Frameworks, Models, and Case Studies

A New Methodology for Studying Conceptual Change in Science and Philosophy

Dissertation von Matteo De Benedetto

München 2022

Frameworks, Models, and Case Studies

A New Methodology for Studying Conceptual Change in Science and Philosophy

(Inaugural-)Dissertation
zur Erlangung des Grades eines Doktors der Philosophie
an der Fakultät für Philosophie, Wissenschaftstheorie und Religionwissenschaft
der Ludwig-Maximilians-Universität München

vorgelegt/eingereicht von
Matteo De Benedetto

aus
Arezzo, Italien

2022

Erstgutachter:	Prof. DDr. Hannes Leitgeb
Zweitgutachter:	Prof. Dr. Leon Horsten
Tag der mündlichen Prüfung:	17.02.2022

Contents

List of Figures

Zusammenfassung

Diese Dissertation beschäftigt sich mit dem Thema des Begriffswandels und mit den damit verbundenen erkenntnistheoretischen und semantischen Problemen. Jene Probleme betreffen die philosophischen Bereiche der Wissenschaftstheorie, der Philosophie des Geistes, der Sprachphilosophie, der Philosophie der Mathematik, der Erkenntnistheorie, der Metaphysik und der Metaphilosophie.

Das Thema wird mit Hilfe eines Vergleichs zwischen vier Arten von Modellen, die den Wandel von Begriffen abbilden, analysiert. Die vier Arten von Modellen, die benutzt werden, sind pragmatische, evolutionäre, unbestimmte und kognitive. Was das pragmatische Modell betrifft, wird besonders mit der carnapianischen Explikation gearbeitet. Für das evolutionäre Modell wird das darwinsche Modell beispielhaft verwendet. Für das unbestimmte Modell wird auf Mark Wilsons Begriffsmodell zurückgegriffen. Und das kognitive Modell arbeitet mit Thagards Modell der Begriffsrevolution.

Der Vergleich erfolgt anhand von neun verschiedenen Dimensionen, die eine Bewertung von Begriffswandeln möglich machen. Diese Dimensionen nehmen jeweils folgende Bereiche in den Blick: Selektionseinheiten, Begriffsontologie, Begriffsstruktur, Arten des Begriffswandels, Normativität, Urteilseffektivität, Darstellung von wissenschaftlichen Begriffswandeln, Darstellung von philosophischen Begriffswandeln und metaphilosophische Hintergründe.

Nach dem Einführungskapitel werden im zweiten Kapitel einige philosophische Vorüberlegungen zu Begriffen und deren Wandel angestellt. Zuerst sollen dabei die wichtigsten philosophischen und psychologischen Theorien zu Begriffsstruktur und -ontologie behandelt werden. Anschließend werden die Hauptprobleme, die in Verbindung mit philosophischen und wissenschaftlichen Begriffswandeln stehen, vorgestellt.

Das dritte Kapitel nimmt die carnapianische Explikation in den Fokus. Diese wird aus historischer und abstrakter Perspektive analysiert. Außerdem wird eine verbesserte und um eine Stufe erweiterte Version von Carnaps Modell des Begriffswandels entwickelt. Danach wird ein mathematisches Modell von der carnapianischen Explikation erstellt, eingebettet in Gärdenfors' geometrische Begriffstheorie. Zum Schluss wird Carnaps Modell mit den obengenannten neun Dimensionen analysiert.

Das vierte Kapitel beschäftigt sich mit den darwinschen Modellen der Begriffsevolution. Zuerst wird die Geschichte der evolutionären Erkenntnistheorie und deren Verbindung zu den darwinschen Modellen beleuchtet. Anschließend wird ein mathematisches Modell, angelehnt an die darwinschen Modelle, für die mathematische Begriffsevolution entwickelt.

Auch hier wird eine Analyse mit Hilfe der neun Dimensionen vorgenommen.

Im fünften Kapitel wende ich mich den unbestimmten Modellen zu. Dazu werden zunächst zwei Modellbeispiele vorgestellt: Waismanns Modell der Porosität von Begriffen und Mark Wilsons Modell für den Begriffswandel. Zu letzterem Modell wird wieder eine Formalisierung vorgenommen, die auf der strukturalistischen Wissenschaftstheorie aufbaut. Wie in den vorangegangenen Kapiteln werden zum Schluss die neun Dimensionen für die Analyse herangezogen.

Das sechste Kapitel beschäftigt sich mit kognitiven Modellen von Begriffswandeln, insbesondere mit Thagards Modell der Begriffsrevolution. Dazu wird ein logisches Modell aufgestellt, das sich auf die Belief Revision-Theorie bezieht. Mit den neun Dimensionen wird auch dieses Modell analysiert.

Im letzten Kapitel dieser Arbeit werden alle vier Arten von Modellen zusammenfassend verglichen, aufbauend auf den Erkenntnissen aus den vorherigen Kapiteln. Mit Hilfe dieses Vergleichs kann ein neues Verständnis des Phänomens des Begriffswandels möglich gemacht werden. Zum Schluss wird ein Ausblick auf Konsequenzen, die dieses neue Verständnis für die Philosophie hervorrufen kann, gewährt. Außerdem sollen Anknüpfungspunkte für zukünftige Forschung in diesem Bereich geliefert werden.

Co-authored and single-authored publications

Section 3.3 is based on the paper "Explication as a Three-Step Procedure: the case of the Church-Turing Thesis", published in *European Journal for Philosophy of Science* (De Benedetto, 2021a)

Section 3.4 is based on the paper "Explicating 'Explication' via Conceptual Spaces", published in *Erkenntnis*, (De Benedetto, 2020).

Section 5.3 is based on the paper "Taming conceptual wanderings: Wilson-Structuralism", published in *Synthese*, (De Benedetto, 2021b).

The content of Section 6.2 and 6.3 is joint work with Sena Bozdag and it is based on the paper "Taking Up Thagard's Challenge: A Formal Model of Conceptual Revision", published in *Journal of Philosophical Logic*, (Bozdag and De Benedetto, 2022).

Preface

I am usually not a big fan of prefaces or acknowledgments pages, especially in academic books. They often read very fake to me. However, the writing of this dissertation occupied four years of my life and I have so many people to thank for the constant support that I received in these years that I decided to write this preface. Hopefully, it will not read very fake.

Despite the current dominant socio-cultural paradigms prescribe to always depict the PhD life as the most horrible and stressful period in the life of a person, in these four years I must admit that I have been very happy and, with the exception of some short time-frames, quite relaxed. Looking back, I owe most of this happiness and relaxation to the professional and human environment that surrounded me during this time. I really feel spoiled when I think about how nice and caring people have been with me these years. I want to thank to all the people that have been nice to me in these four years.

I will try now to thank personally a proper subset of people that have been particularly important to me during the writing of this dissertation. I must stress that, writing these words, I have the horrible feeling that I will omit someone very important to me. To the omitted people, I want to say: I am very sorry about it, please forgive me. To my partial defence I can bring three factors: my terrible memory, certain recreational activities in which I indulged a little bit too much in my teenage years, and the fact that I am writing this preface one day before the submission deadline.

Obviously, the first person I want to thank for making this dissertation possible is my supervisor, Hannes Leitgeb. I really find hard to describe how good is Hannes as a supervisor. It would be ridiculous to stress how good is Hannes as a philosopher. Everybody who knows him, knows it. Similarly, it should be evident to anyone who reads this thesis how much my work has been influenced by Hannes. What I want to stress, instead, is how fun, human, and caring Hannes has been with me all these years. All the hours of supervision that I had with him were incredibly enjoyable, up to the point that I always looked forward to the next supervision appointment. Moreover, Hannes always helped me with every problem that I faced during my PhD, he supported me in every stupid project I wanted to work on, and he always cheered me up when I was frustrated by the PhD life. Hannes was particularly important for me these years, because he provided a huge counterweight to the mass of despicable, annoying, presumptuous people that fill the ranks of analytic philosophy. If today I believe that it is possible to be a good analytic philosopher and a decent human being, it is mostly because of Hannes. Thanks Hannes.

Next, I want to thank three people that, in different times and in different ways, have acted as my shadow-supervisors: Norbert Gratzl, Lavinia Picollo, and Lorenzo Rossi. Despite not being forced to, Norbert, Lavinia, and Lorenzo helped me a lot with small and big practical problems of my PhD journey. They also repeatedly improved my research plans and products, as well as they always listened to my rants, wisely advising me in lots of important and non-important choices that I made these years. Moreover, they are three outstanding logicians to whom I always looked up and, more importantly, they are three amazing people, the company of whom I enjoy a lot. Thank you to all three of you.

Then, I want to thank Sena Bozdag for being the best academic sister that one could possibly have. Sena and I started the PhD together and we always shared not only the office, but far more importantly the ups and the downs of the PhD life. She always helped me with all the professional and personal problems that annoyed me these years. I really cannot imagine doing my PhD or even working in an office without Sena nearby. Thanks Sena.

Another person that was paramount for the completion of my PhD is Ursula Danninger. Ursula was the first person to welcome me at the MCMP and she was relentless in her help with every challenge that the Bavarian bureaucracy put in my way. Moreover, despite the lack of progresses, she never gave up on my German, being the best German conversation partner I ever had. She was also always full of excellent suggestions for German movies and series to watch. Thank you Ursula.

In general, the MCMP as a whole has been an incredibly stimulating and kind environment for me in these years. I met so many incredible people in the MCMP and I had the luck of spending a lot of time with some truly amazing human beings. I had a lot of nice talks, nice lunches, nice dinners, nice coffees, and nice beers in these years. All these social activities contributed a lot to my aforementioned happiness and relaxation. A partial subset of MCMPeers (very broadly constituted, past members, external members, and visitors included) that I want to personally thank for such lovely activities are: Marianna Antonutti Marfori, Ivano Ciardelli, Benedict Eastaugh, Martin Fischer, Gianluca Grilletti, Michal Hladky, Bruno Jacinto, Dominik Klein, Elio La Rosa, Alessandra Marra, Matias Osta Velez, Thomas Schindler, Tom Sterkenburg, Damian Szmuc, Marta Sznajder, Diego Tajer, Borut Trpin, and Naftali Weinberger. Thank you all.

Speaking of lovely activities, I want to thank also all the people that have been close to me these years. As I stressed before, I have been very happy in the past four years. I want to stress how important was (and still is!) to me to have in my life so many wonderful people that made me forget (and critically downsize the importance of) work-related problems. If I have an imitation of what business people call life-work balance, it is mostly because of you. Specifically, I want to thank my Munich friends (Roberta, Allie, Melanie, Michele, Flavia, Claudio, Caterina, Lara, Alberto), my friends back home (Toni, Vodiz, Spiz, Kob, Leo, Chef), my friends from the university years (Jan, Alessandro, Andrea, Maria, Nic, Carlo, Carmela), and of course my lovely flatmates (Stefan, Miriam, Anas, Franco, Nora). Thank you dear friends, you are amazing and I feel very privileged for knowing you.

My family has been a constant source of love and support throughout all my life. I feel very spoiled to have such a family. In particular, I want to thank my Dad, my Mum, my

Uncle, and my Sister. I love you all.

Finally, I want to thank Miriam who edited parts of this thesis and, more importantly, my life. Thank you Miri.

Chapter 1

Introduction

The topic of this thesis is conceptual change, broadly understood as the many philosophically interesting ways in which our concepts (in the intuitive sense of the term) change and the many epistemological and semantic problems connected with these changes.

As it will become clear later, this thesis cannot be easily boxed in one of the main standard sub-fields of analytic philosophy. This is partly the fault of the chosen topic. Philosophical problems connected with conceptual change encompass in fact several philosophical subfields, such as general philosophy of science, philosophy of cognitive science, philosophy of language, philosophy of mathematics, epistemology, metaphysics, metaphilosophy, and the history of philosophy and science.

Conceptual change is traditionally considered primarily a topic in general philosophy of science, due to the worries that the existence of conceptual change prompts for ideas of scientific progress and objectivity. Concepts are also usually considered to be the cognitive substrata of several important higher-cognitive tasks such as memory, abstraction, categorization, and inferential behavior. As such, the mechanisms behind conceptual dynamics have been, and are, heavily debated in cognitive science and its philosophy. Moreover, concepts are of course closely connected with linguistic predicates and, as such, changes in concepts often correspond to philosophically interesting changes in the related linguistic practices. Conceptual change has also been a source of philosophical discussions in philosophy of mathematics, especially in connection with the birth and development of the philosophy of mathematical practice. The extent and the nature of scientific and philosophical conceptual change has been (and still is!) at the center of debates between contrasting epistemological, metaphysical, and metaphilosophical pictures. Finally, conceptual change is a central component of many important episodes of scientific and philosophical change, making it a well-studied phenomenon in the history of science and philosophy.

It would be unfair to blame just the topic of this thesis for the thesis quirks, though. Part of why this work is not easily boxed in one of the aforementioned subfields of analytic philosophy is the specific way in which I approached its subject-matter. Conceptual change is analyzed in this thesis, in fact, primarily via the development of improved models of conceptual change and a new methodology of how to assess them, compare them, and judge them. My analysis will be carried out on four prominent kinds of model of conceptual

change. For each of these four kinds of model, I will develop a formal improved model of conceptual change that builds on a paradigmatic specimen of the kind of model under focus. The methodological analysis of these four kinds of model will be carried out through a novel meta-framework for judging models of conceptual change that I will call the Toolbox framework. This meta-framework consists of nine different evaluative dimensions with respect to which models of conceptual change can be judged, assessed, and compared.

The methodology that I will use in this work is non-standard within analytic philosophy. My judgments on conceptual change and the best ways of modeling this phenomenon will not be mainly based on arguments. I will not start defending or attacking specific conceptions of my subject-matter, nor will I break the phenomenon of conceptual change into a list of theses and sub-theses on its specific aspects.

My main activity in this thesis will be model-building, understood as the development of abstract and idealized representations of a given phenomenon. I will build several models of conceptual change, focusing on the formal or informal improvement of one prominent model for each of the four types of model that will appear in this thesis. In order to formally improve and characterize specific models of conceptual change, I will employ a vast array of formal tools from contemporary logic, mathematics, and cognitive science. This activity of building improved models of conceptual change is the first, lower, layer of my methodology. The second layer consists of reconstructing significant historical episodes within a given model. I will use the reconstruction of case studies as a way of testing the adequacy and the fruitfulness of a given (type of) model of conceptual change. A good model of conceptual change ought to adequately reconstruct significant episodes of its subject-matter from the history of science and philosophy. The third, and more abstract, layer of my methodology is the collective assessment, comparison, and judgment of several types of model of conceptual change. This third layer will be based on the two lower levels of the methodology, i.e. on the adequacy of a given type of model in reconstructing historical episodes of conceptual change. The results of this collective assessment, comparison, and judgment will offer a general conception of conceptual change as a philosophical phenomenon. Despite the evident lack of orthodoxy of my methodology, at least for the canon of analytic philosophy, the strong focus on abstract model-building and idealized reconstructions of case studies is reminiscent of certain sub-tradition of analytic philosophy such as Logical Empiricism and (some kind of) Historicism. In particular, two philosophers whose methodologies have strong similarities with the present work are Carnap and Lakatos. Not surprisingly, their influence permeates all the parts of this thesis and their work has been a constant source of inspiration for me since the beginning of my studies.

The four types of model of conceptual change that will be analyzed in this thesis are pragmatic, evolutionary, indeterminate, and cognitive models. These four types of model will be judged and compared along the nine evaluative dimensions of my Toolbox framework, by virtue of which a given type of model of conceptual change can be analyzed with respect to its units of selection, concept ontology, concept structure, kinds of conceptual change, normativity, effectiveness of normative judgments, picture of scientific conceptual change, picture of philosophical conceptual change, and metaphilosophical background.

In order to judge these four different types of model of conceptual change, I will focus

primarily on a single paradigmatic instance of each model. I will focus on Carnapian explication for what concerns pragmatic models, while my specimen for evolutionary models will be Darwinian models based on conceptual populations. As a prominent indeterminate model of conceptual change I will take Mark Wilson's frameweork of patches and facades, while Thagard's model of conceptual revolution will be my main example of cognitive models. In addition to these four specimen, several others specific models of conceptual change will be presented and briefly analyzed. These models include Campbell's selective variation model, Popper's active learning model, Toulmin's conceptual population model, Hull's model of conceptual evolution, Lakatos' concept-stretching, Mormann's model of axiomatic variation, Waismann's open-texture, Andersen's, Barker's, and Chen's neo-Kuhnian model of scientific revolution, Kornmesser's and Schurz's theory-frame model, and Gärdenfors' and Zenker's model of scientific change based on conceptual spaces.

In order to analyze and compare the conception of conceptual change provided by these different models, I will rely on several historical reconstructions of episodes of scientific conceptual change. The historical episodes of scientific change that will figure in this work include the emergence of the morphological concept of fish in biological taxonomies, the development of scientific conceptions of temperature, the Church-Turing thesis and related axiomatizations of effective calculability, the history of the concept of polyhedron in 17th and 18th century mathematics, Hamilton's invention of the quaternions, the history of the pre-abstract group concepts in 18th and 19th century mathematics, the expansion of Newtonian mechanics to viscous fluids forces phenomena, and the chemical revolution.

I will also present five different formal and informal improvements of four specific models of conceptual change. I will first present two different improvements of Carnapian explication, a formal and an informal one. My informal improvement of Carnapian explication will consists of a more fine-grained version of the procedure that adds an intermediate, third step to the two steps of Carnapian explication. I will show how this novel three-step version of explication is more suitable than its traditional two-step relative to handle complex cases of explications. My second, formal improvement of Carnapian explication will be a full explication of the concept of explication itself within the theory of conceptual spaces. By virtue of this formal improvement, the whole procedure of explication together with its application procedures and its pragmatic desiderata will be reconceptualized as a precise procedure involving topological and geometrical constraints inside the theory of conceptual spaces. My third improved model of conceptual change will consist of a formal explication of Darwinian models of conceptual change that will make vast use of Godfrey-Smith's population-based Darwinism for targeting explicitly mathematical conceptual change. My fourth improvement will be dedicated instead to Wilson's indeterminate model of conceptual change. I will show how Wilson's very informal framework can be explicated within a modified version of the structuralist model-theoretic reconstructions of scientific theories. Finally, the fifth improved model of conceptual change will be a belief-revision-like logical framework that reconstructs Thagard's model of conceptual revolution as specific revision and contraction operations that work on conceptual structures.

At the end of this work, a general conception of conceptual change in science and philosophy will emerge, thanks to the combined action of the three layers of my methodol-

ogy. This conception takes conceptual change to be a multi-faceted phenomenon centered around the dynamics of groups of concepts. According to this conception, concepts are best reconstructed as plastic and inter-subjective entities equipped with a non-trivial internal structure and subject to a certain degree of localized holism. Furthermore, conceptual dynamics can be judged from a weakly normative perspective, bound to be dependent on shared values and goals. Conceptual change is then best understood, according to this conception, as a ubiquitous phenomenon underlying all of our intellectual activities, from science to ordinary linguistic practices. As such, conceptual change does not pose any particular problem to value-laden notions of scientific progress, objectivity, and realism. At the same time, this conception prompts all our concept-driven intellectual activities, including philosophical and metaphilosophical reflections, to take into serious consideration the phenomenon of conceptual change. An important consequence of this conception, and of the analysis that generated it, is in fact that an adequate understanding of the dynamics of philosophical concepts is a prerequisite for analytic philosophy to develop a realistic and non-idealized depiction of itself and its activities.

In Chapter 2, I will present several philosophical preliminary discussions on concepts and conceptual change broadly understood. I will first survey the philosophical and psychological literature about concepts, discussing the main views of concept ontology and structure. Then, I will present the main problems that the phenomenon of conceptual change poses for scientific progress, objectivity, and realism. I will focus especially on the different kinds of models of scientific conceptual change that have been developed in the related literature. I will then discuss the problems connected to philosophical conceptual change and the related metaphilosophical debates over conceptual analysis and conceptual engineering. In the light of some recent debates over externalism in semantics and the possibility of changing concepts, I will offer a defense of the reality of conceptual change together with a plead for meaning pluralism and meta-semantic plasticity in matters of conceptual dynamics. This plead will also constitute a philosophical justification of the novel methodology developed in this work. At the end of the chapter, I will present what I will call the Toobox framework, i.e. a meta-framework for assessing and comparing different models of conceptual change along nine evaluative dimensions.

Chapter 3 will be focused on pragmatic models of conceptual change and in particular on Carnapian explication. I will first give an in-depth historical analysis of the development of Carnapian explication and Carnap's related metaphilosophical background. Then, I will analyze the procedure of Carnapian explication from an abstract epistemological point of view, focusing especially on the desiderata that a certain explicatum must possess and on related recent philosophical debates on the shortcomings of explication as a model of conceptual change. Next, I will present my refined three-step version of Carnapian explication that adds to the traditional two steps a third, intermediate step dedicated to the semi-formal sharpening of the clarified explicandum. This new step allows my refined version of Carnapian explication to carefully distinguish between different explications of a given concept even when two explications clarify the same explicandum in the same way. I will demonstrate the fruitfulness of my three-step version of Carnapian explication with the help of a detailed case study on the Church-Turing thesis and related explications of

our concept of effective calculability. Then, I will present my general explication of the concept of explication itself within the theory of conceptual spaces. I will illustrate how my proposal is able to frame the procedure of explication and its desiderata in the formal framework of conceptual spaces, showing how specific readings of explication desiderata can be framed as topological constraints on the related conceptual spaces. I will also argue that my framework is able to defend the usefulness of Carnapian explication against some recent critiques. In order to build my case and to show the virtues of my proposal, I will reconstruct two paradigmatic cases of explication in my framework: the emergence of the morphological concept of fish and the development of the scientific concepts of temperature. Finally, at the end of this chapter, I will analyze Carnapian explication in the Toolbox framework, evaluating this prominent pragmatic model of conceptual change along the nine dimensions of my meta-framework.

In Chapter 4, I will focus on evolutionary models of conceptual change and in particular on Darwinian models. I will first describe the historical background behind the ideal of evolutionary epistemology and I will explain the connection with Darwinian models of conceptual change. Then, I will present four different influential Darwinian models of scientific change: Campbell's blind variation and selective retention model, Popper's goal-driven trial and error model, Toulmin's model of conceptual populations, and Hull's model of scientific evolution. I will briefly summarize the debate about the viability of such evolutionary models of scientific change, defending the need of more specific and precise, historically testable evolutionary models of scientific change. After that, I will present a novel formal model of conceptual evolution specifically designed to tackle mathematical conceptual change. My framework will be built upon Mormann's evolutionary model of mathematical conceptual change and Godfrey-Smith's population-based Darwinism. I will show how this framework, centered around the contrasting notions of Lakatosian and Euclidean populations, as well as the spatial tools of what I will call the Lakatosian space, is able to adequately model the plurality of conceptual evolutions that historical episodes of mathematical conceptual change exhibit. I will also show how my framework is able to give a normative assessment of the rationality of a given mathematical conceptual history in terms of mathematical selection or drift. In order to show the usefulness of my framework, I will apply it to three different episodes of conceptual change in mathematics: the history of the concept of polyhedron, the invention of the quaternions, and the development of the pre-abstract group concepts. At the end of the chapter, I will analyze Darwinian models of conceptual change such as my framework for mathematical conceptual evolution in the Toolbox framework. I will thus judge such evolutionary models of conceptual change along nine evaluative dimensions.

The focus of Chapter 5 will be on what I will call indeterminate models of conceptual change. I will first describe what I mean with this label and then I will present two paradigmatic example of this kind of models: Waismann's open-texture model and Mark Wilson's patches and facades framework. Next, I will offer a formal reconstruction of Wilson's indeterminate model of conceptual change within a modified version of the structuralist model-theoretic reconstruction of scientific theories. Specifically, I will show how my modified structuralist framework, i.e. what I will call Wilson-Structuralism, is able

to explicate Wilson's patches and facades as Theory-Elements and Wilson-Theory-Nets. I will also show how, within my framework, one can give a precise semantic understanding of many conceptual wanderings described by Wilson as specific set-theoretic relationships between Theory-Elements. In order to demonstrate the faithfulness of my reconstruction of Wilson's framework, I will reconstruct in Wilson-Structuralism one of Wilson's main case studies of conceptual wanderings, i.e. viscous fluids forces in classical mechanics. Finally, at the end of this chapter, I will analyze indeterminate models of conceptual change along the nine dimensions of the Toolbox framework.

In Chapter 6, I will focus on cognitive models of conceptual change. First, I will describe three influential instances of this kind of models: Thagard's model of conceptual revolution, Anderson's, Barker's and Chen's frame-based Neo-Kuhnian model of scientific revolutions, and Gärdenfors' and Zenker's model of scientific change based on conceptual spaces. Then, I will present a logical reconstruction of Thagard's model of conceptual change that offers belief-revision-like revision and contraction operations that work on conceptual structures. I will show how this conceptual revision framework, by working at the conceptual level of abstraction, is able to model almost all the kinds of radical conceptual change described in Thagard's model. In order to substantiate the adequacy of my logical reconstruction of Thagard's model of conceptual change I will reconstruct one of Thagard's main example of a conceptual revolution, i.e. the chemical revolution, as a series of revision and contraction operations within my conceptual revision framework. Finally, I will assess cognitive models of conceptual change through my Toolbox framework.

Finally, in the Conclusions chapter, I will collectively analyze all four types of model of conceptual change discussed in the preceding chapters. I will combine all the specific analyses given in the different chapters into a general study of the conceptions of conceptual change corresponding to these four different types of model. I will first show the general results of this combined analysis by presenting a chart of how the different models face along the nine dimensions of the Toolbox framework. Then, I will break this chart row-by-row focusing on one dimension of the Toolbox framework at a time in order to give a more fine-grained analysis of a specific aspect with respect to which the different models conceptualize conceptual change. I will end up with a general conception of the phenomenon of conceptual change as it emerges from this combined analysis. At the end of this last chapter, I will draw some general morals of philosophical interest that this work arguably supports and I will briefly mention some related directions for future work.

Chapter 2

Concepts and Conceptual Change

In this chapter, I will prepare my analysis of models of conceptual change in science and philosophy, focusing on some preliminary issues. More specifically, I will survey the philosophical literature on concepts and conceptual change that will constitute the background for my discussion. I will also stress and defend the significance of the phenomenon of conceptual change for philosophy and science. Furthermore, in order to make more precise my analysis of models of conceptual change, I will present a meta-framework in which models of conceptual change can be compared along nine dimensions.

In Section 1, I will survey the philosophical and psychological literature about concepts, fixing my terminology and briefly presenting the main positions on the ontology and the structure of concepts. In Section 2, I will describe the philosophical problem of conceptual change in science and in philosophy. First, I will focus on how conceptual change in science became a central issue for any adequate notion of scientific progress, realism, and rationality, presenting the main alternatives for modeling conceptual continuity between scientific theories. Then, I will analyze whether the main metaphilosophical views in analytic philosophy are compatible with the philosophical significance of conceptual change in philosophy. In Section 3, I will stress how recent metaphilosophical debates prompt us to question the compatibility between conceptual change and certain externalist and essentialist views about meaning. I will then defend the philosophical significance of conceptual change, stressing how this alleged incompatibility stems out of two mistaken assumptions about how our language works: meaning monism and metasemantic finality. This defense will also constitute a philosophical justification of the novel methodology for studying conceptual change developed and employed in this work. Finally, in Section 4, I will present a meta-framework that I will call the *Toolbox framework*, composed of nine dimensions along which models of conceptual change can be judged and compared.

2.1 Theories of Concepts

Concepts are usually understood as the units of thought (Murphy, 2002; Margolis and Laurence, 2019), i.e. the entities that allow us to perform several higher cognitive abilities such

as inference, categorization, abstraction, memory, learning, and the like. More specifically, concepts are usually considered the atomic units of thought, i.e. the most basic entities from which beliefs and other mental entities are constructed. Concepts are also usually considered to be in a special relation with the linguistic entities we call 'predicates'. Concepts are, in fact, usually put in a one-to-one correspondence with predicates and sometimes even identified with their meanings. Concepts are moreover usually considered an essential part of our mental and worldly life, because of their pivotal role in the aforementioned higher cognitive abilities, in our thinking about the past and the future, and in our everyday acts and plans. Without concepts, we will be in the unenviable position of Borges' *Funes*, who lived in a "teeming world of (...) only details" (Borges, 1998, p. 136).

The traditional philosophical picture of concepts involves at least four different entities: ideas, properties, meanings, and extensions. These four fundamental philosophical notions are directly related to (and often identified with) concepts. Ideas are usually considered to be the psychological components of our mental attitudes, i.e. the atomic mental entities that allow us to successfully engage in intentional and unintentional psychological thinking processes. Properties (or categories or kinds) are instead usually understood as the abstract entities that our linguistic and thought activity tries to pick out, i.e. classes of entities in which the external reality can be divided. Meanings and extensions are instead the two building blocks of semantics. Meanings are usually understood as the entities expressed by words that allow us to successfully engage in any linguistic activity. Extensions are instead usually thought to be the entities to which our words refer to, i.e. the external entities that ground the reference of our linguistic activity. The traditional philosophical conception of conceptual activity is something like the following scenario: thinking about a given entity, say about dogs, involves a given subject possessing (and usually intentionally using) an idea of dogs, closely connected with the property of being a dog, the meaning of the word dog and the actual dogs out in the world.

The above characterization of concepts (and of ideas, properties, meanings, and extensions, respectively) is of course only a functional one, understanding concepts as whatever enable us to carry out certain cognitive processes. It leaves completely open fundamental questions such as the ones on the ontological status of concepts and their structure. These two kinds of questions are at the center of the philosophical and psychological literature on concepts. Different answers to these questions give very different pictures of what concepts are and how are they connected with ideas, properties, meanings, and extensions.

Let us survey the most common answers in philosophy and psychology to these two questions. First, I will briefly describe the most common alternative to question over the ontological status of concepts. Then, I will analyze in far more detail the main answers that philosophers and psychologists gave to question of how concepts are structured. My main references for presenting these alternatives will be (Smith and Medin, 1981; Margolis and Laurence, 1999; Murphy, 2002; Margolis and Laurence, 2015).

2.1.1 The ontology of concepts

The four main positions regarding the ontological status of concepts are closely connected with the four different entities, related to concepts, that we briefly introduce above: ideas, properties, meanings, and extensions. For each of these entities, there is an ontological view of concepts that identifies concepts with that specific entity involved in conceptual activity. I will dub the view that identifies concepts with ideas the *psychological* view and the one that identifies concepts with properties the *abstract* view. The ontological view that identifies concepts with meanings will be instead dubbed the *linguistic* view and the view that concepts are extensions will be dubbed the *worldly* view. Let us survey these four ontological views about concepts, then.

The *psychological view* claims that concepts are mental entities, i.e. ideas. According to this position, concepts are then some kind(s) of basic mind-dependent entities, from which more complex mental entities (correlated to propositional attitudes) such as beliefs are constructed. The psychological view is the default position in cognitive science and also between philosophers of mind. This position is also historically connected with the empiricist tradition in modern epistemology, in particular with Locke's (Locke, 1690) and Hume's (Hume, 1739) theories of ideas. Philosophers holding the psychological view are usually (but definitely not always) also supporters of a general empiricist attitude in epistemology, a cognitivist approach to semantics, an anti-realism about kinds and categories, and a close alignment with the related psychological literature.

The *abstract view* claims that concepts are abstract entities, i.e. properties. According to this position, concepts are some kind of abstract entities, external to and independent from the mind. The abstract view is a common position in philosophy of mathematics and metaphysics, where often concepts are equated with the related abstract properties. This position is historically connected with the rationalist tradition in modern epistemology and Platonism in metaphysics and philosophy of mathematics. Paradigmatic examples of the abstract view are Plato's forms (Plato, 399-395 BC) and Frege's senses (Frege, 1892a). Philosophers holding the abstract view are usually (but definitely not always) supporters of a general rationalist attitude in epistemology, a realism about kinds and categories, and a close alignment to mathematics and other allegedly a priori disciplines.

The *linguistic view* claims that concepts are linguistic entities, i.e. meanings. According to this position, concepts are inter-subjective linguistic entities. The linguistic view is commonly assumed in philosophy of language, where often concepts are equated with the intensions of the related words. Historically, this position is connected with the strong focus on linguistic analysis typical of philosophical methodologies prominent in early analytic philosophy. Some proponents of the linguistic view often stress the know-how aspect of conceptual knowledge, identifying concepts with certain kinds of (broadly) linguistic disposition or abilities. Examples of philosophers that can be interpreted as holding a linguistic view of concepts are Dummett (Dummett, 1993), Brandom (Brandom, 1994, 2000), and Wittgenstein (Wittgenstein, 1958).

The *worldly view* claims that concepts are worldly entities, i.e. extensions. According to this position, concepts can be simply equated with the worldly entities that the related

linguistic term refers to. The worldly view is often assumed by philosophers of language that equate concepts with the extensions of the related words. Historically, this position is connected with the rise of semantic externalism, i.e. the (meta)semantic position that stresses how the meaning of many parts of our language is determined (at least in part) by external factors. Some remarks of Kripke's (Kripke, 1972) famous externalist plead can be interpreted as suggesting a worldly view of concepts.

The description I just gave of these four positions is of course a very rough summary of them and by no means the few lines above do justice to the complex and vast philosophical debate about the ontology of concepts. As we will see in the next sections and chapters, many interesting philosophical theories about conceptual ontology involve a subtle mixture of these four macro-views of the ontological status of concepts.

2.1.2 The structure of concepts

If the debate over the ontological status of concepts can be roughly summarized describing few main (bundles of) standpoints, the same thing cannot be easily done for the debate over conceptual structure. In philosophy and psychology, in fact, a plethora of different theories about conceptual structure have been proposed. In what follows, I will focus on eight influential views about how concepts are structured: definitional theories, functional theories, prototype theories, exemplar theories, atomic theories, theory theories, ability theories, mixed theories.

Definitional theories. The most influential conception of conceptual structure is without a doubt the one depicted by definitional theories, also known as the classical picture or the received view of concepts. According to definitional theories, (most) concepts have definitional structure, i.e. our conceptual knowledge is structured in a hierarchical way, where a given concept is obtained combining the (usually) necessary and sufficient properties that simpler concepts are identified with. For instance, the concept of a bachelor, according to these theories, is made of the definition 'unmarried man', where the concept of unmarried and the concept of man are simpler concepts. The definitional structure of concepts explains also how they allow us a vast range of cognitive abilities such as categorization, learning, inferences, and the like. We categorize via realizing that a given individual satisfies the definition of a given concept, we learn concepts through learning how they are defined, and we use definitions for drawing analytic inferences from our conceptual knowledge. We furthermore form new concepts from simpler ones via combining their definitions.

Definitional theories have been common throughout the whole history of philosophy, from Plato to twentieth century philosophy, and they were also the received view in the first half of the last century in psychology and linguistics. Definitional theories have been defended together with all the four aforementioned ontological views about concepts, thus forming psychological, abstract, linguistic, and worldly definitional theories of concepts. As stressed by Fodor and his co-authors (Fodor et. al., 1980) in their famous critique of definitional theories, the explanatory power of definitions come from being entangled with

a classical set of epistemological and metaphysical assumptions that see in definitions the glue that holds ideas, properties, meanings, and extensions together, thus connecting our thought, our language, and our world.

However, in the last century psychologists have become more and more dissatisfied with definitional theories, up to the point that virtually no contemporary cognitive psychologist is a defender of such an account. This is because, starting from the early seventies, a series of experiments started to accumulate an increasing amount of evidence against the classical theory, culminating in what in Kuhnian terms we can call the Roschian revolution in cognitive psychology. More specifically, the classical theory of concepts gives a flat and sharp depiction of concepts. Since concepts are organized via necessary and sufficient definitions, either something falls under the definition of a given concept or not. No vague cases are allowed. Furthermore, in the classical picture all instances of a given concept exemplify the concept in the same, perfect, way. There are no better or worse examples of a bachelor, since all bachelors are unmarried men. This sharp and flat conception of conceptual structure has been strikingly disproved by the so-called typicality and borderline effects. A robust series of studies (e.g. Rosch 1973, 1975; McCloskey and Glucksberg 1978; Hampton 1979) showed that people's use of concepts involves a very non-flat and non-sharp way of assigning conceptual membership, where objects can be more or less typical instances of a concept and several cases can fall in-between membership or non membership[1]. These findings were completely alien to the rigid hierarchical definitional structure of the classical theory of concepts and thus prompted psychologists to develop alternative views of conceptual structure.

If in psychology the classical theory has fallen out of fashion, in philosophy one can instead easily find, still nowadays, theories of concepts assuming a definitional structure. Despite very influential philosophical critiques to definitional theories (e.g. Quine 1951; Wittgenstein 1958), in fact, definitions can still play a role in a theory where concepts are seen as abstract entities. Abstract definitional theories of concepts are usually built around so-called external definitions, i.e. definitions independent and usually often unknown to the subject[2]. External definitions are usually coupled with a certain degree of essentialism about kinds and externalism about meanings (Kripke, 1972; Putnam, 1975) that detach concepts and related entities from the human mind, thereby shielding definitions from contrasting psychological evidence. Accounts of concepts based on such external definitions are Putnam's and Rey's conceptual cores (Putnam, 1970; Rey, 1983), and Peacocke's implicit definitions (Peacocke, 1992).

[1]The existence and the significance of borderline cases in conceptual and linguistic activity is also a central theme of the philosophical literature on vagueness and the sorites paradox (cf. Williamson 1994; Keefe and Smith 1996; Keefe 2000). For a description of how the philosophical and the psychological literature on vagueness relate to each other, see (Égré, Ripley, and Verheyen, 2019).

[2]There are also more classical abstract definitional theories of concepts, such as Zalta's object theory (Zalta, 2001).

Functional theories. Another, very influential, traditional picture of conceptual structure in analytic philosophy is the one described by functional theories of concepts[3]. According to functional theories, concepts are entities akin to functions in mathematics, i.e. they are mappings from one domain of entities to another one (usually from objects to truth-values). As such, concepts are definable through the specific mapping with which they are identified. Concepts are thus seen by functional theories are closely connected with the related predicates and, just like them, as incomplete entities (one could say unsaturated, using Frege's chemical metaphor, cf. Dummett 1973), usually defined thanks to their domain and their co-domain. Two examples of (very influential) functional theories of concepts are Frege's own theory of language (Frege, 1891, 1892a,b) and Carnap's theory of intensions (Carnap, 1947). Functional theories of concepts, and their depiction of concepts as abstract functions, are closely connected with the rise of formal semantics, and especially of intensional semantics in philosophy of language and linguistics.

Functional theories are naturally connected with the abstract and the linguistic views of concepts. The core idea of functional theories, i.e. that concepts are entities akin to functions, takes concepts to be non-mental entities, closely connected with the related predicates. Thus, concepts have been identified by functional theorists either as abstract properties or as the (intensional) meanings of the related predicates. Psychological and worldly views on the ontology of concepts seem instead incompatible with functional theories of concepts. As such, functional theories have encountered the peculiar fate of being amongst the received views of concepts in philosophy of language and formal semantics (cf. Katz 1972), while receiving little discussion in philosophy of mind and psychology.

Prototype theories. If definitional theories were discredited in psychology by typicality and vagueness effects, prototype theories were explicitly developed with these effects in mind. According to prototypical theories, (most) concepts have prototypical structure, i.e. our conceptual knowledge is stored via a summary representation of the properties that instances of a concept tend to possess. The pivotal difference with definitional theories is then the step from a summary of necessary and sufficient properties to a statistical summary of the properties that a given instance is most likely to have. This tendency-based structure is able to straightforwardly explain the typicality and graded structure effects that caused so much trouble to the classical picture. Categorization is, in fact, according to prototype theories, achieved via measuring the similarity of a given instance with the prototypical structure of a given concept. Instances of a concept can then be more or less similar to the prototype, i.e. having more or less prototypical properties, resulting in a more or less typical exemplification of the concept. Similarity calculations allow the possibility of borderline cases, i.e. instances that are almost equally similar to

[3]As it should be clear, what I dubbed here the functional theory of concepts should not be confused with the strain of conceptual pluralism that it is sometimes called in philosophy of mind conceptual functionalism (Lalumera, 2010) and that I will treat amongst the mixed theories of concepts. The two theories have nothing in common. The functional theory of concepts understands in fact concepts as entities akin to functions in mathematics, while conceptual functionalism thinks instead that concepts have no common structure and thus are a functional kind.

prototypes of different concepts. The assumption of a prototypical conceptual structure is able to explain several cognitive abilities related to concepts. If categorization, as already mentioned, is done via judging the similarity of a given object with the prototype of a given concept, learning is achieved via the abstraction of a prototype, while prototype combination allow us to form new concepts from simpler ones (Smith and Osherson, 1984; Smith et al., 1988; Rips, 1995). Prototype theories can easily model various conceptual inferences, such as category-based induction (i.e. inferring the properties of an instance from the concept it belongs to) and other probabilistic inferences, thanks to the statistical aspect of prototypical representation (Rips, Shoben, and Smith, 1973).

Prototype theories are deeply entrenched in the psychological view of concepts. They were first developed by Rosch and her colleagues (Rosch, 1973, 1975; Rosch and Mervis, 1975; Rosch, 1978) for explaining psychological data about typicality effects and graded structure and they remain the default option for any psychological model of cognitive abilities related to concepts. In philosophy, prototypical theories have been most famously used by Wittgenstein (Wittgenstein, 1958), who is referred as a direct inspiration for the development of prototype theories by Rosch, and by Kuhn (Kuhn, 1974, 1976, 1990, 1991), who in his late work spelled out his crucial notions of paradigm and kind in increasingly prototypical terms (cf. Hoyningen-Huene 1993; Andersen et al. 1996, 2006). Wittgenstein's *ante litteram* prototype theory could be interpreted as a linguistic prototype view. The only ontological views with which the idea of a prototypical structure seems *prima facie* incompatible are thus the abstract and the worldly view.

If the idea that concepts have some kind of prototypical structure is almost uncontroversial in contemporary cognitive science, the way in which this structure is exactly spelled out has been the center of many controversies since Rosch's seminal work. Starting from Rosch's original formulation of the prototype theory, psychologists disagreed on what exactly the prototype of a concept is, forcing Rosch to explain many times what is precisely implied by assuming that concepts have a prototypical structure. Different specific models of prototypical structure have been proposed, differing in how many prototypes they assume and how is the prototype structured. Examples of these models include feature list models (Hampton, 1979), dimensional models (Smith and Medin, 1981, Ch. 5), and holistic models (Smith and Medin, 1981, Ch. 6). Another topic of heated discussion has been how to exactly measure the similarity of a given object with the prototype of a concept, a debate in which several mathematical measures of similarity were proposed and discussed (Tversky, 1977). The similarity-based categorization model of prototypes theory has been also criticized for not taking into account other determinants of conceptual structure such as ideals and base-rates (Hampton and Gardiner, 1983; Barsalou, 1985). More generally, many psychologists argued that prototype theories need further structure in order to account for the whole spectrum of cognitive abilities in which concepts are involved. This critique cause the appearance of many enriched prototype models of concepts such as schemata (Lakoff, 1987a,b), frames (Minsky, 1975; Jackendoff, 1992), and spaces (Gärdenfors, 2000, 2014). These enriched models can be considered the default way of representing conceptual knowledge in cognitive science and computer science nowadays.

Exemplar theories. Another type of conceptual structure developed to account for the empirical flaws of definitional theories is the one championed by exemplar theories. According to these theories, (most) concepts have exemplar structures, i.e. they are represented via a set of their specific instances. Exemplar theories take then a more radical departure from the classical picture of concepts than prototypes accounts, denying that conceptual knowledge is stored via an abstract conjunction of properties possessed by (all or many) instances of the concept. According to exemplar views, we do not build concepts summarizing properties, we do that by storing specific instances of them. Evidence for the exemplar view is provided via the so-called exemplar effects, i.e. cases where single instances of a concept influence our categorization and induction skills. These effects are explained by exemplar theorists with the assumption that conceptual competences like categorization and category-based induction are performed via the retrieval of specific instances of the concept. According to this view, when we categorize a dog in the street as an instance of the concept dog, we retrieve a specific dog from our long-term memory (say, our own dog) and, by judging the similarity between the two dogs, we realize that the dog on the street is indeed a dog. Analogously, category-based induction is performed via default reasoning mechanisms based on the similarity with stored exemplars of concepts. Conceptual learning and combination are also performed thanks to the specific instances of concepts that are stored in our mind. Exemplar theories can also easily account for the typicality and borderline cases effects that prompted the development of prototype theories, since their categorization models are also based on similarity judgments.

Exemplar theories, just like prototype ones and perhaps even more, are strongly connected with a psychological view of concepts. They were first developed by Medin and Schaffer (Medin and Schaffer, 1978), few years after the publication of Rosch's revolutionary studies. In comparison with prototype theories, exemplar models have not been very popular amongst philosophers, but they enjoy the stable status of a valid alternative to prototype-based models in cognitive psychology, especially in studies related to categorization abilities (Nosofksy, 1984).

Just like for prototype models, there have been several different proposals of how the exemplar structure of concepts is exactly spelled out. Specific exemplar models of concepts differ in the number and the nature of the exemplars, how the exemplars are chosen amongst all the instances of a concept we experience, the exact similarity measure involved in categorization judgments (Smith and Medin, 1981, Ch. 7). Exemplar models have also been criticized for giving an extremely similarity-focused picture of conceptual abilities, giving rise to enriched exemplar models that often involve also some kind of prototypical structure (Nosofksy, 1992).

Atomic theories. A radical conception of conceptual structure is the one given by atomic theories (or conceptual atomism). According to these theories, in fact, concepts have no semantic structure whatsoever. The many cognitive, semantic, and epistemological abilities related to concepts can be explained via mechanisms other than a concept internal structure. The most prominent advocate of conceptual atomism is Jerry Fodor

(Fodor, 1975, 1998, 2008), who developed and repeatedly defended an atomistic view of concepts against the many critiques that this radical position attracted. Most of Fodor's reasons supporting conceptual atomism are not in the form of positive evidence for it, but they are arguments against the viability of all the other accounts of conceptual structure (e.g. Fodor et. al. 1980; Fodor and Lepore 1996). Roughly speaking, the central point of Fodor's dissatisfaction with non-atomist theories of concepts is that they cannot fully explain conceptual learning, since they all presuppose a basis of primitive concepts the acquisition of which cannot be explained by any account of the internal structure of a concept. The acquisition of these primitive concepts has then to be explained either with some kind of innatism or with some learning mechanism external to the structure of concepts. Fodor argues then that there is no reason to postulate a conceptual structure, when a combination of these two external explanations (i.e. innatism and externalism about conceptual content) suffices to justify all kinds of conceptual knowledge. It is important to stress that, just like the classical picture of concepts thrived in combination with related common epistemological assumptions, the explanatory power of Fodor's conceptual atomism is best understood in the light of his computationalist view of the human mind and his language of thought hypothesis (cf. Fodor 1975, 2008; Crane 2015). Seen in this perspective, Fodor's conceptual atomism offers a formal model of conceptual abilities as syntactic properties of the atoms composing the language of thought consistent with old-fashioned computationalism in philosophy of mind.

Atomic theories of concepts are naturally connected with a psychological view of concepts. Apart from Fodor's own picture of the human mind, conceptual atomism has been a very popular position in philosophy, especially as a conceptual counterpart of semantic externalism and the so-called new theory of reference. Fodor's attacks against non-atomic theories of concepts are actually structurally analogous to Kripke's critique of internalist theories of meaning and references (Kripke, 1972). Moreover, atomic theories of concepts are often coupled with causal and teleological theories of mental content (Millikan, 2000). Both kinds of theories, when held together, strengthen one another, since conceptual atomism offers a deflationary way of talking about conceptual abilities to externalist account of mental content, while causal and teleological accounts of mental content offer an externalist way of connecting atomic concepts with the world (cf. Fodor 1990; Millikan 1998). Despite the philosophical success, conceptual atomism has never been equally popular amongst psychologists. Especially after the emergence of the so-called embodied theories of cognition and the consequent fall of the old computationalist paradigm, atomic theories of concepts give a depiction of the human mind significantly different from the one depicted by contemporary cognitive science (Clark, 1993; Shapiro, 2004).

Theory theories. We saw how both prototype and exemplar theories, having both of them similarity-centered models of categorization, have been criticized for not sufficiently taking into account contextual factors such as ideals and base-rates. In a similar fashion, similarity-based models of categorization have been criticized for neglecting the so-called knowledge-effects, i.e. the influence of general knowledge on our categorization. In order

to account for these effects, some scholars argued that we have to change the whole picture of conceptual structure offered by prototype and exemplar theories (Murphy and Medin, 1985; Barsalou, 1987). This change gave rise to the so-called theory theories (or knowledge theories) of concepts. According to these theories, (most) concepts are structured like (simple) scientific theories. So, for instance, our concept of dog, for the theory theorists, is structured like a simple theory of why dogs are dogs, including several kinds of biological, morphological, and social explanations that contribute to our idea of dogs. In other words, theory theorists claim that a given concept is a summary of the different kinds of knowledge that we have on a given category of entities. Theory theories account for the cognitive abilities related to concepts by considering them as chapter in our epistemological account of theoretical knowledge, especially of scientific one. Categorization is then a simplified version of scientific hypothesis testing, where we check whether certain empirical data, i.e. the properties of a given individual, are consistent with the theory, i.e. the concept under which (we assume) it falls. Analogously, conceptual inferences are for the theory theorists analyzable with the models through which philosophers of science understand related scientific inferences. The relationship between theory theories of concepts and philosophy of science is even stricter in regards to model of developments, where theory theories models of concept learning and development are directly inspired by models of theory change in philosophy of science (Carey, 1985; Gopnik and Meltzoff, 1997).

Theory theories are, since their appearance, primarily developed within a psychological view of concepts. These theories have become quite popular in many subfields of cognitive psychology, especially in the ones studying how children acquire and develop concepts (Carey, 1985; Keil, 1989; Carey, 2009). Many holistic accounts of meaning in philosophy of language, such as Quine's (Quine, 1960) one for instance, can be seen as possible blueprints for a linguistic theory view of concepts.

Different models of theory-based conceptual structures have been proposed, differing in their understanding of theory and in the mechanisms of acquisition and change used. A common problem of all theory theories is the exact specification of the theory-like way in terms of which concepts are allegedly structured. In philosophy of science, in fact, there is no commonly accepted account of how scientific theories are structured. Moreover, theory-theories may also face a circularity-problem (similar to the one stressed by Fodor et al. against definitional theories of concepts, cf. Fodor et. al. 1980), since any reasonable way of spelling out the atomic components of the (alleged) theoretical structure of concepts seems to involve some kind of non-theory-like structure. Many proponents of the knowledge view acknowledge (to a certain extent) these two issues and take a more moderate departure from similarity-based account of conceptual structure, claiming that their approach should be considered an addition and not a replacement of prototype or exemplar accounts of conceptual structures (Murphy and Medin, 1985).

