*;
3. all events of type *C* (i.e. events similar to *c*) are regularly followed by (or invariably occur simultaneously with) events of type *E* (i.e. events such as *e*) (Psillos 2002, 19; 2009, 131).

According to regularity theory, cause is based on a range of non-causal facts – namely, on the existence of contiguity in space-time between the events, an order of cause and effect in time and the constant regularity of occurrence. There is thus no "power" or necessary connection binding the events (facts) that we call "causes" and those we call "effects".

For example, saying that the aerobic exercises Alexander did caused his (subsequent) increased heart rate is equivalent to saying that (i) Alexander's aerobic exercises were contiguous in space-time with the event of his increased heart rate,

(ii) that his increased heart rate followed the exercises time-wise and (iii) whenever someone performs aerobic exercises, their heart rate will increase.

If we look at the key concepts in the definition of cause provided by regularity theory, some of them are clearly not precise enough for us to apply this approach to any example of causation. For example, when is event *c* contiguous with event *e*? Just how spatially close do *c* and *e* have to be to satisfy the first condition? We can illustrate this in the following way. Assume that someone has stated, "The conflict in Syria between Bashar Assad's government forces and their opponents (with support from foreign powers on both sides) is the cause of the migration of ordinary Syrians". Can we say this statement is a true causal statement using the regularity theory of causation? That depends on (among other things) whether we view "the conflict in Syria between Bashar Assad's government forces and their opponents" and "the migration of ordinary Syrians" as two particular events. If we do, we then need to know whether they are *contiguous*. With regard to the first question, regularity theory gives no direct answers as to what kinds of entities can be considered events. Although Hume originally thought of causation in terms of phenomena that are observable in space-time (such as a billiard ball hitting another), his formulations provide no direct hint as to the granularity of our thinking about events. For this reason, we may suppose that our example involves (macro) events, even though both the conflict in Syria and the migration of Syrians may represent processes that take place over a longer interval of time. This means that answering the first question may not be straightforward. This is because we may not be able to determine the precise borders of the conflict and of the migration process, and so we cannot definitively say whether we can evaluate these events as cause and effect using regularity theory.

We encounter a similar problem if we want to decide whether the second condition is satisfied. When considering the third condition, we came across another problem. In the third condition, type *C* events are always followed by type *E* events. This assumption can never be verified as true, since it also involves events (type *C* and *E*) that have not occurred yet and therefore by definition always remain inaccessible to us. Thus, even though certain metaphysically "loaded" terms, such as *power* and *necessary connection* are avoided in the regularity theory

of causation, it still requires the metaphysical assumption that all events of a certain type are always followed by events of some other type. As we have pointed out, this assumption cannot be verified.

For a regularity theorist, evidence plays an eliminative role. It helps eliminate those causal hypotheses for which it has been shown – at least once – that even though a type *C* event occurred, a type *E* event did not follow.

Apart from the problems involved in evaluating the conditions posited by the regularity view, this approach also faces other challenges. We shall at least mention some of these.

The first problem relates to the inability of regularity theory to distinguish between cases in which one event simply leads to another event, and cases where a single event leads to two consecutive phenomena. More specifically, we may discover that events *c* and *e* satisfy the regularity definition of causation, and thus identify event *c* as the cause of event *e*. However, both *c* and *e* could be the effects, following one another in time, of the earlier event *c** that leads to them. For example, lightning is contiguous with the acoustic phenomenon of thunder: thunder occurs after lightning, and whenever lightning strikes, thunder follows. Nevertheless, lightning is not the cause of thunder. Both phenomena are the effects of an earlier cause: an electrostatic discharge in the atmosphere.

This case can be generalized. The regularity approach cannot distinguish between cases in which, by pure coincidence, two events appear as if they are space-time contiguous (in other words, occur "on the same spot") and so formally satisfy the regularity definition, but nonetheless we still have reasons to believe that the first is not the cause of the other (based on our pre-theoretical or theoretical knowledge), and cases in which one event is the cause of another event (and the regularity definition is also satisfied).

Finally, the regularity approach cannot correctly identify causes in situations where the effect of one cause preempts the effect of another cause. Let us assume that person *X* ingests a critical dose of cyanide, and so will die within 24 hours. However, before the cyanide takes effect, *X* suffers a fatal accident causing them to die. Applying causation as regularity to this case will not lead us to correctly identify the cause of *X*'s death because both the ingestion of cyanide and lethal

accidents have regular effects. Despite the person having ingested cyanide *and* suffered a lethal accident, we would view only the accident as being the cause of death.

Despite these problems, the regularity theory of causation lies behind some of the causal statements we encounter in science. Knowing some of the difficulties with this approach also enables us to critically assess causal statements that use this concept of causation, even if implicitly. These difficulties also led to the emergence of some other approaches, and it is to these which we now turn.

5.2.2 Causes as INUS conditions

John Mackie, in his *The Cement of the Universe* (see Mackie 1974, Chapter 3), put forward a version of the regularity theory of causation that analyzes causes using the concepts of *necessary* and *sufficient* conditions.[41] This approach builds on John Stuart Mill's analysis of causal statements. Specifically, Mackie makes use of Mill's idea that causes (usually) appear as sets of complex conditions. He also notes that a certain kind of event (an effect) may have different causes. More precisely, an effect (such as a fire) may have multiple causes. In turn, each of these causes can comprise a cluster of factors, none of which on its own is sufficient for E to occur. However, taken together, these factors constitute a condition that is sufficient for E to occur.

We can illustrate these two ideas using a simple example. Suppose a house catches fire. This particular fire (singular event e) is a general type of event – a house fire (E). Fires can have various causes. Let us suppose, though, that in this case the cause was a short circuit. If we say that the short circuit caused the fire, we are using – or so Mackie says – the concept of cause in a specific sense: "The short circuit (A) is one of the factors (say, BCD) that together cause fires (E)". If the short circuit caused the fire, this does not mean that the short circuit was *a necessary condition* for the fire to occur. After all, house fires can (gener-

[41] Brennan (2017) offers a synoptic overview of the contemporary debates around the concepts of *necessary* and *sufficient conditions*. In the Slovak context, Gahér (2011; 2012) put forward a critique of the standard interpretations and a revision of both concepts.

ally) be caused by a lit candle (and other factors) placed near a flammable object catching fire. Neither does it mean that the short circuit was *a sufficient condition* for the fire. Hence, fire E could have been caused by a short circuit (factor A), but that alone is not enough for a fire to start. For the short circuit to result in a fire, other factors must be present – the house was built using flammable materials (B), there was a source of oxygen near the short circuit (C), the house was not equipped with a sprinkler system (D) etc. The fact that these four (or more) factors are present can be denoted as (the conjunction of factors) $ABCD$. Taken together, these factors comprise a sufficient condition for the fire (E). Factors $ABCD$ are not, however, a necessary condition for E, since there is another set (conjunction) of factors X that represents an alternative sufficient condition for E. These could be the circumstances under which the fire spreads from the candle (or fireplace) to an area where there is flammable material (the presence of oxygen) etc.

Mackie argues that in cases such as that of the house fire, we use a concept of cause that identifies *the cause* with a certain *part of a complex condition* – in our case with factor A. To say that the short circuit A was the cause of fire E actually means that factor A was a non-redundant part of complex condition $ABCD$, which is in itself a sufficient (although not necessary) condition for E. However, there may also be another set of facts (FGH), where each of the factors is a non-redundant part of FGH, which, as a whole, are a sufficient (though not a necessary) condition for E.

According to Mackie, a causal regularity expressing the causes of house fires would have the following form:

$$[(ABCD) \vee (FGH) \vee X] \leftrightarrow E$$

Mackie argues that causal statements of this form express the fact that all cases of $ABCD$ or FGH or X are followed by E's (type E events), and all E's are preceded by factors $ABCD$ or FGH or X (see Mackie 1974, 62).

In this sense, saying that *A caused E* amounts to saying that A is *an Insufficient but Non-redundant part of an Unnecessary but Sufficient condition* for E, or its acronym INUS. Event A (a type A event) is thus an INUS condition.

However, Mackie does not think *exclusively* about events as INUS conditions. He argues that if we think about causes in terms of necessary and sufficient conditions, then causes are *at the very least* INUS conditions. However, other cases (such as these below) are also compatible with analyzing causes as INUS conditions:

1. $A \leftrightarrow E$
2. $AX \leftrightarrow E$
3. $(A \vee Y) \leftrightarrow E$

The first case corresponds to a situation in which A is both a sufficient and necessary condition for E. In the second case, A is an insufficient but necessary part of a condition that, when taken as a whole, is both necessary and sufficient for E. The last case represents the possibility of A being a sufficient but not a necessary condition for E.

Mackie was not satisfied with his own approach to causes as INUS conditions and later replaced it with a variant of the theory of causal mechanisms. Indeed, there are several problems with his original version. Let us briefly review just two of them.

The first relates to the apparent sensitivity of Mackie's approach to the way in which we describe complex conditions. The philosopher Jaegwon Kim came up with the following example: Imagine a regularity of the form $(A \vee (\neg A \wedge B)) \leftrightarrow E$. Here, factor B is an INUS condition. But note that the complex condition

$$(A \vee (\neg A \wedge B))$$

is logically equivalent to the condition

$$(B \vee (\neg B \wedge A)).$$

This means that the former can be substituted with the latter to produce an equivalent regularity $(B \vee (\neg B \wedge A)) \leftrightarrow E$. However, in this logically equivalent variant, the INUS cause is suddenly $\neg B$, and not B as is the case in the original version (see Kim 1971). Identifying causes as (at least) INUS conditions can produce

results that vary significantly depending on how the complex circumstances are described.

This problem indicates that if we analyze causes *exclusively* in terms of the necessary and sufficient conditions, we cannot reliably identify the contributing causal factors, which had been the original intent of Mackie's project. However, there is the possibility of adding further requirements to this approach that would eliminate these problems (e.g. by introducing the temporal priority of certain factors or an explanatory function tied only to certain complex conditions, not their logically equivalent variants).

The second problem, pointed out by Mackie himself, arises in connection with a feature of regularity approaches noted in the previous section: Namely that when we analyze the causes as an INUS condition, we cannot distinguish the *real* causes from cases in which certain phenomena are the effects of a single common cause. Psillos (2009, 152) illustrates the general structure of these cases in the following way:

Assume that there are two effects, E_1 and E_2, whose common INUS condition is C. For example, (CX or Y) is a necessary and sufficient condition for E_1, whereas (CZ or W) is a necessary and sufficient condition for E_2. It follows that the complex condition $(E_1 \wedge \neg (Y \wedge Z))$ is sufficient for E_2, while the complex condition $((E_1 \wedge \neg (Y \wedge Z)) \vee W)$ is necessary and sufficient for E_2. Thus, according to the definition of an INUS condition, E_1 is an INUS cause of E_2.

As in the first problem, this case shows that analyzing causes as INUS conditions is insufficient. Nevertheless, it is possible to use the INUS approach to discover and identify causes. When thinking causally we usually draw on rich background knowledge which does not restrict us to searching for factors that would be part of complex conditions. Therefore, the concept of cause as (at least) an INUS condition can be used, along with other theoretical information, to search for and identify the circumstances that led to the given events. Naturally, there may be situations in which we are unable to identify the conditions that could (at least hypothetically) represent the causes, no matter which other tools we bring into play. However, none of the other approaches to causation avoid these problems either.

By way of conclusion, we may state that the concept of cause as an INUS condition, when used in parallel with theoretical background knowledge, can be a valid tool for analyzing the phenomena whose causal factors we are trying to uncover.

5.2.3 The counterfactual approach

The counterfactual approach to causation employs a concept of cause (of some event) that can be defined using another, more basic concept – the concept of counterfactual dependence. Like the regularity account, it is also a reductionist approach. It views causes as factors responsible for differences in states of affairs – in other words, as factors responsible for whether one or another state of affairs obtains.