Ability theories. All the theories of conceptual structure we have seen so far understand conceptual knowledge as a certain kind of propositional knowledge, a know-that. What I will call ability theories, instead, is a bundle of theories of conceptual structure

that share an understanding of conceptual knowledge as a certain kind of know-how. In other words, according to ability theories, concepts are identifiable as certain kinds of skills, epistemologically similar to knowing how to play a sport or a musical instrument. These theories claim that conceptual structure cannot be identified with any set of properties, instances, theories, or causal relationships, but it is instead a kind of disposition to perform certain cognitive abilities related to concepts. Paradigmatic examples of ability theories are conceptual neo-empiricism (Barsalou, 1999; Prinz, 2002; Shapiro, 2004) and inferentialism (Carnap, 1934; Dummett, 1991; Brandom, 1994, 2000; Peregrin, 2011). Neo-empiricism, in its various form, is a modal form of ability theory that sees conceptual knowledge as embodied in the perceptual and motor cognitive systems. Building upon the aforementioned embodied mind turn, neo-empiricists claim that concepts are certain kinds of (simulated) perceptual or motor abilities, the existence of which derives from the bodily aspect of human experience. Inferentialism, instead, understands concepts primarily as inferential tools. According to inferentialists, a concept is not determined by its internal structure, but by its inferential relations with other concepts. Inferentialism is thus a holistic approach to conceptual knowledge that identifies concepts with their role in inferential practices. Historically, inferentialism is closely connected with the rise of the field of proof-theory (Gentzen, 1934/35; Prawitz, 1965) in logic and the related inferential role or proof-theoretic semantics (Schroeder-Heister, 2018) for logical systems.

Ability theories seem compatible with most of the main ontological views about concepts. Neo-empiricism is, in fact, usually coupled with a psychological view of concepts, while inferentialism mostly comes together with a linguistic view. Moreover, Peacocke's (Peacocke, 1992) aforementioned neo-classical theory of external definition, with its strong focus on dispositions and implicit rules, can be also considered an abstract ability theory. The only ontological view of concepts that seems incompatible with ability theories is the worldly view. Both in psychological and philosophical literature, ability theories continue to remain popular alternatives to more traditional theories of conceptual structure.

Mixed theories. The seven theories of conceptual structure we have seen so far share a common assumption on how concepts are structured, namely, that all (or at least most) kinds of concepts have the same kind of structure. This assumption has not been unchallenged in psychological and philosophical debates. I call mixed theories the group of theories of conceptual structure that hold that concepts do not have a single general structure. I will briefly describe four different mixed theories: dual theories, hybrid theories, conceptual pluralism, and conceptual eliminativism.

Dual theories claim that there are two different entities behind the intuitive notion of a concept. Usually, these two entities belong to a different ontological realm and they are responsible for different parts of conceptual knowledge. For instance, Rey's theory of conceptual cores (Rey, 1983) is made of a mental prototype that is used for psychological categorization and related cognitive abilities and an abstract conceptual core, constituted by an external definition, that is used for metaphysical categorization and identification. Another interesting dual theory of concepts is Fodor's mature formulation of conceptual atomism

(Fodor, 1990), where structureless, causally identified, atoms of language of thought are coupled with prototype-like non-semantic contents used for psychological categorization.

Hybrid theories (Vicente and Manrique, 2016) claim instead that concepts have a complex structure that is made of two (usually psychological) kinds of structure, such as (for instance) a prototype and some exemplars. The difference with dual theories is that hybrid theories claim that the multiple structure of a concept is used in combination for performing the same kinds of tasks. Both dual theories and hybrid theories claim that a single concept has more than one kind of structure, but they still assume that all concepts share the same kinds of (complex) structure. Recently, philosophers have started to question whether the class of concepts is such that all its members share a certain kind of structure.

Conceptual pluralism (Weiskopf, 2009; Lalumera, 2010) claims that concepts are a functional kind, i.e. a class of entities identified by performing a certain function, not by sharing a certain kind of structure. According to this position, the class of concepts is structurally similar to kind of things such as means of transportation, whose members share only the fact that all of them perform the function of transporting people. Conceptual pluralists think that different kinds of concepts might have different kinds of structures and even the same concept can be instantiated, depending on the context, by different conceptual structures (cf. Wilson 2006; Haueis 2021). Evidence for conceptual pluralism can be given by the well-known fact that different theories of concepts often give an excellent account of certain specific kinds of concept or certain functions of concepts, but they seem inadequate to describe different kinds or different functions.

Finally, conceptual eliminativism (Machery, 2009) claims that concepts are not a natural kind and thus the term 'concept' should be eliminated from our scientific image of the world. Eliminativists claim in fact that there is robust empirical evidence supporting different theories of conceptual structures and that furthermore there is robust evidence that all these different structures are related to different cognitive processes behind intuitive conceptual abilities such as categorization or conceptual combination. For these reasons, eliminativists propose that psychologists should abandon any talk of concepts, conceptualizing conceptual knowledge only at the sub-level of these different kinds of conceptual structure. Machery's book-length plead for eliminativism prompted a very heated discussion about which kinds of things are concepts and which kind of function they serve in contemporary psychology and philosophy (cf. the many interesting replies to Machery's summary of eliminativism in Machery 2010).

2.2 The Problem of Conceptual Change in Science and Philosophy

In the last section, we surveyed the philosophical and psychological literature on concepts, focusing on the main proposals about which kind of entities concepts are and how are they structured. We have now an idea of how we can statically understand concepts. In this section, I will instead focus on the dynamical aspect of concepts, i.e. how these enti-

ties change. Specifically, I will debate the epistemological problems related to conceptual change that made it a central topic in philosophy of science and meta-philosophy.

First, I will focus on the problems that conceptual change in science poses to philosophers of science, describing why conceptual change is paramount to our understanding of scientific progress, scientific realism, and scientific rationality. Then, I will describe the meta-philosophical debate over the existence of conceptual change in philosophy and its relations with our ideas regarding the goals, the scope, and the methods of philosophical activity.

2.2.1 Conceptual change in science

The fact that our scientific image of the world changes with time is uncontroversial. Our best scientific theories in current times are different from the ones that were considered the best ways of describing the same phenomena two hundreds years ago. This difference makes us think that even future scientific theories would probably be different from the current ones. Scientific theory change, by itself, would be unproblematic if we could be completely sure that every new theory is just a more precise and more complete version of the theory it replaces. If all scientific changes were of this kind, we could see in scientific theory change just a cumulative series of extensions and improvements of old theories.

Unfortunately, even the most optimistic enthusiasts of scientific activity cannot seriously believe this idyllic picture of science history. Just a quick and superficial glimpse at the actual history of science reveals in fact that far more radical and epistemologically worrisome changes took place. Extensions and improvements are just very specific cases of scientific theory change. New theories replacing old ones often drastically change the image of the world given by the replaced theories, modifying important aspects of old theories such as laws, explanations, ontological assumptions, and (most importantly for the present work) concepts. The most radical and important macro-changes in scientific theories have been dubbed, in analogy to political revolutions, scientific revolutions (Kuhn, 1970; Nickles, 2017). The frequency and the extent of scientific revolutions prompted philosophers to question over-optimistic accounts of scientific progress, rationality, and realism[4]. How can we be sure that science progresses towards truth if lots of experimental knowledge is lost during a revolution? How can scientific activity be rational if there is no way to objectively compare two radically different theories? How can we trust the scientific description of reality if in revolutionary times new entities are postulated and old ones disappear? A complete survey of the answers that philosophers have given to these enormous questions is out of the scope of the present work and, most likely, impossible to squeeze in any single book[5].

In what follows I will focus on the specific problems that conceptual change causes

[4]Most famously, the theoretical discontinuity of scientific theories is at the heart of both Laudan's 'pessimistic induction' (Laudan, 1981, 1984a) and Hesse's 'principle of no privilege' (Hesse, 1976), two influential historical arguments against scientific realism.

[5]For general surveys of philosophical debates over scientific progress, scientific objectivity, and scientific realism see (Niiniluoto, 2019; Reiss and Sprenger, 2020; Chakravartty, 2017; Psillos, 2018).

in regards to these bigger questions. Conceptual change is in fact at the heart of the most worrisome byproduct of scientific revolutions: incommensurability between different (bundles of) scientific theories. The mathematical notion of incommensurability, i.e. the lack of a common measure, was made a common term in philosophy of science by the influential work of Kuhn (Kuhn, 1970) and Feyerabend (Feyerabend, 1962) who applied this term to the (alleged) breakdown of rational communication in scientific revolutions. In particular, radical conceptual change is central to one pivotal component of Kuhn's complex notion of incommensurability, i.e. what is known as *taxonomic incommensurability* (Hoyningen-Huene, 1993; Sankey, 1997; Sankey and Hoyningen-Huene, 2001). Taxonomic incommensurability denotes the fact that different theories can have a completely different understanding of a given scientific term and its related meaning. This conceptual kind of incommensurability challenges supporters of scientific progress, scientific rationality, and scientific realism to explain the continuity in goals, roles, value, and ontological import of these incommensurable concepts. In other words, what is needed to defend science is a model of the dynamics of scientific concepts, i.e. a model of conceptual change.

This is how Kuhn's extremely influential notion of incommensurability caused the problem of conceptual change to become one of the main topic that philosophers of science discussed in the second part of the last century. Let us survey the main types of model of conceptual change that have been proposed in the literature, then. I will organize my presentation of these models via clustering them in five categories, differing in how these models conceptualized the continuity between scientific concepts. I will talk respectively of syntactic, semantic, cognitive, pragmatic, and evolutionary models of conceptual change[6]. Before starting my analysis of models of scientific conceptual change, I need to specify what I mean with the term 'models' in this context, in order to avoid misunderstandings. I use the term 'model' in this work in its naive scientific sense, i.e. to denote an abstract, idealized, and simplified representation of a given phenomenon. I do not assume any specific ontological, epistemological, or semantic theory about models[7]. I just make the basic assumption that models are fruitful ways of describing and studying a phenomenon like conceptual change. So, when I will talk about a given model of conceptual change, it should be considered a lightweight short expression for denoting an abstract, idealized, and simplified representation of conceptual change[8].

[6]It should be noted that many of these models of conceptual change are often a part of a more general model of scientific theories and theory change. Usually, a given understanding of conceptual continuity is inscribed in a theory that understands scientific theories in the same way. Usually, but not always, since there are examples of models of conceptual change that have conceptualize conceptual continuity in a given way inside a differently characterized view of scientific theory. Kitcher's semantic model of conceptual change lies for instance within a practice-based pragmatic view of scientific theory (Kitcher, 1995). I will focus on conceptual change and thus characterize the models exclusively according to how they understand conceptual change.

[7]For a general survey on the philosophy of models and the many ontological, epistemological, and semantic theories regarding them, see (Frigg and Hartmann, 2020).

[8]An exception will be when, in Section 3 of Chapter 5, I will talk about models within the model-theoretic perspective of Structuralism in Philosophy of Science.

Syntactic models of conceptual change Syntactic models of conceptual change track the continuity of scientific concepts with syntactic logical tools such as (various kinds of) translations, reductions, and definitions. According to this approach, even if different scientific theories understand a given concept differently, their differences can be analyzed and bridged by mutual translation and interpretation. Old concepts belonging to old theories can be defined in newer theories by translating them (together with the related parts of the theory) via proof-theoretic means. The analysis of scientific episodes of conceptual changes becomes then an exercise in reconstruction and translation.

This approach to conceptual change can be traced back to the kind of philosophy of science championed by early logical empiricism. Syntactic models of conceptual change are in fact naturally connected to two pivotal components of (early) logical empiricist philosophy: the statement-view of scientific theories (Carnap, 1934; Hempel, 1952) and the methodology of epistemological reduction (Carnap, 1928b; Hempel, 1966). Roughly put, the statement-view claims that scientific theories are best understood (or more exactly reconstructed) as a logically structured bundle of statements. The methodology of epistemological reduction denotes instead the technique of recursively defining a given notion into epistemologically simpler terms, until only epistemologically basic terms are obtained. The reconstruction of external reality within a phenomenal constitution system contained in Carnap's *Aufbau* (Carnap, 1928a) can be considered a paradigmatic example of this kind of technique. The combination of epistemological reduction and the statement-view of scientific theories allowed early logical empiricist philosophy of science to naturally understand conceptual dynamics in terms of logical reductions between theories. More generally, analogous models of theory change were the received view of scientific theory change against which Kuhn and especially Feyerabend stressed the possibility of incommensurability. Feyerabend (Feyerabend, 1962) specifically argued that incommensurability makes inadequate any application of linguistic models of conceptual change to scientific revolutions.

Due to the influence of Kuhn's and Feyerabend's work and to even more influential critiques to the logical empiricists epistemology and to its (allegedly) reductionist methodology (cf. Quine 1951), linguistic models of conceptual change were heavily criticized. More generally, the so-called historicist turn in philosophy of science and the contemporary rise of sociology of science pushed linguistic reconstruction of scientific theories and scientific concepts at the corners of philosophy of science. Nevertheless, influential syntactic models of scientific change were still developed in the second part of the last century, such as Quine's explication (Quine, 1960) or Sellars' analogy-based model (Sellars, 1963, 1973). Both these models are holistic syntactic models of conceptual change, built around the interrelationships between (groups of) concepts (cf. Brown 2007). Furthermore, in recent years, thanks also to a renewed historical scholarship on logical empiricism (e.g. Friedman 1999; Carus 2007), the usefulness of linguistic models of conceptual change and their close relationships with other, allegedly different, models of scientific change has been re-appraised (cf. Lutz 2012, 2014; Andreas 2014; Schurz 2014a).

Semantic models of conceptual change Semantic models of conceptual change understand the dynamics of scientific concepts as specific changes in their semantic content. According to these models, despite the different linguistic frameworks in which scientific theories are constructed, continuity can be found in the underlying non-linguistic semantic entities. The semantic content of old concepts and theories can then be connected with the one of the newer ones via specific changes in the related semantic structures.

The rise of semantic models of conceptual change is closely connected with the development of the non-statement view of scientific theories. As the name indicates, non-statement views defined themselves in opposition to the aforementioned statement view of scientific theories, claiming that theories are not best reconstructed as a bundle of statements, but as semantic entities. Even if certain aspects of the logical empiricists understanding of scientific theories foresaw a semantic conception of scientific theories, the paradigmatic example of non-statement view is structuralist philosophy of science as developed by Suppes, Suppe, Sneed, and Stegmüller (Suppes, 1967; Suppe, 1977; Sneed, 1979; Stegmüller, 1976; Suppe, 1989; Suppes, 2002). This structuralist research program championed the reconstruction of scientific theories as set-theoretic entities (Balzer et al., 1987; Balzer and Moulines, 1996). The structuralist way of understanding conceptual change is then to use fine-grained model-theoretic tools to capture the continuity in semantic content between (set-theoretic reconstructions of) concepts of subsequent theories. The other main kind of non-statement view in philosophy of science, often also labeled structuralism, is the so-called state-space approach (van Fraassen, 1989; French and Ladyman, 1999; da Costa and French, 2003; French, 2017). According to the state-space approach, just like for supporters of model-theoretic structuralism, the diachronical continuity of science is reconstructed as specific changes in the semantical structures related to scientific theories. The difference with the model-theoretic approach is in how these semantical structures are conceptualized. In the state-space approach, scientific theories are best reconstructed as state-spaces, i.e. abstract spaces having dimensions corresponding to the relevant variables of the theory and points corresponding to possible states of a real system.

Another, influential, type of semantic model of conceptual change is composed of referential models of conceptual change. These models focus on the reference of scientific terms, understanding conceptual change as a specific kind of overlapping in the reference of the related scientific terms. Referential models of conceptual change are usually coupled with a general externalist (meta)semantic attitude over meaning that stresses the worldly component of the process by virtue of which the reference of scientific terms gets fixed. A seminal example of an externalist referential model of conceptual change is due to Putnam (Putnam, 1973). Other examples of referential models are the so-called causal descriptive models (Lewis, 1984; Kroon, 1985). Two influential, broadly causally descriptive, referential models of conceptual change in science are Psillos' (Psillos, 1999) core-causal description model and Kitcher's reference potential model (Kitcher, 1995).

Cognitive models of conceptual change Cognitive models of conceptual change understand the dynamics of scientific concepts as specific changes in the cognitive structures

underlying scientific theories. According to these models, scientific theories are best understood as cognitive architectures, representable with one of the many ways of representing conceptual knowledge developed in cognitive science[9]. Diachronic conceptual change in science is then seen as a specific kind of change in related cognitive structures, where the structures representing newer concepts are obtained from previous structures corresponding to old concepts by specific rule-governed transformations.

Cognitive models of conceptual change are naturally connected with the rise of knowledge representation models in cognitive science and artificial intelligence. Different models use different cognitive structures as a background framework, but they all share the general understanding of scientific change as a kind of cognitive change just described. The most common kind of background framework for this kind of model is given by frames. Frames have been used in many cognitive models of conceptual change such as Kornmesser's and Schurz's theory-frame models (Kornmesser and Schurz, 2018) and Andersen's, Barker's, and Chen's neo-Kuhnian approach (Andersen et al., 2006). Other cognitive frameworks for representing conceptual knowledge that have been used to models scientific change are conceptual systems (Thagard, 1992), conceptual spaces (Gärdenfors and Zenker, 2011, 2013; Zenker, 2014; Zenker and Gärdenfors, 2015a; Masterton, Zenker, and Gärdenfors, 2017) and schemata (Giere, 1988, 1999).

Pragmatic models of conceptual change All three types of model of conceptual change presented so far understand change in scientific concepts as a kind of transformation involving primarily the syntax or the semantics of scientific theories. Supporters of pragmatic models of conceptual change claim instead that a pivotal part in scientific conceptual change is played by pragmatic factors. More accurately, pragmatic models understand the dynamics of scientific concepts as a change driven by the values and goals of the scientists. In contrast to syntactic, semantic, and cognitive models, pragmatic models do not analyze conceptual continuity by means of reductions or transformation rules, but by meta-conceptual frameworks. The replacement of old scientific concepts with new ones can then be understood, according to pragmatic models of conceptual change, inside a framework where the reasons for the scientist choices can be analyzed in relation to their value-laden rationality or their interests.

Pragmatic models of conceptual change are closely connected with the increasing interest of philosophers of science for the topic of values in science, together with the modification of the idea(l) of scientific objectivity from a value-free to a value-laden conception (Reiss and Sprenger, 2020). Pragmatic models of conceptual change can focus on different pragmatic factors and different values in their analysis, relying on different meta-frameworks for judging scientific conceptual change. Examples of pragmatic models that focus on the so-called epistemic values are Carnap's mature method of explication (Carnap, 1950b), Kuhn's own model of theory choice (Kuhn, 1977), Lakatos' scientific research

[9]It should be noted that cognitive models of conceptual change could be considered also a sub-kind of semantic models of conceptual change, since many supporters of them argue that the cognitive architectures related to a given scientific theory constitute a cognitive semantic for it.

programs methodology (Lakatos, 1970), Friedman's dynamics of reasons (Friedman, 2001), and Wilson's Machian explication (Wilson, 2006, 2012a). Other pragmatic models are instead more focused on non-epistemic values, giving a more practice-oriented conception of conceptual change (e.g. Bloor 1976; Hacking 1983; Shapin and Schaffer 1985; Galison 1987; Pickering 1995).

Evolutionary models of conceptual change Evolutionary models of conceptual change understand the dynamics of scientific concepts as a kind of evolution analogous to the one undertaken by biological entities. According to these models, the change of scientific theories and concepts is best understood as a kind of selection process akin to natural selection. If in evolution by natural selection the selection process is guided by the fitness of the individuals, this selection-guidance role is usually played in evolutionary models of scientific change by bundles of epistemic values or abstract notions of fitness with reality. The rise and fall of scientific theories and concept becomes then a kind of cultural evolution obeying specific rules and patterns.

Even though applications of evolutionary thinking to scientific change were already common, the most influential evolutionary account of scientific products is without a doubt Popper's one (Popper, 1972a, 1984). Popper, in fact, proposed a general model of scientific change built around an analogous Darwinian selection mechanism, making explicit the commitment to an evolutionary approach to epistemology. Other examples of evolutionary models of conceptual change are Toulmin's conceptual populations (Toulmin, 1967, 1970, 1972), Campbell's selective retentions model (Campbell, 1960, 1974b), and Hull's selective processes model (Hull, 1988a). The appearance of these evolutionary models of conceptual change prompted a philosophical debate over the possibility and the role of a truly evolutionary epistemology for philosophical activity *tout court* (cf. Campbell 1974a; Bradie 1986, 1994). Moreover, several influential philosophers of science, such as Kuhn (Kuhn, 1970, 1991) and Lakatos (Lakatos, 1970), have repeatedly used a vast range of evolutionary metaphors in their models of scientific change. The extent to which they therefore subscribe to evolutionary models of scientific change is a controversial topic in historical scholarship (Renzi, 2009; Reydon and Hoyningen-Huene, 2010; Hacking, 1979; Kadvany, 2001).

2.2.2 Conceptual change in philosophy

If the significance of conceptual change in science is a rather uncontroversial observation in philosophy of science, the philosophical relevance of the same phenomenon in philosophy is very much disputed. Many conceptions of what philosophy is deny, in fact, much philosophical significance to the dynamics of philosophical concepts. To be sure, nobody denies the mere historical fact that philosophical ideas about a given topic have changed in the history of philosophy. This is rather uncontroversial. What is controversial is whether conceptual dynamics have much significance for philosophical activity outside historical interests or related enterprises in the history of ideas. Does philosophical conceptual change matters for philosophical progress and philosophical methodology? Several popular conceptions of

philosophy have given a negative answer to this question, identifying proper philosophical methods with absolute, time-independent methodologies according to which present and past alternative conceptions of a given issue are of little philosophical significance.

An important example of a philosophical methodology that downsizes the philosophical significance of conceptual change is *conceptual analysis* as traditionally understood in analytic philosophy. Since its foundational linguistic (Rorty, 1967) or conceptual turn (Williamson, 2007), in fact, analytic philosophy has largely understood philosophy as the logical or linguistic analysis of abstract entities such as concepts, propositions, and intuitions. These object of philosophical inquiry were subjected to a transformative analysis that would reveal their true logical or linguistic form. Especially in early analytic philosophy, most strongly in Frege and in the logical atomist phase of Russell and Wittgenstein, the goal of analytic philosophy was identified with giving a definite logical analysis of a given notion. Philosophical concepts are then, in this metaphilosophical view, the passive and static objects of such a descriptive analysis, while past and present alternative conceptions of the same subject are completely irrelevant to the analysis. Thus, the dynamics of philosophical concepts, past or present, are more of interest for the historian than for the philosopher. The same negative assessment of the philosophical significance of conceptual change is given by different conceptions of analysis in early analytic philosophy, such as the linguistic and connective notions of analysis championed by the ordinary language movement. For philosophers like Ryle or Austin, philosophical analysis was still the method through which static and passive objects of philosophical inquiry were clarified. Even when the idea(l) of finding a definite analysis of statements got criticized and (eventually) overcome, the linguistic and conceptual analysis of philosophical concepts remained a central topic of analytic philosophy, as it drastically exemplified by the seemingly infinite literature on the analysis of knowledge prompted by Gettier's famous paper (Gettier, 1963; Ichikawa and Steup, 2018).

There are of course many philosophical methodologies outside analytic philosophy that give a very different picture of the significance of conceptual change in philosophy. All kinds of historicist methodologies, for instance, conceive the re-appraisal of past proposals as a necessary step in the understanding of a given philosophical problem. One does not need to hold an historicist perspective on philosophical activity to appreciate the significance of conceptual change in philosophy, though. Even within analytic philosophy, one can find an example of a meta-philosophy that puts conceptual dynamics at the center of its philosophical methodology. Logical empiricism, in fact, understood philosophical activity not as a kind of descriptive analysis, but as a kind of constructive enterprise. For (most of) the logical empiricists, the proper methodology of philosophy is *rational reconstruction*, i.e. a "redescription and reorganization of a (purported) body of knowledge or conceptual scheme or set of events that exhibits the logical (or rational) relations between its elements" (Beaney, 2013, p. 253). Rational reconstruction merges together the constructivist Neo-Kantian view of epistemology and the logical tools of analysis typical of early analytic philosophy (cf. Friedman 1999, 2000; Carus 2007). Like logical or linguistic forms of analysis, rational reconstruction involves in fact the transformation of the objects of philosophical inquiry via the use of logical or linguistic tools, but it crucially under-

stands this transformation as an active construction of the philosopher and not as a mere discovery or reveal of an external phenomenon. The most paradigmatic example of this constructivist spirit is Carnap's use of rational reconstruction. From his early writings up until his last works, Carnap always understood his philosophy as the incessant construction of logical and linguistic frameworks in which scientific and philosophical concepts and theories could be rationally reconstructed (Carus, 2007, 2012b). Philosophical concepts, for a methodology like rational reconstruction, are not just passive and static object of philosophical analysis, but they are the subjects of an open-ended engineering (Richardson, 2013). Concepts are reorganized and reconstructed in different formal and linguistic frameworks, making conceptual change a central topic for philosophical inquiry. Carnap increasingly stressed the centrality of conceptual change in his philosophy, by replacing in his mature writing the talk of rational reconstruction with the notion of *explication* (Carnap, 1950b), which as we will see is a specific kind of rational reconstruction that takes concepts as its main units of construction.

In recent years, a kind of constructivist philosophical methodology has been increasingly popular also amongst analytic philosophers outside the logical empiricist tradition, namely the bundle of metaphilosophical positions that goes under the name *conceptual engineering* (Cappelen, 2018; Cappelen, Plunkett, and Burgess, 2020) and *conceptual ethics* (Burgess and Plunkett, 2013a,b). Conceptual engineers propose the substitution of the traditional methodology of conceptual and linguistic analysis in analytic metaphysics and epistemology with what they call conceptual engineering, broadly understood as the "enterprise of assessing and improving our representational devices" (Cappelen, 2018, p. 3). In the last twenty years, in fact, conceptual analysis and related assumptions on its transparency and its epistemological significance have been heavily criticized, together with other traditional assumptions about the goal and the scope of philosophical activity (e.g. Knobe and Nichols 2008; Machery 2017). These critiques of traditional philosophical methods such as conceptual analysis have prompted many analytic philosophers to focus more on metaphilosophical problems. What is the correct self-image of analytic philosophy? What are its methods? Do we have to change something in the traditional way of analyzing philosophical problems? Are traditional philosophical methods and philosophical concepts defective? According to conceptual engineers, many of our traditional philosophical concepts are very likely to be defective (cf. Cappelen 2020; Scharp 2020). The list of alleged defects of our traditional concepts involves epistemic defects such as vagueness, ambiguity, and inconsistency, as well as pragmatic and lexical effects undesirable for social and political reasons. If many of our concepts have these defects, it seems that any descriptive conceptual analysis would just reveal these defects and will not offer us any way of solving these issues. Philosophers will be left with the unenviable work of using defective tools to tackle complex problems. To remedy this dystopian view of philosophical activity, conceptual engineers propose to radically change (what is allegedly considered) the central methodology of philosophy, replacing conceptual analysis with conceptual engineering. After this methodological switch, from a descriptive to a inherently normative methodology, philosophers will have the meta-conceptual tools for assessing any defectiveness of our traditional philosophical concepts and, when needed, to normatively choose better concepts.

Not surprisingly, this proposed radical revolution in philosophical methodology caused the quick appearance of many different takes on whether traditional analytic philosophers should give up conceptual analysis and start engineering philosophical concepts, on what exactly conceptual engineering entails for philosophical activity, and on how it can be implemented in actual philosophical practice. A (proper) part of this metaphilosophical debate is particularly relevant for our more general topic of conceptual change, because it debates the nature and the possibility of any significant change in meanings, concepts, and kinds. Let us turn to this discussion, then.

2.3 Defending Conceptual Change

We saw that conceptual engineers propose to embrace a normative methodology in philosophy that focuses on the improvement of our concepts. But what exactly does it mean to improve a concept? Intuitively, any reasonable way of spelling out how a concept can be improved will depend on the particular theory of concepts assumed. As we saw in Section 1 of this chapter, many different kinds of entities have been identified by philosophers and psychologists with concepts. To improve, say, a prototypically structured mental representation would arguably be a very different matter than to improve, say, an abstract external definition. Moreover, the possibility of improving concepts is also dependent on ontological and metaphysical conceptions of concept identity. For instance, if someone holds a very abstract view of concepts, such as the one identifying concepts with entities outside space and time, it seems very difficult for her view to allow concepts to be improved in any meaningful sense. Similarly, if someone holds a very radical view of conceptual identity according to which concepts are not the kinds of things that can undergo change, she cannot meaningfully assert that concepts can be improved.

Given these complications, which kind of entities conceptual engineers want to engineer? There is not much agreement on this important matter in the growing literature on conceptual engineering. Supporters of conceptual engineering have supported theories that take the engineering units to be psychological (Scharp, 2013; Machery, 2017), linguistic (Richard, 2020; Nado, 2019), abstract (Sawyer, 2018, 2020), and worldly entities (Cappelen, 2018). There is also no agreement on the structure and the identification of these units. As a matter of fact, most of the proposals do not even clearly spell out which theory of conceptual structure they assume.

A surprisingly popular characterization of conceptual engineering wants the objects of engineering to be worldly creatures, i.e. entities independent from any subject. A paradigmatic example of this kind of worldly conceptual engineering is Cappelen's (Cappelen, 2018) so-called *Austerity Framework*. According to Cappelen, conceptual engineering is and should be about changes in meaning, specifically about changes in the extensions of words determined by changes in their intensions. Intensions and extensions are understood by Cappelen in an externalist way, i.e. as strongly determined by the external world. So, to engineer a given concept is for Cappelen to successfully change its extension by changing its intension. This change is thus a worldly one, an actual change in the status of the ex-

ternal reality. Metaphysically speaking, according to Cappelen, changing a given concept is more similar to changing your clothes than to change your opinion on a given matter. Cappelen's picture of what conceptual engineering is and should be is surprising because, as he acknowledges, it makes conceptual engineering a very difficult enterprise. Non-trivial changes in the extensions of words are in fact difficult to obtain[10]. From an externalist point of view, in fact, since the extension of a word is determined mostly by the world, parochial changes in the use or in the stipulated intentional meaning of a given word from a (group of) speaker(s) are (usually) not sufficient to determine a change in the extension of that word. Moreover, externalist extension change is not only difficult, but also an inherently non-transparent phenomenon. Cappelen stresses with his three "Corollaries of Externalism: Inscrutability, Lack of Control, and Anti-luminosity" (Cappelen, 2018, pp. 72-78) that we often do not control, nor we could, when and how the reference of our linguistic practices change. Reference change is in fact according to his strongly-externalist view mostly determined by worldly factors outside of our knowledge and thus we are often not able to judge whether a change has actually occurred in the extension of a given word. Thus, in Cappelen's Austerity Framework and similar views, conceptual engineering is a mostly uncontrollable and untraceable worldly phenomenon. Nevertheless, he believes philosophers should engage with it, given the aforementioned inherently defectiveness of many of our philosophical concepts.

The tension between the popularity of Cappelen's Austerity Framework and its far from optimistic depiction of the prospects of conceptual engineering caused the debate about conceptual enginering to focus on the possibility of intentional meaning change. Enthusiasts of conceptual engineering have stressed the possibility of indirect form of collective control over reference change (Koch, 2021) or the necessity of a more psycho-linguistic approach to meaning change (Scharp, 2020; Koch, 2020) in order to have a more favorable conception of the prospective success of any wannabe conceptual engineer. Critics of conceptual engineering have instead stressed that the lack of control and knowledge over meaning change is far more problematic than what Cappelen himself acknowledges, thereby causing any implementation of the conceptual engineering project to be bound to fail (cf. Deutsch 2020).

This discussion about the possibility of having a viable conception of conceptual engineering within an externalist metasemantics raises also a similar question for the more general phenomenon of conceptual change. If, in fact, changes in meanings are often beyond our control and difficult to even detect, what sense does it make to study conceptual change? In other words, given the popularity of externalist frameworks in philosophy of language and epistemology, conceptual change and especially the intentionally designed kind that corresponds to conceptual engineering seem to be inherently confused and mysterious phenomena. The situation is even worse if one couples together with an externalist (meta)semantics, a certain kind of essentialism about kinds, i.e. the view that there are

[10]As Cappelen himself notes, trivial changes in the extension determined by the appearance and disappearance of entities (such as an animal dying, for instance) are (not and should) not considered forms of conceptual engineering.

some properties that members of a given kind possess in all possible worlds. A certain kind of origin essentialism is in fact often coupled with externalist views since the seminal works of Kripke (Kripke, 1972) and Putnam (Putnam, 1970, 1975). Then, if meanings change beyond our control and our knowledge and many kinds possess essential properties that cannot be changed and that often are also not transparent to us, in which sense can concepts and meanings change in a philosophically interesting way?

In short, my reply to these worries is that this is a pseudo-problem, caused by some unhealthy philosophical attitudes. In fact, a brief look at the history of science, philosophy, or any other conceptual human activity shows that significant change and growth in our conceptual tools occurs indeed. Concepts and meanings, in whatever understanding you have of them, are not fixed and stable entities, but they change consistently with our human agendas. Only an overly-abstract philosophical attitude could negate the existence of such an ubiquitous phenomenon. More specifically, I will diagnose the root of this mistaken negative attitude towards the existence conceptual change in two philosophical theses commonly assumed in these discussions: meaning monism and metasemantic finality. I will call *meaning monism* the thesis that an adequate explanation of the meaning of most of our ordinary and scientific terms can be given relying on just one kind of component. What I will call *metasemantic finality* is instead the thesis that the factors that ground the meaning of linguistic entities are fixed and ascertainable in advance in their role, influence, and nature. I will show how both these two theses, despite they are implicitly assumed in contemporary metaphilosophical debates over conceptual engineering, are in stark contrast with a vast philosophical literature related to scientific conceptual change. As such, I will argue that an adequate account of conceptual change should embrace, instead of meaning monism and metasemantic finality, two opposite theses that I will call *meaning pluralism* and *metasemantic plasticity*.

In order to fully understand my defense of the existence of conceptual change, we first have to take a look at Kripke's and Putnam's seminal works on externalist semantics and at externalist solutions to the problem of conceptual change in philosophy of science.

2.3.1 Externalism, essentialism, and conceptual change

Kripke's and Putnam's works are often cited as clear evidence of why we should take a externalist approach to meaning. Moreover, their frameworks compose the blueprint upon which virtually all contemporary externalist (meta)semantics are based. Let us take a closer look at these works, then.

The first thing to note about Kripke's and Putnam's approach to meaning change is that, although they are both externalist and (to some extent) essentialist about meaning, they both clearly stress the possibility of various kinds of conceptual and meaning change.

Kripke stresses how linguistic aspects of meaning such as identifying marks or operational criteria (i.e. the descriptive part of meaning according to him) change consistently with our knowledge about a given kind, in a process directed (possibly) towards the reveal of the kind-essence (Kripke, 1972, pp. 128-133). Moreover, in discussing Evan's famous example of 'Madagascar' reference shift, from referring to a part of the African continent to

denoting the island we identify it with today, Kripke (Kripke, 1972, p. 163) explains that his picture of meaning is consistent with the fact that radical shifts of speakers' intentions in a community determine a shift in the reference of the related term.

Putnam agrees with Kripke in his explanation of Evans' 'Madagascar' example, leaving open the possibility of radical reference shifts. He also, far more than Kripke, stresses the significance of conceptual and intersubjective linguistic elements in the complex entity that is the meaning of a given term (cf. Putnam 1970, 1973). In Putnam's meaning vector (Putnam, 1975, p. 269), in fact, the mental stereotype (understood as a kind of prototype-like representation of significant features possessed by the members of a given kind), together with the syntactic and semantic markers, are (amongst) the intra- and inter-subjective components that shape how a community use and understand a given term.

Putnam's semantics allows us also to understand how significant meaning change can co-exist with an externalist and essentialist semantics. Putnam in fact stresses that both the externalist and the essentialist components of his semantics are a matter of degree. How essential a given semantic marker is in the meaning of a given term is in fact, according to Putnam, often not an all-or-nothing matter. The set of semantic markers of a given term can be understood as (partially, perhaps) ordered in terms of how essential they are to the meaning of the term. At one extreme of this order, one finds semantic markers that are completely contingent and can be changed without any impact on a term's overall meaning, whereas at the other end lie the markers that are pivotal element in the essence of a given term. In between these extremes, we can find a variety of semantic markers more or less important for the overall meaning of the term. Similarly, the partial determination of meaning by external factors can be understood as a gradual opposition between purely subjective and purely objective factors that compose the two poles of Putnam's meaning vectors. In between the purely psychological and the purely worldly components of the vectors, representing respectively the purely subjective and the purely objective meaning-determining factors, we can find (depending on the specific account of meaning involved) a variety of components that can be (partially) ordered from the most subjective and internal to the most objective and external.

The semantic significance of conceptual change can thus be understood in Putnam's semantics as the gradual change in the components of the meaning vector, where the change is more significant and more radical when it involves the change of more semantically entrenched (i.e. essential) and more objective (i.e. external) components of the vector. In such a picture, then, individual's use and control of meaning varies dependently on the structure of linguistic labor in a community and on the actual history of a given term. Charitable cooperation between speakers' intentions, uses, and interpretations allows inter-community (and inter-theoretical) identities and successful meaning change (Putnam, 1973, 1995).

Moreover, it should be noted that both Kripke's and Putnam's work must be understood as a reaction against the overly internalist and overly individualist semantics that constituted the received view in philosophy at the time. Thus, they strongly stress the external and communal character of some part of our language because that was the part

of linguistic phenomena related to meaning that, according to them, it was neglected by philosophers of their time. Putnam is very clear on this aim of his semantics:

> "Grotesquely mistaken views of language which are and always have been current reflect two specific and very central philosophical tendencies: the tendency to treat cognition as a purely individual matter and the tendency to ignore the world, insofar as it consists of more than the individual's 'observations'. (...) Traditional philosophy of language, like much of traditional philosophy, leaves out other people and the world; a better philosophy and a better science of language must encompass both" (Putnam, 1975, p. 271).

Thus, we must not interpret Kripke and Putnam as claiming that external, communal, and essential components of meaning are the only philosophically relevant components of related linguistic phenomena. Kripke and Putnam showed instead that meaning is a multi-faceted and complex phenomenon, where communal and external contributions are not negligible by any philosophical or linguistic theory that aims at giving an adequate description of language.

2.3.2 Meaning Pluralism in Philosophy of Science

This re-appraisal of Kripke's and Putnam's externalism as involving a plurality of meaning-determining factors, and therefore open to the possibility of significant meaning and conceptual change, is consistent with the development of externalist approaches to conceptual change in science. Externalist semantics have in fact been studied and heavily discussed in philosophy of science in connection with the aforementioned problem of conceptual change in scientific theory change (cf. Section 2.1). As already stressed by Putnam (Putnam, 1973), meaning externalism can explain, better than internalist semantics, the referential continuity of scientific terms. If internalist theories of meaning have to bridge the difference between the theoretical languages of two scientific theories with a complex syntactic translation, the world-based determination of reference championed by externalists provides an easy explanation of how different scientific theories can refer to the same entity. Even radical changes in the scientific description of a given theoretical term, such as the one common in scientific revolutions, do not pose a problem for the externalist. The sameness of reference is, in fact, held fixed by the worldly causal relationship between the radically different scientific descriptions and the natural phenomenon that they intend to describe (cf. Hardin and Rosenberg 1982).

Meaning externalism makes it easy, then, for different scientific terms to refer to the same natural phenomenon. Perhaps too easy, though. Externalist approaches to the problem of conceptual change in science have in fact been accused of making referential continuity trivial (cf. Laudan 1984b; Psillos 1999, 2018). If the burden of fixing the reference is put solely on worldly factors, then almost all scientific theories proposed in history of science successfully refer to the same natural phenomena referred to our best scientific theories. No matter how bad or conceptually misguided the description given by

a scientific theory of a certain phenomenon is, such a theory would correctly refer thanks to the hidden properties of the related external phenomenon. Laudan (Laudan, 1981, 1984b) argued that such a purely externalist conception of how scientific terms refer to the world gives us a completely unbelievable depiction of scientific activity. Even scientific terms that have completely disappeared from the scientific image of the world without any recognizable heir, such as the famous case of phlogiston, successfully referred according to the purely externalist picture thanks to the causal relationship between oxygen and the intended baptism of phlogiston theorists. So, phlogiston theorists, while they were trying to prove that oxygen was not a fundamental element of reality, successfully referred to the world thanks to the causal relationship between their theories and the very element they were disproving the existence of. This depiction of scientific theories reference is, even for defenders of scientific realism like Psillos (Psillos, 1999), utterly absurd, making referential success an entirely trivial matter completely independent from actual scientific activity.

Note that this easiness of referential continuity in purely externalist semantics is the same exact phenomenon behind the difficulty of conceptual engineering stressed by Cappelen and its fellow externalist conceptual engineers. Intensionally changing a concept extension is difficult because referential continuity is incredibly easy and vice versa. So, the same historical arguments that Laudan and Psillos gave against overly externalist account of scientific term reference can be applied to Cappelen-like pictures of conceptual engineering and conceptual change. These pictures make the history of science absurd, locking the reference of failed scientific theory of the past to phenomena completely unknown to (and even explicitly denied by) them and as such they should be abandoned.

Luckily, if philosophy of science provides strong arguments for the failure of extremely externalist conception of meaning, it can also give us some possible solution to the problem of understanding conceptual change within an externalist (meta)semantics. A natural solution to this problem is, consistently with our previous re-assessment of the original pluralistic aim of Kripke's and Putnam's externalism, to hold a more inclusive view of meaning and reference where a multiplicity of components determines how scientific term refer to the world. As Psillos (Psillos, 1999) stressed, in fact, even strongly externalist approaches to scientific meaning have to take into considerations some kind of theory-laden structural component in the process of fixing the reference of a natural kind term. If in the case of proper names original baptism seems a transparent way of fixing reference, the reference-fixing process for natural kind terms is often dependent on some kind of theoretical framework, i.e. it happens inside a given theoretical picture of the world. Negating this descriptive aspect of the reference-fixing process would lead to the trivial depiction of referential success incompatible with the history of science that we have criticized before. In order not to make referential success too easy, then, reference-fixing must include a descriptive component. This is the main insight of the so-called causal-descriptive theories of reference (Lewis, 1984; Kroon, 1985). Referential success is seen by these theories as the combined product of external causation and theory-laden causal explanations. Psillos' own version of such causal-descriptive theory of reference crucially involves the notion of kind-constitutive properties (Enç, 1976) and it is particularly apt to show how the addition of a descriptive component allows externalists to have a non-trivial view of referential

continuity in science.

Psillos' (Psillos, 1999) theory of how scientific terms refer to the world is centered around the notion of a core-causal description associate with a term, i.e. the set of properties through which a theory explains the kind-constitutive properties by virtue of which the referent of the term it is supposed to play a given causal role. A scientific term successfully refers to a given entity when the kind-constitutive properties of the entity correspond to the ones postulated by the core-causal description that a given theory associates with the scientific term. In this way, in Psillos' theory, the reference of a term is jointly determined by an external causal element (i.e. the causal origin of the information that ultimately fixes the reference) and by a descriptive element (i.e. the theory-laden core-causal description). Consistently, referential continuity involves two conditions, the sameness of the causal role played by the putative referents of the terms together with the identity of the core-causal description associated with the terms. This required identity of core-causal descriptions ensures that there is a substantial overlap between the properties through which two co-referring theories explain the attributed causal role of a given term. In this amended externalist semantics, then, referential continuity in science is not at all a trivial matter. As Psillos shows, his theory allows scientific theories that share the central part of their causal description of a given phenomenon to co-refer, while it forbids theories that give radically opposite description of a given situation to refer to the same phenomenon. In this way, causal-descriptive theories of reference can distinguish historical episodes of referential continuity between subsequent theories, such as the case of luminous ether in 19th century optics (Psillos, 1999, pp. 125-139, 282-287), and cases of reference change, such as the aforementioned case of the chemical revolution.

The example of Psillos' causal-descriptive theory of reference for scientific term teaches us then the same moral than our previous re-assessment of Kripke's and Putnam's seminal takes on externalism: meaning monism ought to be abandoned in favor of *meaning pluralism*, i.e. the recognition that a plurality of meaning-component is needed in order to have an adequate metasemantic theory. In other words, a healthy attitude in (meta)semantics must recognize a multiplicity of meaning-determining components in order to achieve an adequate account of referential success, stability, and change. This is the first lesson that traditional debates over conceptual change in philosophy of language and philosophy of science can teach to metaphilosophical contemporary debates over conceptual change and conceptual engineering.

2.3.3 Metasemantic plasticity and the conceptual change locks

After we saw how the philosophy of science literature related to scientific conceptual change prompts us to abandon meaning monism in favor of meaning pluralism, it is now time to see how the same literature gives convincing evidence for abandoning another metasemantic assumption common in metaphilosophical debates over conceptual engineering and conceptual change. At the beginning of this section, I claimed that contemporary discussions over conceptual engineering are relying on a common yet completely unjustified assumption that I dubbed *metasemantic finality*, i.e. the statement that the factors that ground

the meanings of linguistic entities are monolithically fixed and ascertainable in advance in their role, influence, and nature.

According to metasemantic finality, then, metasemantic philosophical theories can ascertain which kind of factors (e.g. internal or external, cognitive or noncognitive, individual or communal, etc.) ground the meaning of a given (type of) element of our language. Furthemore, these kind of theories postulate a given metasemantic kind of meaning and reference for entire classes of elements of our language, such as proper names or natural kinds, leaving up no space for individual differences in how specific terms of our language are significant and refer to the world. This is the kind of overgeneralizing tendencies in philosophy that Putnam's semantical work reacted to (cf. Putnam 1970). More importantly, this fixed conception of metasemantical facts is not argued for and arguably wrong.

The fixity and uniqueness of metasemantical facts stands in fact in stark contrast with the long-recognized plasticity and context-dependency that many scientific terms exhibit in how they are used in our best scientific theories (cf. Cartwright 1983; Batterman 2001; Wilson 2006). Any reasonable account of scientific terms meaning and reference has to take into account the complex bundle of semantic and pragmatic factors that allow our best scientific theories to refer (perhaps directly or perhaps indirectly) to the world (Kuhn, 1976, 1990; Chang, 2012). A given term can have a very different meaning in the various contexts in which it is used, complex semantic architectures cause terms to refer to different things in different applications of the same theory, a certain kind of localized holism is inevitably observed in many different parts of science (Wilson, 1982, 1994, 2017). Long story short, synchronic and diachronic meaning change is omnipresent in scientific practice. Contemporary semantics for scientific theories and terms semantic structure have no place for the metasemantic finality that it is assumed in aforementioned debate over externalism and conceptual engineering.

Instead of metasemantic finality, we must recognize a *metasemantic plasticity* in conceptual affairs, i.e. we have to acknowledge that the factors grounding the meaning of our scientific terms are often not general in nature, they are mostly bound to change according to the practical need of science, and they are only gradually ascertainable via a case-by-case painstaking analysis of the history of how a certain term was (and is) used in all the related scientific contexts.

In order to understand better the implications of the metasemantic plasticity in conceptual affairs that I am proposing, I will use an engineering metaphor inspired by Simon's seminal discussion of complexity (Simon, 1981). Assume that the meaning of a given term is very generally understood, like Putnam did, as a kind of vector, the components of which are all the factors that contribute to the meaning such as the related concepts, beliefs, syntactic markers, semantic markers, speaker's meanings, community meanings, indexicals, contexts, stereotypes, extensions. Dependent on the quantity of meaning components that one identifies, the meaning (understood, again, in the most general sense of the term) of a term could be equated with a n-tuple $m = \langle c_1, \ldots, c_n \rangle$ where $c_1, \ldots, c_n$ are the different meaning components. Now, replace every component of the vector with the set of all possible entities that might figure as that component in the meaning vector of that term (in contemporary metaphysical terms, the domain of the variance thesis for that component).

So that, instead of each meaning component c_i, we now have a (possibly infinite) set of possible components $C_i = \{c_1, \ldots, c_i, \ldots\}$. Call the vector $m_C = \langle C_1, \ldots, C_n \rangle$ so construed the variance meaning vector of a given term. This is the vector having as components the sets of all the possible components of the meaning vector of the given term.

Thus, all the possible changes in the meaning of a given term are represented by all the meaning vectors that can be construed by taking a member from each set of possible components C_i. Some of these meanings would be actually instantiated, some would be only possibilities, others again would arguably be very likely to be never instantiated because of their impracticality (e.g. incompatible intensions and extensions). Now, imagine to print all the members of each set of possible components C_i on the wheel of a (quite big!) lock that has n-wheels. Call this imaginary lock the *conceptual change lock*. Imagine that this lock is programmed to open only if (one of) a certain (set of) combination(s), i.e. a certain possible meaning vector for the term under focus, is reached. Think about this sets of combinations as the favored meanings (again, in the general sense of the term) that are optimal given a certain normative account of conceptual change for that term. The nature of this normative account is not at issue here, pick your favorite one from the existing literature on conceptual change and conceptual engineering, be it a metaphysical, an epistemic, a pragmatic, or a social-political one. Of course, these combinations would be constrained by a variety of semantic and meta-semantic, internal and external, factors. Again, the nature of these combinations and how are they constrained is not at issue here. What is important is that conceptual change and conceptual engineering can then be thought as the activity of resolving this conceptual change. More specifically, the general phenomenon of conceptual change corresponds to any resolution of this lock, while conceptual engineering corresponds to specific intentionally designed resolutions.

Now, lock-puzzles can be of different kinds, corresponding to different levels of complexity. Simon (Simon, 1981) distinguishes for instance between complex locks, i.e. locks where each cog solution is completely independent from the ones of the other cogs, and simple locks, i.e. locks where the solution of a given cog is constrained by the ones of other cogs. Wimsatt (Wimsatt, 1986) proposed a mixed lock that he call the developmental lock, i.e. a lock where solution are constrained only in one direction, as the lock more adequate to describe the development of science. In general, different types of lock depend on whether and how the solution to each wheel is dependent on the correct solution of other wheels.