The terms "counterfactual" or "counterfactual statement" refer to statements of the form "If *C* had been the case, then *E* would have been the case" or "If *C* had not been the case, then *E* would not have been the case". Thus, statements such as "If I had struck the match, it would have lit" or "If the stone had not hit the window, the window would not have broken" are examples of counterfactual statements. In general, a counterfactual is a conditional statement describing how the state of affairs would differ from the actual state if the condition expressed in the sentence had been satisfied (assuming that the condition is not actually satisfied).

One of the main proponents of the modern version of this approach was the American philosopher David Lewis (1941–2001). Lewis analyzed causation in terms of causal dependence. The latter is then defined using the concept of counterfactual dependence (see Lewis 1986b, but also Lewis 1973).

The basic framework of Lewis' theory of causation is provided by thinking about "possible worlds". In general, possible worlds can be viewed as theoretical pendants of the logical possibilities that represent the ways in which the world we live in could differ from the way it is. For Lewis, each of these possibilities is physically just as real as the world we live in; however, his theory of causation can be described without having to go into his theory of possible worlds in more detail. Suffice it to say that we will be thinking about these possible worlds as

various, mutually differing but logically consistent views of what the world could look like. In other words, if two possible worlds differ in at least one fact (e.g. in one world it is true that David Lewis is a philosopher, but in the other it is false), or in at least one law of nature (e.g. in one world the speed of light c cannot exceed 300 000 km/s, while in the other $c > 300\,000$ km/s), then the two possible worlds are different.

The concept of possible worlds is key to Lewis' counterfactual approach. Lewis analyzes the truth conditions for counterfactuals by comparing our actual world (its facts and laws) with other possible worlds. He assumes that possible worlds can (in principle) be compared. Moreover, he believes that, in general, we can list the features that make one of the possible worlds more similar to our actual world than it is to other possible worlds. If one possible world is more similar to our world than it is to other possible worlds, Lewis says it is *closer* to our world. Let us illustrate these general theses using an example.

Imagine I am holding a box of matches in my hand but have no intention of striking a match now or at any later moment. Assume I now declare, "If I had struck this match, it would have lit". Is this statement true? It appears to be. Lewis would explain the truth of this counterfactual in the following way. Let us denote our world using the symbol "@". In @, the statement "I had struck this match" is false. However, let us imagine a set of all the possible worlds A in which that sentence is true. We will call them A-worlds. We can now define the truth conditions for our counterfactual as follows:

> The statement "If I had struck this match, it would have lit" is true in the actual world @ if and only if those A-worlds in which the statement "This match lit" is true are closer to @ than are those A-worlds in which the statement "This match lit" is false.

To put it differently, the statement can be evaluated as true because those possible worlds in which the statement's condition is satisfied and in which the struck match will light, are more similar to our actual world (and therefore closer to it) than those worlds in which the condition is satisfied but struck matches do not light.

Generally, truth conditions for a counterfactual of the form $A \boxdot\!\!\rightarrow E$ ("If A had occurred, then E would have occurred") can be defined as follows:

> The statement "$A \boxdot\!\!\rightarrow E$" is true in the actual world @ if and only if those A-worlds in which E is true are closer to @ than those A-worlds in which E is false.

These truth conditions for counterfactuals are an important part of Lewis' counterfactual theory of causality. The truth valuation of counterfactuals is a basic step toward defining counterfactual (in)dependence. Let us now take a closer look at the main components of Lewis' approach (see Psillos 2002, 92–101).

Let's assume we want to find out what the cause of event e was. According to Lewis, causation is the relation between particular events. Therefore, the cause of e will be a particular event c_i. Counterfactual dependence, the concept of which Lewis uses to define causation, is defined on statements (or propositions). In order to move away from speaking about events and towards speaking about statements, we need to introduce the labels "$O(c)$", "$O(e)$" as the respective abbreviations of the statements "Event c occurred" and "Event e occurred". Similarly, "$\neg O(c)$" and "$\neg O(e)$" will be used to denote the statements "Event c did not occur" and "Event e did not occur".

We can now group the statements "$O(c)$" and "$\neg O(c)$" into a single set $\{O(c), \neg O(c)\}$. Similarly, we group the statements "$O(e)$" and "$\neg O(e)$" into another set $\{O(e), \neg O(e)\}$. We may define counterfactual dependence between these sets as follows:

> **COUNTERFACTUAL DEPENDENCE**
> A set of statements $\{O(c), \neg O(c)\}$ *counterfactually depends* on a set of statements $\{O(e), \neg O(e)\}$ if and only if the counterfactuals $O(c) \boxdot\!\!\rightarrow O(e)$ and $\neg O(c) \boxdot\!\!\rightarrow \neg O(e)$ are both true.

As we can see, this definition assumes that (i) we *understand* when counterfactual statements are true and (ii) we *know* whether they are true. The second condition is clearly far from trivial and in some cases, we may even doubt whether it is pos-

sible to state exactly what the truth conditions are for the given counterfactual. However, we shall not engage with these complications here.

Lewis then uses the concept of counterfactual dependence to define *causal dependence*:

> **CAUSAL DEPENDENCE**
> Event *e causally depends on* event *c* if and only if the set of statements $\{O(e), \neg O(e)\}$ counterfactually depends the on set of statements $\{O(c), \neg O(c)\}$.

In other words, Lewis defines the causal dependence of events by means of the counterfactual dependence of sets of statements that describe the occurrence of these events. The causal dependence of one event on another is a sufficient condition for causation. If it is true that *e* causally depends on *c*, then *c* is the *cause* of *e*. However, causal dependence is not a *necessary* condition for causation.

Before completing his definition of causation, Lewis adds the definition of a *causal chain*:

> **CAUSAL CHAIN**
> The sequence of events $\langle c, d, e, \ldots \rangle$ is a causal chain if and only if *d* is causally dependent on *c*, *e* is causally dependent on *d*, etc.

Using all these concepts, Lewis formulates his definition of a cause as follows:

> **CAUSE**
> Event *c* is the *cause* of event *e* if and only if there is a causal chain leading from *c* to *e*.

The concept of cause as an event that grounds the counterfactual dependence of the respective statements also plays a role in some theories of scientific explanation (see e.g. Woodward 2003).

Let us look briefly at a particular kind of problem of which there are several versions in the literature (early pre-emption, late pre-emption and the problem of overdetermination – see Menzies 2017.)

This problem involves a situation where a particular cause c_1 of event e is not the only (potential) cause of e that is present. More specifically, in addition to c_1, there is another cause c_2 that leads to event e. Let us assume that it was in fact c_1, and not c_2, that caused e. Nevertheless, it still holds that the fact that e occurred does not counterfactually depend on the fact that c_1 occurred. For even if c_1 had not occurred, e would still have occurred owing to the fact that c_2 occurred. Therefore, c_1 cannot be identified as the cause of e using the counterfactual approach – much as we would like it to – because e was not *causally dependent* on event c_1.

As we have noted, there are several variants of this problem, which we shall not deal with. In addition, the plausibility of the counterfactual approach to causation depends on the definition of truth conditions for the counterfactuals and on testing the latter to establish whether they are true. Again, we refer the reader to the existing literature (see Paul 2009, Menzies 2017 or Zeleňák 2008).

5.2.4 Probabilistic theories of causation

The starting point for probabilistic theories of causation is the idea that causal thinking – thinking about the causes of the phenomena around us – is not limited to the deterministic notion of causation. Causes need not be the kind of events that *always* (or *necessarily*) lead to effects of some type. The alternative to this deterministic notion is the idea that causes are kinds of events (or their instances) that *increase* or *decrease* the probability of some other kind of event (or their instances) occurring. According to this approach, the causes are not the factors that explain the difference between an event happening or not happening. They are the factors (conditions, circumstances) that are responsible for the *difference* in the probability of an event (phenomenon) occurring or not occurring. In general, probabilistic theories of causation view causes as *kinds of events*, but in some versions (due to the way probability is interpreted), probabilities are ascribed to particular instances of events. In our explanation, we shall limit ourselves to attributing probabilities to kinds of events.

Chance and the numerical expression of chance using the (concept of) probability lie in the background of the many processes and events that are the object of empirical science. Hence, the idea that causation also relates to events that only occur in a certain percentage of all cases fits with the role that probability plays in research in several disciplines.

Probabilistic theories of causation have appeared in the work of a number of philosophers – from Reichenbach's *The Direction of Time* (see Reichenbach 1956) and Good's two-part paper "A Causal Calculus (I), (II)" (see Good 1961a; 1961b) to the work by Suppes (1970), Cartwright (1979) and Eells (1991). In this section, we will outline a basic scheme of probabilistic causation which is common to a number of authors (with slight modifications).

Since the probabilistic approach to causation is based on a view in which causes are seen as factors that increase or decrease the probability of a kind of event occurring, we can start by defining a *prima facie positive* cause and a *prima facie negative* cause (see Illari – Russo 2014, 79), which we will then amend later on:

PRIMA FACIE POSITIVE CAUSE

Event C_t is the *prima facie positive cause* of E_{t*} if and only if

(a) $t < t*$

(b) $P(C_t) > 0$

(c) $P(E_{t*} \mid C_t) > P(E_{t*})$

Condition (a) in this definition expresses the temporal direction of events that are considered to be the causes of other events. The symbols "t" and "$t*$" represent the moments in time (or intervals) in which the respective events take place. Hence, the first condition expresses the fact that moment in time (interval) t precedes moment in time (interval) $t*$. Condition (b) expresses the assumption that the probability of a type C event occurring at time t is greater than zero. Finally, condition (c) states that the probability of event E occurring at $t*$, supposing that C took place at time t, is greater than the probability of E occurring at time $t*$. In other words, the occurrence of the factor or event C that precedes the factor or event E increases the probability of E occurring. This definition is the first step

toward delineating the concept of positive cause. Before continuing, let us look at the definition of a *prima facie negative cause*:

PRIMA FACIE NEGATIVE CAUSE

Event C_t is the *prima facie negative cause* of E_{t*} if and only if

(a) $t < t*$

(b) $P(C_t) > 0$

(c) $P(E_{t*} \mid C_t) < P(E_{t*})$

The only difference between this and the previous definition lies in condition (c). A negative cause is a factor or event C that precedes a factor or event E and decreases the probability of it occurring (see Illari – Russo 2014, 77).

An increase or decrease in the probability of event E conditioned on event C is not sufficient for defining the probabilistic concept of cause. Let us take a simple example: let E denote the occurrence of a storm and C the fall in the level of mercury in a barometer. If the barometer is working normally, then the probability of a storm occurring, given that the level of mercury has fallen, is greater than the probability of a storm occurring without the mercury level having dropped. Based on the definition of a *prima facie* positive cause, we would have to view the decrease in the level of mercury in the barometer as the cause of the storm. However, this would clearly be a mistake. Despite the correlation between the storm and the decrease in the mercury level, the decrease is not the cause of the storm. Both are the effects of a common cause – a drop in atmospheric pressure. Moreover, in this case, both the relationship $P(E \mid C) > P(E)$ and the converse relationship $P(C \mid E) > P(C)$ obtain. Thus, the occurrence of the storm increases the probability of the level of mercury falling. We would not be willing to accept, though, that the storm was the cause of the level of mercury dropping.

To avoid such issues, we need to distinguish between "false" factors of probability and real ones. We can complete our definitions of positive and negative probabilistic cause using the following definition of screening-off (see e.g. Kutach 2014, 106):

SCREENING-OFF
Event C screens off event E_2 from event E_1 if and only if $P(E_1 \mid E_2 \,\&\, C) = P(E_1 \mid C)$.

Equivalently, event C screens off event E_2 from event E_1 if and only if the occurrence of E_2 neither decreases nor increases the probability of E_1 occurring, assuming that C had occurred. Coming back to our example, we may say that the decrease in atmospheric pressure (C) screens off the decrease in the level of mercury (E_2) from the occurrence of the storm (E_1). This is because the probability of a storm occurring, assuming there was a drop in the atmospheric pressure, *is equal to* the probability of a storm occurring based on the assumption that there was a drop in the atmospheric pressure *as well as* a fall in the level of mercury.

Now that we have the concept of screening-off at our disposal, we can formulate the following amended definitions of positive and negative probabilistic cause:

POSITIVE CAUSE
Event C_t is the *prima facie positive cause* of E_{t*} if and only if

(a) $t < t*$

(b) $P(C_t) > 0$

(c) $P(E_{t*} \mid C_t) > P(E_{t*})$

(d) There is no event $C*_{t0}$ that precedes C_t in time and screens off C_t from E_{t*}.