Which kind of lock is then instantiated by the conceptual change lock? Strong externalism might be thought as assuming that the conceptual change lock is a lock where the wheel corresponding to the extension determines all the other right combinations, while strong internalists might assume that the wheels corresponding to the internal components of meaning (such as the concept or a speaker's meaning) determine all the others. In general, metasemantic positions over the possibility of conceptual engineering and conceptual change can be translated as claims over which type of lock the conceptual change lock instantiates. The conceptual change lock helps us model also claims of degrees of control and knowledge in conceptual change and engineering. The degree of control over a given meaning component might be represented by the degree of manual control in turning the relate wheel, with the possibility of perfectly tuning a given wheel representing the

complete control on that component, while the complete lack of any intentional rotation represents an extreme lack of control over the related meaning component. Intermediate degrees of control over a meaning component are then represented by imperfect control in the intentional tuning of the related wheel. The degree of knowledge about a given meaning component can then be represented by the amount of printed possible components in the related wheel. Total (modal) knowledge over a component can be represented by a wheel where all the possible components are printed, while complete absence of such knowledge might be modeled by a blank wheel that tells us nothing about which component are we choosing. Intermediate cases of incomplete knowledge can spelled out as wheels where only a proper subset of the possible meaning components are printed.

So, which kind of lock is the conceptual change lock? How do its wheels work? Which wheel depends on which wheel? Which amount of control on its wheel do we wave? What is the amount of printed meaning components we can read? My answer to these questions is that there is no reason to assume a single answer for all the possible conceptual change locks. Different kinds of concepts can instantiate different kinds of locks. Take formal mathematical concepts, for instance. Such concepts are likely to instantiate very internalist and controllable kind of locks, since worldly factors have virtually no role in determining their meaning. Natural kind concepts, instead, are likely to be instantiated by more externalist and not fully controllable locks, since their meanings are dependent also on how the world is actually structured. Moreover, even amongst concepts of the same kind, locks for different terms would correspond to different kinds of locks, differing in the complexity type, the functioning, the degree of control and knowledge that we can have. My approach to the problem of conceptual change locks is that in order to know how a given conceptual change lock works there is no substitute to actually trying to solve it. Only by fiddling with it we can learn its exact nature. In plain words, there is no substitute to actually study the history and the uses of a given term. No two terms are absolutely the same. Even linguistic entities with similar properties and uses may hide significant differences in the nature and the working of the factors that compose their meanings. This is why the aforementioned metasemantic finality commonly assumed in recent debates over conceptual engineering is a mistaken way of looking at the problem of conceptual change. Only by studying a given conceptual history we can understand which kind of changes a concept supported in the past and which kind of metasemantical facts grounded and ground its meaning. Only after this necessary, painful and time-consuming, step, we can make an informed judgment on whether and how it is possible to intentionally change it. This is why any adequate theory of how concepts change must embrace metasemantic plasticity.

Consistently with my approach to conceptual change based on meaning pluralism and metasemantic plasticity, the novel methodology that I sketched in the Introduction and that I will use in this work will allow lots of flexibility and context-dependency in analyzing conceptual change. In the many models and historical episodes of conceptual change that I will analyze, I will use a kind of charitable interpretation of what scientists and philosophers involved in the episode claim to be happening in all the aspects of the meaning of the term under focus. Thus, when scientists or philosophers spoke of conceptual change or of two

different concepts, I will try to give an analysis as consistent as possible with the linguistic practice of the people involved. This may cause, for readers supporting metasemantic finality in conceptual affairs, a dose of semantic and metasemantic promiscuity. I pledge guilty as charged, with the confidence that, on top of the aforementioned inconsistency of metasemantic finality with our best accounts of scientific language use, the sum of the case studies contained in this work will show that this promiscuity is an inevitable assumption of any study of conceptual change in science and in philosophy that wants to be adequate to the history of science and philosophy. Moreover, as I will show in the Conclusions chapter, the collective analysis of the many different models of conceptual change and related case studies carried out in this work will demonstrate how, despite this semantic and metasemantic promiscuity, a general conception of conceptual change will emerge.

2.4 The Toolbox Framework: A Meta-Framework for Evaluating and Comparing Models of Conceptual Change

In this section, I will present a normative meta-framework for evaluating and comparing models of conceptual change along nine dimensions. These dimensions are taken from the components and the differences that we saw in this chapter as determining different opinions in the philosophical and psychological literature about what concepts and conceptual change are and why and how they are important to science and philosophy. I will call this meta-framework the *Toolbox framework* and I will use it for assessing, judging, and comparing all the four kinds of model of conceptual change that we will see in this thesis. The Toolbox framework is made of nine evaluative dimensions: units of selection, concept ontology, concept structure, kinds and degrees of conceptual change, degree of normativity, effectiveness of normative judgment, assumptions and consequences for conceptual change in science, assumptions and consequence for conceptual change in philosophy, metaphilosophical assumptions and implications. Let us survey these dimensions one by one.

Units of selection Along this dimension, models of conceptual change are judged, assessed, and compared by the level of abstraction at which they identify conceptual entities as meaningful units of change. Options can range between very fine-grained level of abstraction, such as the one where single concepts or even single conceptual parts are taken as units, to very coarse-grained levels where models work semi-holistically on macro-units of change like clusters of interrelated concepts or entire theories. Amongst these two extremes, several abstraction levels present themselves, differing in whether they use linguistic and conceptual types or tokens as their units and in how they identify the meaningful structures of change. As we will see, some models can also be thought as having implementable variants that can work at different levels of abstraction.

Concept ontology This dimension focuses on the compatibility of a given model of conceptual change with the different philosophical positions on the ontology of concepts. As we saw in Section 1.1., the four major bundles of ontological positions are the psychological, the abstract view, the linguistic, and the worldly view. As we will see, some models of conceptual change are more compatible with a psychological or an linguistic view of concepts, while others are instead more easy to be implemented together with an abstract or a worldly view of conceptual entities. Other models are instead quite neutral on this dimension, not imposing (almost) any constraint on which ontological view of concepts they are coupled with. I will spell out for each model how a suitable implementation with a given ontological view of concept would look like.

Concept structure This dimension focuses instead on how a given model of conceptual change assumes the structure of concepts to be constituted. As we saw in Section 1.2, there is a vast range of philosophical and psychological theories of conceptual structure. Analogously to the conceptual ontology dimension, I will use this dimension to assess the compatibility of a given model with the bundle of theories of concepts I presented in the first section of this chapter. We will see that some models assume a specific theory of conceptual structure as a necessary background, while others are compatible with a variety of positions on what exactly the structure of concepts consists of.

Kinds and Degrees of conceptual change This dimension focuses on the kinds and degrees of conceptual change that a given model of conceptual change identifies. We will see that some models postulate a fine-grained hierarchy of kinds (or/and degrees) of conceptual change, ordering these kinds of change by the severity of the change involved, while other models treat very different episodes of conceptual change as examples of the same phenomenon. Other models specifically target instead only certain types of conceptual change, due to certain background assumptions on their epistemological and semantic significance.

Degree of normativity This dimension tracks the extent to which a given model of conceptual change is more or less normative in judging episodes of conceptual change. Extremes can range between models that do not give us any way of normatively assessing conceptual histories, confining themselves to a descriptive approach, and models that instead come equipped with very normative selection mechanisms thanks to which we can judge a given historical change to be more or less rational. We will see that many models strike a middle ground between these two extremes, leaving a small place for a kind of normativity heavily dependent on values and goals.

Effectiveness of normative judgment This dimension focuses on how effective the normative judgment of a model of conceptual change is (in the case of models having such a normative component). Different models give normative judgments that vary a lot in their effectiveness, from pragmatic heuristic procedures for preferring a given concept to its

competitors to truly algorithmic procedures of how a given episodes of conceptual change should (have) go(ne). We will see that many models assign different normative kinds of judgment to the different kinds of conceptual change they identify.

Assumptions/consequences for conceptual change in science In this dimension, I will assess the assumptions and the consequences of a given model of conceptual change in relation to the problems that scientific conceptual change poses in philosophy of science. As we have seen in Section 2.1, conceptual change in science poses several issues for scientific progress, objectivity, and rationality. Different models of conceptual change have then different implications for these issues, being more or less compatible with different solutions.

Assumptions/consequences for conceptual change in philosophy In this dimension, I will assess the assumptions and the consequences of a given model of conceptual change in relation to the problems that philosophical conceptual change poses in philosophy. As we have seen in Section 2.2, different philosophical methodologies put different degrees of significance in the phenomenon of conceptual change in philosophy. Different models of conceptual change have then different implications for these issues, describing conceptual change in philosophy as a more or less frequent and philosophically relevant phenomenon and consequently being more or less compatible with different philosophical methodologies.

Metaphilosophical assumptions and implications In this dimension I will focus on the metaphilosophical background that a given model of conceptual change has. More specifically, I will analyze which conception of philosophical activity lies behind a given model and whether different metaphilosophical positions are compatible with it. I will also discuss whether and how a given model of conceptual change could be an adequate basis for the implementation of a conceptual engineering program.

Chapter 3

Carnapian Explication

This chapter will focus on Carnapian explication, i.e. the pragmatic model of conceptual change central to Carnap's mature philosophical thought. After receiving a limited amount of attention for years, the method of explication has been re-appraised, together with the originality of Carnap's later writings, by a more recent and more informed historical scholarship (e.g. Coffa 1991; Friedman 1999, 2000; Carsten and Awodey 2004; Friedman and Creath 2007; Carus 2007; Wagner 2012). In recent years, the procedure of explication has also been discussed by philosophers outside the Carnapian tradition, thanks to the aforementioned (cf. Chapter 2, Section 2.2) re-appreciation of normative methodologies in analytic philosophy in the debates about conceptual analysis and conceptual engineering in philosophy.

In this chapter, I will analyze Carnapian explication both from a historical and from an abstract epistemological point of view. For what concerns the history of Carnapian explication, I will trace back the development of the method of explication in Carnap's philosophical methodology, from the early focus on rational reconstruction to the later explicit use of explication. More specifically, I will show how the ideal of explication fully embodies five central ideals of Carnap's metaphilosophy, the influence of which can be traced throughout Carnap's whole intellectual life. From an abstract epistemological perspective, instead, I will present the features of Carnapian explication as a method of conceptual change, focusing on the metaphilosophical debates about its features and its alleged limitations. I will then propose two, complementary, improvements of Carnapian explication that aim at broadening its scope and making it a more exact and useful model of conceptual change. I will first propose a more fine-grained, three step version of Carnapian explication, suitable to analyze more complex episodes of conceptual change from the history of science. My second modification of Carnapian explication will consist instead in an explication of the concept of explication itself. Specifically, I will show how the procedure of explication can be formalized within the theory of conceptual spaces.

In Section 1, I will present the historical roots of Carnapian explication, focusing on the development of Carnap's philosophical methodology and showing how this development can be understood from the perspective of five central ideals in Carnap's metaphilosophy. In Section 2, I will present the procedure of Carnapian explication from an abstract episte-

mological perspective, discussing its steps, its scope, and its aims. I will specifically focus on recent discussions about the desiderata of Carnapian explication and on general critiques of the usefulness of explication as a method of conceptual change and engineering. In Section 3, I will present a modification of Carnapian explication that adds a mid-level step to the procedure. With the help of a very detailed case study centered around the Church-Turing Thesis and the different axiomatizations of the intuitive concept of calculability, I will show how my three-step version of explication is able to adequately treat complex cases of explications for which the canonical two-step version of explication is not fine-grained enough. In Section 4, I will instead present a formal model of the whole procedure of Carnapian explication built inside the theory of conceptual spaces. With the help of multiple examples and two case studies, I will show how my formal explication of 'explication' is able to make precise the procedure of explication and (many of) its desiderata, defending it from critiques about the alleged narrowness and difficulty of its application. Finally, in Section 5, I will assess the features of Carnapian explication as a general method of conceptual change in science and in philosophy, analyzing it along the nine dimension of my Toolbox framework.

3.1 The Development of Carnap's Methodology: from Rational Reconstruction to Explication

In this section, I will analyze the history of Carnap's explication within the general development of Carnap's (meta)philosophical views. In recent years, the history of Carnap's philosophical development and the originality of his mature thought has been narrated many times (e.g. Carus 2007; Leitgeb and Carus 2020). The aim of this section will not be a full analysis of Carnap's intellectual development, but a more specific focus on how the concept of explication has evolved from Carnap's early method of rational reconstruction together with certain specific changes in Carnap's overall philosophy.

I will analyze this evolution in a step-wise manner, focusing on three different moments in Carnap's metaphilosophical development, broadly corresponding to three time periods in which Carnap's intellectual biography can be divided. I will first focus on Carnap's method of rational reconstruction, biographically corresponding to Carnap's early works and paradigmatically exemplified by the *Aufbau* (Carnap, 1928a). The second metaphilosophical moment in Carnap's development that I will treat is the transition phase between rational reconstruction and explication, broadly corresponding to the Vienna and Prague years of Carnap's life paradigmatically exemplified by his views as expressed in the *Logical Syntax* (Carnap, 1934). Finally, I will describe, as the third metaphilosophical moment, Carnap's explicit presentation and use of explication, biographically corresponding to Carnap's time in the United States and intellectually exemplified by his later works on probability (Carnap, 1950b).

We will see how the description of these three different moments in Carnap's metaphilosophical development will show how Carnap's explication is the byproduct of the more

general evolution of Carnap's philosophy and metaphilosophy. It will also be clear how, in the passage from rational reconstruction to explication, the central characteristics of Carnap's metaphilosophy stay fixed, while their philosophical and technical implementation changed. So that we can trace in Carnap's explication the output of how Carnap's ideal of rational reconstruction and its related metaphilosophical stance evolved due to the environmental pressure of the technical and philosophical problems they faced.

3.1.1 A primer in Carnap's metaphilosophy

Carnap's approach to philosophy, since his early works, exhibits a metaphilosophical stance radically different from the ones common in philosophy at the time (or even nowadays). I will describe his distinctive metaphilosophy focusing on five central ideals that guided his approach to philosophy: *constructivism*, *positivism*, *logicism*, *structuralism*, and *pluralism*.

Constructivism In direct contrast to most of philosophers, Carnap conceived philosophical activity mainly as a constructive effort. Philosophical problems were never for Carnap meant to be merely discussed or analyzed, but they were instead meant to be solved. Moreover, this solution was not supposed to come only through philosophical reflections, but it crucially involved the engineering-like activity of constructing tools for solving the problem at issue (Creath, 1991; Richardson, 2013). This constructivist effort can be paradigmatically seen in Carnap's incessant construction of formal linguistic frameworks devoted to the resolution of philosophical and scientific problems, from the constitution systems in the *Aufbau* (Carnap, 1928a) to the systems of inductive logic that he developed up until his last years (Carnap, 1950b, 1952; Carnap and Jeffrey, 1971).

Positivism A lasting influence on Carnap's metaphilosophy was Neo-Kantianism, especially in the form of the scientifically-minded type championed by the so-called Marburg school (cf. Friedman 2000; Carus 2007). The list of Neo-Kantian aspects of Carnap's philosophy is long, but a central one is Carnap's central preoccupation with "the fact of science". This preoccupation is closely connected with Carnap's everlasting positivism. Throughout his developing views, in fact, Carnap always held the empirical sciences as the benchmark of knowledge, from which philosophical reflection must always start and ultimately defer, and empirical confirmation as the ultimate tribunal of any philosophical or scientific theory. Carnap's positivism is also exemplified by his lasting quest for overcoming metaphysics (at least in its traditional pseudo-scientific form) (Friedman, 2012) or the voluntarism that permeates much of Carnap's philosophical attitude (Jeffrey, 1994). Carnap's positivism can be spotted throughout his philosophical activity, from the early quest for revealing metaphysical issues as pseudo-problems (Carnap, 1928b) to the later program of translating quasi-syntactical sentences into the formal mode of speech (Carnap, 1934).

Logicism Together with Neo-Kantianism, the other main component of Carnap's philosophical heritage, as it is customarily presented, is constituted by Frege's (Frege, 1884) and Russell's (Russell, 1914; Whitehead and Russell, 1910-1913) logically-minded philosophies and the related technical advances in mathematical logic. The influence of Frege and Russell is exemplified by the central role that logical methods play in Carnap's philosophy and the exceptional status of logic in Carnap's general epistemological views. Logical tools are the main technical tools through which Carnap pursues his aforementioned activity of constructing formal linguistic frameworks. Moreover, analyzing language through the lens of formal logics is for Carnap a central epistemological task in any tentative resolution of philosophical and scientific problems. We can then speak of a broadly logicist attitude in Carnap's philosophy for describing the centrality of logic and logical tools in Carnap's overall philosophical projects[1]. Examples of this kind of logicism can be found in almost every major work of Carnap, from the pivotal role of logical abstraction in the *Aufbau* (Carnap, 1928a) to the technical efforts towards developing a satisfying intensional semantics or towards clarifying the syntactic and semantic structure of logical languages in his later works (e.g. Carnap 1934, 1947).

Structuralism Another ideal of Carnap's metaphilosophy closely connected to Carnap's logicism is his structuralism. Carnap, in fact, understood his logical reconstruction of theories as making evident their structural content. Carnap's focus on the structural content of theories pervades his works in epistemology, philosophy of mathematics, and philosophy of science. In epistemology, Carnap reconstructed cognitive phenomena via logically abstracting their structural relations, understanding this logical structure as the cognitive component essential for epistemological purposes (Carnap, 1928a,b). In philosophy of mathematics, Carnap gave several logical analyses of the structural content of mathematical theories, trying to spell out this structural content via axiomatic definitions, logical constructions, and definitions by abstraction (Schiemer, 2020b). In philosophy of science, Carnap repeatedly tried to devise logical frameworks in which the logical structure of scientific theories could be analyzed, championing a double view of the language of scientific theory that foresaw some central aspects of structuralist approaches in philosophy of science (Carnap, 1956, 1961, 1966).

Pluralism Finally, the fifth ideal of Carnap's philosophy can be identified in his pluralism on philosophical and scientific matters. In his career, Carnap always tried to mediate between different philosophical and scientific positions, promoting a pluralist view of philosophy and science in which multiple approaches to a problem could and should be pursued.

[1]Note that Carnap's logicist ideal takes also the form of a logicist position in philosophy of mathematics in some of his early work in philosophy of mathematics (cf. Goldfarb 1996). As we will see, however, Carnap after the tolerance turn will assume a neutral position towards the foundational debate and as such it cannot be considered anymore a logicist in this strict sense. Therefore, the broad logicism of Carnap's metaphilosophy stressed here should not be equated with a logicist position in philosophy of mathematics.

He also often tried to achieve a neutralist perspective on a given issue, showing the common assumptions and problems of rival views. This pluralist approach is evident in many specific Carnapian theses, from his neutralist stance on the ultimate basis of our knowledge in the *Aufbau* (Carnap, 1928a), to the famous proclamation of the principle of tolerance in logic (Carnap, 1934) and his related meta-ontological pluralist attitude (Carnap, 1950a).

As we will see, these five ideals remained a constant guide for Carnap in his philosophical development and they underlie all the different steps in philosophical methodology that I am going to analyze in the rest of the section. However, as it will be clear in what follows, how exactly the philosophical implications of these five ideals are spelled out varies in the three different methodologies that we are going to see, consistently with the technical and philosophical problems at issue and with the overall evolution of Carnap's philosophical views.

3.1.2 The method of rational reconstruction

The first step in my analysis of Carnap's metaphilosophical development concerns the method of rational reconstruction. This is how Carnap himself explicitly conceptualized his approach to philosophical problems in his early works.

As stressed by many scholars (Friedman, 2000; Carus, 2007; Beaney, 2013), the philosophical background of rational reconstruction can be traced back to Carnap's double heritage of Neo-Kantianism and Logicism. From the Neo-Kantian distinction between the genesis and the validity of knowledge, Carnap's method of rational reconstruction takes the "as-if" approach to understanding the rationality of a given phenomenon. In rationally reconstructing a certain process, in fact, Carnap does not aim at a faithful actual description or history of it, but he wants instead to reconstruct its rational structure, understood as a fictional construction able to justify how the actual process could have taken place. From Logicism, instead, Carnap's rational reconstruction takes the formal tools through which the reconstruction is performed, such as the fundamental technique of logical abstraction. It is only through logical analysis, for the early Carnap, that the structural rational content of a given phenomenon can be adequately reconstructed.

The methodology of rational reconstruction consists then, for the early Carnap, in the fictional reconstruction of a given phenomenon in a logical language, aimed at making evident the rationality of its structural components. In Carnap's own words, rational reconstruction is "a schematized description of an imaginary procedure, consisting of rationally prescribed steps, which could lead to essentially the same results as the actual (...) process" (Carnap, 1963a, p. 15).

The paradigmatic example of rational reconstruction is the reconstruction of our construction of reality that Carnap attempted in the *Aufbau* (Carnap, 1928a)[2]. More specifically, Carnap wanted to show how all our concepts can be reduced by logical operations to

[2]As it is well known, Carnap's attempted reconstruction project failed. However, in more recent years, scholars have showed how weaker versions of Carnap's reduction problem can be carried out (Mormann, 1994; Leitgeb, 2007, 2011)

a given basis of knowledge, thus logically justifying the possibility of knowledge itself. This rational reconstruction is framed in terms of what Carnap calls 'constitution systems', i.e. logical conceptual frameworks where concepts are hierarchically ordered in terms of their logical complexity, from the basic concepts to all the other concepts that can be logically constructed from them.

The rational reconstruction of cognitive phenomena via constitution systems attempted by Carnap in the *Aufbau* shows how the methodology of rational reconstruction embodies the aforementioned five ideals of Carnap's metaphilosophy: constructivism, positivism, logicism, structuralism, and pluralism.

Carnap's constructivist ideal is exemplified by Carnap's constitution systems, one of the first forms in which Carnap's linguistic engineering manifests itself. Moreover, Carnap's engineering attitude towards philosophical problems can be seen by the fact that the actual construction of the phenomenalist constitutional system, carried out in detail, occupies a significant part of the *Aufbau* pages (Carnap, 1928a, Part IV). Carnap does not just offer a programmatic sketch of a constitution system, he actually constructs one, aiming to show how the alleged construction of reality envisaged by Russell (Russell, 1914) can be carried out. Carnap's positivist ideal is exemplified by Carnap's general philosophical motivations in the *Aufbau* (Carnap, 1928a, pp. 5-11), namely the desire of establishing a scientific epistemology and the related strive to purge philosophical discussion from any metaphysics and ambiguity in methods and language (which, according to the early Carnap were two inherently connected phenomena, cf. Carnap 1928b). The rational reconstruction of cognitive phenomena attempted in the *Aufbau* is then Carnap's way of replacing traditional epistemological debates with a scientifically minded epistemology. Carnap's logicism can be seen in the centrality of logical methods in Carnap's (re)construction project. Constitution systems are, in fact, nothing but logical hierarchies of types strongly inspired by Whitehead's and Russell's works in mathematical logic (Whitehead and Russell, 1910-1913). Moreover, the construction of complex concepts from basic ones within the phenomenalist constitution system carried out in the *Aufbau* crucially uses Frege's and Russell's methodology of logical abstraction and explicit definitions in logically reducing objects of a given type to objects of a simpler one. The structuralist ideal is also central to the *Aufbau* project. Carnap's reconstruction of cognitive phenomena is in fact explicitly focused on reconstructing the structural content of knowledge processes, understood by Carnap as the essential epistemological component of our cognition that logical tools can reveal. Finally, in Carnap's neutralist stance towards traditional debates over the ultimate basis of knowledge and in the related plurality of possible constitution systems we can see the exemplification of Carnap's pluralism. Even though, in the *Aufbau*, Carnap actually constructs only a constitution system with a phenomenalist basis, he explicitly mentions the possibility and viability of constructing constitution systems with a physicalist and even a cultural basis.

We have then seen how the methodology of rational reconstruction, as exemplified in the *Aufbau*, embodies the five aforementioned ideals of Carnap's metaphilosophy. Rational reconstruction provides, for the early Carnap, a way of replacing actual phenomena with their reconstructed logical structure, allowing a resolution of traditional philosophical prob-

lems that can be seen, from the perspective of rational reconstruction, as stemming from the ambiguity and confusion of the traditional language and methodology of philosophy.

3.1.3 From rational reconstruction to explication

My second step in the analysis of Carnap's metaphilosophical development will be devoted to a transition phase in Carnap's methodology between the early focus on rational reconstruction and the mature explicit embrace of explication. Biographically, this transition phase broadly correspond to the Vienna and Prague years of Carnap (roughly from 1926 to 1934) and to the heyday of the Vienna Circle and Carnap's involvement in it.

The history of Carnap's intellectual development in the Vienna years (and the related rise of the Vienna Circle movement) has been narrated many times in philosophical and historical literature (cf. Carus 2007; Stadler 2015) and I will not recount it here. My narrower focus will be instead on Carnap's evolving philosophical methodology in the Vienna years and especially on how some technical problems and philosophical tension prompted a gradual change in Carnap's method of rational reconstruction, the output of which was his later procedure of explication. As a paradigmatic example of Carnap's evolving methodology, I will take Carnap's efforts in philosophy of mathematics contained in the *Logical Syntax* (Carnap, 1934).

The main philosophical aim of the *Logical Syntax* is the construction of a canonical meta-language for science, where logical analysis of a given scientific theory or phenomenon can be carried out. More specifically, Carnap wanted to show how such a logical meta-language for science could be constructed by purely syntactical means. The main technical part of the book is devoted to the construction of two such formal languages, Language I and Language II, devoted to the reconstruction of mathematical theories and, according to Carnap, broadly corresponding to (respectively) intuitionistic and classical mathematicians stances on the foundations of mathematics. The history and the actual aims and scope of Carnap's *Logical Syntax* are quite complex and have been at the center of much historical and philosophical debates (e.g. Friedman 1999; Awodey and Carus 2003, 2007; Wagner 2009; Creath 2012). For our present aim, i.e. the appreciation of Carnap's evolving philosophical methodology, there are two crucial metaphilosophical ideas in the *Logical Syntax*: the principle of tolerance and the distinction between the material and the formal mode of speech.

The principle of tolerance, which has been the center of a lot of philosophical discussion (e.g. Awodey and Carus 2009; Creath 2009; Yap 2010; Steinberger 2016), states a(n almost) complete pluralism in matters of formal language construction. This pluralist position was in total contrast with the (at the time) heated disputes on which features of the language of science should or should not be admitted, such as for instance the discussion on the viability of non-constructive methods in mathematics (Mancosu, 1997) or the so-called protocol-sentence debate within the Vienna Circle (Uebel, 2007). Carnap's revolutionary solution to these disputes is to transform intransigent positions over the correct language of science into different linguistic proposals that might be useful for different purposes. In a slogan, according to Carnap, in philosophy and science we ought to pass from prohibitions

to conventions:

> "*Principle of Tolerance: It is not our business to set up prohibitions, but to arrive at conventions.* (...) *In logic, there are no morals.* Everyone is at liberty to build up his own logic, i.e., his own form of language, as he wishes. All that is required of him is that, if he wishes to discuss it, he must state his methods clearly, and give syntactical rules instead of philosophical arguments" (Carnap, 1934, pp. 51-52).

The other main metaphilosophical idea that Carnap presented in the *Logical Syntax* is the distinction between the material and the formal mode of speech (Carnap, 1934, p. 239). In a nutshell, the material mode of speech involves reference to extra-linguistic objects and relations, while reference in the formal mode of speech is exclusively intra-linguistic. So that, for instance, the sentence 'five is a number' is in the material mode of speech, but it can be rendered into the formal mode of speech as "five' is a number-word'. This distinction is a crucial part in Carnap's aforementioned plan of substituting philosophy with logical analysis in the canonical meta-language of science. In the logic of science, in fact, seemingly metaphysical statements (called by Carnap 'pseudo-object' or 'quasi-syntactical' sentences Carnap 1934, pp. 284-285) such as existence statements about abstract entities can be rationally reconstructed as purely syntactical statements about the syntactical structure of the related linguistic entities in the language of science.

The combination of the principle of tolerance and the distinction between the material and the formal mode of speech has then to be understood as a central part of Carnap's general dream of a canonical meta-language for science and the related logic of science. This cluster of ideas exemplifies the metaphilosophical stance at work in Carnap's transition from his earlier method of rational reconstruction to his later ideal of explication. The significance of this transition is clearer when we look at how, in the transition phase, Carnap's metaphilosophical ideas of constructivism, positivism, logicism, structuralism and pluralism are exemplified.

Carnap's constructivism can be seen in the transition phase by his quest for a canonical meta-language of science, exemplified by the actual construction in the *Syntax* of the two languages Language I and Language II. Carnap's constructivism is also expressed by his many actual examples of translations from the material to the formal mode of speech, of which part V of the *Syntax* is full (Carnap, 1934, pp. 277-333). Carnap's positivism is the central ideal behind the *Syntax* explicit aim: the construction of a canonical meta-language for science in which the logic of science can be carried out by purely syntactical means. Such a syntactic logic of science would be the scientific successor of philosophical activity. In the logic of science, there is no place for traditional metaphysics disputes, the pseudo-problems of which are understood as misunderstanding of the logical structure of language, whose real syntactic form can be revealed by translating them (when possible) in the formal mode of speech. Carnap's logicism transpires from the exceptional epistemological status of logical analysis, still considered to be the main tool by which the rational content of scientific theories and phenomena can be revealed. Even if Carnap in the *Syntax* abandoned his

early logicist positions in philosophy of mathematics, due to the new tolerance principle on foundational matters, logical analysis (especially the one conducted by syntactic means) is still the central tool of his methodology. Carnap's structuralism is, just like in the earlier rational reconstruction period, exemplified by Carnap's focus on the logical structure of mathematical theories, as it is clear from Carnap's construction of Language I and Language II in the *Syntax*. In comparison to the *Aufbau*, Carnap develops new logical methods for specifying the structural content of mathematical theories (Schiemer, 2020b). Finally, Carnap's pluralism in the *Syntax* phase is of course crystallized in the principle of tolerance and its radical statement of freedom in the construction of logical languages. The neutralism with respect to metaphysical positions of the *Aufbau* period is strongly widened towards a neutralism even with respect to different meta-scientific disputes such as the foundation one in philosophy of mathematics. However, if Carnap achieves such a radical pluralism in philosophical and scientific matters in the *Syntax*, the resolution of pseudo-disputes must still be a purely algorithmic matter, bound to be resolved internally to the syntax of science. There is not, at this point in Carnap's metaphilosophical ideas evolution, already space for the pragmatics and semantics of science typical of his later views (Richardson, 2012; Uebel, 2012, 2018).

We have then seen how Carnap's philosophical methodology, in the transition from rational reconstruction to explication, stays faithful to these five metaphilosophical ideals, while gradually changing the specific way in which these ideals are exemplified in Carnap's philosophical projects. As our look at the metaphilosophical ideas contained in the *Logical Syntax* showed, Carnap widened the scope of his pluralism, championing an almost complete freedom in how the meta-language of science is constructed. Moreover, Carnap's "formal mode of speech" program of translating (some) pseudo-problems of metaphysics into (meta)syntactical statements within the logic of science shows, in comparison to Carnap's earlier methodology of rational reconstruction, a less radical opposition to traditional philosophical discourse and a more open-ended idea of conceptual change. At the same time, Carnap's insistence on a purely syntactical logic of science, inside which all scientific and philosophical questions must be translated and resolved, is indeed in tension with the radical freedom that the principle of tolerance prescribes. A truly tolerant metaphilosophy needs a truly tolerant methodology. Carnap's development of such a methodology, i.e. a procedure able to correspond to the freedom expressed by the principle of tolerance, will be the focus of the next subsection.

3.1.4 The ideal of explication

The third and last step in my analysis of Carnap's metaphilosophical development will focus on Carnapian explication, i.e. the philosophical methodology with which Carnap explicitly identifies his later work. Biographically, the explication period corresponds to Carnap's life in the USA. Intellectually, the explication methodology corresponds to Carnap's increasing focus on inductive logic (Sznajder, 2018), it is introduced for the first time in *Meaning and Necessity* (Carnap, 1947), and it is paradigmatically exemplified by the work contained in the *Logical Foundations of Probability* (Carnap, 1950b).

In the *Logical Foundations of Probability*, the first chapter is explicitly devoted to presenting the procedure of explication, of which the work on the concept of probability contained in the rest of the book is an example. I will analyze the procedure of explication in full detail, from an abstract epistemological point of view, in the next section. Here I will just give a brief account of its significance for Carnap's overall metaphilosophical development. Carnap presents explication with the following words:

> "By the procedure of *explication* we mean the transformation of an inexact, prescientific concept, the *explicandum*, into a new exact concept, the *explicatum*. Although the explicandum cannot be given in exact terms, it should be made as clear as possible by informal explanations and examples." (Carnap, 1950b, p. 3. Original emphases)

Explication involves then the transformation of a inexact concept into a more exact one. More accurately, as Carnap makes clear in his reply to Strawson's critique of explication (Strawson, 1963; Carnap, 1963b), the exactness of a concept has to be understood relative to a certain task or goal. Explication then replaces a certain concept, inadequate for a certain task, with another, more adequate concept. This dependency on a given task or goal is the first crucial difference between explication and its predecessors, i.e. rational reconstruction and the translation from the material to the formal mode of speech. Carnap's earlier methodologies had in fact a more absolute character, replacing a certain concept (or statement or question or theory) with its scientific substitute in the language(s) of science. The replacement performed by rational reconstruction and the translation into the formal mode of speech were also both understood as one-way processes, in which the replaced concept had no role in philosophy or science once the replacement had taken place. The procedure of explication, instead, has a dialectical, open-ended character (Stein, 1992; Carus, 2007, 2012b; Uebel, 2012). The concept that gets explicated, i.e. the explicandum, is not replaced away from its successor, i.e. the explicatum, but it can always be the starting point of other explications.

Moreover, the explicandum plays also a crucial role in the assessment of an explication overall success. Explication is in fact an inherently pragmatic procedure, i.e. its adequacy is not a matter of right or wrong, but of what is more or less satisfactory for the task that the explicator has in mind. Judging this adequacy is then never an all or nothing matter. The explicator has always a certain degree of freedom in choosing the explicatum for substituting a given concept. In Carnap's (Carnap, 1950a) late terminology, as Stein stressed, questions about explication adequacy are thus external questions:

> "The explicatum, as an exactly characterized concept, belongs to some formalized discourse – some 'framework'. The explicandum (...) belongs ipso facto to a mode of discourse outside that framework. Therefore any question about the relation of the explicatum to the explicandum is an 'external' question; this holds, in particular, of the question whether an explication is adequate." (Stein, 1992, p. 280).

The adequacy of an explication is thus an external question, bound to be pragmatically judged outside the linguistic framework of the explicatum. This is the crucial difference between explication and Carnap's earlier methodology of translating pseudo-syntactical sentences into the formal mode of speech. The adequacy of an explication is not an algorithmic matter to be decided within a language of science by purely syntactical means. The adequacy of a given explication has to be discussed outside scientific frameworks, relative to a given goal and context, with the normative tools of instrumental rationality (Carus, 2007, 2017). In assessing this adequacy, then, the original explicandum works as a central measure of the satisfactoriness of the explicatum. Just like in engineering sciences, the satisfactoriness of a certain tool can be judged only with respect to its goals, its predecessors and its alternatives. This centrality of the explicandum in the assessment of the overall success of an explication allows what Carus (Carus, 2007, 2012b) calls the "feedback-relation"between evolved and constructed languages in Carnap's mature (meta)philosophy. Formally constructed languages, in fact, offer replacements (i.e. explicata) for particular parts (i.e. explicanda) of evolved ones, which are judged externally to the constructed frameworks by the pragmatic mode of discourse typical of evolved languages.

The procedure of explication can then be seen as a bridge between different (types of) linguistic and conceptual frameworks. Explication bridges different frameworks in an inherently pluralist and goal-dependent way, connecting parts of different languages that can perform a similar function with respect to a specific problem at hand. As such, the methodology of explication connects and justifies all Carnap's incessant construction of linguistic and conceptual frameworks as the development of possible explications for our philosophical concepts. We can now see how explication fully embodies Carnap's metaphilosophical ideals of constructivism, positivism, logicism, structuralism, and pluralism.

The ideal of explication clearly embodies Carnap's constructivism in its engineering-like view of philosophical activity and progress. The main activity of philosophers, according to the explication ideal, ought to be the construction of multiple frameworks from which explicata of our concepts can be obtained. Philosophical progress is then seen as the repeated explication of our philosophical and scientific concepts, an advancement akin to technological progress in which more reliable and more specialized tools are constantly produced. The figure of the philosopher becomes then in the later Carnap an engineer figure involved in a kind of conceptual metrology (Richardson, 2013). This engineering-like view of philosophical progress exemplifies, of course, also Carnap's positivism, according to which a better, more scientifically-minded philosophy needs the development of better conceptual tools. The positivist ideal permeates also Carnap's stress of a paradigmatic kind of explication, which is the step from qualitative to quantitative concepts (Carnap, 1950b, pp. 8-15). The replacement of qualitative concepts with quantitative ones achieved by (almost) all natural sciences is in fact for Carnap the paradigmatic example of the conceptual progress around which explication is centered. Carnap's logicism can be similarly seen in Carnap's late efforts in explication, such as the work on intensional semantics (Carnap, 1947) or the work on inductive logic (Carnap, 1950b, 1952; Carnap and Jeffrey, 1971), where the explicata are all framed within formal logical languages. Even though Carnap (Carnap, 1963b, p. 935) explicitly stresses the possibility of purely informal explications, the paradigmatic

example of conceptual progress remains for him the formalization of concepts and theories in logical languages. Carnap's pluralism is fully embraced in the explicit reject of an absolute right or wrong judgment in matters of explication adequacy. Carnapian explication embraces then the full radical pluralism, proclaimed already in the *Logical Syntax* by Carnap's principle of tolerance, of not setting (almost) any prohibition to philosophical and scientific freedom. As long as the explication proposal is clearly stated, every possible proposal for transforming a given concept should be welcomed and discussed. Consistently with this wider scope of Carnap's later pluralism, the tools for discussing explication matters are not bound to logical or syntactic strict rationality, but they belong instead to the goal-dependent, pragmatic kind of rationality known as instrumental rationality.

Carnap's structuralism, at a first glance, is not so evident in his explication methodology like the other four metaphilosophical ideals. In contrast to rational reconstruction, in fact, a given concept is not necessarily replaced by another concept exhibiting its logical structure. An explicatum is always dependent on the particular goal and context of a given explication. Nevertheless, there is an inherently structuralist component in the ideal of explication. It is, however, a kind of methodological structuralism, akin to the one of the Erlangen program (Schiemer, 2020a). The bridge-function that explication performs in connecting different linguistic frameworks is in fact akin to the transfer-principle methodology common in nineteenth-century projective geometry. As exemplified by the Erlangen program, projective geometers increasingly focused on the geometrical (and later topological) invariants under abstract transformations. This quest for increasingly abstract geometrical invariants was carried out with the method of transfer-principles (Schiemer, 2020a), i.e. analytically defined mappings between different geometrical domains that preserve the relevant projective properties of chosen configurations. A transfer-principle, then, is able to transfer certain geometrical or topological properties from one geometrical domain to the other, allowing geometers to preserve certain structures and related functions in a changing domain. It has been argued (Schiemer, 2020a) that the methodology of transfer-principles embodies two key structuralist ideas, i.e. the indifference to the nature of primitive spatial elements and the emphasis on structures-preserving mapping across domains. Carnap's methodology of explication can then be thought as a metaphilosophical analogous of the transfer principle method. If in the context of the Erlangen program, transfer principles were used for preserving structural properties across different geometrical domains, in Carnap's metaphilosophy explication is used for preserving functional and conceptual properties across linguistic frameworks. Explication is then a kind of transfer-principle for conceptual utility or (if one is a functionalist or an inferentialist about concepts) conceptual structure, a method for preserving a certain conceptual role or function in a different linguistic context. The set of all possible explicata of a given philosophical explicandum can then be considered the set of its possible counterparts across philosophical domains. The ideal of explication embodies Carnap's structuralism as a methodological structuralism, exemplified by its indifference with respect to the nature of the particular linguistic framework in which a given concepts is framed and in its emphasis on function-preserving mappings that transfer conceptual virtues across different frameworks.

We have then seen how the ideal of explication fully embodies Carnap's metaphilosoph-

ical ideals of constructivism, positivism, logicism, structuralism, and pluralism. Moreover, we saw how the procedure of explication is the result of the gradual evolution of Carnap's philosophical methodology that, while staying faithful to his metaphilosophical stance, changed through the years consistent with the philosophical and scientific issues faced by Carnap. The result of this evolution is the ideal of explication, as an example of which the whole philosophical activity of Carnap can be retrospectively assessed. The ideal of explication prescribes in fact to the wannabe philosopher a constant engineering quest to develop better tools for making philosophical discussions advance to a more exact and more scientifically-minded phase. What Carnap did throughout all of his philosophical activity can be perfectly identified in these terms.

3.2 The Procedure of Carnapian Explication

After having traced the development of the notion of explication in the more general progression of Carnap's philosophical views, I will analyze Carnap's method of explication from an abstract epistemological point of view.

We saw in the previous section that explication is a procedure involving two concepts. On one side, there is the explicandum, belonging to natural language (or more generally an evolved language), the scope of which thus contains arguably an amount of vagueness and ambiguity. On the other side, there is the explicatum, belonging to a (more) precise language, the scope of which is rigidly characterized by explicit and precise rules of use.

Explication is traditionally seen as a two-steps procedure. First of all, one has to clarify the explicandum, trying to explicitly state the intended meaning of the concept that one wants to explicate. Since the explicandum is still expressed in a natural language, an exact definition is not required. What Carnap (Carnap, 1950b, pp. 3-5) requires from the explicator, instead, is to state some positive and negative instances of the explicandum, together with some description or (partial) rules of use.

This step clarifies and (if necessary) disambiguates the concept that one seeks to explicate. It is, in fact, possible that in trying to clarify the explicandum, the explicator realizes that there are two or more different concepts that are ambiguously grouped in natural language within a single notion. A famous example of this phenomenon occurred in Carnap's (Carnap, 1950b) explication of probability. In clarifying this concept, Carnap in fact realized that behind the intuitive understanding of probability lie two different notions, the logical and the frequentist concepts of probability. In clarifying the explicandum, the explicator also freely chooses the context of the explicandum that she wants to explicate. It is, in fact, often the case that a given explication wants to replace only some contexts or uses of the intuitive notion. A classical example of this decision of context are Tarski's opening remarks (Tarski, 1933) before his explication of the concept of truth, where he states that he is interested in explicating the context of truth-assertions like "'snow is white' is true" and not in explicating uses such as "you are a true friend".

Then, there is the second step of the explication, i.e. the formulation of the explicatum in a certain target theory via an explicit definition or by stating its rules of use (Fig. 3.1).

Figure 3.1: The two-step structure of Carnapian explication.

The purpose of explication is the substitution, relative to a specific function-context, of a less satisfactory concept with a (more) satisfactory one. As we saw in the last section, the adequacy of a given explication can never be absolutely correct or wrong, but it is instead a matter of relative satisfactoriness with respect to the explicator's goals. Even though explication is not a matter of right or wrong, one can still judge whether an explication is a good one or a bad one. In fact, external questions for Carnap can still be objects of rational discourse, although of the pragmatic kind of rationality that is often called instrumental. Relative to a specific purpose or function, one can state certain pragmatic meta-principles that a concept has to respect in order to qualify as a good explicatum for a certain explicandum. Carnap (Carnap, 1950b, pp. 5-8) stated four desiderata that a good explicatum has to respect:

- Similarity: to the extent to which the other desiderata allow it, the explicatum ought to be as much as similar to the explicandum (exact similarity, i.e. identity, is explicitly not required).
- Fruitfulness: the explicatum ought to be connected with other scientific concepts, in order to make as many generalizations as possible expressible within the theory in which it is framed.
- Exactness: rules of use of the explicatum ought to be stated in an exact form (e.g. definitions, axioms).
- Simplicity: the explicatum ought to be as simple as the other desiderata allow it to be.

These principles give a hint of the virtues that a good explicatum has to possess, but they are intrinsically pluralist in their intent. Carnap, in fact, stresses how it is always possible to have different explicata that are equally adequate in respect to a given explicandum. In matters of explication, there is no absolutely correct answer to the problem of capturing an informal notion with a (more) formal one.

A problem central to much of recent debates about explication as a philosophical procedure is the extreme inexactness of these desiderata stated by Carnap. The four principles above hint, in fact, at the theoretical virtues that a good explicatum must have, but they are too vague and ambiguous to constitute a practical guide for explicating a certain concept. Carnap never attempted to further develop these criteria. He instead developed various practical examples of what he considered good explicata for fundamental philosophical concepts, e.g. his works on logical probability as an explication of confirmation

(Carnap, 1950b) or his efforts towards explicating our concepts of modality and analyticity (Carnap, 1947). Using these and other examples, in science and philosophy, of formal notions that have replaced informal ones, various scholars have proposed refined and more precise versions of these. Let us survey this debate, then.

3.2.1 Discussing explication desiderata

To better structure the discussion, let us treat Carnap's four desiderata, one by one.

Similarity. This is perhaps the desideratum that has most attracted the attention of scholars, due to its pivotal role in distinguishing explication from other methods of conceptual change. Since, as we have seen, the explicandum is normally a vague, informal concept and the explicatum is instead a (more) precise, formal notion, exact similarity is not required[3]. If, then, an explicatum is allowed to have a different extension than the explicandum, the question at issue is to which degree an explicatum has to be similar to the related explicandum.

Hanna argued that the explicatum has to agree with the explicandum in all clear-cut cases where the latter can be applied (Hanna, 1967, pp. 34-36). This strict reading of the similarity requirement makes explication, as Hanna himself acknowledges, just a procedure for eliminating any vagueness from our informal concepts and it thus makes the explicatum a precisification of the explicandum. This can be clearly seen in Hanna's formal explication of 'explication' where the (formal notion that seeks to explicate the) explicatum is technically a precisification of the (formal notion that seeks to explicate the) explicandum (Hanna, 1967, pp. 37-38).

Another strict reading of the similarity requirement is Quine's "synonymy in favored contexts" i.e. synonymy with respect to all the contexts where the use of the explicandum is clear and precise (Quine, 1961, p. 25). Both these readings seem in direct contrast with Carnap's own examples, e.g. the explication of the concept *fish* (Carnap, 1950b, p. 6), where he allows explicata to be concepts that explicitly reject clear, non-defective, positive instances of the explicandum. They also appear too narrow for any general procedure of conceptual engineering for science and philosophy. Often, in fact, as even Quine himself acknowledged later (Quine, 1960, pp. 258-260), scientists change meanings and uses of pre-theoretical concepts for purely theoretical reasons, despite how clear a certain use of an explicandum originally is[4].

Brun has recently argued for a more liberal reading of the similarity requirement, which he understands as requiring the explicatum to preserve all the context-dependent instances of the explicandum (Brun, 2016, pp. 1218-1219). The context is freely decided by the explicator in relation to the purpose for which the explicatum is expected to substitute the

[3]Famously, this lack of exact similarity is the core of Strawson's famous "subject-change" critique of explication as a philosophical method in (Strawson, 1963).

[4]It should be noted that Quine developed also his own version of the explication procedure, quite different from Carnap's one, understanding it as a form of transformative conceptual analysis (Gustafsson, 2013).

explicandum. As we already saw, Carnap himself stressed that in clarifying the explicandum, the explicator must decide the intended context of the explication, just like Tarski did in his aforementioned explication of truth. Thus, according to this interpretation, an explicatum is allowed to diverge from the scope of the explicandum even in clear-cut cases of application of the latter if they are not within the specific context freely chosen by the explicator.

Brun also proposed, in a more recent work, to split the similarity requirement into two steps. The first step consists in adjusting the extension of the explicandum (implicitly fixing the context), thereby obtaining a sharpened '$explicandum_2$'. The second step requires what Goodman calls 'extensional isomorphism', i.e. an injection from the extension of $explicandum_2$ to the extension of the explicatum (Brun, 2020, pp. 11-13). This two-step reading is connected with Brun's more general proposal of merging explication with (a particular interpretation of) Goodman's method of reflective equilibrium. Brun acknowledges that the injection-requirement seem trivial for a single concept, but stresses its significance for explicating a system of concepts and overcoming some limitations of (what he takes to be) the linear-monoconceptual Carnapian picture of explication (Brun, 2020).

Fruitfulness. Carnap vaguely described this desideratum in terms of relations to other concepts and generalization-power. He distinguished the generalizations that a fruitful explicatum ought to produce between two cases, i.e. "empirical laws in the case of a non-logical concept, logical theorems in the case of a logical concept" (Carnap, 1950b, p. 7). This seems indeed a necessary condition for a good explication of certain kinds of concepts, but as a general rule, it seems not really useful (by itself). After all, every formal concept whatsoever can produce an infinity of generalizations and truths[5].

Dutilh Novaes and Reck proposed to read fruitfulness as the improvement of the pragmatic and epistemic situation of an agent. They claimed that a fruitful explicatum has to make our reasoning more effective and more reliable, thereby proving itself to be a better cognitive tool (for a certain purpose) than the explicandum (Dutilh Novaes and Reck, 2017, pp. 205-211).

Shepherd and Justus took fruitfulness to be, like similarity, a context-dependent desideratum, relative to the type of concept the explicandum is and the purpose that the explicatum has to perform (Shepherd and Justus, 2015, pp. 395-400).

Exactness. Here the main question is whether exactness means (a certain level of) formal rigor. Formal frameworks were considered by Carnap, even after his tolerance turn, the benchmark of exactness and rigor. If one looks at his own efforts in explicating concepts like analyticity or confirmation, one finds always the explicatum defined in a formal framework. Should we therefore understand the exactness requirement simply as the request of formulating the explicatum within a formal framework? Hanna believes that this is not enough. He, in fact, claimed that a certain explicatum has always to have a perfectly

[5]Dennett humorously stressed this point in his general critique of (some) contemporary analytic philosophy (Dennett, 2006).

clear extension, thereby (together with his aforementioned extensional reading of similarity) making the explicatum a complete precisification of the explicandum (Hanna, 1967, p. 36).

These strictly-formal readings of the exactness requirement seem too narrow for a general procedure of conceptual engineering, especially considering the fact that Carnap explicitly warned us against a strictly formal reading of the exactness requirement:

> "The use of symbolic logic and of a constructed language system with explicit syntactical and semantical rules is the most elaborate and most efficient method. For philosophical explications the use of this method is advisable only in special cases, but not generally" (Carnap, 1963b, p. 935).

Therefore, the exactness requirement has been understood in a comparative way, relative to the explicandum (Dutilh Novaes and Reck, 2017, p. 201), allowing a certain open-endedness in the offspring of the procedure of explicating a certain concept. This comparative reading of exactness can be easily cashed-out in terms of vagueness, by requiring the explicatum to be less vague than the explicandum.

Brun favored a weaker reading of comparative exactness, i.e. the explicatum should not be vaguer than the explicandum. Using as evidence Carnap's aforementioned fish example, he argued that in some cases the explicatum is as vague as the explicandum (Brun, 2016, pp. 1220-1221). He also stresses that Carnap hinted at an additional aspect of the exactness requirement in discussing the temperature example, namely, that usually quantitative concepts are preferable over qualitative and comparative ones because they allow us to make more fine-grained distinctions (Brun, 2020, p. 1220).

Simplicity. Last and explicitly least, this desideratum is often left aside in the discussion about explication. Carnap stresses that simplicity is only a last resort when the explicator has to choose between different explicata that fulfill to the same degree the other, more important, desiderata (Carnap, 1950b, p. 7). Therefore, there is not much debate over what the simplicity requirement means.

To my knowledge, only Brun's treatment of explication includes a discussion of simplicity. He stresses that simplicity has to be understood not in the ontological occamian sense (i.e. ontological parsimony) but as a syntactical-logical requirement on the definition of the explicatum and perhaps also on the overall structure of the target theory in which the explicatum is defined (Brun, 2016, p. 1221).

Of course, one may add to these four desiderata other theoretical virtues that a good explicatum should possess. Perhaps, other possible desiderata could be general pragmatic and theoretic virtues that good scientific theories embody, such as explanatory power, predictive power, novelty, unification power, trans-theoretic coherence, and so on. Philosophers of science have discussed at length the role of these epistemic values in scientific activity and whether they can be sharply separated from non-epistemic values (e.g. Kuhn 1977; Longino 1990; Solomon 2001; Haack 2003; Douglas 2009; Okasha 2011).

Another, interesting, possible desideratum is due to Karl Menger, to whom Carnap (Carnap, 1950b, p. 7) acknowledged a certain debt in developing the idea of explication, who in discussing geometrical definitions stresses that a good explicatum "should extend the use of the word by dealing with objects not known or not dealt with in ordinary language" (Menger, 1943, p. 5).

3.2.2 A note on recent critiques of explication

The discussion about the desiderata that a good explicatum has to respect has also sparked a more general methodological discussion about the viability of explication as a procedure of conceptual engineering. Apart from the aforementioned well-known critique of Strawson, new critiques have emerged[6].

For instance, Dutilh Novaes and Reck have recently argued that explication (and, more generally formalization) is an inherently paradoxical enterprise. Explication is paradoxical, according to them, because there is a tension between two of its most important desiderata, namely fruitfulness and similarity: similarity allegedly calls for a close relationship with the explicandum, while fruitfulness pushes the explicatum towards a more radical departure from the explicandum. They named this phenomenon "the paradox of adequate formalization" (Dutilh Novaes and Reck, 2017, p. 211), claiming that it is nothing but another form of the well-known "paradox of analysis" (Beaney, 2021).

Reck in another work stresses, in a Strawsonian fashion, that Carnapian explication has some blind spots, such as its strong focus on formal aspects, its strive for exactness, its unwillingness to take into account methodologically different alternatives (Reck, 2012, pp. 106-114). Reck acknowledges that these blind-spots are far from being impossible to be mitigated by a more pragmatic and liberal theory of explication, but he argues that such a theory would lead us back to philosophical disputes of the very kind that explication was meant to overcome.

Other two limitations of explication as a general procedure for conceptual engineering are stressed, instead, by Brun, who argues that Carnapian explication is heavily limited by its focus on individual concepts (i.e. it does not take into account more complex entities such as systems of concepts) and by its linear structure that seems to describe a no-turning-back triumphant engineering from the explicandum to the explicatum, hiding thus the complexity of the dialectics between the two parts of explication[7]. Brun (Brun, 2020) argues that these two points can be mitigated via a more liberal recipe-approach to conceptual engineering, merging explication with Goodman's reflective equilibrium.

[6]For Strawson's original take, see (Strawson, 1963). Carnap responded in (Carnap, 1963b). For more recent responses see (Maher, 2007; Justus, 2012).

[7]See his discussion in (Brun, 2016, pp. 1229-1232). A possible line of response, already stressed by Brun, can be found in Carus' remarks about the dialectic between evolved and constructed languages in explication. See (Carus, 2007, pp 273-284).

3.3 Explication as a Three-Step Procedure: Computability as a Case Study

We have then seen how, in recent years, there has been a renewal of interest in Carnapian explication as a philosophical method. In recent debates over explication, most of the focus has been put onto the explicandum and the explicatum, i.e. the starting and the ending point of the procedure of explicating a given concept. In contrast, only few scholars have tried to analyze what happens during the transformation of a given explicandum into an explicatum (Shepherd and Justus, 2015; Quinon, 2019).

In this section I will focus on the breakdown of explication as the step-by-step transformation of an intuitive concept into a more precise one. In order to do that, I will apply the method of explication to (some specific versions of) the Church-Turing Thesis (CTT), i.e. the statement according to which our intuitive notion of effective calculability is captured by the formal notion of mechanical computability[8]. More specifically, I will apply the method of explication to two recent axiomatic approaches to CTT: Sieg's axioms for computors (Sieg, 2013) and Dershowitz's & Gurevich's axioms for computations (Dershowitz and Gurevich, 2008).

In recent years, building upon Robin Gandy's seminal work(s) in the 1980s, Wilfried Sieg has proposed axioms for human and mechanical computers. In 2008, Dershowitz & Gurevich, coming from a completely different background, proposed another axiomatic characterization of computability. Following Gurevich, I will refer to these axiomatizations as "Foundational Analyses of Computability" (Gurevich, 2012). Carnapian explication has been previously applied to two classical explications of effective calculability, namely Turing computability (Floyd, 2012) and general recursiveness (Hanna, 1967; Quinon, 2019), but not to more recent models of computability[9].

We will see that, from the perspective of Carnapian explication, these two foundational analyses of computability differ in how they clarify and restrict the boundaries of the intuitive notion of computability. These two different ways of analyzing the notion of effective calculability can be traced back to two pioneers of computability, Turing (Turing, 1936) and Kolmogorov (Kolmogorov, 1953). I will argue that the main conceptual difference between them lies in the different outputs of their respective informal analyses of effective calculability, i.e. two informal axiomatizations of what is effectively calculable. These axiomatizations implicitly define two semi-formal notions of computability: *computorability* and *algorithmability*. I will then argue that, in order to adequately capture the conceptual differences between these two semi-formal notions, the classical two-step picture of explication is not enough. In order to overcome this problem, I will present a more fine-grained three-step version of Carnapian explication. I will call the new mid-level step of this refined

[8] I use 'mechanical computability' as a neutral term of art to denote all the extensionally equivalent notions for which one can state a version of CTT, e.g. general recursiveness, Turing computability, λ-definability, etc. 'Effective calculability' denotes instead the intuitive notion of calculability.

[9] Considering the robustness and the success of the concept of mechanical computability in capturing an intuitive notion, one may wonder why Carnap does not mention CTT as a paradigmatic example of explication. For an analysis of this issue and a possible answer to it, see (Quinon, 2019).

version of explication the *semi-formal sharpening of the clarified explicandum*.

I will show how my three-step version of Carnapian explication is able to explain the differences between the Turing-Gandy-Sieg and the Kolmogorov-Dershowitz-Gurevich groups of explications, allowing a better conceptual understanding of foundational analyses of computability. More generally, this case study will demonstrate that the three-step version of Carnapian explication is a better tool than the original two-step version for treating complex cases of explications with several explicata for a given explicandum. By adding the new mid-level step of the semi-formal sharpening of a clarified notion between the clarification of the intuitive concept and the formulation of a new one, this refined version of Carnapian explication allows more fine-grained distinctions between different explications of a given concept. Even two explications that clarify the same intuitive concept in the same way (i.e. they output the same clarified explicandum after the clarification step), such as Turing's and Kolmogorov's "analyses" of effective calculability, can be semi-formally sharpened in different ways (i.e. they output two different concepts after the semi-formal sharpening step). The new mid-level step also provides further evidence for the importance of the intermediate steps in the evolution of an explication (cf. Quinon 2019). These intermediate steps have often been overlooked in philosophical discussions about explication, which have often focused only on the explicandum and the explicatum, but the case of CTT shows that without analyzing these steps it is sometimes impossible to properly understand the conceptual differences between two explicata.

3.3.1 The Turing-Gandy-Sieg explications of effective calculability

In this subsection, I will focus on the first group of explications of effective calculability that I will analyze through the lens of Carnapian explication in this case study, namely, the Turing-Gandy-Sieg explications.

In introducing Sieg's foundational analysis of computability, I said that he presented axioms for calculability. More exactly, he gave axioms for calculators, both human and mechanical ones. Why did he state his explication in terms of calculators? In Sieg's own words, "to investigate calculations is to analyze symbolic processes carried out by calculators; that is a lesson we owe to Turing" (Sieg, 2002b, p. 390). Then, in order to properly understand Sieg's work one has to grasp the significance of concepts such as human calculator, mechanical device, and symbolic process.

Turing's analysis and Gandy's principles for mechanisms

In the 30's, the idea(l) of CTT arouse from the conceptual offspring of the so-called foundational crisis. The main reason why a (group of) formal equivalent(s) of effective calculability was sought and developed is usually considered to be the growing skepticism for a positive solution to the *Entscheidungsproblem* (i.e. the problem of finding an effective procedure for deciding the validity of first-order logical formulas) and the consequent wait for a proof of its undecidability. Moreover, Gödel's incompleteness results called for a generalization

via a notion capable of explicating what a suitable formal system was. Thirdly, a formal concept of decision procedure was needed in order to negatively solve some open mathematical problems of the time, such as the Diophantine equations problem and Thue's word problem for semi-groups.

It was precisely this nest of related problems that caused the praised "confluence of ideas" (Gandy, 1988) of 1936. Already in 1934, two soon-to-be extensionally equivalent instances of mechanical computability had been already developed: general recursiveness and λ-definability[10]. In 1935, Church boldly stated what is now known as Church's Thesis, i.e. the proposition that our intuitive concept of effective calculability is adequately captured by the formal notion of general recursiveness.

However, before Turing, with the important exception of Post's unpublished drafts (Post, 1941), one cannot find any (significant) occurrence of the term 'computor' or 'calculator' in any seminal work on computability. Even though the scope of their explications was human effective calculations, pre-Turing explications of calculability, such as Church's (Church, 1936), Gödel's (Gödel, 1934), and Kleene's (Kleene, 1936) do not take in account any sort of calculator[11]. Nowadays it seems natural to us to think about calculability in terms of an abstract computor, but this only shows the impact of Turing's work on our perception of computation. Before Turing, the problem of adequately defining mechanical computability was centered around the possible ways in which numeric functions can be deduced in a suitable formalism. The attention and the work of the scholars focused almost entirely on the formal side of CTT, leaving the intuitive concept of effective calculability as something of which nothing meaningful can be said.

An important exception is Church's (Church, 1936, p. 101) so-called 'step-by-step' argument. It has the form of a division into cases argument. According to Church, an effectively calculable function is evaluated either by the application of an algorithm or by the method of 'calculability within a logic', namely, for a certain function F, by deriving in a certain "logic" a theorem $f(\mu) = \nu$ that holds if and only if $F(m) = n$, where f is an expression of the language of that "logic" and μ and ν are expressions that respectively represent arbitrary positive integers m and n. Church, then, shows how in both ways is involved a series of steps, which he assumed to be recursive. Being all the steps involved in the computation recursive, the function itself is recursive.

From a epistemological point of view, however, this argument has a major problem. As it was stressed by Sieg, Church gives no justification for imposing the recursiveness condition on the computational steps, a limitation that is clearly pivotal to the cogency of the argument (cf. Sieg 1997). Moreover, as an analysis of our intuitive concept of calculability, Church's reflections lack any appeal to common intuition, paying attention only to the calculation process of numerical functions in a formal systems of some sort.

It could be said, then, that the problem of pre-Turing analyses of computation was to motivate the recursiveness of the computational steps of numeric functions. How does

[10]For an historical account of the development of these two notions see (Kleene, 1981; Gandy, 1988).

[11]Following Gandy, I take the term 'computor' to denote an idealized human calculator. I use the terms 'computer' and 'calculator' to refer to any kind of computing agent, being it a human or a mechanical device of some sort.

Turing solve this problem? The short answer is that he changed the level of analysis. Considering computable numbers, Turing understood that in order to analyze the intuitive notion of effective calculability, one has to focus not on the superficial calculus of the numerical function, but instead on the symbolic processes underneath:

> "The real question at issue is 'What are the possible processes which can be carried out in computing a number?'" (Turing, 1936, p. 135).

Changing the question to address allowed Turing to put the computor at the center of his analysis of effective calculability. Roughly speaking, Turing's take on calculability is that a function is computable if and only if it can be computed by an idealized human being working in a clerical fashion. Restrictions on the calculation process can then be stated as bounds on the possible actions of the abstract computor, motivated by human physical-cognitive limitations[12].

Turing imagined a computor working on a one-dimensional paper, something like 'a child's arithmetic book', printing only a finite number of symbols. Infinity is forbidden because "if we were to allow an infinity of symbols, then there would be symbols differing to an arbitrary small extent" (Turing, 1936, p. 135). Ruling out any form of ingenuity, the behaviour of the abstract computor is fully determined by the symbols at which he is looking at the moment, together with his state of mind. Actions of the Turing computor have, then, to meet some natural, physical and cognitional, limitations: only a fixed number of symbols can be observed at one glance by the computor, only a fixed number of states of mind can be involved in the calculation, all the actions of the computor can be divided into elementary operations, and so on.

In Sieg's reconstruction of Turing's analysis, these limitations are summarized by three general bounds[13]:

- "(B) (Boundedness) There is a fixed bound on the number of configurations a computor can immediately recognize.
- (L) (Locality) A computor can change only immediately recognizable (sub-) configurations.
- (D) (Determinacy) The immediately recognizable (sub-)configuration determines uniquely the next computational step (and id)." (Sieg, 2002b, pp. 249-250).

It is easy to see that these informal axioms suitably represent the several bounds informally imposed by Turing on symbolic processes in his analysis. Then, Turing argues that every function computable by a computor, who respects these restrictive conditions,

[12]It is important to stress that Turing, even though his bounds on calculations are motivated by human limitations, did not have what is sometimes called a cognitivist approach to computation. In other words, *pace* Gödel (Gödel, 1972) , Turing never claimed that his analysis captured the richness of human procedures nor did he want to state something about human cognitive abilities. See the analysis contained in (Copeland and Shagrir, 2013).

[13]Sieg, following Post, imposes these bounds on instantaneous descriptions (ids). For my analysis, this makes no difference.

is Turing computable. This claim is splittable in two sub-parts. First, we have what Sieg calls *Turing's Central Thesis*, i.e. "computations carried out by a computor satisfying the boundedness and locality conditions can be directly simulated by a string machine" (Sieg, 2002a, p. 397). The determinacy condition is achieved by the sum of boundedness and locality and it is therefore not necessary. Then, we have the mathematical proof that computation by a string machine can be simulated by a letter machine.

Thus, Turing's analysis can be seen as an informal two-steps argument. First, he imposes several bounds on the vague notion of effective calculability sharpening it into the notion of computability by an idealized human calculator. Then, he argues that computations by such a computor can be faithfully represented by what we now call Turing machines. Only after this last step, we have reconstructed the whole of Turing's analysis. Such a detailed analysis, as Sieg himself recognizes, makes Turing's assumptions, upon which the analysis is built, fully evident:

> "The separation of informal conceptual analysis and mathematical equivalence proof is essential for recognizing that the correctness of Turing's Thesis (taken generically) rests on two pillars; namely, on the correctness of boundedness and locality conditions for computors, and on the correctness of the pertinent central thesis. The latter asserts explicitly that computations of a computor can be mimicked directly by a particular kind of machine." (Sieg, 2002a, p. 399).

One of Sieg's explicit aims is then to sharpen Turing's informal argument via the axiomatic method. This is why he stated his foundational analysis in terms of calculations by abstract calculators. The other explicit aim of Sieg's axioms is to axiomatically characterize another, more general class of calculators, i.e. mechanical devices. Where in the case of computors his work builds upon Turing's analysis, in the case of mechanical devices Sieg attacked the problem by improving and simplifying Gandy's treatment of machine computability. In fact, even though it is sometimes claimed in the literature (especially the popularized one) that Turing proved something about machines, Turing's 1936 analysis pays attention only to human calculators. More generally, it is safe to say that the scope of the entire confluence of 1936 was human effective calculability and that therefore CTT has nothing to do with machines. It was Gandy (Gandy, 1980) who firstly addressed the problem of 'Thesis M', namely the proposition stating that what can be computed by a machine is Turing computable. In order to address Thesis M, Gandy followed the example of Turing's analysis of effective calculability and imposed some limitations upon our intuitive concept of mechanical device.

The term machine is understood by Gandy "with its nineteenth century meaning" (Gandy, 1980, p. 125). More specifically, he thinks that machines are discrete deterministic mechanical devices. With this wording, he excludes analogue machines and other devices the calculations of which cannot be described in discrete terms or that are not entirely deterministic. Examples of machines in Gandy's restricted sense are Turing machines, Von Neumann's crystalline automata and John Conway's 'the game of life'. According to Gandy, the main difference between a computor and a mechanical device is that the

latter is able to act in parallel, performing a finite though unbounded number of bounded computations at the same time. Any machine of this kind must satisfy four principles.

The first principle states how one should be able to describe a machine. Gandy chooses (isomorphic classes of) hereditarily finite sets to suitably describe each state of the machine. From a given machine state, a structural operation gives us the next state of the machine. After having stated the form of description of his machines, Gandy imposes three different bounds on their working. These limitations should be understood as aiming to avoid the possibility of an omniscient device. Principle II expresses the requirement that the hierarchy of structures describing the machine has a maximum height. Principle III instead requires that any device can be uniquely reassembled from parts of bounded size. Gandy himself remarked that almost any kind of machine (in the general sense of the term) can be described in a way for which Principle II and III are satisfied. The most important principle is then the fourth, the Principle of Local Causality, which contains Gandy's core idea of how a machine computes:

> "Principle IV. (Preliminary version [of the Principle of Local Causality]) The next state, Fx, of a machine can be reassembled from its restrictions to overlapping 'regions' s and these restrictions are locally caused. That is, for each region s of Fx there is a 'causal neighborhood' $t \subseteq TC(x)$ of bounded size such that $Fx \uparrow s$ depends only on $x \uparrow t$."(Gandy, 1980, p. 135).

This principle forbids globally instantaneous signal propagation in the machine, allowing it only for locally determined regions from which a given state depends. This principle mirrored, in a more general way, both the locality and the boundedness conditions imposed by Turing on human calculations. Where in the case of human computors those restrictions were motivated by an appeal to human memory and intellectual natural limitations, Gandy motivates his fourth principle appealing to physics, specifically to the ban of instantaneous action at distance contained in General Relativity theory. This impossibility, together with Gandy's second physical assumption, namely that there is a lower bound on the size of distinguishable atomic components of the machine, provides the pivotal finiteness of the causal neighborhood on which the machine at every step of the computation operates[14].

We have now everything that we need in order to properly understand Sieg's axioms for calculators. Let us turn to them, now.

Sieg's axioms for computors and the notion of computorability

Sieg has presented his axioms in various works throughout the last two decades (Sieg, 2002a,b, 2009). Throughout the years, he changed the presentation of his axioms. Technically speaking, he repeatedly simplified his set-theoretic framework by changing some secondary definitions of operations and sets. Philosophically speaking, the latest presentations of his axioms place more emphasis on Turing's symbolic view of calculation than on

[14]For a more detailed discussion of Gandy's constraints and his underlying physical assumptions, see (Sieg and Byrnes, 1999).

the boundedness and the locality conditions, thereby making Turing's ideas appear closer to Post than they are usually considered to be[15]. However, the conceptual core of his work has remained steadily the same for almost two decades and this is what matters for my analysis.

For the computor case, Sieg reformulates Turing's informal limitations, emphasizing the pivotal role of working with finite symbolic configurations, à la Post. According to him, the following aspects of the computors' work are what characterize them:

- (i) "they operate deterministically on finite configurations
- (ii) they recognize in each configuration exactly one pattern (from a bounded number of different kinds of such)
- (iii) they operate locally on the recognized patterns
- (iv) they assemble the next configuration from the original one and the result of the local operation." (Sieg, 2009, p. 587).

These four semi-formal bounds axiomatically define the notion that I call *computorability*, i.e. the mid-level notion of computability obtained via Turing analysis.

The mathematical formulation of Sieg's explicata on suitably defined systems uses the concept of 'discrete dynamical system', i.e. a system described by a class of syntactical configurations D together with an operation $F : D \to D$. States are presented within the same set-theoretic framework of Gandy's (Gandy, 1980) work, as hereditarily finite sets. These sets are obtained from a potentially infinite set U of primitive elements via a finite iteration of the power set operation. This hierarchy of sets has the remarkable properties of being cumulative and the $\in$-relation on its members being well-founded. This allows one to define operations on HF trough recursion on $\in$-relations. Abstracting from specific representations, following Gandy, states of the computational processes are represented by structural classes. A *structural class* S is a class of states closed under $\in$-isomorphisms. In order to formally define it, Sieg introduces the transitive closure of x ($Tc(x) := x \cup \bigcup\{Tc(y)|y \in x\}$; $\forall a \in U\ Tc(a) = \{a\}$) and the support of x ($Sup(x) := Tc(x) \cap U$). Two states x, y are, then, said to be $\in$-isomorphic if and only if there is a bijection $F : Tc(x) \to Tc(y)$, such that $\forall z \in Tc(x), \forall w \in Tc(z)$ we have $w \in z \leftrightarrow F(w) \in F(z)$ and $\forall z \in Tc(x), \forall r \in Sup(z)$ we have $r \in z \leftrightarrow F(r) \in F(z)$. For any state x, the corresponding structural class S_x is defined to be the equivalence class for the equivalence relation of being $\in$-isomorphic. This class is called *à la* Gandy the *stereotype of* x.

In order to achieve a real abstract description of the system, Sieg makes all operations G work structurally. Two states x, y are isomorphic over (the support of) a certain state z ($x \cong_z y$) if and only if there is a permutation π on U (that can be extended to the universe of all sets) such that $\pi(x) = y \wedge \forall r \in Sup(z)(\pi(r) = r)$. Then, an operation G on

[15]That said, it should not be thought that Sieg equates Turing's and Post's takes on computability. See, in this respect, (Sieg, Szabó and McLaughin, 2016).

a structural class S ($G : S \to S$) is called *structural* if and only if for all permutations π and all $x \in S$: $G(\pi(x)) \cong_{\pi(x)} \pi(G(x))$.

Thus, a computor can be described with a pair $\langle S, G\rangle$, where S is a structural class and G a structural operation on S. In order to express the ability of the computor to assemble a new state from the actual state, a process we believe to be philosophically crucial in Sieg's take on computability, and the two Turing's bounds imposed on the actions of the computor, Sieg introduces two important notions: causal neighborhood (Cn) and determined regions (Dr). A causal neighborhood is a member of a fixed finite class of isomorphism types, such that there does not exist another member into which is $\in$-embeddable. More precisely, y is a *part for* x ($y <^* x$) if $y \neq x \wedge y \neq \emptyset \wedge y \subset \{v|(\exists z)(v <^* z \wedge z \in x)\} \cup \{r|r \in x\}$. Let T be a fixed finite class of stereotypes. Then, y is a *T-part for* x if $y <^* x \wedge y \in T$. A T-part y for x is a *causal neighborhood for* x ($y \in Cn(x)$) if there is no T-part y^* for x such that y is $\in$-embeddable into y^*.

The determined regions of a state z ($Dr(z, x)$) are, instead, a set of states who are, roughly speaking, isomorphic over a certain causal neighborhood of a given state x and their new atoms structurally correspond to each other. Formally, $v \in Dr(z, x)$ if and only if $v \subseteq z \wedge \exists y \in Cn(x)(G(y) \cong_y v \wedge Sup(v) \cap Sup(x) \subseteq Sup(y))$.

We can finally see the definition of a Turing computor, suitably representing Turing locality and boundedness conditions:

> "$M = \langle S; T, G\rangle$ is a *Turing computor on* S, where S is a structural class, T a finite set of stereotypes, and G a structural operation on $\cup T$, if and only if, for every $x \in S$ there is a $z \in S$, such that:
>
> (LC.0) $(\exists! y) y \in Cn(x)$
>
> (LC.1) $(\exists! v \in Dr(z, x))v \cong_x G(Cn(x))$
>
> (A.1) $z = (x \setminus Cn(x)) \cup Dr(z, x)$." (Sieg, 2009, p. 589).

Together with this definition of a Turing computor, Sieg presents a dependent notion of Turing computability. Namely, a certain function F is Turing computable if and only if there is a Turing computor M who computes the values of any argument of F.

The extension of this set-theoretic framework to the notion of a mechanical device is pretty straightforward. The main problem that the parallel computation of Gandy machines poses to an axiomatic characterization of this kind is the phenomenon of overlapping (determined) regions. This simultaneous appearance of new atoms in bounded regions can cause an ambiguity on whether a certain state depends on certain determined regions. This is obviously understandable because it technically underlies what is considered by Gandy the core difference between a computor and a mechanical computer, namely, that a Gandy machine is able to do (certain kinds of) parallel computations.

In order to avoid overlapping ambiguity in the process of assembling the next state from determined regions, Sieg introduces another structural operation, say G_2. This operation must satisfy the same restrictions of the original G, "except that regions determined by it need only be unique up to isomorphism over x" (Sieg and Byrnes, 1999, p 160). In

this way we can isolate isomorphic stereotypes over x, avoiding ambiguity in overlapping regions. Together with G_2, we have also suitable T_2, Cn_2, Dr_2. We refer to the original notions, writing G_1, T_1, Cn_1, Dr_1. These are needed to assure us that for every given state x, the subsequent state $F(x)$ can be uniquely assembled from the determined regions of G_1 and the (structurally) determined regions of G_2. Sieg also defines, for any two states z and x, $A(z, x) := (Sup(z) \setminus Sup(x))$. With the aid of this new set of notions, we have the following definition of a Gandy machine:

> "$M = \langle S; T_1, G_1, T_2, G_2\rangle$ is a *Gandy machine on* S, where S is a structural class, T_i a fine set of stereotypes, G_i a structural operation on T_i, if and only if, for every $x \in S$ there is a $z \in S$, such that:
>
> (L.1) $(\forall y \in Cn_1(x))(\exists! v \in Dr_1(z, x))v \cong_x G_1(y)$;
>
> (L.2) $(\forall y \in Cn_2(x))(\exists v \in Dr_2(z, x))v \cong_x G_2(y)$;
>
> (A.1) $(\forall C)(C \subseteq Dr_1(z, x) \wedge \bigcap\{Sup(v) \cap A(z, x) | v \in C\} \neq \emptyset \rightarrow (\exists w \in Dr_2(z, x))(\forall v \in C)v <^* w)$;
>
> (A.2) $z = \bigcup Dr_1(z, x)$." (Sieg, 2009, p. 591).

As in the case of Turing computability, Sieg defines a dependent notion of computability in parallel or Gandy computability, together with a representation theorem that states that any such Gandy machine is computationally equivalent to a two-letter Turing machine.

One can now see how Sieg's explication of effective calculability achieves its two intended aims. He manages to axiomatically characterize the informal bounds imposed by Turing in his analysis of the intuitive concept of effective calculability. The boundedness and the locality conditions are expressed through a uniqueness restriction to the causal neighborhood of the state at which the computor is looking and to the determined regions considerable by the computor in the process of assembling the next state. In the case of a Gandy machine, these conditions are loosened up in order to allow the computer to perform an unbounded number of bounded computations at the same time. This looseness shows how the concept of a discrete deterministic mechanical device generalizes the concept of a computor.

Representing a simple Turing machine in Sieg's and Gandy's formal framework

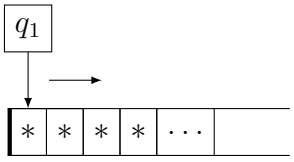
In order to better understand how Sieg's definitions work, let's see how an easy example of a Turing machine can be represented in Sieg's formal framework. For historical pleasure, I take Turing's (Turing, 1936, p. 119) first example of a Turing machine; namely, the Turing machine that alternatively prints 0 and 1 (leaving every time a blank square between them).

The alphabet of this Turing machine is composed by three different symbols $\Sigma = \{*, 0, 1\}$, where $*$ represent a blank state. The machine is able to perform three different operations $C = \{P_0, P_1, R\}$, i.e. respectively to print zero, print one, move right. The tape contains a finite number n of squares $T = \{1, \ldots, n\}$, which we assume to be all blank. The tape can be potentially infinitely extended with an unbounded number of blank squares.

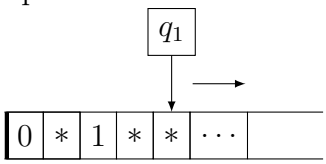
The machine has six different internal states $Q = \{q_1, \ldots, q_6\}$. The program of the machine, and thus the Turing machine, consists of the following 6 instructions[16]:

1. $q_1 * P_0 q_2$
2. $q_2 0 R q_3$
3. $q_3 * R q_4$
4. $q_4 * P_1 q_5$
5. $q_5 1 R q_6$
6. $q_6 * R q_1$

Every instruction of the form $q_i \sigma c_k q_j$ means that if the machine at the internal state q_i scans the symbol σ, then it performs the action c_k changing its internal state into q_j. I assume that the machine starts by scanning the leftmost square 1:



Then, the machine will start printing alternatively zero and one and, for instance, after 6 steps the situation will be this one:



Then, we can represent this machine in Sieg's formalism and check that its set-theoretic representation is a Turing computor. We have already seen that Sieg represents Turing machines *à la* Post, i.e. via instantaneous descriptions (ids) of the form $\alpha q_i \sigma_j \beta$, where α and β are possibly empty strings of symbols, q_i is the current internal state of the machine and σ_j the symbol currently scanned by the machine head. Then, every instruction becomes a rule for obtaining a new id from a given one: $\alpha q_i \sigma_j \beta \Rightarrow \gamma q_l \sigma_k \delta$. Set-theoretically an id is represented as the union of three sets of ordered pairs $ID := Tp \cup Ct \cup St$. $Tp := \{\langle b, b\rangle, \langle b, a_1\rangle, \ldots, \langle a_n, e\rangle, \langle e, e\rangle\}$ represents the tape of the machine, where a_1 is the leftmost square, a_n its rightmost one, b and e special atoms signaling, respectively, the beginning and the end of the tape. $Ct := \{\langle \sigma_{j(0)}, a_1\rangle, \ldots, \langle \sigma_{j(n-1)}, a_n\rangle\}$ represents the tape content, where any σ_j is a certain symbol in the machine alphabet. $St := \{\langle q_i, a_r\rangle\}$

[16]Turing, in his original exposition, allows the machine to print a symbol and then to move right within a single instruction, thereby getting a machine program consisting only of 4 different instructions. I preferred to stick to the limit of one atomic operation for instruction.

represents the current internal state of the machine and the square that is currently scanned by the machine head.

The structural set of states S is obtained as the $\in$-isomorphic closure of all possible ids. Causal neighborhoods are uniquely determined by the current scanned symbol σ_j and the current internal state q_i, that uniquely denote a certain instruction. Then, causal neighborhoods, i.e. the local part on which the structural operation G operates, have, for some squares a_r, some internal state q_i and some symbol σ_j, the form $\{\langle q_i, a_r\rangle, \langle \sigma_j, a_r\rangle, \langle a_{r-1}, a_r\rangle, \langle a_r, a_{r+1}\rangle\}$.

Thus, we have all that we need now to represent the aforementioned Turing machine. Let's take for instance its initial state, when the machine is in the situation sketched in the first picture. The (stereotype of the) initial id is represented as $\{\langle q_1, a_1\rangle, \langle *, a_1\rangle, \ldots, \langle *, a_n\rangle, \langle b, b\rangle, \ldots, \langle e, e\rangle\}$. Thus, the causal neighborhood on which G operates is $\{\langle q_1, a_1\rangle, \langle *, a_1\rangle, \langle a_1, a_2\rangle\}$. $G(x)$, i.e. the next state of the computation (obtained by instruction 1), is then $\{\langle q_2, a_1\rangle, \langle 0, a_1\rangle, \langle a_1, a_2\rangle, \ldots\}$. From this state, another application of G yields $\{\langle q_3, a_2\rangle, \langle 0, a_1\rangle, \langle *, a_2\rangle, \langle a_1, a_2\rangle, \ldots\}$, and so on. All the other instructions can be represented in a similar way. When the computation reaches the nth square and it needs a new atom in order to move right, the next state can be easily assembled introducing a new blank square, say a_p, like $G(\{\langle q_i, a_n\rangle, \langle \sigma_j, a_n\rangle, \langle a_n, e\rangle, \ldots\}) = \{\langle q_k, a_p\rangle, \langle \sigma_j, a_n\rangle, \langle *, a_p\rangle, \langle a_n, a_p\rangle, \langle a_p, e\rangle, \ldots\}$.

It is, then, quite evident that our Turing machine, set-theoretically represented in this way, satisfies all Sieg's axioms for computors and it is thus an example of a Turing computor. In fact, we noticed before that, for any stereotype x, its causal neighborhood is uniquely determined by (the ordered pairs containing, together with the currently scanned square) the current internal state q_i and the currently scanned symbol σ_j. Therefore, at a certain state $x = \{\{\langle q_i, a_r\rangle, \langle \sigma_j, a_r\rangle, \langle a_r, a_{r+1}\rangle, \ldots\}\}$, $LC.0$ is satisfied by $y = \{\langle q_i, a_r\rangle, \langle \sigma_j, a_r\rangle, \langle a_r, a_{r+1}\rangle\}$. Then, as we have sketched, all the possible six different applications of G to y maintain this uniqueness for the determined region v that is $\in$-isomorphic over (the support of) x to $G(y)$, thereby satisfying $LC.1$. For example, if $i = 4$ and $\sigma_j = *$, we have that $v \cong_x G(y) = \{\langle q_5, a_r\rangle, \langle 1, a_r\rangle, \langle a_r, a_{r+1}\rangle\}$. Finally, as $A.1$ requires, the next state z can always be uniquely reassembled via the union of the complement of y in x and v: $z = \{\langle q_5, a_r\rangle, \langle 1, a_r\rangle, \langle a_r, a_{r+1}\rangle, \} \cup \{\langle \sigma_j, a_1\rangle, \ldots, \langle \sigma_j, a_{r-1}\rangle, \langle \sigma_j, a_{r+1}\rangle, \ldots, \langle \sigma_j, a_n\rangle, \langle b, b\rangle, \ldots, \langle a_{r-1}, a_r\rangle, \langle a_{r+1}, a_{r+2}\rangle, \ldots, \langle e, e\rangle\}$.

Now, we can better appreciate how the concept of assembly plays a pivotal role in Sieg's axioms. It is in fact this concept that, throughout the restrictions on causal neighborhood and determined regions, technically represents Turing's way of thinking about effective calculability in terms of the possible actions that a computor can perform on symbolic processes. Furthermore, these restrictions can quite naturally be widened in order to account for the additional freedom of action that a machine calculator is allowed to have.

The two-step reconstruction of the Turing-Gandy-Sieg explications

I can then reconstruct this group of explications of effective calculability in terms of Carnapian explication. The pivotal first-step, what is usually known as Turing analysis, is the

clarification of the explicandum. The intuitive notion of effective calculability gets clarified and disambiguated, the context of uses and the scope of the concept becomes clearer. The output of this clarification is what is defined by Sieg's semi-formal axioms for computors, what I have called computorability. This was firstly achieved informally by Turing, recognized (and used as a basis for the case of machine calculability) by Gandy, and finally made explicit by Sieg. Then, from the notion of computorability, different explicata can be formulated, such as Turing machines or the Gandy-Sieg set-theoretical notions of calculators. Here is a diagram of the conceptual structure of this group of explications (Fig. 3.2):

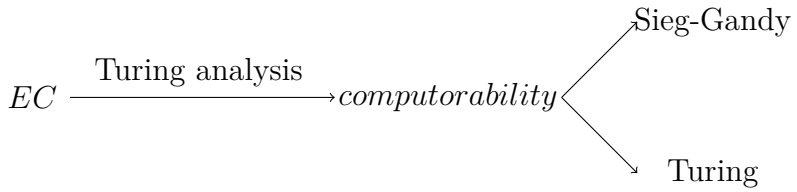


Figure 3.2: The two-step structure of the Turing-Gandy-Sieg explications of effective calculability

I have thus presented Sieg's foundational analysis of computability, together with the related works of Turing and Gandy, showing how it can be understood as a Carnapian explication. I will now turn to Dershowitz's & Gurevich's alternative foundational analysis.

3.3.2 The Kolmogorov-Dershowitz-Gurevich explications

As in Sieg's case, Dershowitz's & Gurevich's axiomatization (Dershowitz and Gurevich, 2008) of the intuitive properties of calculability is the result of decades of work on computability. I will thus recall Gurevich's previous work on computability, specifically focusing on the history, the ideas, and the aims of the ASM project[17], the results of which led to Dershowitz's & Gurevich's axiomatization.

According to Gurevich's own reconstruction, the original idea behind the ASM project was to provide an operational semantics for algorithms by elaborating on what he calls the "implicit Turing thesis" (Gurevich, 1991, p. 267), i.e. the proposition stating that every algorithm can be simulated by an appropriate Turing machine. It was the aim of finding a more efficient simulation of algorithms, something closer and more faithful to computer scientists' actual view of algorithms, that moved him to start the ASM project.

In order to achieve a better simulation of algorithms, the main conceptual idea is to work at an arbitrary abstraction level, thereby considering any algorithm as an independent entity, without imposing a specific data-representation level on its simulation. This conceptual approach to explicating effective calculability can be traced back to Kolmogorov's take on computability. Gurevich repeatedly highlighted the significance of Kolmogorov's

[17]ASM stands for Abstract State Machines, previously known also as Evolving Algebras. As the name tells us, these are idealized machines designed to emulate various classes of algorithms.

work for his own approach to computability. I take the connection between Gurevich's and Kolmogorov's work to be crucial to understand Dershowitz's & Gurevich's axiomatization of calculability. Since Kolmogorov's take on computability is not so well known, I will briefly present his ideas on computability.

Kolmogorov's hidden analysis

If Turing's approach to the problem of explicating the notion of effective calculability was to look at the possible actions of the computor performing the symbolic calculation process, Kolmogorov instead tried to impose restrictions onto the evolution of the computation in a different way. Together with his own student Uspenski, he seemed to be conceptually dissatisfied with the approach of Turing and the other pioneers of the 1936 confluence to computability. He wanted to find a new approach, trying to apprehend the concept of an algorithmic process in all its generality (Kolmogorov and Uspenski, 1958, pp. 62, 68) (Uspenski and Semyonov, 1993, pp. 253, 258).

Uspenski tells us that a fundamental idea in Kolmogorov's work on computability was to view calculation steps as objects, analyzable independently from any computer whatsoever (Uspenski and Semyonov 1993, pp. 253-254, Uspenski 1992, p. 395). Komogorov tried to outline the structure of an arbitrary algorithm, without reference to the specific agent computing it[18].

Kolmogorov proposed what Uspenski calls a "philosophical scheme" (Uspenski, 1992, p. 394), i.e. a semi-formal axiomatic characterization of what an algorithm is:

> 1. "An algorithm Γ applied to any 'condition' ('initial state') A from some set $D(\Gamma)$ ('domain of applicability' of the algorithm Γ) gives a 'solution' ('concluding state') B.
>
> 2. The algorithmic process may be subdivided into separate steps of apriori bounded complexity; each step consists of an 'immediate processing' of the state S (that occurs at this step) into the state $S^* = \Omega_\Gamma(S)$.
>
> 3. The processing of $A^0 = A$ into $A^1 = \Omega_\Gamma(A^0)$, A^1 into $A^2 = \Omega_\Gamma(A^1)$, A^2 into $A^3 = \Omega_\Gamma(A^2)$, etc., continues until either a nonresultative stop occurs (if the operator Ω_Γ is not defined for the state that just appeared) or until the signal saying that the 'solution' has been obtained occurs. It is not excluded that the process will continue indefinitely (if the signal for the solution's appearance never occurs).
>
> 4. Immediate processing of S into $S^* = \Omega_\Gamma(S)$ is carried out only on the basis of information on the form of an apriori limited 'active part' of the state S and involves only this active part." (Kolmogorov, 1953).

[18] Gurevich reported that Leonid Levin, in private conversation, told him that Kolmogorov viewed computations as physical processes. See (Gurevich, 2012, p. 6) and (Gurevich, 2015, p. 197). For my analysis, what matters is that Kolmogorov saw computations as objects that can be treated independently from any computer whatsoever, besides them being either abstract or physical entities.

At first glance, the philosophical originality of Kolmogorov's take is not evident. Kolmogorov's philosophical scheme seems another definition of what I called computorability, i.e. the output of Turing's praised analysis. After all, does not Kolmogorov, as Turing did first, sharpen effective calculability by imposing a boundedness and a locality condition? Is his proposal, then, just another example of Turing's praised way of thinking about computability? With the help of the previous philosophical considerations, a second look at Kolmogorov's axiomatization shows that this is not the case.

Terminological similarities between the two formulations should not trick one into equating these two informal characterizations. In fact, if Turing framed the two bounds by imposing constraints on the possible actions of the computor, in Kolmogorov's axiomatization what is bounded and local is every step of the computational process, which is seen as the structural evolution from the input to the (possible) output. We can better appreciate Kolmogorov's originality by noting that no specific set of initial states or elementary operations is singled out by his definition of an algorithm. The restrictions imposed by these four intuitive considerations are structural limitations on the computational process. Specifically, all the steps of the algorithm must be bounded in their complexity and each of these steps must consist of an elementary transformation that depends only on a limited active part of the state under consideration. Both the boundedness and the locality constraint, then, are understood in a completely different way than in Turing's analysis. Thus, in stating his informal bounds, Kolmogorov focuses on the computational process and not, as Turing did, on the actions of the computor. From the point of view of the computational process, then, effectiveness is achieved by restricting our attention to those processes whose steps satisfy the bounded complexity and the limited active part requirements.

This focus on the computational process is connected with Kolmogorov's aforementioned programmatic aims of characterizing the general structure of an arbitrary algorithm. Kolmogorov and Uspenski stressed this point by stating that the Kolmogorov thesis is stronger than the Turing thesis, claiming that Kolmogorov did not want to just reduce computations to his definition, but also to capture their actual structure (Kolmogorov and Uspenski, 1958, p. 74) (Uspenski, 1992, p. 396) (Uspenski and Semyonov, 1993, p. 251). Coherently with this more abstract aim, Kolmogorov and his students repeatedly stressed how Kolmogorov's approach to computability is more general than previous ones.

I will then refer to the hidden process that led Kolmogorov to his 1953 philosophical scheme as *Kolmogorov analysis*. Despite not being explicit, we have seen that there is evidence in the writings of Kolmogorov and his own students that Kolmogorov analysis is philosophically different from Turing's one. Kolmogorov analysis wants to capture the structure of an arbitrary algorithm, stating the non-trivial restrictions characterizing classical computability on the general evolution of the computational process.

Note that here I stress a generality of Kolmogorov's approach to computability that, coherently with the philosophical perspective of my analysis, is purely conceptual and lies in its semi-formal axiomatization of an arbitrary algorithm. Kolmogorov's work on computability can also be said to be technically more general than Turing's, because Kolmogorov machines work on bounded graphs and are known to be able to implement more algorithms than classical Turing machines. This is a historically interesting fact, but should

not be confused or merged with the conceptual generality stressed above. Also, Sieg and Byrnes have shown that Turing's approach to computability can be suitably generalized to bounded graphs, achieving with the so-called K-graph machines the same level of technical generality as Kolmogorov machines (Sieg and Byrnes, 1996).

Kolmogorov achieved a semi-formal axiomatization of effective computability which is conceptually different from Turing's one. In what follows, I will show how Dershowitz's & Gurevich's axiomatization of computability recovers the conceptual approach of Kolmogorov's work on computability.

Dershowitz's and Gurevich's axioms for computations and the notion of algorithmability

Dershowitz & Gurevich presented their characterization of effective calculability (Dershowitz and Gurevich, 2008), adding a restriction to the initial states allowed by Gurevich's postulates for sequential algorithms (Gurevich, 2000)[19]. Gurevich thinks about algorithms in terms of objects independent from any computer: "in our view, rather common in computer science, algorithms are not humans or devices; they are abstract entities" (Gurevich, 2014, p. 38). What Gurevich wants to stress with this wording is not the (quite trivial) fact that algorithms should not be identified with their computing agents, something that every computability pioneer would obviously agree with. Gurevich, following Kolmogorov's footsteps, stresses instead that his explication of calculability treats algorithms and effective computations independently from their calculators. This philosophical standpoint justifies the technical motto of the ASM project of simulating algorithms at an arbitrary abstraction level, which can be traced back to Kolmogorov's and Uspenski's aforementioned programmatic aims.

According to Gurevich, then, it is possible to informally characterize any classical (i.e. sequential) algorithm by three non-trivial constraints on the evolution of the computational process, called the *Sequential Postulates*. For any sequential algorithm A:

> "POSTULATE I (Sequential time). An algorithm is a state transition system. Its transitions are partial functions." (Dershowitz and Gurevich, 2008, p. 313).

The first postulate tells us that a sequential algorithm, taking this wording in its vague acceptation, is a sequence of discrete computational steps. Gurevich excludes from the scope of his analysis continuous (analog) processes, transfinite computation sequences (involving limits), nondeterministic transitions, and nonprocedural input-output specifications.

> "POSTULATE II (Abstract state). States are structures, sharing the same fixed, finite vocabulary. States and initial states are closed under isomorphism. Transitions preserve the domain, and transitions and isomorphisms commute." (Dershowitz and Gurevich, 2008, p. 317).

[19] Classically, a sequential algorithm (sometimes called sequential-time algorithm) is an algorithm that executes determined, isolated computations bounded step after bounded step.

The second postulate tells us how algorithmic states are represented. It states that any algorithm, on his native level of abstraction, can be represented by (and actually, according to Gurevich, is) a (series of) first-order logical structure(s). There is obviously a certain degree of speculation in this thesis and Gurevich did not present any kind of justification for this proposition except for a classical argument by example: "logic experience shows that any kind of static mathematical situation can be adequately described as a first-order structure" (Gurevich, 1999, p. 10). This representational tenet is shared also by Uspenski (Uspenski, 1992, p. 395).

Moreover, this postulate tells us that neither the vocabulary nor the base set of the algorithm states change during the computation. The evolution of the computational process is thereby given by suitable changes in the values of the functions of a given state[20]. The isomorphic closure is then imposed on the states and transitions of the algorithm for the same reasons adduced by Gandy and Sieg.

> "POSTULATE III (Bounded exploration). Transitions are determined by a fixed finite 'glossary' of 'critical' terms. That is, there exists some finite set of (variable-free) terms over the vocabulary of the states, such that states that agree on the values of these glossary terms, also agree on all next-step state changes" (Dershowitz and Gurevich, 2008, p. 319).

The third postulate expresses the effectiveness requirement, a fundamental limitation imposed on transitions. It can be seen as a generalization of Kolmogorov's bounded complexity and limited active part ideas. This restriction is pivotal to characterize sequential algorithms and, a fortiori, effectively calculable functions. Specifically, this postulate bounds the number of terms that have to be considered in order to make a transition from a given state to the next one. Unbounded searches and infinitary rules are therefore forbidden.

Philosophically speaking, this postulate is informally justified by what Gurevich calls the "Accessibility Principle", i.e. the assertion that, from the point of view of process-evolution, the only way in which a given algorithm A can access an element a of a given state X is to produce a ground term suitably evaluating that element. Then, the finite set T, which this postulate is about, is nothing but the set of all these terms.

The accessibility principle shows how Dershowitz's & Gurevich's postulates conceptually differ from Sieg's. This principle makes explicit a philosophical tenet of Kolmogorov analysis, namely the idea of explicating effective calculability independently from any calculator. If the central concept in Sieg's formal treatment is the notion of assembly (together with the related notions of causal neighborhood and determined regions), i.e. the formal representation of the actions that the computing agent can perform during the process of assembling a new state from a given one, in Dershowitz's & Gurevich's account the computing agent is completely out of the conceptual picture. Technically, the absence of a specific

[20]This encompasses also formulas usually expressed by relations. In fact, in Gurevich's formal framework relations are represented by functions and boolean values. Therefore first-order logical formulas are expressed in a quantifier-free way.

data-representation level makes it impossible to say something about how the computation process is carried out by the computing agent. Gurevich stresses this inability:

> "Imagine yourself being an executor of A at a state X. Since you cannot take advantage of a particular representation of the elements, the only way to access elements of X is to use the basic functions of X. *Essentially you use X as an oracle*"[21] (Gurevich, 1999, p. 14, my emphasis).

Gurevich presented these three postulates, proving a representation theorem between processes satisfying these axioms and a particular type of abstract transition system, called *abstract state machine* (ASM) (Gurevich, 2000). ASMs are a particular version of static algebras, i.e. first-order structures without relations, including in their language three Boolean values as primitive (true, false, undef), as well as the usual Boolean operations. Basic transition rules are recursively constructed by two constructors working on basic update instructions. The problem of representing the addition of new atomic elements, something common in many sequential algorithms, is solved by using a Reserve universe, a naked set from which an appropriate constructor can take an element when it is needed. A sequence of rules is called a program. An ASM is then defined in the following way[22]:

> "An *abstract state machine* (ASM) is given by: a set (or proper class) S of algebraic states, closed under isomorphism, sharing a vocabulary F; a set (or proper class) $I \subseteq S$ of initial states, closed under isomorphism; a *program* P, consisting of finitely many *commands*, each taking the form of a *guarded assignment*: if p then $t := u$ for terms t and u over F and conjuction p of equalities and disequalities between terms" (Dershowitz and Gurevich, 2008, p. 321, original emphases).

Gurevich proved a representation theorem, known as the "ASM Theorem" (Dershowitz and Gurevich, 2008) or the "Main Theorem" (Gurevich, 2000), ensuring that every algorithm that satisfies the three sequential postulates is step-by-step equivalent to a certain ASM program.

In order to adequately characterize computable functions, Dershowitz & Gurevich added another postulate to this characterization. In fact, one cannot be sure that a process satisfying the sequential postulates computes only calculable functions, because initial states are not restricted to computable ones and the other bounds work at a high level of abstraction. Recall that the computing agent is conceptually out of the picture and it technically acts as a kind of oracle. If, then, the initial state of a certain process is an oracle-like non-computable input, the resulting output of the computation can be outside the scope of Turing computable functions. Thus, Dershowitz & Gurevich need to impose, differently from Sieg's axiomatization, an explicit limitation on the set of possible initial states:

[21] Note that I have changed the symbols used by Gurevich to denote an algorithm and its state in order to make the quote coherent with the present terminology.

[22] For a full technical presentation of ASMs see (Gurevich, 1995) or (Gurevich, 2000).