Similarly:

NEGATIVE CAUSE
Event C_t is the *prima facie negative cause* of E_{t*} if and only if

(a) $t < t*$

(b) $P(C_t) > 0$

(c) $P(E_{t*} \mid C_t) < P(E_{t*})$

(d) There is no event $C*_{t0}$ that precedes C_t in time and screens off C_t from E_{t*}.

The concept of negative cause is a great advantage to theories of causation. In disciplines such as medicine, pharmacology or psychology, researchers are particularly interested in ascertaining the role certain factors play in reducing certain kinds of effects. For example, we might be interested in finding out whether eliminating excess fat from the diet diminishes the risk of diabetes or whether regular exercise reduces the probability of cardiovascular disease.

Alas, there are also various problems with probabilistic approaches to causation. One is how the concept of probability is interpreted in the various definitions. Some theorists work with the concept of probability as relative frequency or with the notion of probability as the physical propensity of certain events to produce certain outcomes. They view probability as an objective property of certain events (or collections of events) and so face the nontrivial question of how it is possible to *know* the specific probabilities related to these events. Other theorists use a subjective (Bayesian) interpretation. In these cases, probability expresses the agent's (scientist's, medical professional's etc.) degree of belief. Here, the frequency with which a given probabilistic statement applies to cases may, but need not be, taken into consideration. However, if probabilities correspond to the degree of belief the agent has, can causation still be viewed as an objective feature of processes taking place "out there"?

A similar problem occurs with respect to the relationship between types of events and the probability of a particular event occurring. These and other issues are debated among the proponents and critics of the probabilistic approaches to defining the concept of causes. (For an overview of the current state of the theories, see Hitchcock 2018.)

5.2.5 Manipulationist accounts

Like the previous theories, manipulationist approaches to causation come in different varieties. There are two main versions: (a) the agency-based approach and (b) the interventionist approach. Proponents of the first approach relate the key

element of this approach – manipulability of some kind – to the agency of "free agents" (see e.g. Collingwood 1940; von Wright 1971; Menzies – Price 1993). Typically, these theories anchor causation in the interventions performed by freely acting humans on variables (conditions, circumstances), affecting the values of other variables (conditions, circumstances). In what follows, we leave this anthropocentric approach to one side and focus on the theorists concerned with the interventionist approach.

In the interventionist approach to manipulationism (e.g. Hausman – Woodward 1999; Pearl 2000; Woodward 2003 and 2009), the concept of intervening with variables is viewed more broadly. More specifically, a change in the values of variables representing causally relevant factors is viewed as resulting from the type of intervention. In general, the concept of intervention used by scholars such as Woodward does not require a human agent (see below). Although these approaches differ in some of their theoretical assumptions, the idea that identifying causes is about being able to subject certain conditions to controlled manipulation and to being able to observe any consequent changes is something they share in common.

To simplify this, if C is a factor or condition that can be changed, then C is a cause of E (an event, a condition, the values of a variable) if and only if the change in E is the result of manipulating (the values of) C. Using a simplified example: Taking penicillin causes a patient (suffering from a streptococcal infection) to recover precisely because the patient's recovery would be achieved by administering penicillin.

The interventionist version of causation as manipulability is actually a variant of the counterfactual approach. The key concepts of the interventionist version – intervention and the manipulation of variables (their values) – are expressed using counterfactual concepts.

C and E are best viewed as variables, the values of each of which are at least two different events (states of affairs, quantities of a given magnitude). In our simple example, variable C can represent either event c (taking the penicillin) or $\neg c$ (not taking the penicillin). Similarly, variable E can represent event e (recovering from a streptococcal infection within n days) or $\neg e$ (the infection continuing).

In the interventionist version of the manipulationist theory of causation, we can identify C as the cause of E because intervention I on (the values of) variable C would change the values (or state) of variable E. In other words, beginning from the state in which the value of C is $\neg c$ and the value of E is $\neg e$, it is possible to achieve a state in which the value of E is e by changing the value of C from $\neg c$ to c by means of intervention I.

To generalize, let X and Y be two variables that range over $x_0, x_1, \ldots, x_m$ and $y_0, y_1, \ldots, y_n$, respectively, where x_i and y_i may represent numerical expressions of the magnitudes X and Y, or may be different events of types X and Y or different states. X is then *the cause of* Y if and only if the changes in the values of Y are (exclusively) due to the manipulation of the values of X.

Moreover, interventionism holds that the relation between the variables X and Y is causal if and only if it remains invariant throughout a large number of interventions on the values of X that influence the values of Y. Woodward, one of the main proponents of the interventionist account, views an intervention as process I for which it holds that:

1. the change in the value of X must be exclusively the result of I
2. I must influence the value of Y only by changing the value of X
3. I itself is not due to a cause that affects Y while circumventing X
4. I must be probabilistically independent of any cause Y that does not lie on a causal route connecting X to Y (see Woodward 2009, 247).

Note that the definition of an intervention by means of conditions 1 – 4 involves causal concepts (corresponding to expressions such as "to be due to", "cause", "causal route" etc.). Since this concept of intervention is also used in two other definitions of this theory (see below), the manipulationist account does not provide a *reductionist definition of cause*. This reference to other causal concepts relates to an issue we shall come back to at the end of this section – the question whether such definitions of causes do not lead to a vicious circle.

Nevertheless, let us first look at how the concept of intervention is used to define the concept of cause. Woodward presents the following definitions of *direct* and *contributing causes* in his work (see Woodward 2003, Chapter 3; 2009, 250–251):

> **DIRECT CAUSE**
> X is the *direct cause* of Y with respect to variable set V if and only if there is a possible intervention I on X that changes Y (or the probability distribution of Y), where all other variables in V besides X and Y are held fixed at some value by additional interventions independent of I.

The set of variables V represents other factors (variables), changes in which may affect the values of Y. Therefore, if we want to say that variable X is the cause of Y, we have to fix the values of these other variables at a particular value. In other words, we *control* for the values of these other variables.

Apart from the concept of direct cause, we can also use the concept of *contributing cause*:

> **CONTRIBUTING CAUSE**
> X is a *contributing cause* of Y with respect to variable set V if and only if
>
> (i) there is a directed path from X to Y, i.e. a set of variables $Z_1,\ldots, Z_n$ such that X is the direct cause of Z_1, Z_1 is the direct cause of Z_2, …, and Z_n is the direct cause of Y
>
> (ii) there is an intervention on X that affects Y when all other variables in V that are not on the path from X to Y are fixed at a particular value.

Therefore, the definition of a contributing cause is linked to the concept of a direct cause. More precisely, contributing causes can be identified within a sequence of variables where the first variable is the direct cause of the second, the second

variable is the direct cause of the third, …, and the next to last variable is the cause of the last (see Woodward 2009, 250–251).

The manipulationist account also uses a concept of intervention that specifies the (counterfactual) conditions describing what would have happened if we had changed some other conditions. What kind of intervention could this be?

Woodward, and some other theorists, considered interventions that are (i) realizable – such as the administering of a drug; (ii) hypothetical – i.e. they could be realized but ethical, psychological or practical reasons may prevent us from doing so, such as making a group of patients smoke cigarettes; or (iii) ideal – i.e. they cannot be physically realized, but our best available scientific theories hold that such interventions are in principle possible – such as modifying the orbit of planet Jupiter (see Illari – Russo 2014, 104).

Critics of this approach have pointed out that the definition of the cause is circular since it uses other causal concepts. They also note the difficulty of testing counterfactual statements. With respect to the circularity issue, proponents of interventionism argue that their goal is not to provide a reductionist definition. Another response would be that the interventionist account of causation reveals a nontrivial aspect of causation, although the latter cannot be defined without reference to other causal concepts. As regards the issue of testing, theorists like Woodward point out that the plausibility (or truth) of counterfactual statements can be derived from the truth of other statements – such as those representing laws of nature.

Again, for a deeper analysis of these problems, as well as potential responses to them, the reader is referred to the work of Woodward (2003; 2009), Illari – Russo (2014, Chapter 10) or Kutach (2014).

5.2.6 Using theories of causation

The theories of causation we have introduced here are by no means all that contemporary inquiry has to offer on the topic. Nevertheless, the approaches to causation selected provide us with some basic tools for analyzing causal statements in science and beyond. Although all these approaches are problematic in some way,

their concepts of cause can be meaningfully applied in various research contexts. To some extent, the plurality of causal concepts reflects the fact that causation is analyzed at different levels of reality within the various disciplines (e.g. at the level of elementary particles, the level of medium sized objects, the level available to our sensory observation), and different ontological and epistemological assumptions apply to each of these levels.

Study questions

1. Define the concept of cause in terms of the regularity approach and provide at least one example of the issues faced in this approach.
2. How does J. Mackie define a cause in terms of INUS conditions? Provide an example from science or everyday life that satisfies this definition.
3. Define the concepts of *counterfactual dependence* and *causal dependence*.
4. Characterize the basic idea behind the probabilistic approach to causation.
5. What is the basic idea of the manipulationist approach to causation?

6 SCIENTIFIC EXPLANATION

Scientific theories are systems of testable (empirical) statements (and their meanings) that systematize and generalize particular pieces of knowledge about a certain area. They thus express *information* that is sufficiently *general* to be used to *explain* or *predict* (or *retrodict*, i.e. predict backwards in time) phenomena. In this chapter, we shall deal with some of the philosophical approaches to ascertaining which conditions information must satisfy for it to constitute a *scientific explanation*. These are known as *models of scientific explanation.* They can be viewed as *ideal* frameworks that enable us to reconstruct *adequate scientific explanations*.

We will use the terms "explanation" and "scientific explanation" to denote the *product* or *result* of cognitive and communicative activity. We will be looking at explanation as "an attempt to render understandable or intelligible some particular event (such as the 1986 accident at the Chernobyl nuclear facility) or some general fact (such as the copper color of the moon during total eclipse) by appealing to other particular and/or general facts drawn from one or more branches of empirical science" (Salmon et al. 1992, 8). Of course, there are also ordinary, extra-scientific explanations that relate to areas of scientific knowledge. However, scientific explanations, unlike ordinary ones, have to conform to higher standards. For example, the information constituting a scientific explanation must in principle be intersubjectively testable. The object of explanation, and the explanatory factors, may be described in the language of a formal theory and involve a considerable degree of abstraction or various idealizing assumptions. These and other aspects are rarely present in extra-scientific explanations.

What follows will be limited to some of the main approaches to scientific explanation. We will focus on approaches devoted to explaining singular, particular phenomena. We refer readers who are interested in a more historical and sophis-

ticated introduction to models of scientific explanation to Wesley Salmon's still unparalleled work, *Four Decades of Scientific Explanation* (see Salmon 1989).[42]

6.1 Models of scientific explanation

Particular models (or theories) of scientific explanation share certain structural elements. They differ in the way these elements are characterized and in what is required of them.

All models of scientific explanation, and hence all explicit reconstructions of a scientific explanation based on a given model, comprise three basic elements:

(i) the *explanandum* = the *object* of explanation or its linguistic representation (a proposition)

(ii) the *explanans* = the *factors* that enable us to *explain the explanandum* (or propositions which describe these factors)

(iii) the *explanatory relation* between the explanandum and the explanans; or, a *set of criteria* applying to the explanandum and the explanans (and the relation between them).

In terms of the *kind of entities* featured in the explananda of scientific explanations, we can distinguish between those models of explanation (or particular explanations) whose objects are *singular* (particular) *events*, *states of affairs* or *singular facts*, and those whose objects are *types of events*, *laws*, *invariant relations* or simply *regularities* (of some kind). The methodological discussion of scientific explanation in the latter half of the 20th century (see Hempel 1942; Hempel – Oppenheim 1948) focused on explanations of *particular events*. Therefore, in what follows, we will concentrate on the logico-methodological features of explanations whose objects are particular events, states of affairs or singular facts.

[42] Skow (2016b) and Woodward (2017) provide an excellent analysis of models of scientific explanation. For analyses in Czech and Slovak, see Jastrzembská (2007; 2009) and Zeleňák (2008), respectively.

6.2 The deductive-nomological model of explanation

Modern methodological discussions of explanation emerged around a basic idea which can be summed up as follows. An event (described in a certain way) can be considered to have been adequately explained by a deductive argument whose conclusion (the explanandum) describes the event to be explained and whose premises (the explanans) include at least one (universal) law without which the conclusion cannot be derived from the premises. This law has to be empirically testable and true, as do the other, singular statements describing the relevant antecedent conditions (or limiting conditions – i.e. conditions in which the law applies). Therefore in order to explain why event *e* took place or why phenomenon *p* occurred, we have to present explanatory information that shows that event *e* (phenomenon *p*) was (or had to be) *expected* given such-and-such a law and such-and-such antecedent conditions. This approach to scientific explanation was first systematically introduced by Hempel and Oppenheim (Hempel – Oppenheim 1948; Hempel 1965). The basic idea behind it found support among many others (see Braithwaite 1953; Popper 2002; Nagel 1961). It is known as the *deductive-nomological* model of explanation ("the D-N model") or the *covering-law* model or the *subsumption* model.[43]

Before delving further into the D-N model, we will give a simple example. Suppose we want to explain why, at time *t*, flagpole *F* casts a shadow that is 13.33 m long. An explanation of this singular fact according to the D-N model has the following form:

(E1) Light rays travel from the Sun to the Earth in straight lines.
At time *t*, the light rays hit the surface of the Earth, and flagpole *F*, at an angle of 42°.
The height of flagpole *F* is 12 m.
13.33 (m) = 12 (m) ÷ tan(42°)

At time *t*, flagpole *F* casts a shadow that is 13.33 m long.

[43] The term "covering-law model" refers to two related models: the D-N model and the inductive-statistical (I-S) model.

The fact that the flagpole casts a shadow of a certain length is thus explained as a logical consequence of testable premises (assumed to be true) that include a law (of the propagation of light) without which it would not be possible to logically derive the conclusion from the other premises.

The D-N model of explaining singular phenomena therefore assumes that: the *explanandum* contains a statement (proposition) about the phenomenon to be explained; the *explanans* contains (apart from any mathematical statements relevant to the explanandum) empirically testable statements (propositions), at least one of which is a law, and these statements are (accepted as) true (with the possible exception of the statement representing the law); and the relation between the explanans and the explanandum is the relation of logical entailment. An additional requirement, that avoids some of the problems pointed out by critics of the D-N model, may be that all of the premises in the explanans have to be relevant – at least in the sense that if any of the premises are removed, the argument would become logically invalid.

The basic scheme of the D-N model of explanation has the following form:

(D-N MODEL)

$$\left.\begin{array}{l} L_1, \ldots, L_n \\ C_1, \ldots, C_n \end{array}\right\} \text{explanans}$$

$$\overline{\qquad}$$

E explanandum

where E is a statement describing the phenomenon to be explained, $L_1,\ldots, L_n$ are statements representing (universal) laws, and $C_1, \ldots, C_n$ are statements describing antecedent conditions, i.e. the conditions (or corresponding magnitudes) under which the laws apply. If we refer to the conjunction of the (relevant) laws $L_1,\ldots, L_n$ as L and if C denotes the conjunction of the (relevant) singular statements $C_1, \ldots, C_n$ about the antecedent conditions, then we can define scientific explanation according to the D-N model as follows (see Hempel – Oppenheim 1948; Salmon 1989, 12–25; Weber et al. 2013, 2):

(D-N EXPLANATION)
The statement (proposition) $L \wedge C$ is an explanation of (singular) phenomenon p described by singular statement E if and only if:

1. L is a strictly universal statement (proposition) representing a (scientific) law (or a conjunction of laws) and C is a singular statement (proposition) or a conjunction of singular statements representing antecedent conditions.
2. $L \wedge C \models E$ (i.e. E is a logical consequence of the conjunction of L and C)
3. $C \not\models E$ (i.e. E cannot be deductively derived from C as such)
4. C is true (i.e. $L \wedge C$ is confirmed)[44]

If the first three conditions are met, but the fourth is not, or if we do not know if it is in fact met, then the argument is a *potential explanation* of phenomenon p.

The explanatory potential of scientific theories is implicitly captured in the first three conditions. To explain the phenomenon described in the explanandum, it is necessary to apply a suitable (scientific) law – an element of a scientific theory that expresses a lawful regularity or an invariant relation between certain states or magnitudes. Moreover, the third condition expresses not just that E cannot be explained without law L, but also that E cannot be explained by a C that in itself logically entails E.[45]

As mentioned above, we may also add a fifth condition guaranteeing that the explanans will only contain premises relevant to the D-N explanation:

[44] The four conditions are a slight modification of the conditions listed in Hempel – Oppenheim (1948). However, both describe equivalent conditions.

[45] For example, if C were identical to E (i.e. $C = E$), then it would be trivially true that $C \models E$, and (due to the monotonicity of logical entailment), for any (irrelevant) A, it would hold that $A \wedge C \models E$.

5. The set of premises $\{L, C\}$ must be relevant to E in the sense that removing any of the premises breaches condition 2, i.e. E would then not be a deductive consequence of the explanans.[46]

This condition was not originally included in the D-N model, but given the objections to Hempel's (and Oppenheim's) original proposal, it is a welcome addition. It guarantees that the set of premises are not only sufficient for the deductive entailment of the explanandum, but are also *minimally sufficient* – i.e. no further assumptions can be added to the explanatory premises that are not needed to deductively derive the explanandum. In other words, due to this restriction, the D-N explanation is a deductive argument that cannot be extended *monotonically*.

There were a number of objections to the original version of the D-N model (as delineated by the first four conditions). We will discuss several of these and subsequently describe situations in which it is methodologically meaningful to use this model of explanation.

One of the objections is that it is possible to formulate *an explanation of phenomenon p* without referring to scientific laws (see condition 1 above). In other words, the requirements placed on a D-N explanation *are not necessary* for providing an adequate scientific explanation. Michael Scriven (1962) illustrated this using the following example:

> The fact that there is a stain on the carpet is explained by other singular facts, namely that my knee hit the table, causing the ink bottle to be knocked over and spilling its contents onto the carpet.

We can therefore provide an explanation for a particular phenomenon (such as the carpet stain), Scriven says, without referring to a scientific law or even formulating the explanation as a deductive argument.[47]

---

[46] Strictly speaking, this condition does not exclude all cases of irrelevance. This will become obvious below, in "hexed salt," the traditional counterexample to the D-N model of explanation. We shall see, however, that proponents of the D-N model can counter this problem by referring to condition 1.

[47] Scriven also thought the fact that Hempel (and Oppenheim) did not distinguish between pro-

In defending the D-N model against this criticism, we might point out (à la Hempel) that the singular statement "My knee hit the table, causing the ink bottle to be knocked over" includes a reference to a cause. Moreover, the concept of cause may require an analysis that cannot be explicitly formulated without using a *general causal* statement (such as a *causal law*). According to Hempel, the explanation proposed above can be characterized as an *explanation sketch* (i.e. a sketch of an ideal explanation) that, when containing a general law, has all the features of a D-N explanation.

Scriven's objection and Hempel's response can be viewed as relevant and correct if at least the following two conditions are met: (i) The object of explanation (explanandum) is described (represented) as the object of *scientific* explanation; and (ii) The concept of cause cannot be analyzed without a concept of *universal causality* or a *type of cause*.

In explaining the carpet stain, we may reasonably doubt whether a *scientific explanation* is actually required (first condition). If it is not, we might argue that Scriven's example is not really a relevant counterexample to the D-N model. After all, its goal is, first and foremost, to represent *scientific explanation*. On the other hand, we may also doubt whether *all* cases of singular causal statements (i.e. statements of the form "Such-and-such a *particular event c* is the cause of such-and-such a *particular* phenomenon *e*) can only be adequately analyzed using a general concept of cause in which the basic component of causation is a *kind* of event or a *type* of cause. Such a concept of cause is expressed in statements of the form "An *event (cause) of kind C* always (or in *x*% of cases) leads to an *event of kind E*". Therefore, we may allow, along with Scriven, that at least in some cases of *scientific explanation*, an adequate and fully scientific explanation need not refer to a universal (or statistical) causal law and have the form of a deductive argument.

We may conclude our discussion of the first criticism of the D-N model thus:

viding an *explanation* and providing *grounds for the explanation* was a defect in the D-N model (see Scriven 1962, 196–201). A similar idea underlies the current causal theory of explanation by Bradford Skow (see his 2016a).

Even though some "counterexamples" (commonly found in the literature) to the D-N model are not relevant, it appears that satisfying the conditions of adequacy of the D-N model *is not a necessary condition* for us to provide an adequate scientific explanation – assuming that at least some singular causal expressions can be adequately analyzed without reference to universal causal laws.

Let us now turn to criticism of the D-N model that comes from a different direction. There are several examples that are intended to show that conditions 1–5 above *are not sufficient* for an adequate scientific explanation (see, e.g. Salmon 1989, 46–50). We will restrict ourselves to three different kinds of objections.

The first concerns (scientific) laws. Since one of the definitional conditions of D-N explanations requires that statement (proposition) *L* express a scientific law, we may ask whether we are, in principle, capable of telling when a universal (true) statement is a lawlike statement, and when it is not. If there is no essential difference between laws (of nature) and other (true) generalizations, or not one we can identify, then there are bound to be difficulties meeting the first condition of the D-N explanation. We may encounter arguments that *prima facie* satisfy all the definitional conditions of the D-N explanation, but nonetheless cannot be considered explanations.

However, we can also look at this objection from another standpoint. The definition of D-N explanation seems merely to *assume* that certain kinds of laws are used in the scientific disciplines. If that is so, then the relation between scientific laws and the applicability of the D-N model can be expressed thus: If scientific laws *are available*, then the D-N model is applicable (once certain requirements are fulfilled). As several empirical disciplines really do have laws at their disposal, the D-N model applies to them.

Of course, the fact that we *lack a distinction* between *scientific laws* and *other true generalizations* indicates we may lack a *correct* (philosophical) *theory of laws* (or the *mechanisms* underlying them). The criticism, then, is that this would mean that arguments can be constructed that apparently satisfy the definition of D-N explanation but that cannot be considered truly explanatory. We could respond by saying that an argument that includes a true universal statement and satisfies the other definitional criteria, but is not viewed as explanatory, can also

be interpreted the other way round. That is, it can be seen as a challenge to the critic's assumption that the premises of the argument do contain a law. In other words, the universal statement in the premises of such an argument might not in fact represent a law of nature (or a social law).

A frequently cited example of this argument (see Salmon 1989, 50) is:

(E2) Any *hexed* sample of table salt will dissolve in water (at room temperature).
This sample of *hexed* table salt was placed in water.

This sample of *hexed* table salt dissolved.

This argument is generally viewed as an example of a D-N argument that *satisfies all the criteria* of (the definition of) the D-N explanation but that is still not considered an explanation. Therefore, it is seen as a correct and relevant *counterexample* to the D-N model. However, accepting that a D-N explanation must include a *scientific law*, it is difficult to see how this applies to the first premise – regardless of whether we have a satisfactory *theory* of (natural or social) *laws*. Certainly, we believe that the D-N model of explanation could only benefit from an adequate theory of laws. But it does not seem reasonable to make the D-N model so dependent on a theory of laws as to make it inapplicable even in cases where natural or social laws are clearly available.

Let us turn to two other types of objections. These are the *problems of asymmetry and irrelevance* of the D-N model (see Salmon 1989, 46–50; Weber et al. 2013, 7–9; Woodward 2017, Section 2.5).

The *problem of asymmetry* can be illustrated using the following example. Suppose we want to explain why flagpole *F* (in example E1 above) is 12 m tall. We could formulate a D-N argument similar to the one in example (E1) above:

(E3) Light rays travel from the Sun to the Earth in straight lines.
At time t, the light rays hit the surface of the Earth, and flagpole F, at an angle of 42°.
At time t, flagpole F casts a shadow that is 13.33 m long.
$13.33\ (\text{m}) \times \tan(42^\circ) = 12\ (\text{m})$

The height of flagpole F is 12 m.