> "POSTULATE IV (Arithmetical State). Initial states are arithmetical and blank. Up to isomorphism, all initial states share the same static operations, and there is exactly one initial state for any given input values." (Dershowitz and Gurevich, 2008, p. 325).

The fourth postulate restricts the set of initial states only to arithmetical blank states[23], thereby restricting the domain of the algorithmic processes allowed only to the numerical ones. A process satisfying both the sequential postulates and this fourth postulate is called an *arithmetical algorithm*. An ASM that satisfies the fourth postulate is called an *arithmetical ASM*. Using the concept of arithmetical ASM the authors proved that every arithmetical algorithm can be emulated by a certain arithmetical ASM and the co-extensiveness of arithmetical ASMs and partial recursiveness.

Then, Dershowitz & Gurevich recovered Kolmogorov's and Uspenski's forgotten conceptual approach to the problem of explicating effective calculability. They gave a new characterization of Kolmogorov's semi-formal axiomatization. Their postulates make explicit the philosophical elements behind Kolmogorov's philosophical scheme, such as the aim of capturing the general structure of an arbitrary algorithm, the emphasis on a more direct simulation of computations, the pivotal restrictions imposed on the evolution of the computing process. I will refer to the semi-formal notion of calculability implicitly defined by the Kolmogorov-Dershowitz-Gurevich postulates as *algorithmability*.

Representing a simple Turing machine in Dershowitz's and Gurevich's formal framework

Just like in the previous subsection, in order to properly understand Dershowitz's and Gurevich's formal definitions, let us look at an example of an algebraic representation of a Turing machine and of an equivalent ASM program.

In this formal framework, an algorithm A is associated with a set $S(A)$ whose elements will be called *states* of A, a subset $I(A)$ of $S(A)$ whose elements will be called *initial states* of A, and a map $\tau_A : S(A) \to S(A)$ that will be called the *one-step transformation* of A. Two algorithms, A and B are then *equivalent* if $S(A) = S(B)$, $I(A) = I(B)$, and $\tau_A = \tau_B$.

States of A are first-order structures, sharing the same vocabulary. A vocabulary is a finite collection of function names, always including the equality sign '=', nullary names 'true', 'false', 'undef', unary name 'Boole', and the names of the usual boolean operations. Terms are inductively defined from nullary functions in the usual way and they are always ground. Then, a structure X of vocabulary Γ is a nonempty set S (called the *base set* of X), together with interpretations of the function names in Γ over S. Nullary functions are identified with their respective values and the equality sign is interpreted with the identity relation on the base set. The value of a term in a given structure is inductively defined as

[23]The authors definition of an arithmetical state is quite long. Roughly speaking, an arithmetical state is a combination of natural numbers, truth values (true, false and undefined), dynamic functions and arithmetical operations. Such an arithmetical state is considered blank when all operations, except the input, have an undefined value.

usual. The one-step transformation τ_A does not change the base set of the algorithm. The set of states is closed under isomorphisms.

If f is a j-ary function name and $\overline{a}$ is a j-tuple of elements of X, then the pair $(f, \overline{a})$ is called a *location* of X. $Content_X(f, \overline{a})$ is the element $f(\overline{a})$ in X. If $(f, \overline{a})$ is a location of X and b is an element of X, then $(f, \overline{a}, b)$ is an *update* of X. In order to execute such update, one has to replace the current $Content_X(f, \overline{a})$ with b. $X + \Delta$ denotes the result of executing a certain set of updates Δ over a state X. For any algorithm A and any of its state X, we define $\Delta(A, X) := \tau_A(X) - X$. We say that two structures X and Y of the same vocabulary Γ *coincide* over a set T of Γ-terms if $\forall t \in T(Val(t, X) = Val(t, Y))$. Then, the bounded exploration postulate assures us that, if A is a sequential algorithm, there exists a finite set T of terms in the vocabulary of A such that $\Delta(A, X) = \Delta(A, Y)$ whenever states X, Y of A coincide over T.

We, then, can easily represent (the computation of) a Turing Machine in this framework. The base set of our structure(s) will be the union of the following three sets, together with the Boolean set. The *Control* set ($Control = \{q_1, \ldots, q_n\}$) contains n elements representing all the possible internal states of the machine. The *Alphabet* set ($Alphabet = \{\sigma_1, \ldots, \sigma_k\}$) contains k different elements representing all the symbols that our machine is able recognize. Finally, the *Tape* set ($Tape = \{a, b, c, \ldots\}$) contains an infinite number of positive integers which represent the (potentially) infinite squares of the Turing machine tape, together with the unary operations *Successor* and *Predecessor* adequately defined on the elements of the set, representing the structure of the machine tape.

The vocabulary of our algebraic representation of a Turing machine contains, together with the logical names, the following function names: the *CurrentControl* nullary function name, shifting values amongst elements of the *Control* set, denoting the current internal state of the machine; the nullary function name *Head*, shifting values amongst elements of the *Tape* set, denoting the square currently scanned by the machine head; the unary function name $Content : Tape \to Alphabet$, representing the symbols contained in every square of the tape.

In order to make as clear as possible the differences between ASMs and Turing computors formalism, I will use the same Turing machine that we represented in Sieg's formal framework: the Turing machine that alternatively prints 0 and 1, leaving every time a blank square between them.

As I did in Sieg's formalism, we assume for simplicity that every square of the tape is blank and that our Turing machine can only move right. Initially, we assume that our Turing machine starts in the internal state q_1, scanning the leftmost square of the tape.

Then, the initial state of the computation of our Turing machine would be represented (up to isomorphisms) by the following structure. The fixed base set of our initial state X_1 is: $\{q_1, q_2, q_3, q_4, q_5, q_6\} \cup \{0, 1, *\} \cup \{a, b, c, \ldots\}$, together with the boolean set. Non-logical functions take the following values: $CurrentControl = q_1$, $Head = a$, $\forall s \in Tape(Content(s) = *)$. Then, changes in the non-logical function values in next state of the computation, X_2, are contained in the update set $\Delta(A, X_1) = \{(CurrentControl, q_2), (Content, Head, 0)\}$. Changes between X_2 and X_3 are, instead, given by the different update set $\Delta(A, X_2) = \{(CurrentControl, q_3), (Head, Successor(Head))\}$, and so on.

We, then, can easily write the program of an ASM simulating step-by-step the computation of our Turing machine, i.e. the update sets for all the states of the computation, nesting do-in-parallel guarded assignments:

$$
\begin{array}{c}
\text{if } CurrentControl = q_1 \text{ and } Content(Head) = *, \text{ then} \\
par\text{: } CurrentControl := q_2; Content(Head) := 0 \\
\text{else if } CurrentControl = q_2 \text{ and } Content(Head) = 0, \text{ then} \\
par\text{: } CurrentControl := q_3; Head := Successor(Head) \\
\text{else if } CurrentControl = q_3 \text{ and } Content(Head) = *, \text{ then} \\
par\text{: } CurrentControl := q_4; Head := Successor(Head) \\
\text{else if } CurrentControl = q_4 \text{ and } Content(Head) = *, \text{ then} \\
par\text{: } CurrentControl := q_5; Content(Head) := 1 \\
\text{else if } CurrentControl = q_5 \text{ and } Content(Head) = 1, \text{ then} \\
par\text{: } CurrentControl := q_6; Head := Successor(Head) \\
\text{else if } CurrentControl = q_6 \text{ and } Content(Head) = *, \text{ then} \\
par\text{: } CurrentControl := q_1; Head := Successor(Head)
\end{array}
$$

The two-step reconstruction of the Kolmogorov-Dershowitz-Gurevich explications

Let me reconstruct the conceptual structure of this group of explications of effective calculability in terms of Carnapian explication. The clarification of the explicandum was implicitly achieved by Kolmogorov in the process that led him to his 1953 philosophical scheme, i.e. Kolmogorov analysis. I called the conceptual offspring of his efforts *algorithmability*, i.e. what is defined by the Kolmogorov-Dershowitz-Gurevich semi-formal axioms for computations. From this notion, different explicata can be formulated, such as the classical Kolmogorov machines or the more recent ASMs. Here is a conceptual diagram of this second group of explications (Fig. 3.3):

Figure 3.3: The two-step structure of the Kolmogorov-Dershowitz-Gurevich explications of effective calculability

I have thus presented Dershowitz's & Gurevich's foundational analysis of computability, together with Kolmogorov's seminal take, explaining how it can be understood as a Carnapian explication. In what follows I will conceptually compare this group of explications with the Turing-Gandy-Sieg ones.

3.3.3 Foundational analyses of computability as three-step explications

We have seen how Sieg's and Dershowitz's & Gurevich's axiomatizations build upon different approaches to computability and how some of the differences between their technical frameworks depend on their different conceptual background. Specifically, Sieg proposed axioms for calculators, building upon Turing's and Gandy's seminal works on computability. The conceptual core of his axioms for computors and mechanical devices is the degree of freedom that the calculator is allowed to have. Dershowitz & Gurevich, following Kolmogorov's footsteps, instead proposed axioms for computations, considering them independently from any calculator whatsoever. The core of their axioms are the abstract state and the bounded work postulates, i.e. non-trivial representational and procedural limitations of the calculation process itself.

These approaches can be traced back, respectively, to Turing's praised analysis of calculability and Kolmogorov's seminal ideas about the intuitive notion of algorithm. In explication terms, it seems *prima facie* that Turing and Kolmogorov clarified the notion of effective calculability in two different ways, arriving respectively at the notions of computorability and algorithmability.

However, where exactly lies the conceptual difference between these two groups of explications? Several possibilities should be considered. These authors could for instance have explicated different intuitive concepts. Alternatively, there could be two possible disambiguations of effective calculability, as in the case of logical and statistical probability (Carnap, 1950b). Another possibility could be that these two groups of explications differ in the way in which they clarify their common explicandum or in the way in which they formulate their explicatum. In other words, where does the branching between these two groups of explications happen? Does it happen at the stage of the clarification of the explicandum or does it happen at the stage of the formulation of the explicatum? My proposal is that this difference lies neither in the first clarification step nor in the final step of the formulation of the explicatum, but rather in an additional step between the two.

The difference between these two approaches to computability cannot lie in the final step of the formulation of the explicatum. After all, there are many other instances of mechanical computability, extensionally equivalent to Turing computability, that do not show any such conceptual difference. Take Post's explication of effective calculability, for instance. Post's notion of normal system generability is indeed a different formalization of the intuitive concept of calculability than Turing's one. According to Post, a function is effectively calculable if and only if it(s symbolic representation) can be generated via the iterative application of certain substitution rules (Post, 1943). This characterization of computability is then by no means conceptually reducible to Turing's one. However, Post's work can be completely subsumed under the conceptual approach of the Turing-Gandy-Sieg explications (cf. Sieg 2018). Even though Post does not specify a semi-formal notion of computability, his work seems to implicitly make use of the notion of computorability. In fact, what matters in Post's systems are the possible moves that one is allowed to make in order to solve the specific symbolic substitution puzzle, i.e. in order to calculate a certain

function. In order to explicate the notion of effective calculability, Post then pays attention to the symbolic processes that can be carried out by an idealized human being working in a clerical way, just as Turing did[24]. Thus, the difference between Turing's and Kolmogorov's approaches must be sought elsewhere.

Does perhaps the difference between these two groups of explications lie in the first step of these explications, i.e. in the clarification of the explicandum? Are these two foundational analyses two distinctive way of disambiguating or clarifying the intuitive notion of effective calculability? Even though it is certainly tricky to give arguments for the uniqueness of an intuitive, vague concept that different people tried to capture, some evidence can be provided. Despite their conceptual differences, Kolmogorov and Turing were trying to explicate the same informal notion: effective calculability. Compare the examples and the initial discussion contained in (Kolmogorov and Uspenski, 1958) with the ones in (Turing, 1936). Moreover, if one looks at the two foundational analyses of computability here treated, both Sieg and Dershowitz & Gurevich explicitly state that the aim of their work is to capture the intuitive notion of effective calculability, as traditionally understood and clarified (Sieg, 2009; Dershowitz and Gurevich, 2008). Turing and Kolmogorov explicated also the same contexts and uses of the term, i.e. the same clarified explicandum. Just like Turing, Kolmogorov was interested in explicating classical calculability in all its symbolic generality, abstracting away from the actual limitations of any computer or any domain. The difference between these two groups of explications is thus not explainable as a difference in the clarification of effective calculability such as the one at play in Church's work on general recursiveness (Quinon, 2019, pp. 22-25).

Where then does the conceptual difference between Turing's and Kolmogorov's approaches to computability lie? It seems that the traditional, two-step version of Carnapian explication is not fine-grained enough to capture the differences between the semi-formal notions of computorability and algorithmability. In what follows I will present a refined three-step version of Carnapian explication that is able to capture the conceptual differences between these two groups of explications.

Explication as a three-step procedure: the semi-formal sharpening of the clarified explicandum

As I repeatedly stressed in this chapter, Carnapian explication is originally a two-steps procedure. I am now going to propose a refinement of the explication with the help of which I will achieve a better analysis of the two groups of explications of effective calculability treated above. This refinement adds another mid-level step, the *semi-formal sharpening of the clarified explicandum*, to the procedure, making it thus become a three-step method.

In the first step of Carnapian explication, i.e. the clarification of the explicandum, the explicator ought to clarify and (possibly) disambiguate the concept that she seeks to explicate, relative to a given context of use. My claim is that after this first, classically recognized step, in some cases the explicator sharpens this notion into a semi-formal one,

[24]In order to better appreciate this connection, see Turing's late take on solvable and unsolvable puzzles in (Turing, 1954).

a mid-level ExplicanDum**. Only after this sharpening, the explicator passes to the final third step of the procedure of explication, i.e. the formulation of the explicatum (Fig. 3.4).

$$ED \overset{\text{clarification}}{\longrightarrow} ED^{*} \underset{\text{sharpened concept}}{\overset{\text{sharpening}}{\longrightarrow}} ED^{**} \overset{\text{formulation}}{\longrightarrow} ET$$

Figure 3.4: The structure of the refined three-step version of explication.

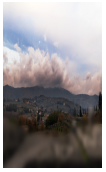
The process of getting from ED^{*} to this semi-formal mid-level notion ED^{**}, i.e. the semi-formal sharpening of the clarified explicandum, is not a theoretically neutral step. Here the explicator has to make a theoretical choice amongst various possible directions in the sharpening. Every sharpening must in fact possess a given theoretical focus, which is given by the core aspects of the concept on which the sharpening focuses. The explicator highlights certain aspects of the concept, while she leaves other aspects at the borders of the conceptual sharpening.

Using a visual analogy, think about the action of shooting a picture of a landscape with a reflex camera. In order to take a good picture of the landscape, one has to focus the reflex camera on certain parts of the landscape, inevitably blurring the rest. It is impossible to focus on every element of the landscape at the same time. In order to take a good picture one has to choose what to focus on. Perhaps one wants to get a clear shot of the background of the landscape or perhaps one wants to put into focus the foreground. You cannot have both at the same time.

(a) ED

(b) ED^{*}

(c) ED^{**}_{1}



(d) ED^{**}_{2}

Figure 3.5: Succession of shots of the same landscape with different focuses, representing the evolution of a given explicandum during the first two steps of the refined three-step version of explication.

This is completely analogous to what happens in the process of semi-formally sharpening the clarified explicandum. The succession of shots of the same landscape in (Fig. 3.5) represents the possible evolution of a given explication. In (a) we can see a completely blurred landscape with unclear boundaries. This represents the starting point of

any explication, i.e. a given intuitive explicandum ED, the scope of which is undefined and that perhaps contains a great deal of ambiguity and vagueness. In (b) we can instead see the same blurred landscape, this time framed inside clear boundaries. This represents the clarified explicandum ED^*, which is obtained from the original explicandum through the clarification and the disambiguation of the uses and the contexts for which the notion is being explicated. Pictures (c) and (d) represent instead two different ways of putting the landscape into focus, respectively highlighting the foreground and the background of the picture, while inevitably blurring the rest of the image. The passage from (b) to one of these two pictures represents the semi-formal sharpening of the clarified explicandum. In this second step, the explicator sharpens the clarified explicandum ED^* onto a semi-formal sharpened ED^{**}. By stating semi-formal axioms or definitions, the explicator freely chooses the parts and features of the clarified explicandum that she wants to highlight. The parts and features to highlight and therefore the direction onto which the clarified explicandum is sharpened is never a philosophically neutral step. As represented by Pictures (c) and (d), multiple choices are always possible[25]. The semi-formal sharpening step thus changes neither the uses nor the context in which the explicandum is explicated, but it makes more precise the notion via a theoretically-laden semi-formal definition.

Finally, there is the last step of the explication, i.e. the formulation of the explicatum. I left this step out of the landscape-picture analogy because this step is different from all the others, due to its focus on the explicatum and not on (a disambiguated or clarified or sharpened version of) the explicandum. The explicatum can be a fully formalized notion or even an informal one in some cases (e.g. the fish-piscis example in Carnap 1950b), but it is always a whole other concept, fully detached from the explicandum.

Let me stress that one should not expect the mid-level step of the sharpening of the clarified explicandum to occur in every explication whatsoever, but only in certain complex cases where many different formal explicata are meant to replace a single explicandum. Apart from the CTT case, other examples of concepts that may exhibit different sharpenings of the same clarified explicandum and for which the three-step explication seems an appropriate tool of analysis are formal theories of truth (Horsten, 2011; Halbach, 2014), different conceptions of set (Incurvati, 2020), theories of informal proofs (Leitgeb, 2009; Sjögren, 2011), notions of logical consequence (Etchemendy, 1990), and mathematical conceptions of infinity (Mancosu, 2009).

In what follows, I will show how this refined three-step version of explication is able to account for the conceptual differences between the two foundational analyses of computability seen in Section 3. Specifically, the step of the semi-formal sharpening of the clarified explicandum will prove itself to be pivotal into adequately capturing the difference between the notions of computability and algorithmability.

[25]Some scholars outside the Carnapian tradition have also highlighted this possibility of sharpening concepts in multiple directions. See for instance (Smith, 2011, pp. 27-29). Shapiro also stressed, in a Waismannian fashion, the possibility of this multiple sharpening in (Shapiro, 2013). However, it should be noted that despite some similarities, my concept of sharpening is a technical term that has to be understood inside the proposal of the three-step version of explication.

Computorability vs Algorithmability[26]

Applying my refined three-step version of Carnapian explication to the two aforementioned groups of explications of effective calculability gives us this conceptual picture (Fig. 3.6):

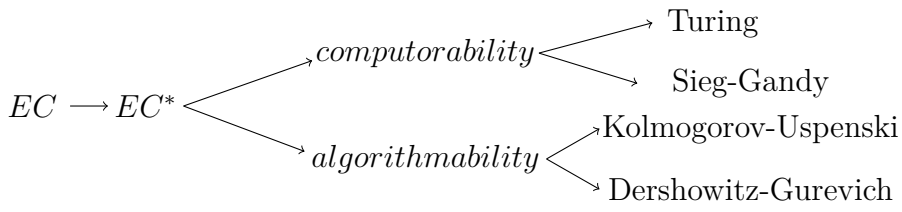


Figure 3.6: The three-step structure of the Turing-Gandy-Sieg and the Kolmogorov-Dershowitz-Gurevich explications of effective calculability

The branching between these two groups of explications happens neither in the first step of clarifying the explicandum nor in the final step of the formulation of the explicatum, but instead on the mid-level. The second-step, the semi-formal sharpening of the clarified explicandum, is where the two foundational analyses of computability differ. From the perspective of this refined version of explication, then, the notion of computorability and the notion of algorithmability are two semi-formal sharpenings of the clarified notion of effective calculability. Despite sharpening the same clarified notion, namely the classical concept of calculability clarified in terms of symbolic processes, they do that in two different ways, paradigmatically exemplified by Turing's praised analysis and Kolmogorov's hidden analysis of computability.

Let me state more precisely what these two semi-formal sharpenings of the clarified notion of effective calculability underlie:

- **Computorability**.

  The sharpening is achieved in terms of the possible actions of the computor. Bottom-up bounds are imposed on the freedom of the agent of the computation. The conceptual focus of the analysis is at the calculation level.

- **Algorithmability**.

  The sharpening is achieved in terms of non-trivial structural features of the process of computation. Top-down restrictions are imposed on the evolution of the computational objects. The conceptual focus of the analysis is at the abstract algorithmic level.

Then, the difference between the notion of computorability and the notion of algorithmability can be understood in terms of two different semi-formal sharpenings of the same clarified explicandum.

[26]It should be noted that both computorability and algorithmability are to be understood as technical terms, as I explain in my definition of these two notions that can be found in this page.

Turing was the first one to sharpen the clarified notion of effective calculability, in what is usually called Turing analysis. Technically, I want to stress that this name is misleading. This is not a conceptual analysis, this is conceptual engineering of an intuitive notion into, first, a clarified notion and then into a sharpened one. Turing merged these two steps in his famous informal exposition of 1936. In his 'analysis', he fixed the use and the context of his explication of effective calculability, abstracting the notion of effective calculability from practical limitations, ruling out infinity and any kind of ingenuity, and focusing on the symbolic processes underneath any calculation process. This disambiguation and clarification of effective calculability belongs to the explication step of the clarification of the explicandum. At the same time, Turing sharpened this clarified explicandum by arguing that what is effectively calculable has to be computable by an abstract human calculator respecting in his actions the bounds that we are now all familiar with, thereby implicitly giving a semi-formal definition of the notion of computorability. This implicit semi-formal axiomatization of effective calculability in terms of actions of a computor belongs instead to the mid-level step of the sharpening of the clarified explicandum.

Turing's approach has been the prevailing way of semi-formally sharpening effective calculability since his 1936 analysis. After Turing, Gandy improved and made (some) assumptions underlying Turing's approach explicit, using them in order to attack the general problem of machine computability and thereby generalizing Turing's analysis of an abstract computor with his four 'principles for mechanisms'. Sieg continued this tradition with a detailed historical analysis of the conceptual background of the 1936 confluence and by axiomatically improving Turing's and Gandy's foundational analyses of computability.

The first occurrence of the Kolmogorov's approach to explicating effective calculability was instead Kolmogorov's own 1953 informal axiomatization of an algorithm. Kolmogorov, like Turing, implicitly clarified and sharpened effective calculability, merging these two steps into the hidden analysis behind his semi-formal definition of algorithmability. Together with Uspenski, he improved the 1953 treatment, arriving at the definition of a Kolmogorov machine, the first explicatum obtained via the sharpened notion of algorithmability. Despite the fruitfulness of Kolmogorov's model of computation (Uspenski, 1992; Sieg and Byrnes, 1996), the originality of his conceptual take has not been equally recognized. Gurevich and the ASM project revitalized Kolmogorov's approach to explicate effective calculability and the notion of algorithmability. As I stressed in Section 3, the ASM project approach to computability was mainly motivated by technical reasons, but nevertheless it recovered Kolmogorov's forgotten conceptual approach. A remarkable offspring of this project is Gurevich's axiomatic characterization of sequential algorithms, which makes evident the originality of the Kolmogorov-focus based view of computation. Together with Dershowitz, Gurevich then restricted his axioms in order to capture effective computations, thereby achieving a foundational analysis of computability deeply rooted in Kolmogorov's approach.

The refined three-step version of explication is also able to explain some debates about these two foundational analyses of computability. Recall that the semi-formal sharpening of the clarified explicandum always implies a theoretical choice. Two different semi-formal sharpenings of the same clarified explicandum are therefore conceptually incompatible.

This can be seen in some remarks made by Sieg and Gurevich on each other's foundational analysis.

Sieg doesn't recognize the originality of Dershowitz's & Gurevich's axioms, objecting to their claim of their approach being more general than the Gandy-Sieg one (Sieg, 2013, pp. 119-121). We have seen that the generality stressed by Dershowitz & Gurevich is of a conceptual nature, being namely the fact that Kolmogorov's semi-formal sharpening of the clarified notion of calculability into algorithmability focuses on the abstract algorithmic level and not on the level of computation. This cannot be captured nor fully understood from the perspective of Turing's sharpening and thus within Sieg's framework.

This conceptual tenet of Turing's way of semi-formally sharpening effective calculability also has a formal correlate in the technical framework of Sieg's explicatum, in its focus on how a certain state of the computation is assembled, i.e. the input-output behavior on the specific level of data-representation at which the calculator works. Symmetrically, in Dershowitz's & Gurevich's treatment there is no place for any concept of calculator whatsoever. There is no trace of a computor in their formal framework or in their postulates, and Gurevich's accessibility principle explicitly forbids any meaningful notion of calculator. The technical correlate of this principle is the abstraction from any specific data-representation that makes the evolution of the state work via an oracle-like ground-term production. The assumption of Kolmogorov's way of semi-formally sharpening effective calculability explains their aforementioned claims of conceptual generality and also their critiques of Gandy's machines as a model of parallel computation. From the perspective of the notion of algorithmability, since the calculator is out of the conceptual picture, a Gandy machine seems then an unnatural layering of mechanical devices (Gurevich, 2012, pp. 271-272). More generally, since both Gandy's and Gurevich's treatment of machine calculability heavily builds on their respective explications of effective calculability, conceptual differences between their formal models of machine computability can be explained by their opposite way of semi-formally sharpening effective calculability.

Another philosophical difference between these two groups of explications that can be explained by my refined three-step version of explication is the one stressed by Smith in the context of his idea of a squeezing argument for effective calculability (Smith, 2013, pp. 357-364). He stressed how Kolmogorov imposed limitations on the concept of effective computation as top-down restrictions, in sharp contrast to Turing's bottom-up bounds on the actions of the computor. This difference is a consequence of the different semi-formal sharpening of the clarified explicandum achieved by these two pioneers of computability. Turing, in order to arrive at the notion of computorability, had to impose bottom-up bounds on the possible actions of the computor, having chosen to highlight the agent of the computation. Kolmogorov, instead, highlighted the abstract structure of an arbitrary algorithm and its evolution process and thus, in order to define the notion of algorithmability, imposed the pivotal limitations as top-down restrictions.

We have then seen how, with the help of my refined three-step version of Carnapian explication, we can adequately capture and conceptualize the difference between algorithmability and computorability in terms of two different semi-formal sharpenings of the same clarified explicandum. More generally, this case study shows the significance of the

intermediate steps in the evolution of a given explication. As we saw in the case of CTT, even small conceptual differences in a crucial mid-level step of the explication procedure can cause heavy formal differences and related conceptual oppositions in the formal explicata. It seems likely, thus, that one could resolve some conceptual oppositions between formal explications of the same notion in the same way. Examples of explications structurally similar to CTT, such as formal theories of truth, logical consequence, informal provability, conceptions of set, and infinity seem promising future applications for the three-step Carnapian explication.

3.4 Formalizing Carnapian Explication in the Theory of Conceptual Spaces

After having focused our attention on the structure of Carnapian explication in the last section, let us go back to the epistemological debates concerning the viability of explication as a philosophical method. In Section 2, we saw that a significant part of the philosophical debate over explication has focused on the desiderata that a good explicatum has to respect. More specifically, many different ways of spelling out the four desiderata singled out by Carnap have been proposed and some critics have stressed the tension between some of them. A problem with these discussions over explication desiderata is that both the specific proposals and the critiques are often difficult to judge due to the aforementioned vagueness and ambiguity of Carnap's four desiderata.

With the hope of improving this situation, in this section I will propose a way of making precise the procedure of explication and its desiderata by means of the theory of conceptual spaces. Specifically, I will show how different readings of these desiderata can be made precise in terms of geometrical and topological constraints over the conceptual spaces of the explicandum and the explicatum. Moreover, I will demonstrate how, thanks to this explication of the concept of explication itself, the specific proposals of desiderata in the philosophical literature can be assessed and compared. I will also argue that my proposal is able to answer the aforementioned critiques of explication as a philosophical procedure by reconstructing Carnapian explication in a pragmatic yet precise meta-framework where one can have more fine-grained readings of explication desiderata.

As a formal background for my explication of 'explication', I will rely on the theory of conceptual spaces (Gärdenfors, 2000, 2014). Conceptual spaces have been successfully applied in different fields, proving themselves to be a powerful tool for representing different types of linguistic and conceptual phenomena, such as concept formation, metaphors, contextual effects, meanings (Gärdenfors, 2000, 2014; Zenker and Gärdenfors, 2015a). In philosophy, conceptual spaces have been used to account for vagueness-related phenomena for classificatory and comparative concepts and to model inductive inferences and other forms of conceptual manipulation (Douven et al., 2013; Decock, Dietz and Douven, 2013; Decock and Douven, 2014; Gärdenfors, 2000; Sznajder, 2016; Osta-Vélez and Gärdenfors, 2020). In philosophy of science, conceptual spaces have been used to model various types of

theory-change in physics as transformation of the related conceptual space(s) (Gärdenfors and Zenker, 2011, 2013; Zenker and Gärdenfors, 2015a; Masterton, Zenker, and Gärdenfors, 2017).

The plan for the rest of the section is the following. I will first present the theory of conceptual spaces, both from a philosophical and a technical point of view. I will focus on some recent technical extensions of the theory, developed in order to treat vague and comparative concepts in the framework of conceptual spaces. Then, I will make methodologically precise the goal of my explication of 'explication', by distinguishing two senses in which one can explicate the concept of explication itself. I will also focus on two pivotal assumptions concerning the concepts and the desiderata involved in the procedure of explication that my proposal requires. After these methodological matters, I will show how the procedure of explication can be made precise inside the theory of conceptual spaces. Specifically, we will see how different readings of explication desiderata presented in the literature can be formalized as topological or geometrical constraints on the conceptual spaces related to the explicandum and the explicatum. I will also make evident how the representation of the explicandum and the explicatum in conceptual spaces allows us to state more fine-grained desiderata for the adequacy of an explication, which arguably show how explication can be successfully defended against some recent critiques. Finally, in order to make clearer my proposal, I will show how two paradigmatic cases of successful explications from the history of science can be represented and assessed in the context of my explication of 'explication': the scientific concept of temperature and the morphological concept of fish.

3.4.1 Conceptual spaces

The theory of conceptual spaces (Gärdenfors, 2000, 2014) has to be understood as a theory of mental representation. According to it, we represent information at three different levels (in order of ascending complexity): subconceptual (e.g. neural networks), conceptual, and symbolic (e.g. Fodor's language of thought). The conceptual level is where the categorization process takes place and where we construct properties, concepts, meanings, and categories. The main tenet of conceptual spaces theory is that we can model what happens at this level geometrically.

Pivotal in the theory of conceptual spaces is the notion of *quality dimension*. The idea is that a quality dimension represents a particular (aspect of a) quality with respect to which objects can be judged as more or less similar. The more similar two objects are with respect to that quality, the closer their related points in that quality dimension. Examples of familiar concepts that can be modeled as quality dimensions include time, weight, size, and brightness. A quality dimension is a dimension in a strict geometrical sense, i.e. every quality dimension is equipped with a specific geometrical or topological structure. Quality dimension often come with a metric, i.e. with a distance function, but also qualitative measures of distances are allowed. Neither dimensions or the metrics are arbitrary, but are usually determined on the basis of a large set of similarity judgments via suitable techniques such as multidimensional scaling or principal component analysis

(Douven and Gärdenfors, 2019, p. 5).

Quality dimensions can be integral or separable. Two dimensions are *integral* iff to assign a value to an object in one of them implies simultaneously assigning a value in the other one. Dimensions that are not integral are called *separable*. Color perception dimensions, such as saturation and hue, are a familiar example of integral dimensions. Dimensions of shape and weight are instead examples of separable dimensions. Quality dimensions appear often related together in stable groups. A set of integral dimensions that are separable from all the other ones is called a *domain*. Examples of domains are the color domain (constituted by the dimensions of hue, saturation, and brightness) and the space domain (height,width, and depth). A *conceptual space* is, then, a collection of one or more domains.

A conceptual space is able to represent objects, properties, and concepts. Objects are represented as vectors. Properties are represented as certain kinds of regions in a domain. Natural properties are hypothesized to be (representable as) convex regions of a domain (Gärdenfors, 2000, p. 71). Concepts, then, are certain kinds of sets of regions in a (possibly open-ended) number of domains. Natural concepts are sets of regions in a number of weighted domains equipped with information about how regions in different domains are correlated (Gärdenfors, 2000, p. 105). It is then possible to distinguish between core and peripheral properties of concepts, by assigning different salience weights to different domains. In a similar fashion, singular dimensions can be weighted in order to account for contextual effects of various kind.

Gärdenfors then takes convexity to be the pivotal feature of regions representing natural properties and concepts. The necessity of convexity as a criterion of naturalness in conceptual spaces has been criticized by various scholars. Mormann highlighted that convexity requires the underlining conceptual space to be metrical or linear and it therefore strongly restricts the possible structure of the conceptual space (Mormann, 1993, p. 220). He instead favored a pluralist approach to naturalness criteria, arguing that in many cases weaker topological notions such as connectedness or closedness are as good as convexity and they do not impose strong restrictions on the underlining structure of the space (Mormann, 1993, p. 226, p. 239).

Recently, Hernández-Conde has strengthen the case against convexity as a naturalness criterion, arguing that this constraint is problematic both from a theoretical and a practical perspective. He claimed that the main arguments that Gärdenfors gave for convexity either require very strong assumptions on the underlining structure of the space or they work also for weaker requirements such as star-shapedness (Hernández-Conde , 2017). Moreover, he showed how convexity appears to be problematic also from the inner perspective of conceptual spaces theory (Hernández-Conde , 2017, pp. 4027-4034). Gärdenfors (Gärdenfors, 2019) replied to these critiques, claiming that it is an open empirical matter which of these criteria is the more adequate one. I remain neutral on whether convexity is the right criterion of naturalness in conceptual spaces. In Section 4.4, I will show how a plurality of geometrical constraints can be used to make precise in conceptual spaces the fruitfulness of an explication, but I will not endorse any reading of explication desiderata as the correct one.

Technicalities

In the theory of conceptual spaces the fundamental notion of a quality dimension has to be understood as a proper geometrical dimension. Axioms and primitive relations of any dimension can be of any kind. Minimal requirements can be defined in terms of the relations of *betweenness* ($B(a,b,c)$) and *equidistance* ($E(a,b,c,d)$). Since the most fundamental task of a quality dimension is the assessment of similarities, what specifically characterizes a certain dimension is the notion of distance with which it is equipped. Dimensions can have either a qualitative (e.g. a notion of equidistance) or a quantitative (e.g. a certain metric) notion of distance. If a certain dimension is equipped with a quantitative distance function, it is then called a metric space. A function $d : S \times S \Rightarrow \mathbb{R}_0^+$ is called a distance function iff $\forall x, y, z \in S$: $d(x,y) \geqslant 0$, $d(x,y) = 0 \leftrightarrow x = y$, $d(x,y) = d(y,x)$, and $d(x,y)+d(y,z) \geqslant d(x,z)$. Examples of metrics, for a n-dimensional space, are the Euclidean metrics ($d_E(x,y) = \sqrt{\Sigma_i(x_i - y_i)^2}$) and the city-block metrics ($d_C(x,y) = \Sigma_i|x_i - y_i|$)[27]. We can easily vary the scales of the different dimensions of a certain conceptual space by putting a weight w_i on the distance function of the dimension i. Similarity is, then, an exponentially decaying function of distance (e.g. Shepard's universal law of generalization $s_{ij} = e^{-c \cdot d_{ij}}$).

In a certain conceptual space S, consisting of a set of domains $\{D_1, \ldots, D_n\}$, each made up of a set of integral dimensions $\{d_1, \ldots, d_m\}$, we can represent objects as vectors $\langle v_1, \ldots, v_j \rangle$. Properties, then, are represented by regions S of a domain. We can define a region of a space as a set of points that respect certain criteria that we impose on the primitive relation: $C(X,Y)$, *X connects with Y* is minimally constrained by symmetry and reflexivity. A possible criterion for defining a region is *connectedness*, i.e. X is connected iff $\forall Y, Z(Y \cup Z = X \rightarrow C(Y,Z))$. A stronger criterion that can be imposed as a definition of region is *star-shapedness relative to a point*, i.e. X is star-shaped relative to a point x_0 iff $\forall z, \forall x \in X(B(x,z,x_0) \rightarrow z \in X)$. An even stronger criterion is *convexity*, i.e. X is convex iff $\forall z, \forall x, y \in C(B(x,z,y) \rightarrow z \in C)$. Concepts are represented, then, as multi-domain bundles of properties $X_1, \ldots, X_n$, together with salience weights w_i on the domains and cross-domain correlations.

(a) (b)

Figure 3.7: A convex (a) and a star-shaped (b) region.

One of the most promising applications of conceptual space is concept formation and categorization. Gärdenfors' account of concept formation and categorization in conceptual

[27]For a survey of other possible metrics, see (Tversky et al., 1971-1989-1990, Volume 2, pp. 51-77).

spaces combines prototype theory (Rosch 1975, cf. Chapter 2, Section 1.2) and the spatial tessellation technique called Voronoi diagrams (Okabe et al., 2000). A Voronoi diagram is a tessellation of a space that, provided with a set of points, divides the space in cells, each cell having as a center one of the points in the original set and containing all the points that lie closer to its center than to the centers of the other cells. More accurately, for any n-dimensional space and any set of pairwise distinct points of S $P = \{p_1, \ldots, p_k\}$, the *Voronoi diagram generated by* P is the set $V(P) = \{v(p_i)|p_i \in P\}$, where $v(p_i)$ is the region $v(p_i) = \{p|d(p, p_i) \leqslant d(p, p_j) \forall j \in \{1, \ldots, k\}\}$ and it is called the *Voronoi polygon/polyhedron associated with* p_i.

(a) (b)

Figure 3.8: A normal (a) and a collated (b) Voronoi Diagram.

The theory of conceptual spaces has been recently extended in order to treat vague (Douven et al., 2013) and comparative concepts (Decock and Douven, 2014; Dietz, 2013; Decock, Dietz and Douven, 2013). The first step for dealing with vagueness in conceptual spaces is to substitute unique prototypes with prototypical areas, thereby making the generator set P become a set of regions. Then, the basic idea is to consider all the possible Voronoi diagrams that can be built from choosing a single point in each generator region. Intuitively, we treat vagueness in a way similar to the supervaluationist account, because every possible Voronoi diagram represents a possible completion of the tessellation of the space and thus a possible way of deciding the borderline cases of the concept involved. Then, we project all these possible Voronoi diagrams onto each other. From the result of this projection, called a collated Voronoi diagram, we can define boundary regions of categorization in order to accurately represent borderline cases of concepts.

More formally, consider the *restricted Voronoi polygon associated with* p_i, i.e. the region made of all points that lie strictly closer to p_i than to the other central points: $\underline{v}(p_i) = \{p|d(p, p_i) < d(p, p_j) \forall j \in \{1, \ldots, k\}\}$. Then, let $R = \{r_1, \ldots, r_k\}$ be a set of pairwise distinct regions and consider the set $\Pi(R) = \prod_{i=1}^{k} r_i = \{\langle p_1, \ldots, p_k\rangle | p_i \in r_i\}$, i.e. the set of all sequences containing exactly one point out of each region of R. Consider, then, the set of all Voronoi diagrams generated by elements of $\Pi(R)$, i.e. $\mathcal{V}(R) = \{V(P)|P \in \Pi(R)\}$; the set of all Voronoi polygons associated with the various points in a region $r_i \in R$, i.e. $\{v(p)\}_{r_i \in R} := \{v(p)|p \in r_i \wedge v(p) \in V(P) \in \mathcal{V}(R)\}$; and the set of all restricted Voronoi polygons associated with the various points in a region $r_i \in R$, i.e. $\{\underline{v}(p)\}_{r_i \in R} := \{\underline{v}(p)|p \in r_i \wedge v(p) \in V(P) \in \mathcal{V}(R)\}$. We can then construct the *collated Voronoi diagram* generated by R, $\underline{U}(R) = \{\underline{u}(r_i)|1 \leqslant i \leqslant k\}$, where each $\underline{u}(r_i) = \bigcap\{\underline{v}(p)\}_{r_i \in R}$ is the *collated polygon associated with* r_i, i.e. the set of all points that lie in the restricted polygon of r_i in all the possible Voronoi diagrams $V(P) \in \mathcal{V}(R)$. Recovering our analogy with the

supervaluationist treatment of vagueness, the notion of the collated polygon associated with a region corresponds to the notion of super-truth in supervaluationism. We also have the *expanded* polygon associated with a region $\overline{u}(r_i) = \bigcup\{v(p)\}_{r_i \in R}$, which is the dual notion of the restricted one and thus it is analogous to the supervaluationist notion of sub-truth. Then, we can define the *boundary region associated with a collated polygon* $\underline{u}(r_i) \in \underline{U}(R)$ as the set $\overline{u}(r_i) \setminus \underline{u}(r_i)$, which is the set of all points that lie in the expanded polygon but not in the collated polygon associated with a given region.

The account of comparative concepts builds upon the vagueness framework, by adding to it an account of graded-membership in conceptual spaces. The informal idea for graded-membership in this account, which traces back to a proposal by Kamp and Partee, is that the degree to which an object falls under a concept is given by the amount of possible completions that group the object with the clear-cut instances of the concept (Kamp and Partee, 1995). Extreme cases of graded membership are, then, objects that always fall under the concept, which receive a degree of membership of 1, and objects that never fall under that concept, which get a degree of membership of 0.

In the conceptual spaces framework, elements of the set $\Pi(R)$ play the role of completions. The simplified idea (for concepts with a finite amount of prototypes) behind the membership function for a given object is to calculate the ratio between the k-tuples of $\Pi(R)$ that generates Voronoi diagrams including the object into the scope of the concept and the number of elements in $\Pi(R)$. The general idea for constructing a membership function for prototypical areas containing an infinite number of prototypical instances is to measure the set of positive completions for a given object in terms of the volume occupied by the related coordinates in the related product space. More formally, we represent each completion by means of a $m \times k$-tuple $\langle x_{1_1}, \ldots, x_{1_m}, \ldots, x_{k_1}, \ldots, x_{k_m} \rangle$ of real numbers, where $\langle x_1, \ldots, x_k \rangle \in \Pi(R)$ and $x_{i_1}, \ldots, x_{i_m}$ are the spatial coordinates of a prototypical instance p_i. Then we can build for any point a, any concept C_i with prototypical area r_i, and any distance function d the proportions of completions $S_{a,i}$ (volume of positive completions), relative to the set $\Pi(R)$:

$$\mu^*(S_{a,i}) = \frac{\mu(S_{a,i})}{\mu(\Pi(R))}$$

Where $\mu(S_{a,i})$ measures the set of positive completions, i.e. :

$$\{\langle x_{1_1}, \ldots, x_{1_m}, \ldots, x_{k_1}, \ldots, x_{k_m} \rangle | d(a, \langle x_{i_1}, \ldots, x_{i_m} \rangle) <$$

$$d(a, \langle x_{j_1}, \ldots, x_{j_m} \rangle) \forall \langle x_{j_1}, \ldots, x_{j_m} \rangle \in r_j \; s.t. \; i \neq j\}$$

and $\mu(\Pi(R))$ measures the set of all possible completions, i.e.

$$\int \mathcal{I}_{\Pi(R)}(\langle x_{1_1}, \ldots, x_{1_m}, \ldots, x_{k_1}, \ldots, x_{k_m} \rangle) dx_{1_1} \ldots dx_{k_m}.$$

We can then define the membership function of an object a relative to a concept C_i as $M_{C_I}(a) = \mu^*(S_{a,i})$. Thanks to this function we obtain a very smooth treatment of comparative concepts defining, for two individuals i, i', *i is C-er than i'* iff $M_C(i) > M_C(i')$.

Comparative concepts of different types, such as 'a is more typically C then b' or 'a is more C-ish then b', can be defined more easily in terms of the Hausdorff distance (a general way of calculating the distance between sets of points) (Decock, Dietz and Douven, 2013, pp. 76-77).

3.4.2 Explicating 'explication'

I stated at the beginning of this section that I want to give an explication of 'explication'. What does it mean, then, to explicate the concept of explication itself? It seems to me that this phrase can be understood (at least) in two different ways.

First, explicating 'explication' could consist in formally or informally giving a specific method for substituting a certain explicandum with a certain explicatum. This is the sense in which Hanna proposed his explication of 'explication' (Hanna, 1967) and Brun recently gave us a recipe for explication (Brun, 2016, 2020). Hanna's explication is a formal procedure, Brun's is stated as an informal method but both try to explicate 'explication' as a specific (formal/informal) procedure for replacing a particular reading of explication and its desiderata. There are of course further differences between the two proposals. As I said, Hanna is clearly explicating a very narrow, and very not Carnapian, sense of explication, while Brun gives a recipe for a very liberal clarification of what explication is. Nevertheless, for our current methodological discussion, they both instantiate the same sense of explicating 'explication'. Let me refer to this sense of explicating 'explication' as the *single-explicatum* sense.

Secondly, a more general sense in which the task of explicating 'explication' could be understood is as the task of providing a precise bridge-theory in which the explicandum and the explicatum could be represented, thereby allowing a (more) precise judgment of the adequacy of explication efforts and a (more) exact representation of the (different readings of the) desiderata. This is what I will try to do using the theory of conceptual spaces in this work and, to my knowledge, is the first attempt of explicating 'explication' in this specific sense[28]. This sense is more general because it does not only explicate a given clarification of a subset of explication desiderata, but it proposes instead some kind of meta-theory in which different readings of various desiderata of explication can be made precise. If the single-explicatum sense amounts to give a practical equivalent of a specific reading of explication, this more general sense of explicating 'explication' amounts to give a theory of explication. With the help of such a general explication of 'explication', external questions about explication adequacy can then be represented in a more precise manner while still remaining subject to instrumental rationality and pragmatical factors. The outcome of this sense of explicating 'explication' is a bridge theory within which (certain kinds of) different readings of explication and its desiderata can be precisely compared and applied to specific cases of conceptual engineering. Let me call this sense the *meta-theoretical* sense of explicating 'explication'.

[28]Some remarks of Kuipers hinted towards a meta-explication of 'explication' in a sense similar to what I will try to do in this section. See his discussion in (Kuipers, 2007, pp. viii-xviii).

In what follows, I will explicate 'explication' in the meta-theoretical sense. I will also show some examples of possible explications in the single-explicatum sense that can be proposed within my framework, but I will not endorse anyone of them as the favored reading of explication and its desiderata. Moreover, my theory will target only certain kinds of explications. More specifically, the applicability of my explication of 'explication' to a given case of explication rests on two pivotal assumptions:

> **Assumption 1**.
> Both the explicandum and the explicatum are representable in conceptual spaces. Moreover, if the explicandum and the explicatum are represented in two different conceptual spaces, a suitable structure-preserving mapping from the conceptual space of the explicandum to the conceptual space of the explicatum is available.

> **Assumption 2**.
> In assessing the adequacy of the given explication, all the desiderata are strictly-conceptual ones, i.e. they impose constraints only on the intrinsic relations between the explicandum and the explicatum.

The purpose of Assumption 1 is to ensure that all the concepts involved in a given explication can be adequately represented in conceptual spaces. The adequacy of conceptual spaces representation of concept formation and manipulation has been empirically tested for many types of concepts (Gärdenfors, 2000; Zenker and Gärdenfors, 2015b), but the exact scope of applicability of the theory is still unclear. It may be that very abstract concepts, such as Truth for instance, whose representational content is dubious, cannot be adequately modeled using conceptual spaces. That said, the many applications of conceptual spaces in different scientific fields arguably show that this assumption is not too restrictive. Furthermore, I will show, in the final part of this section, how my explication of 'explication' by means of conceptual spaces theory is applicable to two paradigmatic cases of explication from the history of science, adding more support to this assumption.

As for Assumption 2, conceptual spaces are a tool for conceptual representation and as such they can represent just the intrinsic relations between concepts. Thus, as I will stress case by case in the next subsection, it would be unclear at the very least how to represent in the context of conceptual spaces theory some desiderata that pose limitations on the target theory in which the explicatum is defined (such as being defined in a consistent theory, for instance) or other more pragmatical meta-theoretical virtues (such as predictive power) the scope of which is not restricted to the concepts involved in the explication. This assumption is required by the very nature of conceptual spaces theory. Nevertheless, I will provide some support to it, by showing how many different readings of explication desiderata proposed in the literature can be made precise by means of conceptual spaces theory.

In order to understand why these two assumptions are pivotal to the applicability of my meta-theoretic explication of 'explication', I will spelled out the procedure for applying

my proposal to a given case of explication. The applicability of my proposal to a given explication is then a three-step procedure:

1. **Representation**: the explicator needs to represent all the concepts involved in the explication in conceptual spaces.

2. **Choice of desiderata**: the explicator needs to choose her favorite group of explication desiderata and represent them in conceptual spaces.

3. **Adequacy assessment**: the explicator needs to check whether the (conceptual spaces representations of the) desiderata are satisfied by the (conceptual spaces representations of the) concepts involved in the explication.

Long story short, one needs to do three things. First, one needs to represent the explicandum and the explicatum in conceptual spaces. Then, one needs to choose one's favorite reading of explication desiderata. Finally, the adequacy of the explication can be mathematically assessed.

It should be clear now why the two aforementioned assumptions are needed for applying my proposal. The first step of this applicability procedure, i.e. the representation step, makes pivotal use of Assumption 1. In fact, in order to represent all the concepts involved in the explication in conceptual spaces theory, the explicatum and the explicandum have to be representable in conceptual spaces. The second step, i.e. the choice of desiderata, requires instead Assumption 2, because (as we will see in the next subsection) arguably only strictly-conceptual desiderata can be surely represented in conceptual spaces as geometrical or topological constraints on the conceptual spaces representations of the two concepts, on their conceptual space(s), and on the transition from the (representation of the) explicandum to the (representation of the) explicatum. Finally, the third step of the adequacy assessment requires Assumption 1 to make sure that, if needed by the concepts and the desiderata, a suitable mapping from the conceptual space of the explicandum and the conceptual space of the explicatum exists.

Now that it is clear what I mean with explicating 'explication' and the methodological assumptions and the applicability procedure of my meta-theoretical explication are spelled out, we can finally turn to the heart of my proposal, namely, the representation of explication desiderata in conceptual space theory.

3.4.3 Explication in conceptual spaces

In what follows, I will focus on the representation of explication desiderata in conceptual spaces. More precisely, I will show how many readings of explication desiderata that we saw in Section 2 can be made precise in terms of geometrical or topological constraints on the conceptual spaces representations of the two concepts, on their conceptual space(s), and on the transition from the (representation of the) explicandum to the (representation of the) explicatum. I will also show how the richness of conceptual spaces representation of

concepts allows us more fine-grained readings of some desiderata that a good explicatum has to satisfy.