However, we would probably not consider (E3) to be a D-N type explanatory argument, even though it uses the relevant laws of physical optics to derive a conclusion that describes the object of explanation. Although some laws (appearing in D-N arguments) enable us to symmetrically derive the value of whichever magnitude from the values of other magnitudes (in our case, the length of the shadow and the height of the flagpole can be derived from the same laws), the explanation assumes a certain *asymmetry* between the magnitudes appearing in the given law. In general, this assumption of *explanatory asymmetry* can be expressed using the following thesis:

(EA) For all X, Y it holds that: If X *is an explanation of why* Y, then Y *is not an explanation of why* X.

If the thesis is true, then the information about the height of the flagpole and the size of the angle of the incidence of the light rays upon Earth can be used to explain the length of the flagpole's shadow. However, it seems the same laws cannot be used to explain the height of the flagpole – even though we have the requisite information about the length of the shadow and the angle of the incidence of the light rays.

Most explanation theorists think (EA) is true. If the thesis is indeed true, then *not all* the laws that *enable* us to derive the (statements of the) *explananda* also *enable* us to derive their *explanation*. This means that the definitional conditions of the D-N model *are not sufficient* to identify explanatory arguments.

The problem of asymmetry does not show us that *symmetrical* laws are not useful to explanations; it shows us that merely deriving the explanandum, by means

of a symmetrical law and other relevant information in the premises of an argument, does not necessarily result in an explanation. One way of guaranteeing the asymmetry of D-N arguments is the requirement that the singular statements *C* appearing in a D-N argument must describe the *causal* conditions relevant to the occurrence of phenomenon *p* that we are attempting to explain. Taking this route enables us to avoid having to produce a number of counterexamples. But it also eliminates some D-N arguments whose explanatory information is not purely causal, but we would still consider them to be explanations. (Typical candidates for non-causal D-N arguments are (some) explanations referring to the dispositional properties, limitations or mathematical properties of certain empirical systems – see Lange 2017.)

The last objection to the D-N model is *explanatory irrelevance*. In general, explanatory irrelevance refers to a situation in which the premises of a D-N argument contain at least one piece of information (a statement) that is irrelevant to the *object of explanation* (i.e. to the *explanandum*). Argument (E2), noted above in connection with the problem of distinguishing laws from other regularities, is the standard argument used to illustrate explanatory irrelevance. We pointed out above the key reason for not considering (E2) as an explanation, which is that its premise which *was supposed* to play the role of a law *is not* in fact a law, and so the argument as a whole does not satisfy the definitional condition of D-N explanation. Does this, however, mean that the definition of D-N explanation is immune to other cases of irrelevance? We cannot completely exclude the possibility there may be more effective counterexamples, but we believe that condition 5 in the definition of the D-N explanation is capable of "filtering out" the known counterexamples as well as potential ones.

The D-N model is quite evidently not the *only* correct model of scientific explanation. Nevertheless, taking into account the problems discussed above and their (partial) solutions, the D-N model can be considered a methodologically functional *model of the ideal reconstruction of a range* of scientific explanations. Questions regarding its *application* become relevant once we come to use a *deterministic* (universally valid) *scientific law* as a *key element in a scientific theory* within a given discipline. The philosophical debates around the D-N model have

also produced a host of *examples* that can be used to *test* alternative theories of scientific explanation.

6.3 The inductive-statistical model

Some of the events (or kinds of events) that are the object of scientific explanation are part of processes and mechanisms that are *probabilistic* in nature. In this context, *probability* refers to either an *objective property* of an empirical system (such as the probability of Polonium Po^{210} emitting an alpha particle) or the *degree of* (our) *epistemic uncertainty* regarding our knowledge of a complex phenomenon (for example, when trying to estimate the effect of a drug on patients with a particular disease).

In this section, we will introduce an alternative to the D-N model that characterizes the *explanans*, the *explanandum*, as well as the *explanatory relation* with reference to *probability*. We will be looking at Hempel's *inductive-statistical* (I-S) *model* of explanation (see Hempel 1962; 1965, 376–412).

Let us begin with an example. Suppose we are looking for a medical explanation of the fact that person *a*, who was diagnosed with prostate cancer fifteen years ago, has fully recovered and is now in good health. The explanation we are seeking could include, for example, the information that men aged 49 and less who were diagnosed with prostate cancer at the primary tumor stage, and underwent a radical prostatectomy (the surgical removal of the entire prostate), have a 0.98 (i.e. 98%) probability of living for another 15 years.[48] Person *a* was 46 years old when the primary tumor was discovered and underwent radical prostatectomy, so it was highly probable that he would live for (at least) another 15 years. The scheme informing this example is known as the *inductive-statistical* (I-S) *model* of explanation. The "father" was, again, Carl G. Hempel.

The I-S model shares two features with the D-N model: (i) an I-S explanation has the form of an argument and (ii) the premises of an I-S explanation must contain at least one (relevant) *scientific law*. Unlike D-N explanations, I-S explanations

[48] The exact value of the probability is not in fact known, but various studies suggest that it is high.

are based on *inductive inference* and the law appearing in the premises takes the form of a *statistical statement* (hypothesis). Let "F" denote the basic *reference class* of objects having property F. We are interested in the probability of these objects also having property G, denoted by the term "G". Let "a" represent the given object. "$P(G \mid F) = r$" represents the probabilistic law (or hypothesis) that the probability of any object having property G, assuming that it has property F, is equal to the *real number* $r \in [0, 1]$. Assuming that $r > 0.5$ (i.e. the probability is greater than 0.5 or 50%), the basic scheme of the I-S model of explanation can be expressed as follows:

(I-S MODEL)

$$\frac{\begin{array}{l} P(G \mid F) = r \\ F(a) \end{array}}{G(a)} \; [r]$$

The I-S scheme thus represents an explanation of the *explanandum* (expressing the fact that object a has property G) that uses an inductive argument. The *explanans* in this argument expresses the information that object a has property F (i.e. it is an element of basic reference class F), and that the probability of any (random) object having property G, assuming that it also has property F, is equal to r, where r is greater than 0.5 (and, ideally, approaches 1, although it is never equal to 1). Let us note that the number r occurs in both the first premise and adjacent to the double horizontal line separating the premises of the I-S argument from the conclusion. This means that the probability r expressed in the statistical law (hypothesis) in an I-S argument transfers to the relation of *inductive support* conferred onto the conclusion by the premises of this argument.

Typically, the premises appearing in an I-S explanation are more complex. Basic reference class F is usually replaced by a *more specific reference class* – a class of objects that have another property (or properties) H besides property F. In that case, the scheme of I-S explanation is richer:

(I-S MODEL*)

$$\frac{\begin{array}{l} P(G \mid F \wedge H) = r \\ F(a) \wedge H(a) \end{array}}{G(a)} \; [r]$$

The example we began this section with could be reconstructed using this scheme in the following way. Let "*a*" denote the person from our original example. "*F*" will denote the property of *being a man of 49 years or less suffering from primary-stage prostate cancer*. Let "*H*" denote the complex property (condition) of *having been diagnosed with prostate cancer at the primary-tumor stage and having undergone radical prostatectomy*. Finally, let "*G*" denote the property of *living for (at least) another 15 years after diagnosis*. To explain, using the I-S model, why *a* lived for another 15 years, we construct an argument that refers to the following facts: (i) *a* was a man who was diagnosed with a prostate tumor before turning 50; (ii) *a* was diagnosed in the primary stage and underwent a radical prostatectomy; (iii) the probability of men living for another 15 years after being diagnosed is 0.98, assuming they were diagnosed before turning 50, the tumor was a primary-stage one and they underwent radical prostatectomy. Taken together, the premises of this argument show that there was *a high probability of* the fact referred to in the conclusion occurring (in our case, 0.98). Therefore, I-S explanations consist in putting forward an argument whose premises demonstrate that the conclusion (describing the object of explanation) was to be expected with a high probability.

The basic version of the *inductive-statistic model* of scientific explanation is expressed in the following definition:

(I-S EXPLANATION)

1. L is a statement of the form $P(G \mid F) = r$ representing a (scientific) statistical law (hypothesis) and C is a singular statement or a conjunction of singular statements representing antecedent conditions (e.g. of object a having property F).
2. $P(E \mid L \wedge C) = r$, where $0.5 < r < 1$.

3. $C \nvdash E$ (E cannot be deductively derived from C alone).
4. L and C are confirmed.

However, the I-S model of scientific explanation as delineated above encounters the problem of *explanatory ambiguity* (see Hempel 1965, 394–397). We can illustrate this using a slightly modified scenario from our previous example. Suppose that apart from having been diagnosed with prostate cancer, person *a* had also been diagnosed five years ago with an operable form of pancreatic cancer. Since pancreatic cancer is an aggressive type of cancer that progresses quickly, the probability of a person living for another five years after diagnosis is only about 5%.[49] Therefore, in our case, the probability of a person living for another fifteen years after being diagnosed with *prostate* cancer, assuming that the same person had also been diagnosed with *pancreatic* cancer five years ago, is only about 0.05. We can now add the following information to the premises containing information (i)–(iii) from the first scenario: (iv) person *a* had been diagnosed with an operable form of pancreatic cancer five years ago; (v) *a* underwent pancreatic cancer surgery followed by radiotherapy and/or chemotherapy; and (vi) the probability that a person will live for another five years after being diagnosed with operable pancreatic cancer, assuming they undergo surgery and radiotherapy and/or chemotherapy, is (approximately) 0.05. The premises (i)–(vi) of the resulting argument make the conclusion of the original argument (i.e., person *a* will live for at least 15 years after being diagnosed with prostate cancer) improbable – with a very low probability of about 0.05. The same fact can be expressed by saying that the probability of person *a not* living for another 15 years is very high: $1 - 0.05 = 0.95$ or 95%.

Schematically, this argument can be expressed as follows:

(I-S MODEL)**

$$\frac{\begin{array}{l} P(\neg G \mid F \wedge H \wedge I \wedge J) = r* \\ F(a) \wedge H(a) \wedge I(a) \wedge J(a) \end{array}}{\neg G(a)} \; [r*]$$

[49] See e.g. Ondruš et al. (2006, 126).

We have thus arrived at two inductive arguments whose premises are consistent, i.e. the information in assumptions (i)–(vi) is all true, but that support two inconsistent conclusions. The first argument confers a high probability on the conclusion that *a* will live for another 15 years after being diagnosed with prostate cancer, while the second argument confers a high probability on the opposite conclusion – that *a* will not live for another fifteen years after being diagnosed with the first type of cancer.

Generally, the *explanatory ambiguity* of I-S explanation applies in cases where there are (at least) two inductive arguments I_1 and I_2 whose premises are consistent (i.e. if we merge the premises of both arguments, the resulting conjunction will be consistent – all of the premises may be true at once), but the premises of I_1 support conclusion T with a great degree of probability, while the premises of I_2 support the logical opposite, i.e. $\neg T$, with a great degree of probability.

Hempel proposed to avoid this problem by adding the *requirement of maximal specificity* to the original definitional conditions (see Hempel 1965, 397–403). This requirement concerns the statistical statement (law, hypothesis) appearing in the explanans of an I-S explanation. It states the following condition:

> If the *knowledge base B* available at time *t* when we seek an explanation for $G(a)$ includes the fact that *a* is an element of a narrower reference class $F*$ that is a subclass of F, then B must also include the statistical statement $P(G \mid F*) = r*, r* = r$ (with the exception of cases where the statistical statement is a theorem of probability theory).

In other words, if we know (or believe) that basic reference class F can be narrowed down to subclass $F*$ that the object of our explanation is an element of, and if it holds that $P(G \mid F) = r \neq r* = P(G \mid F*)$, the premise $P(G \mid F) = r$ cannot be used to I-S explain why $G(a)$. Conversely, if $P(G \mid F) = r = r* = P(G \mid F*)$, then the premise $P(G \mid F)$ can be used (assuming the other conditions are met) to probabilistically explain $G(a)$.