Then, let the ExplicanDum be a given concept (in the intuitive sense of the term), represented in a conceptual space CS_{ED} by a certain concept (in the technical sense of conceptual spaces teory) $C_{ED} = \{r_{ED_1}, \ldots, r_{ED_k}\}$. Assume also that any region r_{ED_i} of the concept is obtained from a prototypical region pr_{ED_i}[29]. Similarly, let the ExplicaTum be represented in a conceptual space CS_{ET} by a certain concept $C_{ET} = \{r_{ET_1}, \ldots, r_{ET_t}\}$. Any given region r_{ET_j} of the concept is then obtained from a prototypical region pr_{ET_j}. Note that in the definitions we have not required that the explicandum and the explicatum are represented in the same conceptual space. As a matter of fact, I will argue that often it is not the case. In order to make precise many explication desiderata we will need a mapping from the elements of CS_{ED} to the elements of CS_{ET}. Assumption 1 guarantees the existence of such an adequate mapping, call it ϕ ($\phi : CS_{ED} \to CS_{ET}$)[30].

In order to structure more clearly my discussion, I will use for every reading of a desideratum the following format. First, when it is needed, I will informally discuss the desideratum and the strategy for representing it in conceptual spaces, then I will give the informal norm behind the desideratum. Then, I will draw a picture of a two-dimensional toy-case of an explicandum and/or an explicatum represented in conceptual spaces in order to show how the desideratum can be understood in conceptual spaces. Finally, as promised, I will present a formalized version of the desideratum. Consider, then, the following representation of explication desiderata in conceptual spaces theory.

Similarity

(S1) *Clear-cut extension preservation* (Hanna).

Hanna (Hanna, 1967) requires the explicatum to preserve the clear-cut extension of the explicandum. In conceptual spaces, the clear-cut extension of a possibly vague concept is given by all the collated polygons (the notion mirroring the super-truth of supervaluationism) associated with its regions. We can then see in the toy-example below (Fig. 3) how the explicatum preserves the clear-cut extension and the anti-extension of the explicandum while deciding part of its borderline region. Since I will use the same format for all the toy-pictures in this section, here is a little guide for understanding them: in a given picture, the inner-polygon represents the clear extension of a concept, the outer polygon (when is present) represents the borderline region of that concept, the rest of the space is the anti-extension of the concept; for simplicity, in these toy-cases I always assume that the explicandum and the explicatum live in the same conceptual space, namely the two-dimensional one represented by the pictures.

[29]For the sake of brevity, I am only considering here the case of categorical concepts. Nonetheless, the different desiderata can also be applied to comparative concepts as I will show in my case study on the concept of temperature.

[30]Note, thus, that the requirements that use the mapping function are not technically norms, but scheme of norms over a given mapping.

Norm: The clear-cut extension of the explicandum ought to be preserved by the explicatum.

(a) (b)

Figure 3.9: The explicatum (b) preserves the clear-cut extension (the inner-polygon) and the anti-extension (what is not in the outer polygon) of the explicandum (a), while deciding the borderline cases (what is in the outer polygon, but not in the inner one).

Formalization: For the simple case in which the set of elements of CS_{ED} is a subset of the one of CS_{ET}, the requirement is simply: $\forall a, \forall i, \exists j : a \in \underline{u}(r_{ED_i}) \rightarrow a \in \underline{u}(r_{ET_j})$ and $a \notin \overline{u}(r_{ED_i}) \rightarrow a \notin \overline{u}(r_{ET_j})$.

For the general case, in which we do not assume this relation between the base sets of the two spaces, we have to rely on the mapping ϕ: $\forall a, \forall i, \exists j : a \in \underline{u}(r_{ED_i}) \rightarrow \phi(a) \in \underline{u}(r_{ET_j})$ and $a \notin \overline{u}(r_{ED_i}) \rightarrow \phi(a) \notin \overline{u}(r_{ET_j})$.

(S2) *Favored-contexts preservation* (Quine).

Quine's (Quine, 1960, 1961) reading of similarity can be represented in the same way of Hanna's, relativizing the requirement to favored (i.e. non-deficient) contexts of the explicandum. Assume that a favor context $r_{FC_{ED_l}}$ is a subset of one of the regions belonging to the clear-cut extension of the explicandum, i.e. $FC_{ED} = \{r_{FC_{ED_1}}, \ldots, r_{FC_{ED_t}}\}$ where $\forall l \leqslant k, \exists i : r_{FC_{ED_l}} \subseteq r_{ED_i}$. We can see, then, in the picture below (Fig.4) how the explicatum preserves the favored-context and the anti-extension of the explicandum, while changing its non-favored clear-extension and the borderline cases.

Norm: The clear-cut extension of the explicandum in favored contexts ought to be preserved by the explicatum.

(a) (b)

Figure 3.10: The explicatum (b) preserves the favored contexts (FC) and the anti-extension of the explicandum (a), while changing some parts of its non-favored extension (inner polygon not in FC) and deciding its borderline region.

Formalization: For the simple case when the set of elements of CS_{ED} is a subset of the

one of CS_{ET}: $\forall a, \forall l, \exists j : a \in \underline{u}(r_{FC_{ED_l}}) \rightarrow a \in \underline{u}(r_{ET_j})$ and $\forall a, \forall i, \exists j : a \notin \overline{u}(r_{ED_i}) \rightarrow a \notin \overline{u}(r_{ET_j})$.

For the general case: $\forall a, \forall l, \exists j : a \in \underline{u}(r_{FC_{ED_l}}) \rightarrow \phi(a) \in \underline{u}(r_{ET_j})$ and $\forall a, \forall i, \exists j : a \notin \overline{u}(r_{ED_i}) \rightarrow \phi(a) \notin \overline{u}(r_{ET_j})$.

(S3) *Extension adjusting + injection* (Brun).

Brun's (Brun, 2016) two-steps reading of similarity requires first that the extension of the freely created mid-level concept (call it *explicandum*$_2$) overlaps with the extension of the original explicandum. Assuming that explicandum$_2$ is represented by the concept $C_{ED2} = \{r_{ED2_1}, \ldots, r_{ED2_j}\}$, we require the intersection of the collated polygons associated with regions of the two concepts not to be empty. Then, Brun requires an injection from the extension of this mid-level concept to the one of the explicatum. The most straightforward way of representing this step of the desideratum would be to require an injective mapping f from the set of clear-cut instances of explicandum$_2$ to the clear-cut extension of the explicatum[31]. It is easy to note that this injection requirement is rather trivially satisfied by almost every case of conceptual engineering that one can imagine[32]. A possible stronger requirement would be that every mapping from explicandum$_2$ to the explicatum be injective.

Norm: A subset of the clear-cut extension of the explicandum must be preserved by the mid-level concept. Furthermore, there ought to be an injection from the extension of this mid-level concept to the extension of the explicatum.

(a) (b)

Figure 3.11: In (a), the explicandum$_2$ (ED2) overlaps with the original explicandum (ED). In (b) there is an example of an explicatum that satisfies the injection requirement of the second step of Brun's reading of similarity.

Formalization: (Step 1) $\underline{u}(r_{ED_i}) \bigcap \underline{u}(r_{ED2_{i'}}) \neq \emptyset$.
(Step 2) There exists an injective function f_{inj} from $\underline{u}(r_{ED2_{i'}})$ to $\underline{u}(r_{ET_j})$.

[31] Assuming, of course, the existence of such a mapping. If one favors Brun's reading of similarity, one has to slightly change our Assumption 1, assuming the existence of a mapping from the conceptual space of the explicandum$_2$ to the conceptual space of the explicatum.

[32] As already mentioned in Section 2.1, Brun acknowledges this triviality, but he claims that the requirement becomes more significant if we take into consideration a system of connected notions instead of a single concept. This alternative requirement can be straightforwardly explicated in the present framework. Alternatively, Brun, in private conversation, suggested that another way of formalizing his similarity requirement could be a combination of Quine's (S2) reading of similarity and a mapping assumption similar to the one contained in my Assumption 1.

(Alternative, stronger, Step 2) All the possible functions f from $\underline{u}(r_{ED2_{i'}})$ to $\underline{u}(r_{ET_j})$ are injective.

(S4) *Contextual quasi-isometry.*

Thanks to the malleability of conceptual spaces and the adoption of a prototypical view of concepts, more fine-grained readings of the similarity desideratum are possible. I would like to propose, as an example, a reading of the similarity requirement which is philosophically particularly interesting. The informal idea behind it is that the similarity requirement is not adequately understood in terms of extension or intension, but should be modeled instead as the preservation of the large-scale conceptual structure of the explicandum. Under this reading, the explicator can thus change quite freely single instances of the explicandum, but she ought to preserve its general conceptual structure. In order to make precise this idea of large-scale structure preservation, I am going to use the concept of *quasi-isometry* (Bridson, 2008, 443-444). A function f from one metric space (M_1, d_1) to another metric space (M_2, d_2) is called a quasi-isometry, let us write f_{QI}, if there exist constants $A \geqslant 1, B \geqslant 0, C \geqslant 0$ such that:

1) $\forall x, y \in M_1 : \frac{1}{A} d_1(x, y) - B \leqslant d_2(f(x), f(y)) \leqslant A d_1(x, y) + B$

2) $\forall z \in M_2, \exists x \in M_1 : d_2(z, f(x)) \leqslant C$.

Informally, condition 1 tells us that the second metric space is allowed to distort sufficiently large distances by (at most) a constant factor, while condition 2 instead consists of a sort of 'quasi-surjection', i.e. it tells us that every element of the second metric space is close to the image of an element of the first one. We can, then, make precise this idea of large-scale preservation by imposing some contextual restrictions on the three constants used in the weak-inequalities of the quasi-isometry. We can, for instance, restrict the constants relative to the diameter $diam(X) : sup\{d(x, y) : x, y \in X\}$, i.e. the maximal distance between two elements of a metrical spaces, of the related conceptual spaces. The intuitive idea behind this restrictions is that the explicatum should not distort too much the conceptual structure of the explicandum, where too much is cashed out in terms of the diameter of the conceptual spaces where the two concepts are represented[33].

Norm: The large-scale conceptual structure of the extension of the explicandum ought to be preserved.

Formalization: There exists a quasi-isometric function f_{QI} from CS_{ED} to CS_{ET} with $A + B \leq sup\{diam(CS_{ED}), diam(CS_{ET})\}$ and $C \leq diam(CS_{ET})$.

Fruitfulness

As we have seen in Section 2.1, the fruitfulness of a given explicatum is often understood in terms of generalization power and connections with other parts of science and philosophy.

[33]Many alternative ways of making this idea of large-scale structure preservation are of course available. For instance, other intuitive ways of restraining the constants of the quasi-isometry would be to require a strict isometry for the prototypical regions or to have graded constraints for different parts of the space. Again, this desideratum, like the others, should be considered just an example of the kind of readings of explication desiderata that conceptual spaces allow.

(a) (b)

Figure 3.12: The large-scale structure of the explicandum (a) is preserved by the explicatum (b).

Under this reading fruitfulness is not a strictly-conceptual desideratum and therefore an explication of its possible readings is outside the scope of the present proposal. That said, I believe that it is possible to propose some strictly-conceptual readings of fruitfulness, looking at some characteristics of the representation of the concept by means of conceptual spaces that make an explicatum a good candidate for being a fruitful notion.

(F1) *Convexity.*

The main idea is to use Gärdenfors' (Gärdenfors, 2000) "criterion P" for natural properties as a normative and theoretical benchmark of (alleged) fruitfulness. Both from a point of cognitive fruitfulness, in the sense of Dutilh Novaes and Reck (Dutilh Novaes and Reck, 2017), and of general conceptual fruitfulness, it seems natural to take as good candidates for fruitfulness concepts the conceptual structure of which resembles the one of our natural concepts. After all, if one takes the engineering metaphor seriously, to require the explicatum to have a conceptual space similar to the ones of natural concepts is just like to require user-friendly products to engineers. We can then require the regions composing the extension of our explicatum to be convex.

Norm: The conceptual-structure of the explicatum ought to resemble the one of our natural concepts.

(a) (b)

Figure 3.13: A non-convex (a) and a convex (b) explicatum.

Formalization: $\forall x, y \in \underline{u}(r_{ET_j}), \forall z : B(x, z, y) \rightarrow z \in \underline{u}(r_{ET_j})$.

(F2) *Star-shapedness relative to a prototype region.*

As mentioned earlier in Section 4.1, there are various reasons for thinking that convexity is too strong as a criterion for natural concept. Thus, one may also want to have a weaker geometrical reading of fruitfulness. The concept of *star-shapedness relative to a point p*,

i.e. convexity relative to a given point, seems to share many attractive feature of convexity without imposing so many restrictions on the underling structure of the space (Hernández-Conde , 2017). We can then require the regions of the explicatum to be star-shaped relative to the prototypes of the concept. Since arguably the explicatum can also have boundaries which are not sharp and thus have not a unique prototype, it seems natural to define the star-shapedness requirement in relation to the set of prototypical instances of the explicatum, i.e. pr_{ET_j}.

Norm: The conceptual structure of the explicatum ought to resemble the one of our natural concepts.

(a) (b)

Figure 3.14: A non-star-shaped (a) and a star-shaped (b) explicatum

Formalization: $\forall x \in \underline{u}(r_{ET_j}), \forall y \in pr_{ET_j}, \forall z : B(x,z,y) \rightarrow z \in \underline{u}(r_{ET_j})$.

(F3) *Connectedness.*

Another, even weaker alternative to convexity that has been discussed in the debate over the right naturalness criterion in conceptual spaces is *connectedness* (Mormann, 1993). We can then use it as another possible reading of fruitfulness, by imposing it as a requirement for the regions of the explicatum.

Norm: The conceptual structure of the explicatum ought to resemble the one of our natural concepts.

(a) (b)

Figure 3.15: A non-connected (a) and a connected (b) explicatum

Formalization: $\forall j, \forall s,t : (s \cup t = \underline{u}(r_{ET_j}) \rightarrow C(s,t))$.

Exactness

(E1) *Clear extension* (Hanna).

A concept with a clear extension is a concept that does not have any boundary case, i.e. without what we have called boundary regions. Thus, we can easily make precise this

reading of the exactness desideratum by requiring the boundary region(s) of the explicatum to be empty.

Norm: The explicatum ought to have a sharp extension with no borderline cases.

(a) (b)

Figure 3.16: A non-sharp (a) and a sharp (b) explicatum

Formalization: $\forall j : \overline{u}(r_{ET_j}) \setminus \underline{u}(r_{ET_j}) = \emptyset$.

(E2) *Vagueness reduction.*

Similarly, a concept is less vague than another one when the first has fewer boundary cases than the latter. For the simple case in which the explicatum is sufficiently similar (i.e. it has the same clear-cut extension) to the explicandum, we can define this requirement in a qualitative way, requiring the boundary regions of the explicatum to be a proper subset of the ones of the explicandum. However, the explicatum, according to various liberal readings of the similarity requirement, can change even the clear-cut extension of the explicandum. Thus, in the general case, we need a quantitative way of comparing the vagueness of the two concepts. What we need is to add a proper measure to the conceptual spaces of the two concepts, thereby technically making them two measure spaces. Of course, according to the peculiarities of the given conceptual spaces, one has to choose an adequate measure. Generally speaking, assuming a non-negative measure μ on both the conceptual space of the explicandum and the one of the explicatum, we require the measure of the boundary regions of the explicandum to be strictly bigger than the one of the boundary regions of the explicatum.

Norm: The explicatum ought to be less vague than the explicandum.

(a) (b)

Figure 3.17: The explicatum (b) has a smaller boundary region than the explicandum (a) and it is therefore less vague.

Formalization: (simple case) $\forall i, \exists j : \overline{u}(r_{ED_i}) \setminus \underline{u}(r_{ED_i}) \supset \overline{u}(r_{ET_j}) \setminus \underline{u}(r_{ET_j})$.
(general case)

$\mu(\{\overline{u}(r_{ED_i}) \setminus \underline{u}(r_{ED_i}) | 1 \leqslant i \leqslant k\}) > \mu(\{\overline{u}(r_{ET_j}) \setminus \underline{u}(r_{ET_j}) | 1 \leqslant j \leqslant t\})$.

(E3) *No addition of vagueness.*

If one wants, following Brun (Brun, 2020), to read the exactness desideratum as the requirement for the explicatum to be not vaguer than the explicandum, it suffices to weaken the precedent desideratum in the obvious way.

Norm: The explicatum ought not to be vaguer than the explicandum.

(a) (b)

Figure 3.18: The boundary region of the explicatum (b) has the same size as the region of the explicandum (a), thereby making the explicatum at most as vague as the explicandum.

Formalization: (simple case) $\forall i, \exists j : \overline{u}(r_{ED_i}) \setminus \underline{u}(r_{ED_i}) \supseteq \overline{u}(r_{ET_j}) \setminus \underline{u}(r_{ET_j})$.
(general case)
$\mu(\{\overline{u}(r_{ED_i}) \setminus \underline{u}(r_{ED_i}) | 1 \leqslant i \leqslant k\}) \geqslant \mu(\{\overline{u}(r_{ET_j}) \setminus \underline{u}(r_{ET_j}) | 1 \leqslant j \leqslant t\})$.

Simplicity and other desiderata

Simplicity, like fruitfulness, seems *prima facie* a desideratum that cannot arguably be expressed in terms of intrinsic relations amongst the concepts used in the explication, i.e. not a strictly-conceptual desideratum. It could be clarified in terms of the simplicity of the syntax of the target theory in which the explicatum is defined or perhaps in terms of parsimony of new formal tools (i.e. cognitive simplicity for scientists or philosophers). Either way, these readings cannot be made precise with the help of conceptual spaces theory alone. Nevertheless, as in the case of fruitfulness, the structure of the conceptual space of the explicatum can indicate the simplicity of that concept and thus allow for a strictly-conceptual reading of this desideratum.

Furthermore, I will present two other possible desiderata that a good explicatum has to satisfy in conceptual spaces theory. Intuitively, it seems natural to require that the conceptual offspring of a good explication must tell us something more than what was contained in the original explicandum. Two ways in which this aspect of the novelty of the explicatum can be made precise are the extension or preservation of the conceptual scope and the augmentation of discrimination power.

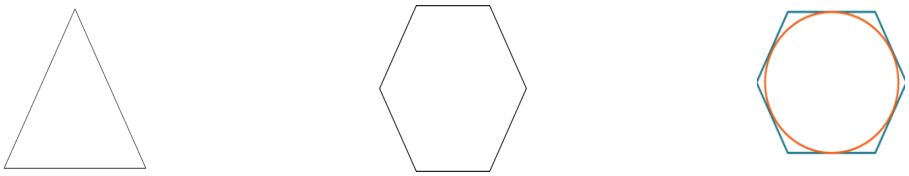
(O1) *Simplicity.*

A concept is represented in conceptual spaces theory as a set of regions and each one of these regions has a certain shape. Similarly to the case of fruitfulness, one can take the simplicity of the regions of a given explicatum as a sign for its overall conceptual simplicity

(as a kind of cognitive economy notion). A simple idea is to count the minimum number of points that are needed to draw the polyhedron π_i the surface of which is (sufficiently) close to the surface of a given region r_{ET_j}, obtaining a positive natural number σ that we can call the *simplicity coefficient* of a region[34].

We can calculate the simplicity coefficient of a given concept $C = \{r_1, \ldots, r_n\}$, by calculating the medium coefficient of its regions: $\sigma(C) = \frac{\sigma(r_1),\ldots,\sigma(r_n)}{n}$. Then, assuming that we have a set of n different explicata $\{ET, ET_1, \ldots, ET_n\}$ such that everyone of them equally satisfies the other (more important) desiderata, we can require our explicatum ET to be the one with the smallest simplicity coefficient.

Norm: Being all the other desiderata equally satisfied, the explicator ought to choose the simplest explicatum.



(a) $\sigma(ET_1) = 3$ (b) $\sigma(ET_2) = 6$ (c) $\sigma(ET_3) = 6$

Figure 3.19: Amongst these explicata, our explicator ought to choose ET_1.

Formalization: $\forall x : ET_x \in \{ET, ET_1, \ldots, ET_n\} \rightarrow \sigma(ET) \leqslant \sigma(ET_x)$.

(O2) *Scope extension.*

Menger (Menger, 1943) stressed that a good explicatum has to be applicable to new cases, thereby having a wider scope than the original explicandum. We can then make this idea precise by requiring the set of clear-cut instances of the explicatum to be strictly bigger than the one of the explicandum, using the same tools that we used for the vagueness-reduction requirement.

Norm: The scope of the explicandum ought to be extended by the explicatum.

(a) (b)

Figure 3.20: The clear-cut extension of the explicatum (b) is bigger than the one of the explicandum (a).

[34]If a given region has already the shape of a polyhedron, we can take directly its shape. If instead, the given region has curved boundaries, we can easily construct a polyhedron whose surface overlaps with the surface of the region everywhere but for a small arbitrary extent.

Formalization: (simple case) $\forall i, \exists j : \underline{u}(r_{ED_i}) \subset \underline{u}(r_{ET_j})$
(general case) $\mu(\{\underline{u}(r_{ED_i})|1 \leqslant i \leqslant k\}) < \mu(\{\underline{u}(r_{ET_j})|1 \leqslant j \leqslant t\})$

(O3) *Scope preservation.*
Just like for the vagueness-reduction case, we can also weaken the scope extension requirement in the following way.

Norm: The scope of the explicandum ought to be preserved by the explicatum.

(a) (b)

Figure 3.21: The clear-cut extension of the explicatum (b) has the same size than the one of the explicandum (a).

Formalization: (simple case) $\forall i, \exists j : \underline{u}(r_{ED_i}) \subseteq \underline{u}(r_{ET_j})$.
(general case) $\mu(\{\underline{u}(r_{ED_i})|1 \leqslant i \leqslant k\}) \leqslant \mu(\{\underline{u}(r_{ET_j})|1 \leqslant j \leqslant t\})$.

(O4) *Further discrimination power.*
Another way in which we can cash-out Menger's (Menger, 1943) idea of novelty is the augmentation of discrimination power. Our engineered conceptual tools must chart the world in a more fine-grained way than their intuitive ancestors. Conceptual spaces offer a natural way of making this idea precise in terms of the similarity function of a given metric space. Since similarity is an exponentially decaying function of distance, the augmentation of discriminatory power implies a weakening of object similarity from the explicandum to the explicatum. Relying on our mapping ϕ, we can then require the similarity between two given objects in the conceptual space of the explicandum to be bigger than the one between their images in the conceptual space of the explicatum.

Norm: The explicatum ought to have a more fine-grained conceptual structure than the explicandum.

(a) (b)

Figure 3.22: The distances in the space of the explicatum (b) are bigger than the ones in the space of the explicandum (a).

Formalization: $\forall x, y \in CS_{ED} : s^{ED}_{ab} > s^{ET}_{\phi(a)\phi(b)}$.

Single-explicatum explications and replies to recent critiques of explication

Now that we have seen multiple examples of different readings of desiderata represented by means of conceptual spaces theory, it is easy to picture various ways of adding them together, thereby creating possible explications of 'explication' in the single-explicatum sense. Generally speaking, any consistent way of mixing these (readings of) desiderata holds a formal explicatum of a particular reading of explication. The aforementioned aim of my proposal is to explicate 'explication' in the meta-theoretical sense, giving a bridge-theory that allows a more precise judgment of explication adequacy. In what follows, I will give a couple of examples of how different desiderata made precise in my framework can be put together to make specific readings of explication precise.

An example of single-explicatum explication that can be made precise within my framework is Hanna's (Hanna, 1967) explication of 'explication'. This specific reading of explication is made precise by putting together the clear-cut extension preservation reading of the similarity desideratum and the clear extension reading of the exactness desideratum:

(Hanna's explication of 'explication') [S1 + E1]:
$\forall a, \forall i, \exists j : a \in \underline{u}(r_{ED_i}) \rightarrow \phi(a) \in \underline{u}(r_{ET_j})$ and $a \notin \overline{u}(r_{ED_i}) \rightarrow \phi(a) \notin \overline{u}(r_{ET_j})$;
$\forall j : \overline{u}(r_{ET_j}) \setminus \underline{u}(r_{ET_j}) = \emptyset$.

As another example, I will add Menger's (Menger, 1943) scope preservation desideratum to some technically interesting readings of the three more important desiderata that Carnap stated, i.e. the quasi-isometry reading of similarity, the convexity reading of fruitfulness, and the vagueness-reduction reading of exactness:

(CCVS explication of 'explication') [S4 + F1 + E2 + O3]:
There exists a quasi-isometric function f_{QI} from CS_{ED} to CS_{ET} with
$A + B \leq sup\{diam(CS_{ED}), diam(CS_{ET})\}$ and $C \leq diam(CS_{ET})$;
$\forall x, y \in \underline{u}(r_{ET_j}), \forall z : B(x, z, y) \rightarrow z \in \underline{u}(r_{ET_j})$;
$\mu(\{\overline{u}(r_{ED_i}) \setminus \underline{u}(r_{ED_i}) | 1 \leqslant i \leqslant k\}) > \mu(\{\overline{u}(r_{ET_j}) \setminus \underline{u}(r_{ET_j}) | 1 \leqslant j \leqslant t\})$;
$\mu(\{\underline{u}(r_{ED_i}) | 1 \leqslant i \leqslant k\}) \leqslant \mu(\{\underline{u}(r_{ET_j}) | 1 \leqslant j \leqslant t\})$.

This explication of 'explication' in the single explicatum sense shows how using conceptual spaces theory allows us to understand explication desiderata in a more fine-grained way. All the readings of the different desiderata make use of the richness of conceptual space representation of concepts. This CCVS explication is also truly Carnapian in spirit: the similarity and exactness desiderata pose liberal but precise constraints on the large-scale conceptual structure of the explicandum and the explicatum, while the fruitfulness and the scope extension require the explicatum to show specific improvements in its extension.

The CCVS explication exemplifies how more fine-grained readings of explication desiderata are able to account for certain (alleged) problems of Carnapian explication as a general methodology for conceptual engineering. Take, for instance, the alleged inherent paradoxical tension between similarity and fruitfulness recently stressed by Dutilh Novaes and Reck, i.e. what they call the paradox of adequate formalization (Dutilh Novaes and Reck,

2017, pp. 211-213). The similarity requirement in the CCVS explication is spelled out as a large-scale constraint on the conceptual structures of the explicandum and the explicatum. Fruitfulness is instead understood as a specific constraint on the conceptual parts of the explicatum. There is no tension whatsoever between these readings of these two desiderata, namely because they have different scopes. If, in fact, the quasi-isometry between the two conceptual spaces requires the explicatum to preserve the large-scale conceptual structure of the explicandum, the convexity requirement calls for a sharpening of the extension of the explicatum. It is true that the explicator has at the same time to carefully preserve the structure of the explicandum and to craft the explicatum to be as fruitful as possible, but that does not mean that this effort is paradoxical. The key to solve this alleged paradox is then to acknowledge that explication is a very fine-grained procedure of conceptual engineering and that the similarity that explication requires between the explicandum and the explicatum is not really about single conceptual instances but it focuses instead on the more general conceptual structure of the concept. Thus, the CCVS explication shows how the geometrical representation of concepts allows us to capture both the large scale and the small-scale structure of the explicandum and the explicatum, thereby giving us the tools to overcome this apparent tension between these two desiderata.

On a more general note, the meta-theoretical explication of 'explication' here proposed can be used to defend explication against the other recent critiques that I presented in Section 2.2. By meta-theoretically explicating 'explication' we are able to make our external discussions on the adequacy of a given explicatum more precise and clear. Thanks to this meta-conceptual engineering we are thus able to have liberal readings of explication desiderata, such as the quasi-isometry reading of similarity, without giving up rigor on the pragmatic altar. This amounts to a way out for the explicator from the impasse described by Reck (Reck, 2012) of having to choose between an implausible strictly rigorous explication and a not-very-Carnapian pragmatic and liberal explication. Moreover, conceptual spaces, thanks to their malleability and their very detailed representation of concepts, seem also a promising tool for modeling any dialectical and multi-conceptual desideratum version of explication desiderata, thereby offering a solution to the limitations of the received view of explication stressed by Brun (Brun, 2020).

3.4.4 Two case studies: temperature and fish

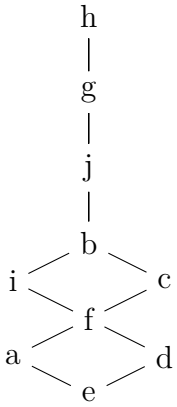
In order to make my framework clearer, I will show how two paradigmatic examples of successful explication can be represented in my framework and can be shown to satisfy the four different desiderata of the CCVS explication. For historical pleasure, I will use as case studies Carnap's examples in (Carnap, 1950b): the scientific concept of temperature and the morphological concept of fish.

Temperature

Let me start with the scientific concept of temperature, seen as an explicatum of our ordinary concepts of warm and cold[35].

As we saw in Section 4.2, the applicability of my explication of 'explication' to a given episode of conceptual change generally involves three steps: representation, choice of desiderata, and adequacy assessment. However, in the two following case studies I will kept assume the desiderata to be the ones of the CCVS explication and therefore the step of the choice of desiderata will not be represented. As such, in these case studies, the applicability of my proposal will involve two steps: first, I will represent the concepts involved in the explication in the theory of conceptual spaces and, then, I will mathematically assess that the desiderata of the CCVS explication are satisfied by these representations of the concepts.

Our first step will be then the representation in the theory of conceptual spaces of our intuitive concepts of temperature. Let us assume, following Carnap, that our intuitive way of talking about temperature uses classificatory concepts like warm and cold, together with the related intuitive comparative concepts 'warmer than' and 'colder than', intuitively understood as 'object a is warmer/colder than object b iff a is perceptually judged warmer/colder than b'. In order to build a conceptual space for these concepts, we will construct perceptive judgments out of a fictional toy-experiment. Assume that a person is asked to compare the warmth of ten buckets of water (alphabetically labeled from a to j), by dipping her arms into two buckets at a time and then judging whether the water contained in one bucket is warmer than the one in the other. Let us assume that, after many of these trials, we can organize the results in the following diagram:



We can read the diagram upwards as a partial order from the coldest sample of water e to the warmest h, with two couples of incomparable buckets (d, a) and (c, i) for which the person's intuitive judgment was not accurate enough to feel any significant difference

[35]Using the scientific concept of temperature as a case study for explication does not mean that I intend to write history nor that I am claiming that explication faithfully represent the actual thoughts and aims of the scientists in the development of this scientific concepts. I only want to stress how, suitably abstracting from history, the scientific concept of temperature can be seen as an example of a good explicatum for the related intuitive concept(s). For an historically informed take, see (Chang, 2004).

in the temperature of the water and thus to discriminate them. We can easily represent this diagram as a simple one-dimensional conceptual space $M_1 = \{E_1, \delta_1\}$ where $E_1 = \{a, b, c, d, e, f, g, h, i, j\}$ is the set of elements and δ_1 is a non-standard graph-theory-like simple metric on the diagram that counts every bottom-up step between two nodes of the graph[36]. Hence, for instance, $\delta_1(e, d) = 1$ because from node e to node d there is only one step, while instead $\delta_1(a, j) = 4$. Note that only bottom-up steps are taken into consideration and thus this metric assigns 0 to the two couples of nodes (a, d) and (c, i), thereby technically mapping them in the same way and signaling that we cannot perceptually discriminate between them. We can define our concept of 'warmer than' in terms of distance from node e: for all $x, y \in E_1$ we say that x is *warmer than* y iff $\delta_1(e, x) > \delta_1(e, y)$. Conversely, we can define x is *colder than* y iff y is warmer than x.

Assume now that the person in the experiment is asked to classify the ten buckets of water using four different categorical concepts: cold, tepid, warm, hot. Assume that the person's judgments are as follows: e is cold; f is tepid; b and j are warm; g and h are hot. Two couples of samples are not categorically judged by the person in the experiment, due to her impossibility to decide to which categorical concept they belong: a and d are both cold and tepid, c and i are both tepid and warm. Then, in the same conceptual space we can easily define also these four intuitive categorical concepts. We can use e, f, b, h as prototypical instances, respectively, of cold, tepid, warm, and hot. Our concepts are then defined in terms of distance to the related prototype, thus tessellating our conceptual space. Let us define for all $x \in E_1$ the concept *cold* $C(x)$ iff $\forall y \in \{e, f, b, h\} : \delta_1(e, x) \leqslant \delta_1(y, x)$. Hence, it follows that $C = \{e, a, d\}$. In the same way, we define the concepts *tepid* $Te(x)$, *warm* $W(x)$, and *hot* $H(x)$ relative respectively to the prototype f, b, h. Note that these definitions respect all the intuitive judgments of the person in the experiment and make the two couples a, d and c, i borderline cases of (respectively) the couples of concepts cold-tepid and tepid-warm, just as we wanted. These four categorical concepts, together with the two comparative concepts previously defined, are then our explicanda.

After we represented our explicanda, we need to represent in the theory of conceptual spaces our explicata. As our explicata, we can take the comparative concepts derived from the Celsius scale and the Kelvin scale of temperature. Following Stevens' theory of scales of measurement (Stevens, 1946), the Celsius scale is an example of an interval scale, while the Kelvin scale is a ratio scale. Interval scales are unique up to all linear transformations, a fact which makes the zero point just a matter of convention, i.e. we can add to it any constant whatsoever without changing the scale. Ratio scales, instead, are unique only up to multiplication, which implies that they have an actual zero point as the absolute zero of Kelvin scale exemplifies. Technically, in theory of measurement, the scales are defined in terms of groups of transformations of relational structures. Interval scales are isomorphic to a real structure $\langle \mathbb{R}^+, \geq \rangle$ whose automorphisms are the affine group: $x \rightarrow rx + s \; r > 0$ (Tversky et al., 1971-1989-1990, Volume 3, pp. 115-126). We can

[36]Note that here (and also in the next subsection) I am using for simplicity a discrete conceptual space and not a continuous one, like the examples in the previous section. My proposal is equally applicable to discrete and continuous spaces.

then straightforwardly represent a subset of this scale in a one-dimensional conceptual space $M_2 = \{E_2, \delta_2\}$, where $E_2 = \{o, a, b, c, d, e, f, g, h, i, j, \ldots, t\}$ is an extension of E_1 isomorphic to a subset of R^+ of 101 elements, totally ordered from o to t[37]. Assigning natural numbers from 0 to 100 to the elements respecting their total order, so that $N(o) = 0$ and $N(t) = 100$, we have the following distance function: $\delta_2(x, y) := |N(x) - N(y)|$. We can then assume that o represents the zero point of our scale. We define the temperature of a certain object as the distance between o and its corresponding point in the conceptual space: $\forall x : T(x) := \delta_2(o, x)$. Related to this conceptual space, we can also define a pair of comparative concepts $warmer^\circ$ and $colder^\circ$, defined as binary relations in terms of higher/lower temperature: $\forall x, y \in E_2$ $warmer^\circ(x, y)$ and $colder^\circ(y, x)$ iff $T(x) > T(y)$.

Then, going back to our toy-experiment, assume that we measure with a celsius-thermometer the temperature of the water contained in each bucket and that this measurement holds the following results: $T(e) = 10^\circ, T(d) = 15^\circ, T(a) = 16^\circ, T(f) = 25^\circ, T(c) = 34^\circ, T(i) = 35^\circ, T(b) = 50^\circ, T(j) = 61^\circ, T(g) = 68^\circ, T(h) = 80^\circ$. We can then represent these judgments in the conceptual space of our explicata, in terms of temperature related to our zero-point o :

$$o - T(e) - T(d) \text{ - } T(a) \text{ — } T(f) \text{ — } T(c) \text{ - } T(i) \text{ —— } T(b) \text{ —— } T(j) - T(g) \text{ — } T(h) - t$$

Of course, we can also tessellate this space with categorical concepts, thereby offering explicata for our four categorical explicanda. We can then define for all $x \in E_2$ the concepts $C^\circ(x)$ iff $T(x) < 20$, $Te^\circ(x)$ iff $20 \leqslant T(x) \leqslant 40$, $W^\circ(x)$ iff $40 < T(x) \leqslant 65$, and $H^\circ(x)$ iff $T(x) > 65$.

After having represented all the concepts involved in the explication within the theory of conceptual spaces, we can now assess the adequacy of our explicata, showing how the different desiderata of the CCVS reading of explication are satisfied[38]. Remember that we have four desiderata that our explication has to satisfy, namely quasi-isometry, convexity, vagueness-reduction, and scope preservation.

In order to fulfill the quasi-isometry requirement we need a function $f : (E_1, \delta_1) \to (E_2, \delta_2)$, for which there exist constants $A + B \leq sup\{diam(E_1), diam(E_2)\}$ and $C \leq diam(E_2)$ such that:

1) $\forall x, y \in E_1 : \frac{1}{A}\delta_1(x, y) - B \leqslant \delta_2(f(x), f(y)) \leqslant A\delta_1(x, y) + B$;

2) $\forall z \in E_2, \exists x \in E_1 : \delta_2(z, f(x)) \leqslant C$.

The diameter of the conceptual space of our explicanda is $\delta_1(e, h) = 7$, while the space of our explicata is $\delta_2(o, t) = 100$. We can then choose as our f the function that maps elements of the first conceptual space to the elements in the second conceptual spaces representing

[37]Note that here, instead of mapping directly the objects to an interval of reals like is customary in measure theory, I am using this 101 elements isomorphic to a subset of an interval as another layer of representation. This is only done for simplicity sake, in order to have a base set that extends the one of the first conceptual space, and it will pay off in the possibility of having qualitative simpler version of (some) CCVS desiderata but it is by no means necessary for my proposal.

[38]Again, note in this case study the application of my proposal has only two steps (instead of the three that are generally involved), due to the default choice of the desiderata of the CCVS explication.

the temperature of the related bucket: $f(x) = T(x)$. The first weak-inequality is always satisfied by choosing, for instance, $A = 10$ and $B = 20$: $\forall x, y \in E_1 : \frac{1}{10}\delta_1(x, y) - 20 \leqslant \delta_2(f(x), f(y)) \leqslant 10\delta_1(x, y) + 20$. For satisfying the second weak-inequality, we have to put $C = 20$ (considering that the maximal distance between an element of the second space and an image of an element of the first one is the one between t and h, which is equal to 20).

For the convexity desideratum note that E_2 is isomorphic to a subset of reals and that we have defined $T(x)$ in terms of the order relation between elements of this subset. Thus, we have $\forall x, y, z \in E_2$: if $B(x, z, y)$ then $T(x) < T(z) < T(y)$ or $T(y) < T(z) < T(x)$. We can then easily see that all the explicata are represented by convex regions in this conceptual space. For instance, our explicatum of cold, $C^\circ(x) := T(x) < 20$, is represented by the region $C^\circ = \{o, \ldots, e, \ldots, d, a, \ldots\}$, the elements of which are the first 20 elements of our base set, ordered in terms of distance and therefore of temperature. We then have $\forall x, y \in C^\circ, \forall z \in E_2 : B(x, z, y) \rightarrow z \in C^\circ$, as requested. The same holds for the other explicata, as it is shown by picture (b).

(a) (b)

Figure 3.23: Representation of the categorical explicanda (a) and the related categorical explicata (b).

For the last two requirements, we have to split the discussion between categorical and comparative explicata. For our two comparative explicata, we have to look at the two product spaces $E_1 \times E_1$ and $E_2 \times E_2$ in which the comparative explicanda and explicata are defined. For the vagueness-reduction requirement, in the first space we have two pairs of elements for which our explicanda 'warmer/colder than' are not defined: $\overline{u}(r_{ED}) \setminus \underline{u}(r_{ED}) = \{(a, d), (c, i)\}$, because $\delta_1(e, a) = \delta_1(e, d)$ and $\delta_1(e, c) = \delta_1(e, i)$. Our explicata are instead sharp, so that we have no equivalent case of vagueness and thus the boundary regions consists only of the empty set: $\overline{u}(r_{ET}) \setminus \underline{u}(r_{ET}) = \{\emptyset\}$. Then, the requirement is trivially satisfied by noting that the empty set is a subset of the boundary region of the first metric space.

As for the scope-preservation requirement, we just notice that all positive instances of our explicanda, $\underline{u}(r_{ED})$, are all positive instances of our explicata, $\underline{u}(r_{ET})$, so that we have (for the 'warmer than' case): $\underline{u}(r_{ED}) = \{(h, g), (h, j), \ldots, (g, j), (g, b), \ldots\} \subset$

$\underline{u}(r_{ET}) = \{(h,g),(h,j),\ldots,(g,j),(g,b),\ldots\}$. For the case of categorical concepts, these two requirements are similarly satisfied. Take for instance our explicandum cold, $C(x)$: its clear-cut extension is $\{e\}$ and its borderline region is $\{a,d\}$. The extension and the borderline region of the related explicatum cold$^{\circ}$ are, respectively, $\{o,\ldots,e,\ldots,d,a,\ldots\}$ and $\{\emptyset\}$. We can then notice that the extension of our explicandum is preserved by our explicatum and that the borderline region of our explicatum, i.e. the empty set, is trivially a subset of the borderline region of our explicandum.

We have, thus, seen how my framework can be applied to a given case of explication. More specifically, we saw how the CCVS explication of explication holds the scientific concepts of temperature as satisfactory explicata for our intuitive categorical and comparative notions of temperature. Conceptual spaces, integrating crucial aspects of measure-theory in their representation of concepts (Zenker, 2014, p. 8), make transparent the conceptual advantages of the scientific concept of temperature in respect to its intuitive counterpart. The scientific categorical concepts, in fact, extend the scope of our intuitive way of talking beyond any everyday possible experience, while simultaneously allowing us more fine-grained, quantitative discriminations. A philosophically significant consequence of this improved power of discrimination is that our scientific notion of temperature-indiscriminability is transitive, in contrast to the arguable non-transitivity of our phenomenal one[39]. Furthermore, the scientific concept makes it possible to define on the temperature scale sharp categorical concepts that make our communication effective and precise.

The same desiderata are equally satisfied by the pair of explicata *warmer** and *colder** (and related categorical concepts), defined in relation to the Kelvin scale in the same way of our six explicata. In addition to them, though, the Kelvin scale allows a further discrimination power and, being a ratio scale, it has the technical advantage of having a smaller class of equivalence and thus to be empirically more testable (Gärdenfors and Zenker, 2013, pp. 1049-1050).

Fish

Let us turn our attention to another example that Carnap gave, namely the scientific concept of fish seen as an explicatum of our intuitive conception of what a fish is[40]. As our explicandum we can take the intuitive concept of fish, understood as "an animal that lives in the water"Às our explicatum we take instead the scientific concept of fish, which we will call (following Carnap) *piscis*, understood as "an aquatic vertebrate with gills and with limbs in the shape of fins" (Helfman et al., 2009, p. 3).

Just like in the previous case study, the application of my explication of 'explication' to the fish-piscis case will involve two steps (instead of the canonical three, cf. Section 4.2), due to the aforementioned default choice of CCVS explication desiderata: first, we will

[39]The non-transitivity of our phenomenal notion of indiscriminability is the center of (Williamson, 1990). For a more recent defense of this position, see (De Clercq and Horsten, 2004).

[40]Again, the same historical disclaimer of the temperature example applies here as well. In what follows, I do not want to write ichthyology or history of biological taxonomy. For a recent complete account of the biological understanding of fishes, see (Helfman et al., 2009).

represent all the concepts involved in the explication in the theory of conceptual spaces and then we will mathematically assess whether these representations satisfy the CCVS desiderata.

In order to build a conceptual space for these concepts, let us construct a fictional toy-example. Assume that a person who knows nothing about biology is asked to classify the animals of a (very) small zoo in order to decide whether a given animal is a fish or not. Assume that the zoo contains the following nine animals: a tuna, a whale, a shark, a mudskipper, a fire-salamander, a crocodile, a zebra, a lion, and a hippo. The person is then instructed to observe the animals for a certain period of time, after which she has to decide whether a given animal is a fish, according to our intuitively ecological explicandum. Thus, the person would look at which animals in the zoo live in the water. Assume that, on a scale from 0 (never in the water) to 10 (always in the water), the results are the following: 0 (zebra, lion), 2 (hippo), 4 (crocodile, fire-salamander), 5 (mudskipper), 9 (whale), 10 (tuna, shark):

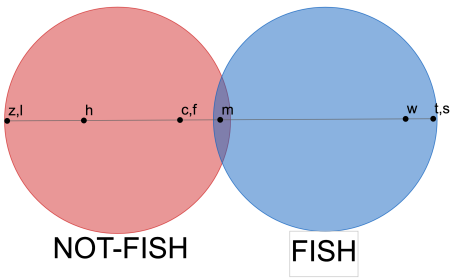


Figure 3.24: The conceptual space of the explicandum *fish*.

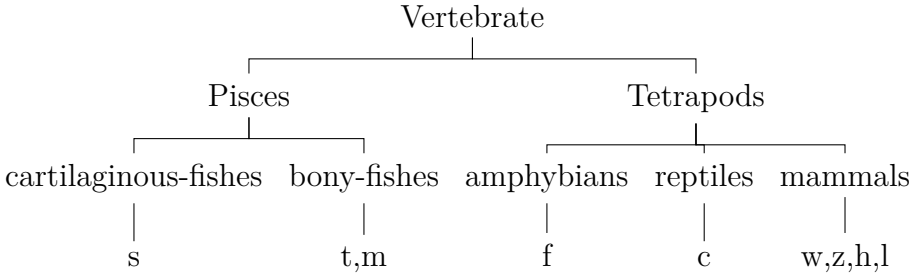
We can then easily represent this diagram in a one-dimensional conceptual space $M_1 = \{E_1, \delta_1\}$ where $E_1 = \{z, l, h, c, f, m, w, t, s\}$ and $\delta_1(x, y) = |R(x) - R(y)|$, $R(x)$ being a positive integer from 0 to 10, i.e. the result of the experiment. Thus, for instance, we have $\delta_1(z, c) = 4$ and $\delta_1(w, m) = 4$. We can then suppose to define the concepts of fish and not-fish using as prototypes t (or s) and z (or l), tessellating therefore the conceptual space in two parts. We then define for all $x \in E_1$ the concept *fish* $F(x)$ iff $\delta_1(t, x) \leqslant \delta_1(z, x)$ and the concept *not-fish* $NF(x)$ iff $\delta_1(z, x) \leqslant \delta_1(t, x)$. Hence, we have $F = \{w, t, s, m\}$ (whale, tuna, shark, mudskipper) , $NF = \{z, l, h, c, f, m\}$ (zebra, lion, hippo, crocodile, fire-salamander, mudskipper). The mudskipper (m) is then a borderline case[41]. We thus have a conceptual space representation of our explicandum *fish*.

We turn now to our explicatum *piscis*. Returning to our toy-experiment, assume now that the same person is asked to classify the animals in the small zoo according to their morphology. After having collected morphological data, the person has to classify again the

[41]The mudskipper presents itself naturally as a borderline case of the intuitive concept of fish due to its ability of surviving out of the water for short periods of time. Fishes like the mudskipper are popularly known as 'walking fishes'. See (Helfman et al., 2009, pp. 60-65).

animals according to our explicatum. Remember that, in order to qualify as a piscis an animals has to be a "an aquatic vertebrate with gills and with limbs in the shape of fins"Then, two major taxonomic changes naturally present themselves in our toy-experiment: the whale and the mudskipper. The whale, who was a clear-cut case of a fish (living exclusively in the water), is not a piscis but instead it is classified as a mammal. Looking closely at the internal and external morphology of the whale, the person in our experiment realizes that it is completely different from the one exemplified by a paradigmatic fish. Whales, in fact, do not have gills, they have lungs, they reproduce like mammals, and so on. The mudskipper, who instead was a borderline case of fish (due to its ability of spending short periods of time outside the water), it qualifies as a clear instance of piscis, due to its internal morphology. Mudskippers have in fact gills and fins, just like other non-amphibious fishes.

Assume then, that the person in the experiment is asked to judge whether the animals in the zoo are pisces or tetrapods, judging by their morphology. The results of this new classification are the following: Pisces (shark, tuna, mudskipper), Tetrapods (fire-salamander, crocodile, whale, zebra, hippo, lion). We can add another sub-level of classification distinguishing pisces between cartilaginous-fishes (shark) and bony-fishes (tuna,mudskipper) and tetrapods between amphibians (fire-salamander), reptiles (crocodile), and mammals (whale, zebra, hippo, lion):



We can, then, easily read this tree as a conceptual space $M_2 = \{E_2, \delta_2\}$, where $E_2 = E_1 = \{z, l, h, c, f, m, w, t, s\}$ and δ_2 is a simple metric that counts the number of nodes of the tree between one element and the other. Thus, for instance, $\delta_2(s,t) = 1$, $\delta_2(t,m) = 0$, $\delta_2(t,f) = 3$, $\delta_2(m,w) = 3$. Using as our prototype of a piscis s (or t or m) and as our prototype of a tetrapod z (or any other tetrapod, for what it matters) we can define for all $x \in E_2$ our explicatum *piscis* $P(x)$ iff $\delta_2(s,x) < \delta_1(z,x)$ and the concept of *tetrapod* $T(x)$ iff $\delta_2(z,x) < \delta_1(t,x)$.

We can now, as we did in the temperature example, assess the adequacy of the CCVS explication of 'explication' by checking whether the conceptual space representation of this example of explication satisfy all its requirements. The first requirement is the existence of a quasi-isometry between the two metric spaces M_1 and M_2: $f_{QI} : (E_1, \delta_1) \to (E_2, \delta_2)$ with $A + B \leq sup\{diam(CS_{ED}), diam(CS_{ET})\}$ and $C \leq diam(CS_{ET})$. The diameter of the first conceptual space is $\delta_1(z,s) = 10$ and the one of the second space is $\delta_2(s,l) = 3$. Considering that the two base sets contain the same elements, the simple trivial mapping

$f(x) = x$ would do the trick. For instance, fixing the constants $A = 2$, $B \geqslant 5$ (in order to account for the drastic change of the classification of whales), $C = 0$, the following weak-inequalities always hold:

1) $\forall x, y \in E_1 : \frac{1}{2}\delta_1(x, y) - 5 \leqslant \delta_2(f(x), f(y)) \leqslant 2\delta_1(x, y) + 5$;
2) $\forall z \in E_2, \exists x \in E_1 : \delta_2(z, f(x)) \leqslant 0$.

The second requirement is instead the convexity of the region representing our explicatum *piscis*: $\forall x, y \in P : B(x, z, y) \rightarrow z \in P$. It is easy to see that the region $P = \{s, t, m\}$ is convex, being it a self-contained part of our taxonomic tree. Our third requirement consists of a reduction of vagueness from the explicandum to the explicatum. In the conceptual space of our explicandum we have only one borderline case: m (mudskipper). In the second conceptual space, our concept *piscis* has instead sharper boundaries and the set of borderline case is empty. Thus, also the vagueness reduction is satisfied. Finally, our last requirement consists in the explicatum preserving the scope of the explicandum. Being our base sets finite, comparing the number of elements that clearly fall under the two concepts will suffice. We can then see how both the clear-cut extension of the explicandum *fish* $F = \{w, t, s\}$ and the extension of the explicatum *piscis* $P = \{s, t, m\}$ have three elements, satisfying the scope-preservation requirement.