This revised definition of I-S explanation thus adds another condition to the original four:

5. The statistical statement L (in the explanans) must meet the *requirement of maximal specificity*.

However, let us return to our two scenarios involving cancer. If we apply the requirement of maximal specificity to the case of the man diagnosed with two forms of cancer at two different stages of his life, then the first I-S argument must be eliminated as inadmissible (since it does not meet the requirement of maximal specificity). We are then left with the second argument. However, choosing the second argument would infringe on the second definitional condition of I-S explanation – namely, that the probability in the statistical hypothesis (and therefore also the probability conferred on the conclusion) should be greater than 0.5. The premises of the second argument confer a (very) *low* probability on the conclusion that person *a* has (so far) lived for 15 years since being diagnosed with prostate cancer. Therefore, in our example, the requirement of maximal specificity is not enough to ensure that an inductive argument with premises (i)–(vi) can explain this fact.

To summarize, the requirement of maximal specificity can "filter out" I-S arguments that are cases of explanatory ambiguity. However, only *some* of the arguments selected using this requirement are indeed I-S explanations. More specifically, these are arguments in which probability value r in the statistical hypotheses (which also expresses the probability of the relation between the premises and the conclusion) is greater than 0.5 (or 50%) This means that phenomena that occur with a probability lower than 0.5 *cannot be* explained using the I-S model.

6.4 Causal models

So far, we have not explicitly touched on the question of the *relation* between the *explanation* of why event *e* (phenomenon *p*) occurred and the *causes* of the event (phenomenon). In the D-N and I-S models the explanation (or, more precisely, its explicit reconstruction) is viewed as an *argument* where the explanatory potential of the *explanans* is represented by a *law*. Both models allow the law featured in the explanans to express a certain *causal relation*. However, as we have seen,

none of the models requires a *causal law* as the necessary condition for adequate explanation. In other words, a proponent of the D-N (or I-S) model may accept the thesis (see Hempel 1965; Psillos 2002, 222–226):

(C-DN) All causal explanations are D-N explanations.

Typically, however, they would not accept the thesis:

(DN-C) All D-N explanations are causal explanations.

Although the D-N and I-S models are consistent with a causal interpretation of explanation, there are also models that are programmatic in that interpretation. These are generally *causal models of explanation* and share the following assumption:

(CE) To explain why a particular event *e* (or a particular *phenomenon* or *fact*) occurred we have to refer to its (partial or complete) *cause c*.[50]

The various causal approaches to explanation differ not only in the *concept* of cause used, but also in the other assumptions and criteria that apply to the components of causal explanation.[51] Thus, some approaches view causes in terms of

[50] We have formulated the CE thesis as generally as possible so that it covers a range of causal theories of explanation. Of course, theorists differ in the criteria placed on the components of causal explanation. See, e.g. Salmon (1984); Lewis (1986a); Woodward 2003; Strevens (2008); as well as Skow (2016a). In this book we do not engage with the often-made distinction between theorists who think that all explanations of particular events (phenomena) are *causal explanations*, and those who focus on causal explanations in their own theories but allow for non-causal explanations of particular events. The difference between the first and the second group of approaches lies in the distinction between *causal theories of explanation* and *theories of causal explanation*. See, e.g. Skow (2016b). Jastrzembská (2009) refers to the first group, of which she is a proponent, as "causal chauvinism".

[51] For example, Strevens (2008, Chapter 2) distinguishes between one-factor causal theories, which admit *any* element that exerts *some* causal influence on event *u* as part of its explanation, and two-factor theories (including his own), which, apart from *causal influence*, also refer to a suitable relation of *explanatory relevance* that selects only those *causal influences* that are *relevant* to the explanation.

counterfactual dependence, while others see them in terms of *probabilistic causation*, *causal processes* and *interactions*, or the *manipulationist* view of causation. There are also models that are not tied to a particular theory of causation.

To illustrate the plurality of causal models of explanation, let us briefly introduce three approaches.

In his earlier work, Wesley Salmon considered the *relation of statistical relevance* (SR) to be explanatory (see Salmon 1971), but he later turned to the model of *causal mechanisms* (CM), in which he characterized causal (or etiological) explanation as one which

> "[...] involves the placing of the explanandum in a causal network consisting of relevant causal interactions that occurred previously and suitable causal processes that connect them to the fact-to-be-explained. [...] Etiological explanations [...] explain a given fact by showing how it came to be as a result of antecedent events, processes, and conditions." (Salmon 1984, 269)

Salmon considers a *process* to be *causal* if it can (continuously) *transmit* energy, information or causal influence – generally what he calls a *mark* – in space and time. (Examples of causal processes are the transmission of sound waves with a certain frequency through space-time, the flight of a ball along a trajectory with a given momentum, and a window that conserves, in space-time, certain physical properties related to its structure etc.) If two causal processes intersect at a certain point in space and time and this intersection *changes* (some of) their properties, we say there has been a *causal interaction* (see Salmon 1984, 146 and 170–171).[52] For example, using Salmon's simple example (see Salmon 1984, 178), the explanation of why the window broke would include information about a baseball bat hitting a ball (a causal interaction), which led to the ball changing its trajectory and momentum (a causal process), and eventually hitting the window (another causal interaction), thus breaking it (i.e. changing some of the window's properties).

[52] For more on Salmon's model of causal mechanisms and a critique of it, see Weber et al. (2013, Section 1.6), Woodward (2017, Section 4) and Zeleňák (2008, Section 6.2).

David Lewis formulated an alternative approach to causal explanation. Central to his account is the idea that

> "to explain an event is to provide some information about its causal history." (Lewis 1986a, 217)

The information provided by an explanation can include either (a) *one of the causes* from the causal history of the given event, or (b) *several of the causes* of this event. In the latter case, the individual causes can be (b.1) events that are more or less simultaneous and *causally independent* of each other, together causing the explanandum event, or (b.2) a *causal chain* of events, in which the first is the cause of the second, the second is the cause of the third etc., with the explanandum event at the end of the chain, or (b.3) a *complex branching* structure of events (see Lewis 1986a, 219). Explanatory information thus comes in various forms and levels of complexity. If we had information about the entire causal history of an event, we would also have a *complete explanation*. However, Lewis thinks this is probably an unattainable goal. Our inability to provide a complete explanation does not particularly worry him though, since we are typically only interested in the immediate or partial causes when explaining events. For example, when explaining why a car went off the road and hit a tree, the information that the road was covered in ice and the car skidded would be explanatorily sufficient. Naturally, if we add the information that the driver was driving under the influence (e.g. of alcohol), a more *complex* causal explanation will result.

Woodward's *manipulationist account* (see Woodward 2003) is another model of causal explanation. It is a complex theory of causal explanation that relies on the (causal) concepts of *intervention*, *variables*, *manipulation*, *counterfactual dependence* and *invariance*. Due to space constraints, we cannot deal with Woodward's model in any detail, but we will provide a general description.

According to Woodward, a causal explanation is obtained by identifying the factors or conditions in which a change leads to another change and that change is the object of the explanation (see Woodward 2003, 10). More precisely,

> "[T]he explanation must enable us to see what sort of difference it would have made for the explanandum if the factors cited in the

> explanans had been different in various possible ways. We can also think of this as information about a pattern of counterfactual dependence between explanans and explanandum [...]" (Woodward 2003, 11)

Causal explanation thus refers to conditions that can be represented as *variables* (of some type). Take the simplest possible schematic scenario. Let variable X represent the conditions or factors referred to in the explanans. In our scenario, X ranges over two different values, x_1 and x_2. Let Y represent the conditions appearing in the explanandum; Y will also range over two different values, y_1 and y_2. Let I represent an *intervention* on X. Suppose that Y currently has the value y_1, i.e. $Y = y_1$, and we ask why that is so. A causal explanation of the fact that $Y = y_1$ would include the following information: (i) $X = x_1$; (ii) if intervention I changed the value of $X = x_1$ to $X = x_2$, then the value of $Y = y_1$ would change to value $Y = y_2$; (iii) the other conditions are *constant*; and (iv) the relation between X and Y (between the changes in their values) is invariant (relative to a set of interventions). According to the manipulationist account a state of affairs can be explained by providing information about the factors in which a change would lead (assuming that the other conditions remain unchanged) to a change in the state of affairs being explained.

The three models of causal explanation discussed above share assumption (CE) as their common theoretical starting point. They differ, first and foremost, in the criteria applied to the *causes*, their *identification* and their *relation* to the explanandum. Moreover, all approaches allow not only for a *strict*, deterministic relation between cause and effect, but also for the *probabilistic version*. Therefore, they are also applicable in contexts that have been the traditional domain of *probabilistic models* of explanation (i.e. the I-S and SR models). An additional advantage of causal approaches lies in the fact that the *asymmetry* between the explanans and explanandum is inherent to their apparatus, since the causal relation is itself asymmetric (i.e. if c is the cause of e, then it is not true that e is the cause if c). These models thus avoid one of the main issues facing the D-N model of explanation. Causal models also eliminate the problem of causally *irrelevant* factors, noted in

our discussion of D-N explanation. Such factors by definition cannot be part of the explanatory information involved in a causal explanation.[53]

Causal approaches thus represent a relatively straightforward solution to most of the problems with the previous models of explanation discussed above. In some causal models (such as Lewis'), the *explanation* of an event may include *less information* than would be required by the D-N (or I-S or SR) model, since a causal explanation may also refer to a *partial* cause, while the D-N model requires information on the (causally) *sufficient* conditions (i.e. on the *complete cause*). On the other hand, in the manipulationist account (Woodward 2003), the *explanatory information* is *richer* than that required in the D-N model, as in addition to the information on the *actual state* (conditions), it has to contain *modal information* about other *possible states* (conditions).

6.5 Concluding remarks

Our brief and selective account of some of the main models shows that a scientific explanation of a particular phenomenon *p* may lie within the *theoretical context* it is part of, represented by suitable *laws* (universal or statistic), or refer to the factors that represent the (partial or complete) *cause* of the phenomenon. Whether a model can potentially be applied to a given case largely depends on the specific conditions of the *theoretical context* of the explanandum. For example, if phenomenon *p* belongs to the domain of a *theoretically developed* discipline and a (sufficiently abstract) description of phenomenon *p* allows us to subsume it under the *relevant law*, then the D-N model may serve as a natural framework for explaining *p*. On the other hand, if the object of explanation is a phenomenon *p* that is described in detail and we can identify its causes using a causal model without referring to a universal or probabilistic law, then a causal explanation will be suitable.

[53] However, as Strevens (2008) points out, not all *causally relevant factors* of event *e* are *explanatorily relevant*. Therefore, a model of causal explanation should select only the minimally sufficient subset of causal factors that have explanatory relevance.

Issues of scientific explanation and its adequate reconstruction, using certain theoretical models, remain a matter of lively debate. We refer the reader to other work that discusses the *unification model* (see Friedman 1974; Kitcher 1989), the *deductive-nomological model of probabilistic explanation* (see Railton 1978) and the *pragmatic* approach to explanation (see van Fraassen 1980).

Study questions

1. Briefly characterize how a particular event can be explained using the D-N model.
2. Pick at least one standard counterexample to the D-N model and briefly characterize it.
3. What constitutes an I-S explanation?
4. State the basic assumption common to a range of causal models of explanation.

REFERENCES

Aliseda, A. 2006. *Abductive Reasoning: Logical Investigations into Discovery and Explanation*. Dordrecht: Springer.

Bartha, P. 2010. *By Parallel Reasoning*. Oxford: Oxford University Press.

Beaney, M. 2015. Analysis. In: Zalta, E. N. (Ed.) *The Stanford Encyclopedia of Philosophy*. Spring 2015 Edition. Available at: <https://plato.stanford.edu/archives/spr2015/entries/analysis/>.

Berka, K. 1983. *Measurement: Its Concepts, Theories and Problems*. Dordrecht: D. Reidel.

Bielik, L. 2018. Explication, HD confirmation and Simplicity. *Erkenntnis*, 83 (5), pp. 1085–1104.

Bielik, L. – Gahér, F. – Zouhar, M. 2010. O definíciách a definovaní [On Definitions and Defining]. *Filozofia*, 65 (8), pp. 719–737.

Bielik, L. – Kosterec, M. – Zouhar, M. 2014a. Model metódy (1): Metóda a problém [Model of Method (1): Method and Problem]. *Filozofia*, 69 (2), pp. 105–118.

Bielik, L. – Kosterec, M. – Zouhar, M. 2014b. Model metódy (2): Inštrukcia a imperatív [Model of Method (2): Instruction and Imperative]. *Filozofia*, 69 (3), pp. 197–211.

Bielik, L. – Kosterec, M. – Zouhar, M. 2014c. Model metódy (3): Inštrukcia a metóda [Model of Method (3): Instruction and Method]. *Filozofia*, 69 (8), pp. 637–652.

Bielik, L. – Kosterec, M. – Zouhar, M. 2014d. Model metódy (4): Aplikácia a klasifikácia [Model of Method (4): Application and Classification]. *Filozofia*, 69 (9), pp. 737–751.

Bogen, J. – Woodward, J. 1988. Saving the Phenomena. *The Philosophical Review*, 97 (3), pp. 303–352.

Braithwaite, R. 1953. *Scientific Explanation*. Cambridge: Cambridge University Press.

Brennan, A. 2017. Necessary and Sufficient Conditions. In: Zalta, E. N.

(Ed.) *The Stanford Encyclopedia of Philosophy*. Summer 2017 Edition. Available at: <https://plato.stanford.edu/archives/sum2017/entries/necessary-sufficient/>.

Broad, W. J. 2008. Hair Analysis Deflates Napoleon Poisoning Theories. Available at: <https://www.nytimes.com/2008/06/10/science/10napo.html>.

Bunge, M. 1996. *Finding Philosophy in Social Science*. New Haven – London: Yale University Press.

Bunge, M. 2005a. *Philosophy of Science, Vol. 1: From Problem to Theory*. New Brunswick – London: Transaction Publishers.

Bunge, M. 2005b. *Philosophy of Science, Vol. 2: From Explanation to Justification*. New Brunswick – London: Transaction Publishers.

Carnap, R. 1962. *Logical Foundations of Probability*. Chicago: The University of Chicago Press, 2nd edition.

Cartwright, N. 1979. Causal Laws and Effective Strategies. *Nous*, 19 (4), pp. 419–437.

Čermák, F. 2001. *Jazyk a jazykověda [Language and Linguistics]*. Praha: Nakladatelství Karolinum.

Černík, V. – Viceník, J. 2011. *Úvod do metodológie spoločenských vied [An Introduction to the Methodology of Social Science]*. Bratislava: Iris.

Childers, T. 2013. *Philosophy & Probability*. Oxford: Oxford University Press.

Čížek, F. et al. 1969. *Filosofie, metodologie, věda [Philosophy, Methodology, Science]*. Praha: Nakladatelství Svoboda.

Cmorej, P. 2000. Úvod do problematiky metodológie vied (III) [An Introduction to the Methodology of Science (III)]. *Organon F*, 7 (3), pp. 326–337.

Cmorej, P. 2001a. *Úvod do logickej syntaxe a sémantiky [An Introduction to Logical Syntax and Semantics]*. Bratislava: Iris.

Cmorej, P. 2001b. Úvod do problematiky metodológie vied (IV) [An Introduction to the Methodology of Science (IV)]. *Organon F*, 8 (1), pp. 79–90.

Cohen, M. R. – Nagel, E. 1934. *An Introduction to Logic and Scientific Method*. New York: Harcourt, Brace and Co.

Collingwood, R. G. 1940. *An Essay in Metaphysics*. Oxford: Oxford University Press.

Crupi, V. 2015. Confirmation. In: Zalta, E. N. (Ed.) *The Stanford Encyclopedia of Philosophy*. Fall 2015 Edition. Available at: <https://plato.stanford.edu/archives/fall2015/entries/confirmation/>.

Douven, I. 2017. Abduction. In: Zalta, E. N. (Ed.) *The Stanford Encyclopedia of Philosophy*. Summer 2017 Edition. Available at: <https://plato.stanford.edu/

archives/sum2017/entries/abduction/>.

Durkheim, E. 2005. *Suicide: A Study in Sociology*. London – New York: Routledge.

Earman, J. 1992. *Bayes or Bust? A Critical Examination of Bayesian Confirmation*. Cambridge, ma: The mit Press.

Earman, J. – Salmon, W. 1999. The Confirmation of Scientific Hypotheses. In: Salmon, M. et al. (Eds.) *Introduction to the Philosophy of Science*. Indianapolis: Hackett Publishing Company. pp. 42–103.

Eells, E. 1991. *Probabilistic Causality*. Cambridge: Cambridge University Press.

Filkorn, V. 1960. *Úvod do metodológie vied [An Introduction to the Methodology of Science]*. Bratislava: Vydavateľstvo Slovenskej akadémie vied.

Filkorn, V. 1972. Pojem metódy [The Concept of Method]. *Filozofia*, 27 (3), pp. 225–244.

Filkorn, V. 1998. *Povaha súčasnej vedy a jej metódy [The Nature of Contemporary Science and Its Methods]*. Bratislava: Veda.

van Fraassen, B. 1980. *The Scientific Image*. New York: Oxford University Press.

Friedman, M. 1974. Explanation and Scientific Understanding. *The Journal of Philosophy*, 71 (1), pp. 5–19.

Gahér, F. 2003. *Logika pre každého [Logic for Everyone]*. Bratislava: Iris.

Gahér, F. 2011. Sú pojmy dostatočná podmienka a nutná podmienka pre empirickú oblasť symetrické? [Are the Concepts of Necessary and Sufficient Condition Symmetric in the Empirical Domain?]. *Organon F*, 18 (3), pp. 331–350.

Gahér, F. 2012. Revízia definícií pojmov dostatočná a nutná podmienka [A Revision of the Definitions of Sufficient and Necessary Conditions]. *Organon F*, 19 (1), pp. 16–37.

Gahér, F. 2016. Metóda ako procedúra [Method as Procedure]. *Filozofia*, 71 (8), pp. 629–643.

Gahér, F. – Marko, V. 2017. *Metóda, problém a úloha [Method, Problem, and Task]*. Bratislava: Univerzita Komenského v Bratislave.

Giere, R. – Bickle, J. – Mauldin, R. 2006. *Understanding Scientific Reasoning*. Belmont: Wadsworth, 4th edition.

Gillies, D. 2000. *Philosophical Theories of Probability*. New York – London: Routledge.

Glavaničová, D. 2017. Definície z pohľadu hyperintenzionálnej sémantiky [Definitions in Hyperintensional Semantics]. *Filozofia*, 72 (1), pp. 15–23.

Good, I. J. 1961a. A Causal Calculus I. *British Journal for the Philosophy of Science*, II (44), pp. 305–318.

Good, I. J. 1961b. A Causal Calculus II. *British Journal for the Philosophy of Science*,

12 (45), pp. 43–51.
Gott, R. – Duggan, S. 2003. *Understanding and Using Scientific Evidence*. Thousand Oaks, CA: SAGE Publications.
Gustason, W. 1994. *Reasoning from Evidence*. New York: Macmillan College Publishing Company.
Halas, J. 2015a. Abstrakcia a idealizácia ako metódy spoločensko-humanitných disciplín [Abstraction and Idealization as Methods of Social Science and the Humanities]. *Organon F*, 22 (1), pp. 71–89.
Halas, J. 2015b. Abstrakcia a idealizácia vo filozofii vedy 1 [Abstraction and Idealization in the Philosophy of Science 1]. *Filozofia*, 70 (7), pp. 546–559.
Halas, J. 2015c. Abstrakcia a idealizácia vo filozofii vedy 2 [Abstraction and Idealization in the Philosophy of Science 2]. *Filozofia*, 70 (8), pp. 633–646.
Halas, J. 2016a. *Abstrakcia a idealizácia [Abstraction and Idealization]*. Bratislava: Univerzita Komenského.
Halas, J. 2016b. Weber's Ideal Types and Idealization. *Filozofia nauki*, 24 (1), pp. 5–26.
Hammack, R. 2009. *Book of Proof*. Richmond: Virginia Commonwealth University. Available at: <http://www.people.vcu.edu/~rhammack/BookOfProof/BookOfProof.pdf>.
Hanzel, I. 2008. Idealizations and Concretizations in Laws and Explanations in Physics. *Journal of the General Philosophy of Science*, 39 (2), pp. 273–301.
Hanzel, I. 2015. Idealizations, Ceteris Paribus Clauses, Idealizational Laws. *Filozofia nauki*, 23 (1), pp. 5–26.
Harman, G. 1965. The Inference to the Best Explanation. *Philosophical Review*, 74 (1), pp. 88–95.
Hausman, D. – Woodward, J. 1999. Independence, Invariance, and the Causal Markov Condition. *British Journal for the Philosophy of Science*, 50 (4), pp. 521–583.
Hawthorne, J. 2017. Inductive Logic. In: Zalta, E. N. (Ed.) *The Stanford Encyclopedia of Philosophy*. Spring 2017 Edition. Available at: <https://plato.stanford.edu/archives/spr2017/entries/logic-inductive/>.
Hempel, C. G. 1942. The Function of General Laws in History. *The Journal of Philosophy*, 39i (2), pp. 35–48.
Hempel, C. G. 1945. Studies in the Logic of Confirmation. *Mind*, 54 (213–214), pp. 1–26; 97–121.
Hempel, C. G. 1962. Deductive-Nomological vs. Statistical Explanation. In: Feigl, H. – Maxwell, G. (Eds.) *Minnesota Studies in the Philosophy of Science, Vol. III.* Minneapolis: University of Minnesota Press. pp. 98–169.

Hempel, C. G. 1966. *Philosophy of Natural Sciencei*. Englewood Cliffs, NJ: Prentice Hall.

Hempel, C. G. – Oppenheim, P. 1948. Studies in the Logic of Explanation. *Philosophy of Science*, 15 (2), pp. 135–175.

Hempel, C. G. 1965. Aspects of Scientific Explanation. In: Hempel, C. G. *Aspects of Scientific Explanation and Other Essays in Philosophy of Science*. New York: The Free Press. pp. 331–496.

Hesse, M. 1966. *Models and Analogies in Science*. University of Notre Dame Press: Notre Dame.

Hitchcock, C. 2018. Probabilistic Causation. In: Zalta, E. N. (Ed.) *The Stanford Encyclopedia of Philosophy*. Fall 2018 Edition. Available at: <https://plato.stanford.edu/archives/fall2018/entries/causation-probabilistic/>.

Howson, C. – Urbach, P. 2006. *Scientific Reasoning. The Bayesian Approach*. Chicago: Open Court, 3 edition.

Hume, D. 1975. *An Enquiry Concerning Human Understanding*. Oxford: Clarendon Press.

Hurley, P. J. 2006. *A Concise Introduction to Logic*. Belmont: Wadsworth.

Hájek, A. 2012. Interpretations of Probability. In: Zalta, E. N. (Ed.) *The Stanford Encyclopedia of Philosophy*. Winter 2012 Edition. Available at: <https://plato.stanford.edu/archives/win2012/entries/probability-interpret/>.

Hájek, A. – Joyce, J. 2008. Confirmation. In: *The Routledge Companion to Philosophy of Science*. New York – London: Routledge. pp. 115–128.

Illari, P. – Russo, F. 2014. *Causality. Philosophical Theory Meets Scientific Practice*. Oxford: Oxford University Press.

Jastrzembská, Z. 2007. *Kauzální aspekty vysvětlení [Causal Aspects of Explanation]*. Brno: Masarykova univerzita.

Jastrzembská, Z. 2009. *Aspekty vysvětlení: Hledání explanačních znalostí [Aspects of Explanation: The Search for Explanatory Knowledge]*. Olomouc: Nakladatelství Olomouc.

Johannson, L. G. 2016. *Philosophy of Science for Scientists*. Dodrecht: Springer.

Jones, M. R. 2005. Idealization and Abstraction: A Framework. In: Jones, M. R. – Cartwright, N. (Eds.) *Idealization XII. Correcting the Model*. Amsterdam: Rodopi. pp. 173–217.