As in the first case study, conceptual spaces, in virtue of the measure-theoretic consideration that they incorporate, allows us to see the conceptual improvements of our explicatum in comparison to the intuitive explicandum. The concept *piscis* allows us to make sharper taxonomic distinctions, offering more objectivity of judgment in comparison to an intuitively ecological concept like *fish*. Not only the concept *piscis* allows us a more fruitful reclassification of many instances of the concept *fish*, such as whales, it also permits more fine-grained discriminations, such as the one between different kinds of pisces. Then, of course, the concept *piscis* has many other non-strictly-conceptual advantages, such as allowing us more generalizations (thanks to its morphological criteria), extending the scope of the concept to non-observable animals (e.g. fossils, etc.), and so on.

Thus, we have seen how the morphological concept of piscis, despite its re-classification of clear-cut cases of the intuitive concept of fish, satisfy all the desiderata of the CCVS explication. Together with the previous example of the scientific concept of temperature, this example shows how we can apply my meta-theoretical explication of 'explication' to paradigmatic examples of explications from the history of science. In both case studies, we saw how the applicability of my proposal is made of three steps: the representation, the choice of desiderata, and the adequacy assessment.

More generally, we saw how my proposal allows a more precise understanding of the subtle differences between different readings of explication desiderata, thereby contributing to dissolve a lot of vagueness and ambiguity often contained in philosophical discussions about explication. Moreover, thanks to the richness of the conceptual spaces representation of concepts, it is possible to define more fine-grained desiderata that arguably allow explication to overcome some of its alleged limitations, such as the so-called 'paradox of adequate formalization'.

After all these steps, one could perhaps ask how satisfactory is this meta-theoretical

explication of 'explication'. We can assess this explication by judging how much it satisfies Carnap's three main desiderata for an explication: similarity, fruitfulness, and exactness. The two case studies contained in this section show how paradigmatic cases of explications are also perfectly satisfactory explications according to the present proposal. It seems, then, that this meta-theoretical explication of 'explication' is sufficiently similar to Carnap's original ideas. My proposal seems also quite fruitful. In fact, we have seen at the end of Section 4.3 how conceptual spaces allow us to have more fine-grained readings of explication desiderata that help us to overcome recent critiques of explication. Moreover, having developed this explication of 'explication' within the framework of conceptual spaces, one can arguably expect many fruitful interactions between the present proposals and the various applications of conceptual spaces in science and philosophy. Finally, the exactness of the present proposal is evident in the way in which different strictly-conceptual readings of explication desiderata discussed in the literature can be made (more) precise as geometrical/topological constraints over the conceptual spaces of the explicandum and the explicatum.

3.5 Assessing Carnapian Explication in the Toolbox Framework

In this final section, I will analyze how Carnapian explication can be classified within the Toolbox framework, i.e. the meta-framework for assessing models of conceptual change that I presented in Chapter 2. More specifically, we will see how the features of Carnapian explication can be assessed along the nine evaluative dimensions of the Toolbox framework: units of selection, concept ontology, concept structure, kinds and degrees of conceptual change, degree of normativity, effectiveness of normative judgment, assumptions and consequences for conceptual change in science, assumptions and consequence for conceptual change in philosophy, metaphilosophical assumptions and implications. Let us survey how Carnapian explication performs in these dimensions, one by one, then.

Units of selection This dimension judges models of conceptual change according to the level of abstraction at which they identify conceptual entities as meaningful units of change. In the case of Carnapian explication, the unit of conceptual change is explicitly set at the level of the single concept or, more accurately, at the level of the diachronic couple of concepts composed by an explicandum and an explicatum. In contrast to Carnap's previous methodologies, such as rational reconstruction and translation to the formal mode of speech, that involved different units of change such as concepts, theories, and statement, explication is explicitly a procedure involving the transformation of a single concept, the explicandum. That said, it seems possible (and actually, according to some scholars, it would be better cf. Brun 2020) to explicate also larger units, such as pairs or sets of concepts. Evidence for the possibility of this extensions can be found in my case study on the concepts of temperature (cf. Sect. 3.4).

Concept ontology This dimension focuses on the compatibility of a given model of conceptual change with the different philosophical positions on the ontology of concepts. In the case of Carnapian explication, consistent with the ideal of pluralism inherent in the methodology, the procedure is not tied to any particular ontological view about concepts. That said, it seems clear that the procedure of explication, with its heavy focus on bridging different linguistic frameworks, goes particularly well with a linguistic view of conceptual entities, a view that is moreover consistent with some remarks of Carnap about concepts (e.g. Carnap 1928a, 1947). Carnapian explication can anyway be implemented together with all the other three main views about concepts, understanding its linguistic focus in a deflationary way as working not directly on conceptual entities, but on the related linguistic predicates.

Concept structure This dimension focuses instead on how a given model of conceptual change assumes the structure of concepts to be constituted. As I stressed in Section 1 of this chapter, Carnapian explication puts a strong focus on the functions and the roles of concepts. As such, the two theories of conceptual structures that seem most compatible with it are functional and inferentialist theories, two theories that identify conceptual structure with (respectively) a concept function and role. As a matter of fact, it can be argued that both theories are traceable in Carnap's writings (Creath, 1994; Peregrin, 2020). Nevertheless, just like for concept ontology, the inherently pluralist spirit of Carnapian explication lends itself to an analogous deflationary reading of its focus on the functions and roles of concepts, that can be seen just as a way of identifying concepts and not necessarily as involving a statement on conceptual structure. This deflationary reading allows Carnapian explication to be implemented together with all other major views of conceptual structure.

Kinds and Degrees of conceptual change This dimension focuses on the kinds and degrees of conceptual change that a given model of conceptual change identifies. Carnap does not explicitly identify multiple kinds or degrees of explications. It is clear by his examples, though, that significant explications are always trans-frameworks mappings where the explicandum and the explicatum belong to different conceptual or linguistic frameworks. This makes Carnapian explication a model of conceptual change designed to target radical episodes of conceptual change of the kind that, in philosophy of science, are often conceptualized as taxonomic incommensurable ones. These changes are the philosophically and scientifically significant one for Carnap and they are thus the ones for which the procedure of explication can offer a suitable methodology.

Degree of normativity This dimension tracks the extent to which a given model of conceptual change is more or less normative in judging episodes of conceptual change. Carnapian explication is indeed a normative model of conceptual change, although its normative assessment of a conceptual history is never an absolute one. Episodes of conceptual change in science and in philosophy can in fact be judged, qua examples of explications,

as more or less satisfactory according to the specific context and goal under focus. This judgment is never the only possible reconstruction, but it does provide, relative to a shared analysis of the context and goal of the explication, a judgment of the rationality of the conceptual change.

Effectiveness of normative judgment This dimension focuses on how effective the normative judgment of a model of conceptual change is. As I stressed throughout all this chapter, a distinctive feature of Carnapian explication is its inherent pluralism. This pluralism is exemplified by the explicit ban of any absolutely correct answer in matters of explication adequacy. The normative judgments of Carnapian explication with respect to the rationality of a given episode of conceptual change is then what Carnap calls an external question, i.e. a pragmatic matter that crucially involves the domain of values and as such it is subject to instrumental rationality.

Assumptions/consequences for conceptual change in science This dimension focuses on the assumptions and the consequences of a given model of conceptual change in relation to the problems that scientific conceptual change poses in philosophy of science. Carnapian explication is evidently a pragmatic model of conceptual change, understanding the continuity between radically different concepts as necessarily involving the domain of values (cf. Carus 2017). If, in fact, Carnap's earlier methodologies appear to defend a syntactical view of conceptual change in science (cf. the aforementioned program of translation in the formal mode of speech), the view of scientific change inherent in the methodology of explication is in fact a fully pragmatic one. From the perspective of explication, scientists engineer our conceptual tools with the goal of having better tools for achieving their research goals and aims (cf. Kitcher 2008; Justus 2012). Scientific progress and scientific objectivity are then, according to the ideal of explication, dependent on the value and goals of the scientific communities. The only absolute value that Carnap repeatedly stresses in his presentation of explication is the clarity of methods and goals, a traditional part of modern science ethos. Explication is also compatible with both realist and anti-realist positions on scientific theories ontological import, depending on how the values and the goal of science are formulated. In general, explication gives a very viable way of defending scientific rationality against incommensurability accusations, albeit one that inevitably involves the value domain and therefore it is incompatible with the old-fashioned value-free ideal of scientific objectivity.

Assumptions/consequences for conceptual change in philosophy This dimension focuses on the assumptions and the consequences of a given model of conceptual change in relation to the problems that philosophical conceptual change poses in philosophy. The picture of philosophical conceptual change that Carnapian explication gives is the one of an activity absolutely crucial for philosophy *tout court*. Philosophical activity, in fact, according to the ideal of explication ought to become an engineering-like activity centered around the development and the critical assessment of new conceptual and linguistic tools.

In this respect, the feedback-relation between evolved and constructed languages around which explication is centered makes the significance and the assessment of a philosophical proposal completely dependent on its predecessors and its alternatives, in stark contrast to traditional methodologies of philosophical analysis (cf. Carus 2012b; Richardson 2012). Explication puts then philosophical conceptual change at the center of philosophical and metaphilosophical activity.

Metaphilosophical assumptions and implications This dimension focuses on the metaphilosophical background that a given model of conceptual change has. As we saw in Section 1 of this Chapter, Carnapian explication fully embodies Carnap's metaphilosophical ideals of constructivism, positivism, logicism, structuralism, and pluralism. As such, Carnapian explication comes equipped with a theory-laden metaphilosophical baggage that we can identify with the ideal behind the methodology of explication. This ideal is nothing less than a radical plan for reforming philosophical activity. According to the ideal of explication, in fact, philosophy ought to become an engineering-like activity centered around the construction and the critical assessment of conceptual tools. As such, as it was stressed by many authors, Carnapian explication is then a paradigmatic, perhaps the most paradigmatic, form of conceptual engineering. Moreover, Carnapian explication presents itself also as one of the best equipped and well-studied methods of conceptual engineering. As we saw in Section 2.2 of Chapter 2, in fact, many alleged conceptual engineering lack a clear positive conception of how their method could be implemented, as well as evidence that such implementation is possible in the case of central philosophical and scientific concepts. As we saw in this chapter, instead, the philosophical and metaphilosophical background of Carnapian explication has been extensively studied for years and there is significant evidence of its applicability to central philosophical and scientific concepts.

Chapter 4

Models of Conceptual Evolution

The focus of this chapter will be models of conceptual evolution (or, alternatively, evolutionary models of conceptual change), i.e. models of conceptual change that understand this phenomenon as a kind of evolution akin to the one that biological entities undergo. More specifically, I will focus on evolutionary models of conceptual change of a Darwinian kind, i.e. evolutionary models centered around a selection mechanism analogous to natural selection.

My analysis of models of conceptual evolution will also involve discussing the ideal of an evolutionary epistemology, i.e. a naturalistic approach to epistemology in which evolutionary considerations take central role. As we will see, in fact, understanding scientific conceptual change as an evolutionary process has often been a central step of a more general philosophical program aimed at an overarching evolutionary model of human knowledge and its many products. I will thus critically present the general program of an evolutionary epistemology with a specific focus on its implications for evolutionary models of conceptual change. As a result this discussion, I will argue that models of conceptual evolution, in order to be more applicable to specific case studies and thus more easily historically testable, need to be narrower and more specifically tailored to a given scientific disciplined. As a first step in this direction, I will present a novel model of conceptual selection for mathematical concepts, specifically tailored to the specific kind of evolution that mathematical concepts and mathematical problems exhibit. Finally, I will analyze evolutionary models of conceptual change, such as my model of conceptual selection, within the meta-framework of the Toolbox framework.

In Section 1, I will generally present the program of evolutionary epistemology, focusing specifically on the part of it that aims modeling the evolution of scientific theories. In Section 1.1, I will analyze this part by presenting the related work of four leading advocates of evolutionary epistemology, i.e. Donald Campbell, Karl Popper, Stephen Toulmin, and David Hull. In Section 1.2, I will present some common critiques to such evolutionary models of scientific change, together with some possible answers to them and a general assessment of the discussion on evolutionary epistemology. In Section 2, I will present a novel selection framework for mathematical concepts, inspired by the recent population-based Darwinian framework of Godfrey-Smith, centered around the notions of conceptual

population and mathematical selection. In Section 3, I will show how my framework can be fruitfully applied to assess the rationality of actual episodes of mathematical conceptual change with the aid of three different case studies from the history of mathematics. Finally, in Section 4, I will assess the specific features of Darwinian models of conceptual change such as the one I presented in this chapter using the nine dimensions of my Toolbox framework.

4.1 Evolutionary Epistemology

As I mentioned above, we can broadly consider an evolutionary epistemology any epistemological approach in which evolutionary considerations take central stage. As it was stressed by several scholars (e.g. Campbell 1974a; Bradie 1986, 1994), evolutionary considerations have played an important role in many modern epistemological works. In this broad sense, then, evolutionary epistemology is hardly a recent phenomenon in the history of philosophy. However, the concept of evolutionary epistemology is in contemporary philosophy associated with a specific sub-kind of evolutionary epistemologies that have a Darwinian character. After the scientific success of (Neo-)Darwinism, in fact, virtually all evolutionary approaches to epistemology seek to apply a kind of natural selection mechanism to the realm of epistemological phenomena[1]. This Darwinian kind of evolutionary epistemology will be the focus of this section.

Evolutionary epistemology, in the narrower Darwinian sense, can then be understood as the cluster of epistemological approaches centered around an adaptationist mechanism somewhat analogous to natural selection. In this form, evolutionary epistemology started to be explicitly conceptualized and recognized as an epistemological movement in the second half of the last century, mostly in connection to the philosophical efforts of Campbell and Popper.

A standard distinction in the related literature (Bradie, 1986) divides evolutionary epistemology in two interrelated but distinct subprograms: the evolution of cognitive mechanisms program (EEM) and the evolution of theories program (EET). The EEM program attempts to explain the development of cognitive mechanisms in humans and animals by extending the scope of evolutionary theory to the related biological substrates of cognition. The EET program, instead, seeks to analyze the growth of knowledge (and in particular the growth of science) by using evolutionary models and analogies drawn from evolutionary biology. If the explanandum of the EEM program are the mechanisms producing knowledge, the EET program targets instead the products of human knowledge. These two programs and their alleged evolutionary explanations are thus logically distinct and as such their value can be assessed independently of one another. If, in fact, many paradigmatic examples of evolutionary epistemology involve both the EEM and the EET program, coupling an evolutionary account of our cognitive apparatus with a selection theory for epistemogical

[1]It should be noted that there are few evolutionary epistemologists that champion a non-adaptationist (and thus not Darwinian) kind of evolutionary approach to intellectual phenomena, e.g. (Ruse, 1986; Rescher, 1990).

phenomena in a hierarchy of selective mechanisms, many critics of evolutionary epistemology have specifically targeted one of the two programs (e.g. Thagard 1980; Fracchia and Lewontin 1999; Renzi and Napolitano 2011).

For the aims of this work, the subprogram of evolutionary epistemology that is most relevant is the EET program, since it explicitly targets the phenomenon of scientific change. Moreover, as we will see, many evolutionary accounts of scientific growth take as their units of selections (sets of) scientific concepts and are therefore classifiable as models of conceptual change.

In the next subsection, I will briefly present the history of the EET program by focusing on four leading examples of it. After this presentation of the EET program, I will, in Section 1.2, present some critiques to the viability of the EET program, critically assessing the related discussion and its implications for evolutionary models of conceptual change.

4.1.1 Evolutionary models of scientific change

In order to present the EET program, I will focus on four paradigmatic examples of it. Specifically, I will briefly present the evolutionary models of scientific change developed (respectively) by Donald Campbell, Karl Popper, Stephen Toulmin, and David Hull. By analyzing these four examples, we will see how evolutionary epistemology and in particular a Darwinian-like selection theory for scientific theories can be spelled out in different ways, while exhibiting a steady core of (meta)philosophical assumptions and ideals. Let us survey these four takes on the evolution of science one by one, then.

Donald Campbell. If asked to give an example of a philosopher engaged in evolutionary epistemology, most people would cite Donald Campbell. The whole of Campbell's psychological and philosophical work can be in fact seen in the light of evolutionary epistemology aims and ideals (cf. Wuketits 2001). Moreover, Campbell's several programmatic papers on evolutionary epistemology (Campbell, 1960, 1974a,b, 1987, 1988, 1997) were historically pivotal to the establishment of evolutionary epistemology as a philosophical and scientific research program. Campbell's work fixed the terminology (even the term 'evolutionary epistemology' itself) and the agenda of evolutionary epistemology, tracing its history and its development by means of an enormous bibliographical study of evolutionarily epistemological efforts in science and philosophy.

Campbell can be seen as a paradigmatic supporter of both the EEM and the EET programs in evolutionary epistemology. He saw in fact the two subprograms as two interconnected parts in his systematic efforts towards establishing an evolutionary science of human knowledge and its scientific products. The center of such an evolutionary science is, according to Campbell, the pervasive mechanism of blind variation and selective survival (cf. Campbell 1960, 1974b). This mechanism, paradigmatically exemplified by the development of biological entities by means of natural selection, is for Campbell at work behind all animal and human knowledge processes. Campbell's appreciation of this mechanism comes from his work in the psychology of perception and vision, where he

gave a blind-variation-plus-selective-survival analysis of several elements of human cognition (Campbell, 1956, 1959, 1966). In its general form, the blind variation and selective survival mechanism is made of three structural features (Campbell, 1959, p. 163): a mechanism producing variation, a selection process according to which certain variations are preserved and others are lost, and a mechanism for maintaining and propagating the surviving variations. These three features (corresponding to the classical Darwinian principles of variation, selection, and inheritance; cf. Lewontin 1970) are then instantiated by different biological and cognitive processes in a way specific to the knowledge process at issue. Campbell (Campbell, 1974a, pp. 422-437) gives us a hierarchy of 10 knowledge processes governed by the mechanism of blind variation and selective survival. He ordered these 10 processes by their level of complexity and development, from the lowest level corresponding to blind trail-and-error problem solving, common even in the least complex members of the animal kingdom, up to the highest level corresponding to the social decision making processes typical of human science.

Science is then for Campbell the most complex and developed form of knowledge. Nevertheless, the mechanism of blind variation and selective survival is still at the heart of how science develops (cf. Campbell 1974b, 1988, 1997). At this level, the variation is not determined by the appearance of biological phenomena, but it is externalized in the appearance of scientific theories and hypotheses. Campbell (Campbell, 1960, p. 384) stresses that at the level of science also the other two features of selection and inheritance are somehow "substituted" by the testing and the propagation of scientific products within scientific communities. Despite this externalization of the mechanism of variation and survival, Campbell defends the clear analogy between biological and scientific development, seeing in the growth and progress of our scientific theories just another example of the progressive adaptation made possible by blind variation and selective survival. So that, Kant's problem of the fact of science and the related wonder of how scientific theories fit the world are for Campbell the same issues that we face in explaining crystal formation[2]. Apparent miracles of fit between our internal and external world are explainable as the simple but powerful action of unjustified variations and selective retention in the dialectic relationship between an organism and its environment (Campbell, 1974b, pp. 142-143).

Karl Popper If Campbell explicitly conceptualized his philosophical works as evolutionary epistemology throughout his whole career, Popper's explicit commitment to an evolutionary approach to epistemological issues can be found, in its fullest form, only in his last published works (e.g. Popper 1972a, 1974b, 1984). If, in fact, already Popper's falsificationist philosophy of science (Popper, 1934, 1963) clearly exhibits Darwinian features in its conceptualization of the evolution of scientific theories as a trial and error process, only in later years Popper starts to explicitly conceptualize his philosophical efforts as

[2]Note that the goal of explaining epistemological puzzles central to Kant's philosophy with considerations taken from evolutionary theory is not unique to Campbell. For instance, Konrad Lorenz repeatedly championed the biologization of Kant's categories as a goal for evolutionary epistemology (Lorenz, 1977, 1982).

efforts towards an evolutionary epistemology.

In his intellectual autobiography, looking back at his work of decades, Popper (Popper, 1974a) stresses that the steady core of his philosophy is a Darwinian theory of learning based on the method of trial and error. Such a non-repetitive learning mechanism is Popper's solution to the problem of induction in epistemology and the problem of demarcation in philosophy of science (Popper, 1963). Both the justification of knowledge and what distinguishes the scientific method are for Popper solvable by realizing that the growth of human knowledge can be explained by a non-inductive theory of learning based on trial and error mechanisms.

According to Popper's Darwinian theory of learning (Popper, 1972b,c), the starting point of every kind of learning can be traced back to the inborn tendencies of organisms in finding regularities and solving problems[3]. Popper sees conscious life as a inherently problem-solving activity that, like science, starts with problems and ends with problems. Learning is then the modification of these inborn expectations of an organism in adapting to the environment and the related survival problems. Consistently, Popper criticizes (what he takes to be) the passive repetitive learning of inductive methods and the related bucket theories of knowledge for failing to see that hypotheses and actions are prior (both logically and temporally) to observations (Popper, 1972b,e). Our knowledge does not grow thanks to a continuous systematization of observations, it grows dialectically through the modification of our educated guesses in the light of our mistakes. This active, evocative kind of epistemology is dubbed by Popper the searchlight theory of knowledge (Popper, 1972e).

The passivity of inductive epistemologies is also for Popper at the heart of the mistakes of induction enthusiasts in philosophy of science (Popper, 1957). What makes a theory scientific is not (the change in) its degree of inductive confirmation, but the possibility of refutation by (quasi-)empirical testing. The conscious attitude towards the attempted falsification of scientific theories is then for Popper (Popper, 1972b, p. 70) what the scientific method, i.e. the critical method, consists of. The critical method of science is, just like for Campbell, just a conscious and externalized version of the trial and error method. The crucial difference between the evolution of scientific theories and analogous biological and psychological processes is given by the fact that in science selection is exosomatic, i.e. it is externalized and directed towards the objects and not the subjects of knowledge (Popper, 1972c,d). Popper stresses that this exosomatic selection is only possible thanks to the argumentative function of language (Popper, 1972c, pp. 236-240) that allows the creation of public intellectual products such as scientific theories that can be falsified or corroborated independently from the survival of their creators.

Popper's efforts in both epistemology and philosophy of science can then be seen as efforts in establishing an evolutionary epistemology centered around a Darwinian theory of learning by adaptation. This active kind of learning by trial and error is for Popper

[3]Note that, as noted by Popper (Popper, 1974a, pp. 44-53 ,72-78) himself, his theory of learning has strong similarities with the ideas of the so-called Würzburger Schule of psychology and in particular to the writings of Bühler and Selz. For a full appreciation of this connection, see (Ter Hark, 2004).

and highly general regularity behind the growth all kinds of biological, psychological, and cultural phenomena that can be thus seen as an interconnected hierarchy of structurally analogous processes (Popper, 1940, 1972d). The same dialectic of hypothetical tendencies and surprising mistakes is at work throughout all the different levels of this hierarchy, "from the amoeba to Einstein" (cf. Popper 1972b, p. 70) .

Stephen Toulmin We have thus seen how Popper and Campbell conceptualized the evolution of scientific theories as driven by a kind of trial and error mechanism where scientific hypotheses get selected by resisting (quasi)empirical testing better than their competitors. Despite its intuitive appeal, this particular version of the EET program in evolutionary epistemology was very soon met with heavy critiques because of its heavily falsificationalist view of scientific method. In fact, Popper's falsificationism and its related image of science as a series of conjectures and refutations governed by the critical method was criticized by many philosophers for depicting a too idealized and wiggish image of the history of science[4]. Not only the history of science and the scientific method seemed impossible to constrain in a neatly-ordered series of tentative hypotheses and attempted refutations, but the overall project of finding a single mechanism responsible for the evolution of scientific theories was increasingly perceived as a pseudo-problem. Together with the so-called historical turn in philosophy of science and the stronger focus on historical adequacy in philosophy of science, also evolutionary epistemology and specifically the EET program underwent an analogous transformation.

A paradigmatic example of this novel, more pluralistic and historically-minded, evolutionary approach to scientific change is Toulmin's "Human Understanding" project. In the homonymous book, Toulmin (Toulmin, 1972) presented a general evolutionary model of scientific rationality where the evolution of scientific theories is seen as a process governed by a plurality of factors and criteria. In comparison with Campbell's and Popper's trial-and-error based evolutionary takes, Toulmin (Toulmin, 1972, pp. 201-260) multiplied the processes behind the variation, the selection, and the propagation of scientific theories and concepts.

Concerning scientific variation, Toulmin's evolutionary model is centered around the notion of a conceptual population (Toulmin, 1970), i.e. a set of conceptual variants that have to face similar scientific problems. Conceptual variants are understood by Toulmin as concepts representing scientific possibilities (Toulmin, 1972, pp. 207-210) , i.e. communal entities showing promise of yielding a recognizable procedure for attacking some outstanding theoretical problem. Scientific concepts are understood by Toulmin as complex, hybrid entities. In his specific hybrid theory of conceptual structure, Toulmin describes concepts as micro-institutions (Toulmin, 1972, p. 166), made of at least three different kinds of conceptual structures corresponding to the language, the representation techniques (i.e. the mathematical apparatus), and the application procedures related to a given scientific

[4]The amount of critiques to Popper's philosophy of science in the philosophical literature is such that listing all of them appears a hopeless task. Every attempt to write such a list should include (at least) the works contained in (Lakatos and Musgrave, 1970).

concept. Toulmin's version of the EET program is thus primarily a model of conceptual change, where concepts are the main actors of the Darwinian selection procedures.

The selective environment of Toulmin's evolutionary model is far broader than in previous evolutionary models of scientific change. It involves not just the specific scientific problems at issue but also the whole social and pragmatic aspects of the scientific discipline to which the conceptual variants belong. The selection mechanism of Toulmin's model are in fact crucially dependent on the shared ideals and criteria of rationality of a given scientific discipline (Toulmin, 1972, pp. 225-227) . The fitness of a given conceptual variant in relation to a given scientific problems can thus be judged only relative to some shared criteria of rationality. As we saw in Chapter 2, this pragmatic encroachment in matters of scientific rationality is typical of pragmatic models of scientific change, a tradition to which Toulmin's model can be definitely associated thanks to its focus on the significance of common values for scientific rationality. The specificity of science evolution is for Toulmin not traceable on a specific method used by scientists in evaluating theories and concepts, but can be seen by realizing that scientific disciplines are mostly compact disciplines (Toulmin, 1972, p. 379) , i.e. human activities organized around a specific set of ideals imposing shared professional and intellectual criteria of adequacy to all their members. Consistently with the pragmatic and pluralistic tendency of his framework, there is for Toulmin a variety of selection mechanisms at work behind the evolution of scientific disciplines, ranging from purely normative and rational ones to professional and sociological mechanisms of the scientific profession (Toulmin, 1972, pp. 488-491).

We have then briefly seen how Toulmin's evolutionary model of conceptual change seeks to combine the EET program and its evolutionary understanding of scientific progress with the emphasis on the historical and pragmatic aspects of science typical of the philosophy of science of his time. By substituting Popper's and Campbell's rigid trial and error evolutionary schema with a broader picture of the intellectual environment and selection at work in science, Toulmin's evolutionary epistemology can be seen as rooted in the kind of anti-essentialist and pluralistic methodology typical of Darwin's population thinking (Toulmin, 1967; Mayr, 1975). According to such a populational approach, the rationality of scientific activity and its product can only be understood as the gradual transformation of conceptual populations regulated by the agreed ideals and criteria of adequacy of a given scientific discipline.

David Hull Another, more recent, example of a pluralistic approach to the EET program can be found in David Hull's model of scientific change (Hull, 1988a). Hull's evolutionary epistemology has lots of features in common with Toulmin's one, such as the centrality of conceptual change, the populational thinking approach to scientific structures, and the attention to the social and professional aspect of scientific activity. What is specific to Hull's work in the evolutionary epistemology literature, especially in comparison to the three examples above, is the level of detail of his proposal and the efforts in specifying and historically testing his ideas about science evolution. A full appraisal of Hull's fine-grained model of scientific conceptual change is out of our present focus, so in what follows

I will focus on three particularly salient features of Hull's evolutionary model of science: the general units of selection processes, the plurality of scientific lineages, and the demic structure of scientific activity.

As we will see in the next subsection, a standard critique against evolutionary models of science is centered around the disanalogies between the units of scientific selections and its (alleged) biological analogues. In order to overcome this critique, Hull (Hull, 1988a,b) developed a very general model of selection processes, of which according to him both biological and intellectual selection are specific instantiations. In this general model of selection, there are two kinds of entities involved in the selection: replicator and interactors. Replicators are entities that pass on their structure largely intact in successive acts of replications, while interactors are defined as entities that interact as a cohesive whole with their environment in such a way that this interaction causes replication to be differential. The interwined activities of interactors and replicators is then nothing but the selection process itself, understood as the process in which the differential extinction and proliferation of interactors cause the differential perpetuation of replicators. An important byproduct of this selection process is what Hull calls lineage, i.e. an entity that persists indefinitely through time either in the same or in an altered state as a result of replication. Thanks to this general model of selection processes, where the roles of replicator and interactors are defined solely in terms of their function in the selection process (and without any reference to specific ontological or structural features that they might exhibit), Hull can adequately describe the evolution of science as a paradigmatic selection process. In scientific selection, the replicators are all the elements of the substantive content of science (e.g. problems, solutions, data reports, goals, meta-beliefs, etc.). These replicators are passed on through a conceptual kind of replication for means of books, journals, individual brains, and other similar means. The main interactors of this selection process are of course the individual scientists, who are explicitly the main agents of Hull's model of scientific evolution.

The selection process at work behind science evolution produces for Hull various kinds of conceptual and social scientific lineages such as scientific theories, research programs, research traditions, schools of thought, and many others. This plurality of lineages does not represent a problem for evolutionary approaches, since it is a consequence of the different ways in which we can conceptualize the elements of a selection process. Similar to Toulmin's anti-essentialism, Hull (Hull, 1976, 1978) championed the use of populational thinking in conceptual matters by repeatedly stressing the centrality of individuality in natural and intellectual selection. In every selection process, replicators, interactors and even lineages are individuals, i.e. transient entities that exist in time and come and go out of existence. Selection processes, qua processes essentially involving individuals, can thus be described in a plurality of ways dependently on the tokens and units of selection that we focus on (cf. the type-specimen method of reference Hull 1988b, pp. 149-154).

The final feature of Hull's model that I want to focus on is his insistence on the demic structure of science. Scientific activity is for Hull strongly based on demoi such as research groups and institutional communities of various sizes. This demic structure is crucial in the selection mechanisms of science that can be divided in two kinds: intra-demic and inter-

demic. Scientific activity involves for Hull a mixture of competitiveness and cooperation that forces scientists to trade off credit for evidential support. As rational agents, scientists can be modeled as if they were trying to maximize their own's conceptual fitness and the one of their own demos. Hull (Hull, 1988a, 1996) developed an economical system of scientific credit and discredit centered around the value of empirical data use that seeks to give an invisible-hand explanation of the success and progress of science.

We have then briefly seen how Hull's evolutionary models of scientific change continues the pragmatic and populational approach to evolutionary epistemology and in particular to the EET program that Toulmin championed twenty years before. Specifically, we saw how Hull proposed a general functional characterization of selection processes applicable to both biological and intellectual kinds of selection and pluralistically open to different characterization of the process itself. Furthermore, Hull expanded in greater detail Toulmin's insights on the evolutionary significance of the social and professional aspects of scientific activity by developing an economical model of scientific credit and discredit based on the demic structure of scientific communities.

4.1.2 The debate on the evolution of scientific theories

The examples we just saw show four different ways in which the EET program of evolutionary epistemology can be carried out. Despite the differences between the particular instantiations of the EET program, we have seen that there is a core ideal common to all such approaches. This core ideal is that the evolution of scientific theories and concepts can be adequately described by a (series of) selection mechanism(s) significantly akin to the one(s) through which Darwinian theories explain the evolution of biological entities. In other words, at the heart of the EET program lies the belief that there is a significant analogy between the evolution of biological entities and the evolution of our intellectual products. The significance of this analogy is precisely what critics of the EET program contest.

Evolutionary epistemology and specifically the EET program have attracted lots of critiques, since their appearance (e.g. Cohen 1973; Skagestadt 1978; Thagard 1980; Lewontin 1982; Fracchia and Lewontin 1999; Renzi and Napolitano 2011). Philosophers and biologists have criticized evolutionary approaches to scientific change on many different grounds, accusing them of misunderstanding the explanatory role and the mechanisms of evolutionary theory or of not realizing what an epistemology of science really amounts to. Amongst the many different critiques that have been raised against the EET program, a central line of argumentation can be discerned. Most critiques of the EET program argued somehow for a disanalogy between biological and intellectual evolution that (allegedly) makes the central analogical ideal of the EET program crumble.

In what follows, I will briefly present some alleged disanalogies between biological and intellectual evolution that have been stressed as serious flaws in the central analogy on which the EET program rests. More specifically, I will organize the presentation of these disanalogies according to the element of the evolutionary process that they target, i.e. whether the (alleged) disanalogy concerns variations, selection, or inheritance mechanisms.

- **Disanalogies in variation:** the most common critique to any analogy between biological and intellectual evolution concerns the blindness of the variations. If, in fact, at least in scientific forums, the blindness of biological variations is an almost universally accepted fact, many philosophers stressed that the same kind of blindness cannot be attributed to scientific theories and hypotheses (cf. Cohen 1973; Skagestadt 1978; Thagard 1980). Attempts to solve scientific problems, so the critique goes, are not blind trials, but they are created with the intended aim of solving a given (set of) problem(s). Such a non-blind variation breaks the central change mechanism of any wannabe Darwinian-like selection process. As such, concludes the critic, the analogy between natural selection and the process by virtue of which our creative products evolve is only a superficial one and cannot have any serious explanatory role. A similar critique to the analogy between biological and scientific variation concerns not so much the blindness of the variation, but their independence from the environment. Some philosophers stressed in fact that the appearance of scientific theories is inevitably intertwined with the scientific problems that are in need of a solution (Cohen, 1973; Thagard, 1980). This intimate relationship between scientific theories and the scientific problems that they are meant to solve allegedly makes the variation inherently coupled with selective criteria. This coupling of variation with the selective environment is, for the critics, another disanalogy between the biological and the intellectual realm that makes them doubt the viability of the EET program.

- **Disanalogies in selection:** If critiques of the EET program that concern variation stress the specificity of scientific variation, critiques that target the selection element of the EET analogy underline the extreme variability behind the alleged selection process of scientific theories and concepts. The choice of a scientific theory or concept arguably involves in fact multiple intellectual, pragmatic, and sociological selection mechanisms and criteria. This complex bundle of interconnected selection processes seem to some critics too cloudy and multifaceted to be described by a single neat Darwinian-like selection mechanism (Godfrey-Smith, 2009, pp 147-151). So that the explanatory role of the EET analogy between biological and intellectual selection is put into question. These doubts are often intertwined with traditional skepticisms in the biological literature about the cogency of extending Darwinian ideas to the social and cultural realms (Lewontin, 1982; Fracchia and Lewontin, 1999). The skepticism of scientists and philosophers towards cultural selection theories (Cavalli-Sforza and Feldman, 1981; Boyd and Richerson, 1985) or efforts towards sociobiology (Wilson, 1980) make thus many scholars doubtful about the EET program, the central analogy of which they see as resting on similar dubious assumptions.

- **Disanalogies in inheritance:** The critiques of the EET program that target the inheritance element in the selection process of scientific products concern instead the viability of such intellectual entities to function as replicators in a selection process (Godfrey-Smith, 2009, 2012). In fact, cultural replication and related imitation and learning dynamics have never been fully accepted by the scientific and philosophical

> community as mechanisms able to create a truly Darwinian process. From the appearance of Dawkins' (Dawkins, 1976) concept of memes, in fact, cultural replication has been accused to exhibit only superficial analogies with actual biological replication. Many critics complained that these superficial analogies do not warrant the extension of Darwinian ideas to the cultural realm, since cultural replication cannot play the fundamental role that biological reproduction has in natural selection. As in the case of critiques concerning selection, then, the EET program inherits a certain degree of skepticism in its use of cultural replicators from previous attempts towards a Darwinian-like theory of cultural evolution.

These general critiques to the central analogy of the EET program are indeed serious enough to cast doubts on the viability and significance on every analogy between biological and intellectual evolution. Nevertheless, supporters of the EET program have many ways to reply to these doubts (e.g. Plotkin 1982; Cziko 1995; Gontier, Bendegem, and Aerts 2006; Charbonneau 2014).

Already in the four examples of EET approaches that we saw in the last subsection, we can see promising lines of defense against these critiques. Popper and Campbell, for instance, spent a lot of effort in defending a certain degree of blindness in scientific variation against commonsense intuition (Campbell, 1974a,b; Popper, 1972a, 1974b). As Campbell (Campbell, 1974b, pp. 152-158) puts it, Darwinian processes do not require complete randomness in the variation, but only a variation that is unjustified and not too goal-directed. Popper's work in philosophy of science can be seen as arguing that the variation of scientific hypotheses is exactly of this unjustified kind (cf. Popper 1957, pp. 66-71). Concerning the doubts about the Darwinian character of the selection and the inheritance mechanisms of science, Toulmin's and Hull's work provides suitable replies. A great deal of Toulmin's (Toulmin, 1972, Section B) evolutionary model of conceptual change is in fact dedicated to give an evolutionary account of the multitude of selection mechanisms and selection criteria at work in the choice of a given scientific theory or concept. Hull's (Hull, 1988a,b) general model of selection processes, instead, and the fundamental distinction between replicators and interactors provide a way in which Darwinian models of science can escape the need of relying on contested models of cultural replication.

The (non)viability of the EET program seems then not ascertainable by virtue of in principle arguments. Many arguments cast doubts on the analogy between biological and intellectual evolution, but several replies are at the disposal of any wannabe evolutionary epistemologist. Moreover, the significance of the analogy at the heart of the EET program rests also on contemporary understanding of both evolutionary theory and scientific activity. Our ideas on both biological evolution and scientific change are of course constantly changing in the normal dialectic of the respective fields. Such as dynamic process naturally determines the emergence and disappearance of positive and negative trends towards the whole enterprise of evolutionary epistemology. So that, just a brief look at the philosophical literature on evolutionary epistemology in recent years shows a variety of positive and negative takes on the EET program determined by the rise and fall of specific scientific and philosophical related approaches. Examples of such approaches that have been argued

to be closely relevant to the viability of the EET program include the extended synthesis in evolutionary biology (Sarto-Jackson, 2019; Pigliucci and Müller, 2010), cultural evolution theory (Mesoudi, 2011; Fadda, 2020), cognitive biology (Kovac, 2000), genealogical approaches to epistemology (Gelfert, 2011; Craig, 1990; Williams, 2002), social epistemology (Goldman and Whitcomb, 2011), and evolutionary game theory (Harms, 1997, 2004; Skyrms, 2010).

Instead of the seemingly neverending dialectic between arguments in favor and against the EET program, a more promising way of assessing the significance of evolutionary models of science seem to be the historical and philosophical testing of the specific models put forward by advocates of evolutionary epistemology. Evolutionary models of scientific change, just like all the others philosophical models of scientific activity, should be historically tested as methodological hypotheses on the epistemological structure of scientific change. In order to do that, though, evolutionary epistemologists should build narrower evolutionary models that target specific selection processes at work in a specific scientific discipline. These more specifically focus models will be easier to apply to related case studies from the history of science and thus easier to historically test. In the next section, I will try to do a first small step towards filling this gap, by presenting a simple formal framework for modeling the mechanism by virtue of which fruitful mathematical concepts get selected.

4.2 An Evolutionary Framework for Conceptual Selection in Mathematics

In this section, I will propose a novel evolutionary framework for conceptual change in mathematics centered around the notion of conceptual populations, the opposition between Euclidean and Lakatosian populations, and the spatial tools that I will call the Lakatosian space.

Before presenting my framework, I must stress that very few examples of evolutionary epistemology or evolutionary models of scientific evolution target mathematics. This lack could be explained by noticing a more general lack of general models of theoretical and conceptual change in the philosophy of mathematics. In fact, despite the plethora of mathematical episodes of conceptual change that have been analyzed in philosophical and historical literature, few general frameworks for modeling this phenomenon have been proposed. In contrast to what happened in the philosophy of natural sciences, where the philosophical debate about conceptual change has centered around contrasting general pictures (e.g. Toulmin 1972; Stegmüller 1976; Thagard 1992; Kitcher 1995; Friedman 2001; Andersen et al. 2006; Wilson 2006), the discussion in philosophy of mathematics has mostly proceeded in a piece-meal fashion (cf. Gillies 1992). An exception to this pattern can be found in Mormann's (Mormann, 2002) proposal to use evolutionary theory to improve Lakatos' (Lakatos, 1976) seminal model of conceptual change in mathematics. Mormann sketched a general Darwinian selection theory for mathematical concepts in

which conceptual variants compete in a world of proof-experiments[5].

I will construct my framework by building upon Mormann's work, proposing a more fine-grained evolutionary framework for conceptual change in mathematics. In order to do this, I will radically change Mormann's evolutionary background theory, using Godfrey-Smith's recent version of the Darwinian selection theory (Godfrey-Smith, 2009) as a conceptual basis for my framework. Godfrey-Smith's population-based Darwinian framework is made of two ingredients: the family of concepts of a Darwinian population and several parameters tracking how much a certain population exhibits paradigmatic Darwinian features such as reliable inheritance mechanisms, abundance of variation, and continuity. The interaction between different Darwinian population concepts and these parameters provides a gradual and pluralist approach to evolution by natural selection. By mirroring much of the structure of Godfrey-Smith's framework, I will show how my proposal achieves an account of conceptual change in mathematics with analogous advantages. My framework will offer an evolutionary model of conceptual change compatible with the diversity of evolutionary dynamics that the history of mathematics exhibits. My framework will also give a novel perspective on whether conceptual change in mathematics is a rational process, distinguishing conceptual histories in mathematics between cases of mathematical selection and cases of evolutionary drift.

As I mentioned above, my framework will be centered around the notion of a conceptual population, i.e. a set of conceptual variants and a set of mathematical problems together with a selection mechanism given by an heuristic power ordering of conceptual variants (relative to a given problem). I will present two ideals of conceptual populations, namely Lakatosian and Euclidean populations, that will represent (almost) opposite evolutionary dynamics. I will augment my framework with four parameters: conceptual variation, reproductive competition, environmental stability, and continuity. These four parameters that track how much a given conceptual population exhibits certain evolutionary features constitute the four dimensions of the Lakatosian space. Depending on how much they exhibit these parameters, conceptual populations can be judged to be more Lakatosian or more Euclidean (or neither of them), occupying different regions of the Lakatosian space.

I will demonstrate how my framework, thanks to the four dimensions of the Lakatosian space, is able to provide a rich understanding of the evolutionary dynamic of a given episode of conceptual change in mathematics. I will show how a mathematical conceptual history can be represented in my framework as a conceptual population and how its evolutionary dynamic can be judged to be a case of mathematical selection or evolutionary drift. The Lakatosian space then becomes a conceptual space for classifying episodes of the history of mathematics in terms of the evolutionary features exhibited by their rationally reconstructed conceptual population. Moreover, I will sketch that the Lakatosian space, augmented with a time-dimension, is able to model also diachronic conceptual histories and related inter-population changes as a series of specific movements along the four dimension

[5]Other three notable general models of conceptual change in mathematics are Wilder's (Wilder, 1953) sketch of an evolutionary account of mathematical concepts and Kitcher's (Kitcher, 1984) and Ferreirós' (Ferreirós, 2015) practice-based frameworks.

of the original Lakatosian space.

In this section I will present my framework in full generality, while in the next section I will show how my framework can be applied to several case studies from the history of mathematics. More specifically, In Section 2.1 I will describe the philosophical background of my framework, i.e. Lakatos' seminal work on conceptual change and Mormann's evolutionary rethinking of it. In Section 2.2, I will present Godfrey-Smith's population-based Darwinian framework, stressing the reasons why I chose it as a background theory for this work. In Section 2.3, I will present my evolutionary framework centered around the notion of a conceptual population, the opposition between Lakatosian and Euclidean populations, and the Lakatosian space. In Section 4, I will show how the Lakatosian space can be augmented with a time-dimension in order to model even inter-population kinds of changes.

4.2.1 Models of conceptual change in mathematics

As I already mentioned, the study of conceptual change in mathematics has mostly proceeded in a case-by-case fashion, through analyses of specific conceptual histories and case studies. In this subsection I will focus instead on surveying frameworks that have tried to understand the phenomenon of conceptual change in mathematics from a more general perspective. Specifically, I will focus on two models that directly inspired my framework, namely Lakatos' (Lakatos, 1976) concept-stretching and Mormann's (Mormann, 2002) selection theory for mathematical concepts.

Lakatos' concept-stretching

Lakatos (Lakatos, 1976) had a more general aim than to give a model of conceptual change in mathematics. He wanted to develop a dialectical philosophy of mathematics, i.e. a philosophy of mathematics focused on "the process by which mathematical argument improves mathematical concepts" (Larvor, 1998, p. 11). This aim was quite unusual in philosophy of mathematics at the time. Early twentieth-century philosophy of mathematics, referred by Lakatos with the umbrella term "formalist school", focused in fact on the (meta)logical properties of formalized mathematical systems and the epistemological and ontological problems connected to their foundations. Informal mathematics and the conceptual histories of mathematical theories were thus not seen as very fruitful objects of philosophical study.

Lakatos directly set out to challenge this status quo. He thought that, in order to understand and rationally reconstruct a given mathematical theory, one cannot neglect the actual history of its main concepts (Hacking, 1979). This is the dialectical component of his philosophy of mathematics. This standpoint led him directly to the issue of conceptual change. In fact, the lack of formal regimentation of informal mathematics makes the concepts used in a certain mathematical field at a certain time vary. This was very much stressed by the 'formalists', who saw the vagueness and ambiguity of mathematical practice and its evolved languages as the reason why philosophers ought to rationally reconstruct

mathematical theories in suitable formal languages (cf. Carnap 1934). Lakatos completely subverted this view, seeing in the inevitable messiness of informal mathematics one of the main engines of mathematical progress. Like the formalist school, Lakatos believed that philosophers should rationally reconstruct this dynamics, but he championed a different form of rational reconstruction. A dynamic process required dynamic tools, which for Lakatos were exemplified by heuristics and not by formal logic[6].

In what it is usually regarded as the first monograph in philosophy of mathematics where conceptual change takes central stage, Lakatos (Lakatos, 1976) presents his model of conceptual change through a rational reconstruction of the notion of polyhedron in connection with Euler's conjecture[7]. In Lakatos' reconstruction, the problem starts with the conjecture that Euler's formula connecting the number of vertices, edges and faces of regular polyhedra ($V - E + F = 2$) holds for any polyhedron whatsoever. This conjecture is supported by Cauchy's thought experiment. In a series of reconstructed steps, the conjecture and its alleged proof get challenged by a series of counterexamples of various kind, after every one of which an attempt to defend or improve the conjecture and its proof is made.

Abstracting from the specific case at issue, Lakatos presented a three-level analysis of the dynamics between proofs and refutations. At the first level we find the (many instances of the) conjecture and the counterexamples to it. Lakatos classified the possible counterexamples that any mathematical conjecture may face. The main distinction is between local and global counterexamples. Local counterexamples refute a particular lemma or step of the tentative proof, while global counterexamples are directed towards the conjecture as a whole. Lakatos then distinguishes between "logical"(i.e. global but not local) and heuristic (local) counterexamples.

Lakatos discusses several methods for dealing with counterexamples, such as lemma-modification, barring-adjustments methods (various ways of refusing to count the counterexample as a genuine one), and the method of lemma incorporation. These methods form the second level of Lakatos' analysis. The method of lemma-incorporation in particular plays a central role in Lakatos' view of conceptual change. It consists in finding a hidden conjecture-lemma (e.g. polyhedra are stretchable onto a plane) refuted by a given counterexample (e.g. the nested cube) and inscribe this 'guilty' lemma into the conjecture as a condition for its applicability. This method saves the conjecture by restricting its domain to a narrower one (e.g. Euler's conjecture for 'simple' polyhedra, i.e. the stretchable ones), thereby connecting the counterexamples, the proof, and the conjecture and thus displaying the "fundamental dialectical unity of proof and refutations"(Lakatos, 1976, p. 39).

The third-level of Lakatos' analysis is made of a set of heuristic rules for applying the above methods to any given counterexample. This list is quite vague and it is not

[6]This distrust of Lakatos for logical reconstructions of dynamic processes is difficult to understand, especially from a contemporary perspective. Dynamic logics and the whole field of belief revision seem natural tools for logically reconstructing virtually any dynamic process whatsoever. Even at Lakatos' time, logicians criticized heavily this unjustified anti-logical stance of Lakatos (e.g. Feferman 1978).

[7]For in-depth analyses of Lakatos' seminal book, see (Larvor, 1998; Kadvany, 2001).

really clear how the different rules interact with each other. These issues notwithstanding, Lakatos' heuristics seems to be centered around the meta-method of proofs and refutations, i.e. a search for heuristic counterexample followed by several applications of the method of lemma-incorporation. Lakatos warns us against the abuse of this meta-method, though. Lemma-incorporation saves the conjecture via restricting its intended domain. If this retreat to a narrower domain is repeated too many times, we may be left with a lack of content in our theorem. This impoverishment of content can be countered trying to replace lemmas that are refuted by heuristic counterexamples with unfalsified ones, thereby increasing the content of the theorem. Another way of countering the decrease of content is the more general deductive guessing for deeper theorems to which given counterexamples do not apply anymore.

Lakatos' rational reconstruction of Euler's conjecture gives us this dynamic taxonomy of conjectures, counterexamples, heuristic methods, and heuristic meta-rules. In this dynamic, the concept of a polyhedron changes consistently with the heuristic agenda. The search for heuristic counterexamples drastically expand the concept of polyhedron out of its intended domain. In this expanded domain, it is not clear how to apply this concept correctly, warranting the use of barring techniques against counterexamples. Lemma-incorporation and deductive guessing then redefine what a polyhedron is, inscribing proof-methods into the definition of the concept in order to shield the conjecture against counterexamples (the former) or to boost its content (the latter). This process creates several proof-generated concepts of a polyhedron, each one of them theoretically stretched by the underlining proof of the conjecture. In the dialectics of proofs and refutations, thus, the concept under focus gets stretched in various directions via the interaction of counterexamples, proofs and heuristic methods. Lakatos refers to these dynamics of conceptual change as *concept-stretching*[8].