Kim, J. 1971. Causes and Events: Mackie on Causation. *The Journal of Philosophy*, 68 (14), pp. 429–441.

Kitchener, R. F. 1999. *The Conduct of Inquiry*. Lanham, MD: University Press of

America.
Kitcher, P. 1989. Explanatory Unification and the Causal Structure of the World. In: Kitcher, P. – Salmon, W. C. (Eds.) *Scientific Explanation. Minnesota Studies in the Philosophy of Science. Vol. xiii*. Minneapolis: University of Minnesota Press. pp. 410–505.
Kitcher, P. 1982. *Abusing Science. The Case against Creationism*. Cambridge, ma: The mit Press.
Kosterec, M. 2016. Analytic Method. *Organon F*, 23 (1), pp. 83–101.
Kuipers, T. A. F. 2007. Introduction. Explication in Philosophy of Science. In: Kuipers, T. A. F. (Ed.) *Handbook of the Philosophy of Science: General Philosophy of Science – Focal Issues*. Amsterdam: Elsevier. pp. vii–xxiii.
Kumar, R. 2011. *Research Methodology*. Thousand Oaks, ca: SAGE, 3 edition.
Kutach, D. 2014. *Causation*. Cambridge: Polity Press.
Lange, M. 2017. *Because without Cause. Non-Causal Explanations in Science and Mathematics*. New York: Oxford University Press.
Laudan, L. 1983. The Demise of the Demarcation Problem. In: Cohen, R. S. – Laudan, L. (Eds.) *Physics, Philosophy and Psychoanalysis*. Dordrecht: D. Reidel. pp. 111–127.
Lewis, D. 1973. *Counterfactuals*. Cambridge, ma: Harvard University Press.
Lewis, D. 1986a. Causal Explanation. In: Lewis, D. *Philosophical Papers, Vol. ii*. Oxford: Oxford University Press. pp. 214–240.
Lewis, D. 1986b. Causation. In: Lewis, D. *Philosophical Papers, Vol. ii*. Oxford: Oxford University PRess. pp. 159–213.
Lipton, P. 2004. *Inference to the Best Explanation*. London: Routledge.
Losee, J. 2001. *A Historical Introduction to the Philosophy of Science*. Oxford: Oxford University Press.
Mackie, J. L. 1974. *The Cement of the Universe: A Study of Causation*. Oxford: Clarendon Press.
Mahner, M. 2007. Demarcating Science from Non-Science. In: Kuipers, T. A. F. (Ed.) *Handbook of the Philosophy of Science: General Philosophy of Science – Focal Issues*. Amsterdam: Elsevier. pp. 515–575.
Mahner, M. 2013. Science and Pseudoscience. In: Pigliucci, M. – Boudry, M. (Eds.) *Philosophy of Pseudoscience. Reconsidering the Demarcation Problem*. Chicago – London: The University of Chicago Press. pp. 29–43.
Materna, P. 2004. *Conceptual Systems*. Berlin: Logos Verlag.
Materna, P. 2007. Once More on Analytic vs. Synthetic. *Logic and Logical Philosophy*,

16 (1), pp. 3–43.

McMullin, E. 1985. Galilean Idealization. *Studies in History and Philosophy of Science*, 16 (3), pp. 247–273.

Menzies, P. 2017. Counterfactual Theories of Causation. In: Zalta, E. N. (Ed.) *The Stanford Encyclopedia of Philosophy*. Winter 2017 Edition. Available at: <https://plato.stanford.edu/archives/win2017/entries/causation-counterfactual/>.

Menzies, P. – Price, H. 1993. Causation as a Secondary Quality. *British Journal for the Philosophy of Science*, 44 (2), pp. 187–203.

Mill, J. S. 1886. *System of Logic Ratiocinative and Inductive*. London: Longmans, Green & Co.

Mistrík, J. 2002. *Lingvistický slovník [A Dictionary of Linguistics]*. Bratislava: Slovenské pedagogické nakladateľstvo.

Müller-Wille, S. 2019. Carolus Linnaeus. In: *Encyclopaedia Brittanica*. Available at: <https://www.britannica.com/biography/Carolus-Linnaeus/>.

Nagel, E. 1961. *The Structure of Science. Problems in the Logic of Scientific Explanation*. New York: Harcourt, Brace & World.

Niiniluoto, I. 2004. Truth-Seeking by Induction. In: Stadler, F. (Ed.) *Induction and Deduction in the Sciences*. Dordrecht – Boston – London: Kluwer Academic Publishers. pp. 57–82.

Nola, R. – Irzik, G. 2005. *Philosophy, Science, Education, and Culture*. Dordrecht: Springer.

Nola, R. – Sankey, H. 2000. A Selective Survey of Theories of Scientific Method. In: Nola, R. – Sankey, H. (Eds.) *After Popper, Kuhn and Feyerabend*. Dordrecht: Kluwer Academic Publishers. pp. 1–65.

Nola, R. – Sankey, H. 2007. *Theories of Scientific Method*. Montreal: McGill-Queen's University Press.

Norton, J. 2005. A Little Survey on Induction. In: Achinstein, P. (Ed.) *Scientific Evidence: Philosophical Theories and Applications*. Baltimore: Johns Hopkins University. pp. 9–34.

Ondruš, D. et al. 2006. *Všeobecná a špeciálna onkológia [General and Special Oncology]*. Bratislava: Univerzita Komenského v Bratislave.

Paul, L. 2009. Counterfactual Theories. In: Beebee, H. – Hitchcock, C. – Menzies, P. (Eds.) *The Oxford Handbook of Causation*. Oxford – New York: Oxford University Press. pp. 158–184.

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge

University Press.
Peirce, C. S. 1992. Deduction, Induction and Hypothesis. In: Houser, N. – Kloesel, C. (Eds.) *The Essential Peirce. Selected Philosophical Writings, Vol. 1*. Bloomington: Indiana University Press.
Pickering, W. S. F. – Walford, G. (Eds.). 2000. *Durkheim's* Suicide. *A Century of Research and Debate*. London – New York: Routledge.
Pigliucci, M. – Boudry, M. (Eds.). 2013. *Philosophy of Pseudoscience. Reconsidering the Demarcation Problem*. Chicago – London: The University of Chicago Press.
Pollock, J. L. 1987. Defeasible Reasoning. *Cognitive Science*, 11 (4), pp. 481–581.
Popper, K. R. 2002. *The Logic of Scientific Discovery*. London – New York: Routledge.
Psillos, S. 2002. *Causation and Explanation*. Montreal: McGill-Queen's University Press.
Psillos, S. 2009. Regularity Theories. In: Beebee, H. – Hitchcock, C. – Menzies, P. (Eds.) *The Oxford Handbook of Causation*. Oxford – New York: Oxford University Press. pp. 131–157.
Quine, W. V. O. 1953. Two Dogmas of Empiricism. In: Quine, W. V. O. *From a Logical Point of View*. Cambridge, ma: Harvard University Press. pp. 21–46.
Railton, P. 1978. A Deductive-Nomological Model of Probabilistic Explanation. *Philosophy of Science*, 45 (2), pp. 206–226.
Reichenbach, H. 1956. *The Direction of Time*. Los Angeles: University of California Press.
Riška, A. 1968. *Metodológia a filozofia [Methodology and Philosophy]*. Bratislava: Vydavateľstvo politickej literatúry.
Salmon, M. et al. 1992. *Introduction to the Philosophy of Science*. Upper Saddle River, nj: Prentice-Hall.
Salmon, M. H. 1995. *Introduction to Logic and Critical Thinking*. San Diego: Harcourt Brace College Publishers.
Salmon, W. C. 1971. Statistical Explanation. In: Salmon, W. C. – Jeffrey, R. – Greeno, J. (Eds.) *Statistical Explanation and Statistical Relevance*. Pittsburgh: University of Pittsburgh Press. pp. 29–87.
Salmon, W. C. 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.
Salmon, W. C. 1989. *Four Decades of Scientific Explanation*. Minneapolis: University of Minnesota Press.
Schurz, G. 1991. Relevant Deduction. *Erkenntnis*, 35 (1-3), pp. 391–437.
Schurz, G. 2014. *Philosophy of Science: A Unified Approach*. New York – London:

Routledge.
Scriven, M. 1962. Explanations, Predictions, and Laws. In: Feigl, H. – Maxwell, G. (Eds.) *Minnesota Studies in the Philosophy of Science, Vol. III*. Minneapolis: University of Minnesota Press. pp. 170–230.
Šefránek, J. 1969. *Logika, jazyk a poznanie [Logic, language, and knowledge]*. Bratislava: Epocha.
Skow, B. 2016a. *Reasons Why*. Oxford: Oxford University Press.
Skow, B. 2016b. Scientific Explanation. In: Humphreys, P. (Ed.) *The Oxford Handbook of Philosophy of Science*. New York: Oxford University Press. pp. 524–543.
Skyrms, B. 2000. *Choice and Chance. An Introduction to Inductive Logic*. Belmont: Wadsworth, 4th edition.
Strevens, M. 2008. *Depth: An Account of Scientific Explanation*. Cambridge, ma: Harvard University Press.
Štěpán, J. 2001. *Klasická logika [Classical Logic]*. Olomouc: Nakladatelství Olomouc.
Suppes, P. 1970. *A Probabilistic Theory of Causality*. Amsterdam: North-Holland.
Tuomela, R. 1987. Science, Protoscience, Pseudoscience. In: Pitt, J. C. – Pera, M. (Eds.) *Rational Changes in Sciences*. Dordrecht: Kluwer. pp. 83–101.
Viceník, J. 2000a. Úvod do problematiky metodológie vied (I) [An Introduction to the Methodology of Science (I)]. *Organon F*, 7 (1), pp. 78–89.
Viceník, J. 2000b. Úvod do problematiky metodológie vied (II) [An Introduction to the Methodology of Science (II)]. *Organon F*, 7 (2), pp. 196–209.
Viceník, J. 2001a. Úvod do problematiky metodológie vied (V) [An Introduction to the Methodology of Science (V)]. *Organon F*, 8 (1), pp. 91–103.
Viceník, J. 2001b. Úvod do problematiky metodológie vied (VI) [An Introduction to the Methodology of Science (VI)]. *Organon F*, 8 (2), pp. 197–213.
von Wright, H. 1971. *Explanation and Understanding*. New York: Cornell University Press.
Weber, E. – Bouwel, J. V. – De Vreese, L. 2013. *Scientific Explanation*. Dordrecht: Springer.
Weisberg, M. 2007. Three Kinds of Idealization. *The Journal of Philosophy*, 104 (12), pp. 636–659.
Woodward, J. 2003. *Making Things Happen*. Oxford: Oxford University Press.
Woodward, J. 2009. Agency and Interventionist Theories. In: Beebee, H. – Hitchcock, C. – Menzies, P. (Eds.) *The Oxford Handbook of Causation*. Oxford – New York: Oxford University Press. pp. 234–262.
Woodward, J. 2017. Scientific Explanation. In: Zalta, E. N. (Ed.) *The Stanford Ency-*

clopedia of Philosophy. Fall 2017 Edition. Available at: <https://plato.stanford.edu/archives/fall2017/entries/scientific-explanation/>.

Zeleňák, E. 2008. *Moderné teórie vysvetlenia a príčinnosti [Modern Theories of Explanation and Causation]*. Ružomberok: Katolícka univerzita v Ružomberku.

Zouhar, M. 2008. *Základy logiky pre spoločenskovedné a humanitné odbory [Basics of Logic for Social Science and the Humanities]*. Bratislava: Veda.

Zouhar, M. 2015a. Logická forma definícií [The Logical Form of Definitions]. *Filozofia*, 70 (3), pp. 161–174.

Zouhar, M. 2015b. Metóda definovania [The Method of Defining]. *Filozofia*, 70 (4), pp. 258–271.

Zouhar, M. – Bielik, L. – Kosterec, M. 2017. *Metóda: metodologické a formálne aspekty [Method: Methodological and Formal Aspects]*. Bratislava: Univerzita Komenského v Bratislave.

SUBJECT INDEX

NAME INDEX

ISBN 978-80-223-4782-2