Mormann's selection theory for mathematical concepts

Thomas Mormann built upon Lakatos' model of conceptual change, sketching a Neo-Lakatosian evolutionary theory of mathematical knowledge (Mormann, 2002). According to him, the fundamental driving force of mathematical evolution is the "axiomatic variation of concepts" (Mormann, 2002, p. 139). He claims that conceptual variation is not restricted to informal mathematics, *contra* (the received view of) Lakatos, and he argues that modern axiomatized mathematics still exhibit conceptual variations, in the form of different axiomatic versions of the same concept (e.g. Hamilton's invention of the quaternions)[9]. Thanks to this ubiquitous conceptual variation, the evolution of mathematical

[8]Some scholars use other terms to refer to Lakatos' model of conceptual change such as 'concept-trafficking' (Mormann, 2002), using concept-stretching to refer only to the particular method of stretching the extension of a concept. I follow Fine (Fine, 1978) in using concept-stretching to refer to the whole model of conceptual change presented by Lakatos.

[9]Note that whether Lakatos regarded conceptual variation to be exclusive of informal mathematics is a *vexata quaestio* in Lakatosian scholarship. For different takes on this question, see (Corfield, 2002; Feferman, 1978; Priest and Thomason, 2007). Moreover, note that here Mormann seems to conflate

knowledge can be understood as a Darwinian evolution of mathematical concepts, based on the competition between variants of the same concept:

> "This competition among conceptual variants may be described within the framework of an evolutionary theory, which conceptualizes the evolution of mathematical knowledge as a selection process of conceptual variants taking place in a 'world' whose challenges are determined by varying theorematical environments" (Mormann, 2002, p. 140).

Mormann takes mathematical concepts to be the main characters of his framework. Specifically, the main unit of selection are conceptual variants, i.e. specific definitions of a given mathematical concept, such as the many different tentative definitions of a polyhedron discussed in Lakatos' fictional classroom. A group/species of conceptual variants is made of several of these tentative definitions competing for the same (or similar) proof-problem(s). In Mormann's framework, proof-experiments constitute survival tasks for conceptual variants. Successful proof-experiments, i.e. valid proofs, constitute a positive outcome for the conceptual variant used in the proof. The more successful a given conceptual variant is, the more it will be used in future proofs, whereas unsuccessful conceptual variants get more and more disregarded by the mathematical community until they often get completely forgotten.

Mormann sketches a selection theory for mathematical concepts in the form of a summary or recipe (Godfrey-Smith, 2007), i.e. a set of general principles for evolution by natural selection analogous to a classical presentation of Darwinism such as Lewontin's (Lewontin, 1970). Mormann's recipe-like selection theory is made of four principles: variation, competition, variation of fitness, inheritance.

The principle of variation asserts the existence, at any stage of the history of mathematics, of several conceptual variants of a given mathematical concept. Variants of the same concept have different properties but they have to face similar proof-problems. Mormann's framework takes conceptual variation as the basic engine of mathematical growth, a dynamics that encompasses also modern axiomatized mathematics via the aforementioned notion of 'axiomatic variation'.

Mormann's principle of competition tells us that this abundance of conceptual variants sharing similar proof-environments forces these variants to inevitably compete against each other for being used in fruitful proofs/theorems. The more a given conceptual variant successfully copes with the proof-problems constituting its environment, the more prominent and higher in ranking becomes in respect to its competitors. Often, very successful variants become the 'accepted definition(s)' of the concept. Unsuccessful variants, instead, gets lower and lower in the ranking, until they are very rarely used and often forgotten. Discharge of a conceptual variant is not as final as biological extinction, though, because variants that have been forgotten can always be reappraised and used again in the future

formalized mathematics with fully axiomatized one. In my own framework the degree of axiomatization of a given mathematical population and its degree of formalization are carefully distinguished and traced via different parameters. However, in this section I will follow Mormann's conflated use of the two aspects.

as it was famously the case for Leibniz's notion of infinitesimals thanks to Robinson's non-standard analysis (Robinson, 1974).

The principle of variation of fitness stresses the fact that the different properties of conceptual variants make them more or less likely to succeed in facing a given proof-problem. Analogously to the biological case, variants can be more or less fit to a given theorematical environment. The degree of fitness of a given variant is understood by Mormann as the extensional generality of the theorems it makes possible to prove. Mormann acknowledges that this way of understanding the fitness of conceptual variants is quite vague, but he thinks that this only shows the need of more work on neo-Lakatosian approaches to conceptual change (Mormann, 2002, p. 148).

The principle of inheritance tells us that new conceptual variants have inevitably to resemble their conceptual ancestors in order to cope with at least some proof-problems faced by the older variants. This implies for Mormann that at least some properties of conceptual variants gets inherited by their successors, ensuring a (partial) historical problem-solving continuity in the history of a given mathematical concept. However, Mormann warns us not to overestimate this continuity, because one of the specific features of the evolution of mathematical concepts is that the proof-problem world that they inhabit can very quickly change drastically. Certain conjectures may become completely obsolete or a new application of a given proof-problem may completely change the domain (and thus the proof-challenges) of a given mathematical concept.

In sum, Mormann generalized Lakatos' idea of concept-stretching to his notion of axiomatic variation that is meant to model also the conceptual dynamics of axiomatized bodies of mathematics. More generally, he sketched a recipe-like selection theory for conceptual variants competing in a world of proof-problems, aiming to describe the dynamics behind the rise and fall of mathematical concepts within an evolutionary framework.

4.2.2 Godfrey-Smith's Darwinian framework

We have seen that Mormann uses a recipe-like selection theory as a background for his model of conceptual change. I will instead use a different kind of evolutionary theory as a background for my framework, namely Godfrey-Smith's population-based Neo-Darwinism (Godfrey-Smith, 2009).

Recipes-like approaches that inspired Mormann's selection theory have been in fact heavily criticized in philosophy of biology (Godfrey-Smith, 2007). These approaches try to give an abstract summary of the evolutionary dynamics in the form of a recipe for a change. Take, for instance, Lewontin's mature formulation of evolution by natural selection:

> "A sufficient mechanism for evolution by natural selection is contained in three propositions:
>
> 1. There is variation in morphological, physiological, and behavioral traits among members of a species (the principle of variation).

2. The variation is in part heritable, so that individuals resemble their relations more than they resemble unrelated individuals and, in particular, offspring resemble their parents (the principle of heredity).
3. Different variants leave different numbers of offspring either in immediate or remote generations (the principle of differential fitness).

All three conditions are necessary as well as sufficient conditions for evolution by natural selection ... Any trait for which the three principles apply may be expected to evolve." (Lewontin, 1985, p. 76)

In this formulation, as well as in other recipes-like ones, variation, heritability and fitness differences are meant to be necessary and sufficient ingredients for producing evolution by natural selection. The problem with these approaches is that it can be shown that these ingredients are neither necessary nor sufficient to cover all the different cases of actual evolution by natural selection. That is, there are cases where all the ingredients are present but change does not occur and cases where change does occur without all the ingredients (Brandon, 1978; Godfrey-Smith, 2007). Godfrey-Smith diagnoses this problem as caused by the attempt of traditional recipes-like approaches to perform two contrasting tasks at the same time. These recipes are, on the one hand, meant to describe all genuine cases of evolution by natural selection and, on the other hand, expected to consist of a simple, causally transparent mechanism for change. Abstract recipes like Lewontin's are then the result of an uncomfortable trade-off between these two tasks, trying to squeeze all the diverse forms in which natural selection produces evolutionary change into one neat, encompassing mechanism.

As an improvement of this situation, Godfrey-Smith proposes a more-fine grained Darwinian framework designed to solve these problems thanks to a gradual and plural approach to evolution by natural selection. Instead of a one-size-fits-all recipe, he proposes a combination of a general set-up together with various specific models thereof. More specifically, Godfrey-Smith's set-up is constituted by the family of concepts of a Darwinian population. We can talk about a Darwinian population in three senses, a minimal, a paradigm, and a marginal one:

- "A *Darwinian population in the minimal sense* is a collection of causally connected individual things in which there is variation in character, which leads to differences in reproductive output (...) and which is inherited to some extent" (Godfrey-Smith, 2009, p. 39);
- A *Darwinian population in the paradigm sense* is a minimal Darwinian population that has reliable inheritance mechanisms, unbiased and slight variation, reproductive competition, reproductive differences highly dependent on intrinsic features of the individuals, and that exhibits continuity[10].

[10]This is a rough summary of this concept. What Godfrey-Smith actually requires from a paradigm Darwinian population is more nuanced and gradient. For a full-account of this notion see (Godfrey-Smith, 2009, pp. 41-59)

- A *Darwinian population in the marginal sense* is a population which does not fully satisfy the requirements for a minimal Darwinian population, but only approximates them.

The minimal concept is supposed to be applicable to very different biological phenomena, requiring only a minimal locality constraint on the members of the population. The members of a Darwinian population in the minimal sense, i.e. the *Darwinian individuals*, must exhibit the three ingredients of recipe-like Darwinism (variation, inheritance, fitness differences) only to some extent. The other two senses in which one can speak of a Darwinian populations are instead designed to stress the extent to which evolution by natural selection is central to the dynamics of a given population. Populations approximating the ideal dynamic of evolution by natural selection are the paradigm ones. These are the Darwinian populations representing significant Darwinian processes, i.e. processes that exhibit all the paradigmatic features of a truly Darwinian process. Paradigm Darwinian populations not only exhibit all the ingredients of recipe-like Darwinism, but they instantiate 'the right kind' of variation, fitness differences, and inheritance. These populations exhibit reliable inheritance mechanisms, slight and unbiased variation, reproductive differences highly dependent on intrinsic individual features and other extra features that contribute to make the perfect scenario for evolution by natural selection. Finally, the concept of a marginal Darwinian population allow one to stretch Darwinian concepts onto biological phenomena whose dynamics are not really Darwinian, but in which one can discern aspects that are partially Darwinian in character.

Godfrey-Smith adds structure to his population-based set-up of evolutionary theory with the aid of the *Darwinian space*, i.e. a space the dimensions of which are parameters tracing how much a population is paradigmatically Darwinian with respect to a given feature. This spatial structure is meant to split into different dimensions the extent to which a given evolutionary process has a Darwinian character, allowing a gradual representation of all the possible types of Darwinian processes. The Darwinian space has five dimensions, representing five different parameters: fidelity of inheritance, abundance of variation, reproductive competition, continuity, dependence of reproductive differences on intrinsic character (Godfrey-Smith, 2009, p. 63). Fidelity of inheritance tracks how much the state of a parent is predictive of the state of the offspring. Abundance of variation measures the amount of variation amongst the individuals of a population at a time. Reproductive competition indicates the extent to which the reproductive success of a given individual reduces the success of others members of the population. Continuity is a measure of the overall extent to which similar members of the populations have similar fitness. Dependence of reproductive differences on intrinsic character tracks how much differences in reproductive output are caused by intrinsic features of the members of the population (and not by extrinsic ones).

Each of these parameters represents an aspect with respect to which a given population can be more or less paradigmatically Darwinian. Different regions of the Darwinian space, i.e. different combinations of these parameters, represent different types of biological phe-

nomena[11]. Paradigm Darwinian populations occupy then the part of the Darwinian space where all five parameters take high values, while marginal Darwinian populations are at the opposite side. Minimal Darwinian populations occupy instead a large portion of the space, including the part where paradigm Darwinian populations are. Moreover, specific regions of the space (representing specific combinations of the parameters) are able to explicate phenomena underlying specific dynamics of populations such as the concept of drift and error catastrophe (Godfrey-Smith, 2009, pp. 59-64).

These spatial tools enrich the family of concepts of a Darwinian population with a more fine-grained structure, enabling Godfrey-Smith's framework to adequately represent the plurality of evolutionary dynamics. The diversity of ways in which evolution by natural selection occurs is not squeezed anymore into a one-size-fits-all abstract recipe, but it is reflected by all the possible combinations of parameters allowed by the Darwinian space. Thanks to this rich structure, Godfrey-Smith's framework is able to account for several issues faced by recipe-like Darwinian accounts, such as the problem of units of selection, the relationship between reproduction and individuality, or the explication of evolutionary drift.

4.2.3 Conceptual populations and the Lakatosian space

We have seen how Godfrey-Smith offers a Neo-Darwinian framework that overcomes several issues faced by recipe-like accounts of evolution by natural selection. In what follows, I am going to propose a novel evolutionary framework for modeling mathematical conceptual change. The framework, like Mormann's, is a selection theory for mathematical concepts, but instead of having a recipe-like selection theory (like Mormann does), I use a population-based framework analogous to Godfrey-Smith's one.

More specifically, I will take the coarse-grained structure of my framework from Godfrey-Smith's presentation of Darwinism, distinguishing two different types of populations of mathematical concepts, namely Lakatosian populations and Euclidean populations. I will structure the relationship between these two types of populations via the addition of a spatial framework, i.e. the Lakatosian space, mirroring Godfrey-Smith's construction of a Darwinian space. The four dimensions of the Lakatosian space are made of four parameters: conceptual variation, reproductive competition, environmental stability, and continuity. These parameters trace how much a given conceptual population is more Lakatosian or more Euclidean (or different from both of them) with respect to a given aspect of its evolutionary dynamic. Different regions of the Lakatosian space will then represent different evolutionary dynamics that conceptual populations in mathematics exhibit.

In comparison to Godfrey-Smith's framework, I will also offer formal versions of several components of my framework. Consistently, with my critical appraisal of the EET program at the end of Section 1.2, I will develop an evolutionary model of scientific change narrower in scope and more precise than the ones we saw in Section 1.1. In order to do this, I

[11]This idea of tracking conceptual similarities via spatial frameworks is reminiscent of Gärdenfors' theory of conceptual spaces (cf. Ch. 3, Sect. 4.1).

will precisely define what a (mathematical) conceptual population is, making explicit the pivotal role of what I will call the heuristic power ordering. The notion of heuristic power plays the role of biological fitness in my evolutionary framework. The heuristic power ordering then classifies the fitness of mathematical conceptual variants and guides the related selection process of a given variant within a given conceptual population. These formal definitions will allow me to give measure for the parameters corresponding to the dimensions of the Lakatosian space and, as we will see in the next section, it will make the model easily applicable to case studies from the history of mathematics.

Conceptual populations

Before presenting my framework, some preliminary definitions and clarifications are needed. Mathematical concepts are the main actors of my framework. I am not relying on any specific theory on conceptual ontology or structure, as well as I am not assuming any epistemological theory about how mathematicians are able to refer to them. I just assume that terms used in mathematical proofs express some kind of entity, with reference to which one can prove mathematical statements. I furthermore assume, following Toulmin (Toulmin, 1972) (see Section 1.1), that mathematical concepts have a somewhat public dimension in which they can be discussed and criticized.

Moreover, my framework does not include time-dependent aspects of mathematical conceptual change such as the emergence and the cultural transmission of concepts in mathematics. Although I hold that any complete evolutionary theory of conceptual change in mathematics would have to take into account how conceptual variants and proofs emerge and reproduce via the activity and the interactions of mathematicians, I will consider them *ex post facto* as time-independent entities detached from their social dimension. I will thus represent conceptual change via a sequence of representative sets (cf. Toulmin 1972, p.p. 201-204). This choice allows me to present a simplified framework that steer clear from the discussed necessity of a truly Darwinian reproduction mechanism in cultural evolution (cf. Godfrey-Smith 2009, 2012; Charbonneau 2014). Despite these limitations, I will show in Section 5 the usefulness of my framework for analyzing mathematical conceptual histories. I will furthermore sketch some directions for extending my framework with a time-dimension in Section 2.5[12].

Central to my framework is Toulmin's (Toulmin, 1970) notion of a conceptual population, understood in my framework as a group of conceptual variants competing for similar (often the same) mathematical problems. Conceptual variants are understood in Mormann's and Toulmin's sense, i.e. as publicly existing specific versions/definitions of a mathematical concept. Mathematical problems are understood more generally than Mormann's proof-problems. They are abstract problems that can be instantiated by many

[12]Note that, despite my explicit commitment to an evolutionary account of conceptual change, my formal framework can also be given a deflationary non-evolutionary reading. Enthusiasts of a Fregean view of mathematical concepts can in fact have a deflationary reading of my framework, re-conceptualizing conceptual variation and evolution as change in the reference of mathematical terms (cf. Schlimm, 2012).

token-like specific questions[13]. Analogously to Toulmin's (Toulmin, 1972, pp. 173-189) pluralistic account of scientific problems, mathematical problems in my framework are not restricted only to searches for proofs but they can be problems of different sorts, such as classification problems (e.g. the search for ordering principles in nineteenth-century geometry), definitional problems (e.g. Hamilton's search for higher-complex numbers with a suitable geometrical and algebraic reading), and many others. Proof-problems such as Euler's conjecture, are only a proper subset of my understanding of mathematical problems[14]. Conceptual variants of the same conceptual population thus live in the same mathematical environment(s), i.e. they have to face similar mathematical problems, competing against each other for starring in successful solution-attempts, i.e. valid solutions.

This interaction between conceptual variants and mathematical problems produces a ranking of conceptual variants of a given conceptual population that I call *heuristic power*. The heuristic power of a given conceptual variant of a given conceptual population tracks the disposition of this variant to successfully interact with its environment, i.e. its propensity to figure in valid solutions of a given mathematical problem. The more promising a variant is, i.e. the higher its heuristic power or ranking in the population, the more likely this variant will become the accepted definition of the concept. Symmetrically, variants with low heuristic power are more likely to appear weird and artificial definitions of a given concept. The heuristic power can be thought as a kind of ordinal fitness ranking amongst variants of a mathematical concept, intuitively understandable as the propensity of a given variant of being used in fruitful solutions by mathematicians working to solve the given mathematical problem[15].

More precisely, any conceptual population (CP) is a triple $CP = \langle C, E, hp \rangle$, where $C = \{c_1, c_2, \ldots\}$ is the set of conceptual variants and $E = \{e_1, e_2, \ldots\}$ is the set of mathematical problems (i.e. the environment). Finally, hp is a heuristic power ordering of pairs of variants and mathematical problems ($C \times E$), representing the propensity of a given conceptual variant to successfully face a given problem. I allow the environment to change, losing old mathematical problems and acquiring new ones[16]. In this way we can define a conceptual history as a succession of conceptual populations $CH = \langle CP_1, \ldots, CP_n \rangle$ such that $\forall i, j$ $1 \leq i, j \leq n$ $C_i = C_j$ and $hp_i = hp_j$. In other words, a conceptual history is a succession of conceptual populations having the same set of conceptual variants and the same heuristic power ordering. The only component that is allowed to change is the set of problems. I call the set of environments of conceptual populations in a given conceptual history an

[13]Corfield, in his proposal of Neo-Lakatosian mathematical research programs (Corfield, 2003), also stresses the difference between abstract problems/questions and token-like proofs as central units of mathematical discovery.

[14]In what follows I will generally talk of mathematical problems (or just simply problems) and related solution-attempts. I will talk of proof-problems and related proof-attempts only in specific cases where it is clear that the mathematical problem in question constitutes a proof-problem.

[15]Even though biological fitness is usually measured on an absolute scale, it has been argued that an ordinal scale would suffice (Okasha, 2018, pp. 168-170).

[16]One could also allow the set of conceptual variants to vary at different stages, thereby having a sequence of changing set of variants. For simplicity, I don't allow it, assuming instead an *ex post facto* omniscience on all the variants appeared in a given conceptual history.

environmental history, i.e. $E_H = \{E_1, \ldots, E_n\}$. I require the environments of a given environmental history to exhibit at every stage a minimal degree of continuity, i.e. $\forall E_i\, 1 \leqslant i \leqslant n-1\, (E_i \cap E_{i+1} \neq \emptyset)$[17].

The heuristic power ordering *hp* is the pivotal component of a conceptual population. This ordering is in fact the purely normative selection mechanism of my evolutionary framework. Conceptual variants can be judged to have more or less heuristic power with respect to a specific mathematical problem, thereby having the propensity of being more or less fit for that environment. I am not going to define a single *hp* ordering that all conceptual population should use, but I will state some constraints that any *hp* ordering of any conceptual population must satisfy. In what follows I interpret these constrains on *hp* orderings as (an incomplete set of) rationality postulates that a mathematical agent ought to satisfy when selecting a given conceptual variant in relation to a given mathematical problem[18]. Formally, I require *hp* to be a partial ordering of pairs of conceptual variants and mathematical problems of a given conceptual population. I will call these pairs of variants and problems *conjecture-pairs*. The *hp* ordering should then be a reflexive and transitive ordering of conjecture-pairs. Not any partial ordering can be a proper *hp* ordering, though. Given two conceptual variants (c_1, c_2) (e.g. two of the many definitions of a polyhedron that Lakatos presents such as 'a surface consisting of a system of polygons' or 'simple polyhedra') and a mathematical problem e (e.g. Euler's conjecture) of a given conceptual population, the *hp* ordering of conjecture-pairs (representing the result of using a specific conceptual variant to try to solve a specific mathematical problem, e.g. Euler's conjecture for simple polyhedra) of the population has to satisfy the following rationality postulates:

- (COUNT): If the conjecture-pair (c_1, e) has strictly fewer counterexamples than the conjecture-pair (c_2, e), then $hp(c_1, e) > hp(c_2, e)$.

- (DOM): Provided that the conjecture-pairs (c_1, e) and (c_2, e) are equal in terms of counterexamples, if the conjecture-pair (c_1, e) has a bigger domain (i.e. more possible instances) than the conjecture-pair (c_2, e), then $hp(c_1, e) > hp(c_2, e)$.

- (REST): Provided that the conjecture-pairs (c_1, e) and (c_2, e) are equal in terms of counterexamples and domain-size, if the conjecture-pair (c_1, e) copes more successfully with the restricted cases of the mathematical problem e than the conjecture-pair (c_2, e), then $hp(c_1, e) > hp(c_2, e)$.

These three rationality postulates are inspired by Lakatos' heuristic strategies (see Section 2.1). They constrain the *hp* ordering and the related selection of conceptual variants

[17]Note that the continuity required here is only a local one, so that two non-successive environments of a given environmental history are allowed to have no proof-problem in common. If one thinks that this is not enough, one can impose a stronger, global continuity requiring the intersection of all environments of an environmental history to be non-empty.

[18]Note that this interpretation of these constraints as rationality postulates is quite natural from the perspective of my framework, but of course is not the only possible one.

of a given conceptual population to all orderings consistent with a notion of rationality inspired by Lakatos' fictional classroom discussion (Lakatos, 1976).

The COUNT principle tells us that the main rationale of the heuristic power ordering of conceptual variants with respect to a given mathematical problem is the number of known counterexamples to the related conjecture. Counterexamples, in line with the propensity reading of the *hp* ordering, should be thought as possible counterexamples, i.e. individual objects that represent possible "refutations" of a given conjecture-pair (i.e. a specific solution-attempt of a given problem for a given conceptual variant) such as the hollow cube in relation to the Euler's conjecture for the naive concept of a polyhedron (Lakatos, 1976, p. 14). In the case of mathematical problems that are not proof-problems, counterexamples are possible objects for which a given solution-attempt of the problem in question does not work (e.g. odd unclassifiable objects in a classification problem). The rationale of the COUNT principle lies behind Lakatos' method of lemma-incorporation and the related heuristics of proofs and refutations.

The DOM principle is inspired instead by Lakatos' appreciation of content-increasing methods. Provided that the number of counterexamples is equal between two conjecture-pairs, the pair with a bigger domain and thus a bigger content should be preferred. With the domain of a conjecture-pair I mean how general a given solution-attempt of a given mathematical problem is. The size of this domain can be measured by looking at the possible instances of the related conceptual variant, i.e. the individual objects in the extension of the specific definition of the concept. The more objects fall under a given conceptual variant definition, the bigger the domain of the related conjecture-pair.

Finally, the REST principle explicates one of the main strategies of Polya's mathematical heuristics (Polya, 2004), i.e. the division of mathematical problems into smaller problems with a restricted domain and the consequent bottom-up solution. Provided that both the number of counterexamples and the size of the domain are equal between two conjecture-pairs, the conjecture-pair that is more successful with restricted cases of the mathematical problem under focus ought to be preferred. A restricted case of a mathematical problem is a version of the problem the domain of which is a proper subset of the domain of the original problem, i.e. what is sometimes called a "special case" of a problem. Success with a restricted case of a problem means having no counterexamples within its restricted domain, i.e. solving a special case of the problem.

These three principles constrain the class of acceptable *hp* ordering of a given conceptual population[19]. Any specific heuristic power ordering of conjecture-pairs in a given conceptual population must then be consistent with the partial ordering(s) based on the number of counterexamples, the domain, and the successes with restricted cases of the problem that conjecture-pairs exhibit. Relative to a given problem, the conceptual variants of a given conceptual populations can be (partially) ordered in terms of how likely they are to figure in successful solution-attempts. The heuristic power ordering provides then a fully

[19]If we allow the set of conceptual variants to change, additional rationality postulates on these changes can be imposed. Examples of possible constraints are an anti-ad-hocness postulate of the kind Lakatos requires in his philosophy of science (Lakatos, 1978) and a menu-independence requirement analogous to the one championed by Sen in rational choice theory (Sen, 1997).

normative selection mechanism for conceptual variants (technically, for conjecture-pairs) of a given conceptual population.

Lakatosian populations and Euclidean populations

I have defined what a conceptual population is and I specified a set of rationality postulates constraining the acceptable class of heuristic power orderings of a given conceptual populations. Relative to a given environmental history, conceptual populations may exhibit different kinds of evolutionary dynamics, both with respect to the set of conceptual variants and the set of mathematical problems. Some kinds of dynamics make conceptual populations approximate Lakatos' ideal of proofs and refutations, while others are typical of populations quite different from Lakatos' examples, populations more closer to the Euclidean ideal of (absence of) change.

I will define then two different kinds of populations that a given conceptual population, relative to a given environmental history, can exemplify, namely Lakatosian and Euclidean populations:

> **Lakatosian Population**: a conceptual population (relative to a given environmental history) with *significant environmental stability*, in which there is *high variation* and *high reproductive competition* between the conceptual variants, which lead to differences in heuristic power *continuously* distributed.

> **Euclidean Population**: a conceptual population (relative to a given environmental history) with *significant environmental stability*, in which there is *low variation* and *low reproductive competition* between the conceptual variants, which lead to differences in heuristic power *discretely* distributed.

These two types of conceptual populations are defined around four notions: environmental stability, conceptual variation, reproductive competition, and continuity in the distribution of heuristic power. These four notions denote four different aspects of the evolutionary dynamics of conceptual populations. The environmental stability of a conceptual population denotes how stable problems of a given conceptual populations are in a given environmental history. Conceptual variation denotes instead the amount of variation amongst conceptual variants of a given population. Reproductive competition denotes the extent to which conceptual variants of a population are competing for the same problem. Finally, the continuity in the distribution of heuristic power denotes, in analogy with the concept of fitness-landscape (Wright, 1932) in evolutionary biology, whether similar conceptual variants of a given population have similar heuristic power. If this condition occurs, I will say that a given conceptual population has a continuous distribution of heuristic power; otherwise I will call that distribution discrete.

Lakatosian populations are then conceptual populations in which significant environmental stability goes together with high conceptual variation, high reproductive competition, and a continuous distribution of heuristic power. This combination of these four aspects makes the evolutionary dynamics of a conceptual population approach the ideal

behind Lakatos' concept-stretching. Examples of Lakatosian populations are Lakatos' own case studies, i.e. the polyhedron population and the continuity population (Lakatos, 1976, 1978). As it was stressed by many critics (Feferman, 1978; Fine, 1978; Corfield, 2003; Werndl, 2009), Lakatos' model of conceptual change seems to implicitly require conditions specific to a certain type of mathematical conceptual histories, of which Euler's conjecture is a paradigmatic example. Lakatos' dance of proof-attempts and counter-examples requires a plethora of different definitions of the concept under focus, competing against each other with the aim of solving the same mathematical problem. This means that conceptual populations must have a specific combination of variation in conceptual variants and mathematical problems in order to approximate Lakatos' ideal of conceptual change. From the perspective of my framework, then, Lakatosian populations are conceptual populations with a lot of different conceptual variants together with a stable, almost fixed, environment. When these two conditions are met I will talk of a conceptual population having, relative to a given environmental history, high conceptual variation and high environmental stability. Furthermore, Lakatosian populations exhibit also high reproductive competition amongst the variants, i.e. the variants of a population are not just competing for similar problems, but for the same one(s). In other words, there have to be many conceptual variants and few problems. Finally, Lakatosian populations enjoy a continuous distribution of heuristic power, i.e. similar conceptual variants have similar heuristic power. This continuity, as we will see in the next section, is connected with a lack of axiomatization of the conceptual histories so represented.

Euclidean populations are instead conceptual populations in which significant environmental stability goes together with low conceptual variation, low reproductive competition, and a discrete distribution of heuristic power. This combination makes the evolutionary dynamics of a conceptual population approach the ideal of Euclidean absence of conceptual change. An example of a Euclidean population could be the concept of natural number, which seems to be one of the most stable concepts in the history of mathematics. A concept whose evolution does not seem to involve any form of conjecture and refutations whatsoever, but just a series of rigorizations and conceptual analyses of a well-understood concept. This stability in conceptual evolution requires a conceptual population to have a low conceptual variation with a very stable environment. Thus, just like Lakatosian populations, Euclidean populations exhibit high environmental stability, but they have a completely opposite environmental dynamics than the Lakatosian ones with respect to the other three aspects under consideration. Euclidean populations have in fact a low conceptual variation and a low reproductive competition. In these population, there are not a lot of variants for the same concept and the existing variants are often meant to tackle different problems, thereby not really competing with each other. Finally, Euclidean populations exhibit a discrete distribution of heuristic power, a typical property of axiomatized mathematics in which small variation may cause significant differences in heuristic power.

Lakatosian populations and Euclidean populations are then two very different kind of conceptual populations, respectively describing almost opposite evolutionary dynamics. In a Lakatosian population lots of conceptual variants with similar heuristic power compete against each other for the same problem(s). In a Euclidean population, instead, few variants

with different heuristic power cope with different problems.

Lakatosian space

I will now add more structure to my framework and specifically to the opposition between Lakatosian and Euclidean populations. I will present four parameters that track the degree to which a conceptual population (with respect to a given environmental history) exhibits one of the four aspects through which I discussed Lakatosian and Euclidean populations, i.e. conceptual variation, reproductive competition, environmental stability, and continuity in the heuristic power distribution. These four parameters can be understood as four dimensions constituting the *Lakatosian space*. Points in this space are possible combinations of these four parameters, representing a possible kind of conceptual population relative to a given environmental history. Conceptual populations exhibiting the same kinds of evolutionary dynamics, i.e. having the same combination of these four parameters, occupy the same point of the Lakatosian space. This additional spatial structure provides my framework with a fine-grained way of measuring to which degree a given conceptual population is Lakatosian or Euclidean with respect to one of these four parameters[20].

Let us survey the four parameters constituting the dimensions of the Lakatosian space, one by one.

Conceptual Variation (CV): This parameter represents the amount of variation amongst the conceptual variants of a given conceptual population. It can be measured tracking the number of elements in the set C, i.e. $CV = |C|$. It classifies conceptual histories based on how many variants of a concept they exhibit. Conceptual populations with high CV are representing (parts of) mathematical conceptual histories in which many possible definitions of a concepts are proposed and discussed. This situation is typical of stages of generalization of accepted concepts, where several properties of the concept in the wider context are open to discussion (cf. Waismann, 1936), such as the case of the quaternions (Hamilton, 1853). Conceptual populations with low CV represent instead (parts of) mathematical conceptual histories in which a (group of) definition(s) is accepted and therefore not truly questioned. This situation is typical of periods in the history of a given mathematical field in which a natural or a very fruitful definition of a concept is found (Tappenden, 2008a,b) and, using a game-theoretic notion, it becomes evolutionary stable against mutations (Weibull, 1995). An example of this situation is the abstract concept of group.

Reproductive Competition (RC): This parameter tracks the extent to which con-

[20]In what follows, I will give possible measures for the parameters on an absolute scale. This could create a measurement problem in applying my framework to historical case studies. Depending on how many conceptual variants and mathematical problems one identifies in reconstructing a given case, the measures of the Lakatosian space parameters may change. In all the case studies I will present in the next section it seems sufficiently clear what the proper choices of conceptual variants and mathematical problems are, so I will freely use absolute measures for my parameters, leaving this measurement problem as an open issue for further work.

ceptual variants have to compete for the same mathematical problems, i.e. how much different definitions of a given concept have to 'fight against each other to survive'. It can be measured by the ratio between the number of conceptual variants and the number of mathematical problems of a given conceptual population: $RC = \frac{|C|}{|\cup E_H|}$. The higher the ratio, the more competitive the conceptual population is. The lower the ratio gets, instead, the less a given conceptual population resembles a 'struggle for existence'. Populations with low RC often have an environment made of several mathematical problems for which 'specialized' variants evolved in parallel, defusing the struggle amongst the variants. This situation is typical of (periods of) bodies of mathematics in which the conceptual variants are very well adapted to specific problems, such as the many pre-calculus 'analysis' techniques. Populations with high RC have instead many conceptual variants competing for the same problem(s). This situation is typical of (periods of) bodies of mathematics centered around a general problem, such as the Newtonian and Leibnizian calculus.

Environmental Stability (ES): This parameter tracks how much historical problem-solving continuity a given conceptual history exhibits, i.e. how stable is the environment faced by a given conceptual population, relative to a given environmental history. As a measure, we can take the ratio between the intersection of all the stages of an environmental history and their union: $ES = \frac{|\cap E_H|}{|\cup E_H|}$. The higher the ratio, the stabler a given environmental history is. The lower the ratio gets, the more dynamic and revolutionary the history of mathematical problems faced by a given conceptual population is. A high degree of environmental stability is common to many different mathematical conceptual histories, up to the point that it is the only feature that Lakatosian and Euclidean populations have in common. Examples of conceptual population exhibiting high ES include very different conceptual histories such as natural numbers, the polyhedron concept, and the quaternions. A low degree of environmental stability is instead usually connected with a highly formalized body of mathematics, due to the de-semantification or topic-neutral effect of formalization stressed by several philosophers (MacFarlane, 2000; Dutilh Novaes, 2012). In a formalized (part of a) conceptual history, concepts and methods can in fact be quickly applied to different problems, giving rise to very dynamic environmental histories. Examples of conceptual populations exhibiting low ES can be found in formal bodies of mathematics such as vector algebra or the study of partial differential equations.

Continuity ($Cont$): This parameter tests whether the distribution of heuristic power amongst conceptual variants of a given population is continuous, i.e. whether similar variants have a similar heuristic power. A possible "measure" for $Cont$ is to see whether it holds in a given population that $\forall c_1, c_2 \in C, \forall e_1 \in E(c_1 \approx c_2 \rightarrow hp(c_1, e_1) \approx hp(c_2, e_1))$, where $x \approx y$ is an intuitive similarity relation between concepts[21]. While the other three dimensions of the Lakatosian space are measured on richer scales, continuity is a boolean parameter, i.e. either a conceptual population exhibits continuity or it does not. In analogy to the fitness-landscape biological metaphor, if a given conceptual population exhibits

[21]This is done for simplicity reasons. There are many possible frameworks for cashing-out a notion of similarity between concepts. For instance, conceptual spaces theory (cf. Ch. 3, Sect. 4.1) could be used.

continuity, then its heuristic power landscape is smooth. Otherwise, the distribution of heuristic power amongst the variants of the population is somewhat discrete, small changes in the definition of a variants can lead to enormous differences in terms of fruitfulness. This continuity (or the lack thereof) in the distribution of heuristic power is connected with the degree of axiomatization of the related body of mathematics. Axiomatized bodies of mathematics constrain in fact the possible choices of conceptual variants to the ones available by the tinkering of the axioms (Schlimm, 2013). Small variations in a given axiom may have then enormous repercussions on the heuristic power of the conceptual variants so defined[22]. Highly axiomatized bodies of mathematics typically exhibit therefore a discrete distribution of heuristic power, while conceptual histories that are not (fully) axiomatized enjoy a continuous one. Examples of the former kind of conceptual populations are the quaternions or the abstract group concept, whereas the pre-abstract group concepts exemplify the latter.

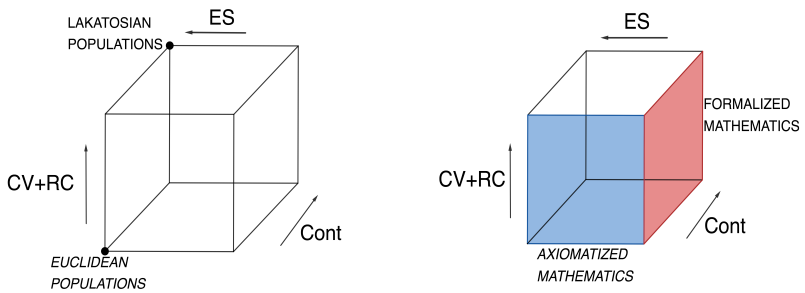
The four parameters then can be understood as the four dimensions of the Lakatosian space. We can then assign to both Lakatosian and Euclidean populations a given region of the Lakatosian space (Figure 1, Panel a).

The Lakatosian space can also offer a new explication (in Carnap's sense of the term, cf. Chapter 3) of concepts related to certain kinds of evolutionary dynamics that conceptual populations exhibit. For instance, I stressed how a low ES is connected to heavily formalized bodies of mathematics, whereas a lack of $Cont$ is a symptom of a high degree of axiomatization. We can then understand the related faces of the Lakatosian space, the one corresponding to low ES and the one corresponding to lack of $Cont$, as extensions of the notion of (respectively) formalization and axiomatization in mathematics (Figure 1, Panel b). Note that usually both axiomatization and formalization are defined relative to some properties of the language in which a given body of mathematics is developed. My framework explicates instead both notions independently from the language, focusing on how these two notions shape the evolutionary dynamics of the related conceptual populations. In this way, understanding formalization as low ES and axiomatization as lack of $Cont$ provides us with a novel perspective on these two concepts.

Inter-population dynamics: modeling the evolutionary history of mathematical concepts

We have seen how, thanks to the four dimensions of the Lakatosian space, my framework is able to distinguish the different evolutionary dynamics of conceptual populations in mathematics. The normative selection mechanism of my framework, based on a suitable heuristic power ordering, can then be applied to explain the mathematical selection inside a given conceptual population. As we will see in the next section, this combination of

[22]This phenomenon is reminiscent of the famous butterfly effect in chaos theory. This is not a coincidence, since one of the well-known consequences of axiomatization is the hierarchical organization of the axiomatized subject into an inter-connected system of knowledge. Inter-connected systems of all kinds are more prone to butterfly effects.



(a) (b)

Figure 4.1: Two three-dimensional representations of the Lakatosian space, where parameters CV and RC are collapsed into one dimension for representational purposes. Panel (a) shows the parts of the space corresponding to Lakatosian and Euclidean populations, while panel (b) shows the parts corresponding to heavily formalized and highly axiomatized bodies of mathematics.

a simple selection mechanism and a plurality of possible evolutionary dynamics makes my framework adequate to the plurality of dynamics exhibited by historical episodes of conceptual change in mathematics.

Until now, we focused on what happens within a given conceptual population, specifically on the normative mechanism behind mathematical selection and on the types of variation and environment. We can refer to these kinds of changes as intra-population changes. As Toulmin stresses, however, such intra-population changes are only a proper part of the changes that an evolutionary model of scientific conceptual change ought to explain. A full evolutionary model of conceptual change has in fact to take into account also the inter-population changes, i.e. the transitions from one conceptual population to another one. In other words, a conceptual population intuitively represents the status of the related scientific concept at a given time of its history. The history of scientific concepts can then be thought as a succession of time-slices, each of which correspond to a given conceptual population. Inter-population changes are thus the transitions. in the history of a given scientific concept, from one conceptual population to the next one.

In order to track these inter-population changes, the four-dimensional Lakatosian space that I presented is not enough. In order to model these transitions in the evolution of mathematical concepts, we have in fact to add another dimension to the Lakatosian space, a time dimension T. I will call this five-dimensional version of the Lakatosian space, the *augmented Lakatosian space*.

The time dimension is different from the others four Lakatosian ones. It does not in fact track variational features of a given conceptual population, but it tracks instead the passage of time in the evolution of a given mathematical concept. It should be thought as a discrete axis, every point t of which corresponds to a given stage in the evolution of the concept under focus. Every stage corresponds to a given conceptual population, representing the environmental dynamics exhibited by a given concept at that point of its

history. I will call this succession of conceptual populations $EV_X = \langle CP_{t_1}, \ldots, CP_{t_n} \rangle$ of a given mathematical concept X the *evolutionary history of* X. Every conceptual population within the evolutionary history of a concept can be more or less Lakatosian (with respect to a given environmental history) in its specific kind of variation, occupying thus a different part of the four-dimensional Lakatosian space. So that a given evolutionary history can be thought as a trajectory in the four-dimensional Lakatosian space. Movement from one conceptual population CP_{t_x} to the next one $CP_{t_{x+1}}$ can be represented as movement along the four dimensions of the original Lakatosian space, each of which can be given an intuitive reading in terms of what possibly happens at the mathematical theories under focus:

- **CV-axis** (Conceptual Variation). Movement along this axis could reflect change in the conceptual variation rate in the history of a given mathematical concept. As Toulmin (Toulmin, 1972, pp. 210-222) stressed, scientific concepts arguably exhibit different rates of conceptual variation at different times. As we saw in Section 2.3.3, the parameter CV of the Lakatosian space tracks the amount of conceptual variation in a conceptual population. The conceptual variation rate seems to be connected to the degree to which a 'natural' definition of a given mathematical concept is accepted. Populations with high CV are then representing stages of the evolutionary history of a concept in which many possible definitions of it are proposed and discussed, while populations with low CV track stages where a (group of) definition(s) is accepted. Change in CV could be correlated to changes in the acceptability of a definition for the concept under focus. Transitions from a conceptual population with high CV to one with low CV could represent the achieved selection of a (group of) natural and/or very fruitful definition(s) of a concept, determining a low variation rate in the subsequent conceptual populations. We can call this type of transition *conceptual stabilization*, in analogy with the analogous concept of evolutionary stability in evolutionary game theory (Weibull, 1995). Opposite transitions, from a conceptual population with low CV to one exhibiting high CV, could instead reflect stages of generalization of accepted concepts, where several properties of the concept in the wider context are open to discussion (Waismann, 1936). We can refer to this second type of transitions as *conceptual generalization*. An example of conceptual stabilization is the step from the pre-abstract group population to the abstract group population, whereas the transition from complex numbers to the quaternion population exemplifies conceptual generalization.

- **RC-axis** (Reproductive Competition). Movement along this axis can be understood directly as change in the degree of reproductive competition of a given mathematical concept and indirectly as change in the degree of conceptualization or abstraction of the related mathematical theory. In fact, the parameter RC of the Lakatosian space tracks the degree of reproductive competition between variants of a given conceptual population. As we saw in presentation of my basic framework, populations with low RC often have an environment made of several mathematical problems for which 'specialized' variants evolved in parallel, defusing the struggle amongst the variants.

Populations with high RC, instead, tend to have many conceptual variants competing for one, general problem. Change in the degree of reproductive competition could thus represent changes in the degree of abstraction of the related mathematical environment. More specifically a transition from a conceptual history with low RC to one with high RC could be due to a re-systematization or a re-conceptualization of the related body of mathematics where several specific mathematical problems are subsumed under a single general problem. We can call this type of transition *environmental generalization*. The opposite type of transition, i.e. one from a conceptual history with high RC to one with low RC, could be due to a deconstruction of the related body of mathematics in which one single general problem gets substituted by several more specific ones. We can call this type of transition *environmental specialization*. A paradigmatic example of environmental generalization is the shift from the very specialized pre-calculus 'analysis' techniques to the Newtonian and Leibnizian calculus (Kitcher, 1984, pp. 230-236).

- **ES-axis** (Environmental Stability). Movement along this axis could represent change in the degree of formalization of the related mathematical theory. We saw in fact that the parameter ES of the Lakatosian space measures the degree of environmental stability of a given environmental history. I already stressed that a low degree of environmental stability is usually connected with an highly formalized body of mathematics, due to the so-called de-semantification effect of formalization (cf. MacFarlane 2000; Dutilh Novaes 2012). Change of ES could then be due to change in the degree of formalization of the related body of mathematics. Specifically, a transition from a conceptual history with high ES to one with low ES could reflect a formalization of a previously non formal mathematical theory. We can refer to such a transition as *formalization*. As a paradigmatic example of formalization we can take the transition from the polyhedron population to the definition of polyhedron in vector algebra discussed also by Lakatos (Lakatos, 1976).

- **Cont-axis** (Continuity). Movement along this axis can be understood as change in the degree of axiomatization of the related body of mathematics. The discrete parameter $Cont$ of the Lakatosian space tracks in fact whether a given conceptual population exhibits a continuous or a discrete distribution of heuristic power amongst its variants. I already stressed how a low $Cont$ is typical of highly axiomatized bodies of mathematics, while a high $Cont$ usually denotes non axiomatized theories. Change of $Cont$ could then be due to change in the degree of axiomatization of the related mathematical theory. Specifically, a transition from a conceptual population exhibiting high $Cont$ to one exhibiting low $Cont$ could reflect an axiomatization of a given part of mathematics which was not yet axiomatized. We can refer to such a transition as *axiomatization*. A paradigmatic example of axiomatization is the step from the pre-abstract group population (which, as we will see in the next section, exhibits continuity) to the population reconstructing abstract group theory (which arguably lacks continuity).

Transitions from a given conceptual history to its historical successor can also involve changes in more than one axis Lakatosian parameters and they can be represented in the augmented Lakatosian space as complex multi-axes movements. In the philosophical literature, such patterns of conceptual changes have been conceptualized in terms of inter-practice transitions (Kitcher, 1984; Ferreirós, 2015). Practice-based frameworks are of course far richer than the present one, including many non-conceptual components of a mathematical practice that are abstracted away in my Lakatosian space such as language, reasoning tools, meta-mathematical views, agents, and others. Nevertheless, inter-practice transitions can be understood as specific kinds of movements in the augmented Lakatosian space along the T axis.

Take for instance Kitcher's very fine-grained classification of inter-practice transitions (Kitcher, 1984, pp. 194-228). He distinguished five kinds of rational inter-practice transitions: question-generation, question-answering, generalization, rigorization, and systematization. Amongst these kinds of transition, the first two could correspond to changes in the mathematical problems composing the environment of a given environmental history. The emergence of a new problem in the mathematical environment could mirror Kitcher's "question-generation" transition, while the disappearance of a problem may correspond to "question-answering". These transitions often imply also changes in the set of conceptual variants of a given Lakatosian population that can be interpreted as symptoms of transformations in the language and statements of the related practice. Regarding the third kind of transition described by Kitcher, in the augmented Lakatosian space Kitcher's generalization transition correspond to what I called conceptual generalization, i.e. a change from a conceptual population exhibiting low CV to one with an high CV[23]. Kitcher's rigorization transition can be instead represented by a movements along the ES-axis of the augmented Lakatosian space. In particular, the specific kind of rigorization corresponding to the formalization of the related body of mathematics can be mirrored by what I have called formalization, i.e. a change from a conceptual population with high ES to one with low ES. Finally, what Kitcher calls "systematization" could be mirrored by movements along the $Cont$ and the RC axes, involving what I have called axiomatization, environmental generalization, or environmental specialization (or a combination of these changes) of the related conceptual populations.

We can now appreciate how my framework is able to represent all five kinds of Kitcher's inter-practice transitions from a purely conceptual point of view. Kitcher's case study on the evolution of analysis (Kitcher, 1984, pp. 229-271) can then be reconstructed as a succession of conceptual populations, from pre-calculus techniques up to fully arithmetized analysis, each of which exhibits a given tuning of the four Lakatosian parameters and therefore occupies a given region of the Lakatosian space.

Let me sketch how such a reconstruction might look like. Kitcher recognize 5 different steps in its (mini-)history of calculus: pre-calculus techniques (Kitcher, 1984, pp.

[23]Note that the hp ordering of a given conceptual population makes possible also to distinguish specific kinds of generalizations, such as single-concept-generalization (where the generalized problem is the application of the selected conceptual variant of the first population into a broader domain) or multi-conceptual generalizations (where more than one variant gets generalized).

230-231), Netwonian and Leibnizian calculus (Kitcher, 1984, pp. 231-241), Euler and 18th century analysis (Kitcher, 1984, pp. 241-245), Cauchy and the early 19th century analysis (Kitcher, 1984, pp. 245-253), and finally Weierstrass' and Dedekind's rigorous analysis of the late 19th century (Kitcher, 1984, pp. 263-268). Each of this step corresponds naturally in my framework to a given conceptual population, containing all the conceptual variants and the mathematical problems at work in that moment of the history of analysis, together with a suitable heuristic power ordering. Kitcher's case study can then be seen as reconstructing the following evolutionary history of analysis: $EV_{an} = \langle CP_{precalc}, CP_{Newt-Leib}, CP_{Euler}, CP_{Cauchy}, CP_{Weier-Dedek} \rangle$, where $CP_{precalc}$ is the conceptual population reconstructing the pre-calculus techniques, $CP_{Newt-Leib}$ the population reconstructing the Newtonian and Leibnizian calculus, CP_{Euler} the population reconstructing 18th century analysis, CP_{Cauchy} the population reconstructing early 19th century analysis, and $CP_{Weier-Dedek}$ the population reconstructing late 19th century analysis. We can assume that these five conceptual population are ordered by the time-dimension T of the augmented Lakatosian space in the following way: $CP_{precalc_{t1}} < CP_{Newt-Leib_{t2}} < CP_{Euler_{t3}} < CP_{Cauchy_{t4}} < CP_{Weier-Dedek_{t5}}$. The full reconstruction of Kitcher's case study would continue by reconstructing completely the five conceptual populations, in all details, and classifying their evolutionary dynamics along the four dimensions of the original Lakatosian space. After this crucial step, Kitcher's inter-practice transitions ,governing the step from one moment of its history of analysis to the next one, could be reconstructed as movements along the axis of the augmented Lakatosian space in the following way:

- Kitcher's 1st transition: the first transition of Kitcher's history of analysis is the one from the pre-calculus techniques to the Newtonian and Leibnizian calculus. Kitcher (Kitcher, 1984, p. 270) conceptualizes this transition as driven by the systematization of previous problems. In my framework, then, we can understand this first transition as an example of what I have called environmental generalization, i.e. a change from a conceptual population exhibiting low RC, such as the pre-calculus one, to a conceptual population with high RC like the population reconstructing the Newtonian and Leibnizian calculus.

- Kitcher's 2nd transition: the second transition in Kitcher's case study is the one from the Newtonian and Leibnizian calculus to the 18th century analysis of Euler's and his contemporaries. Kitcher (Kitcher, 1984, p. 241,242,270) describes this transition as the result of a generalization in the calculus technique, together with the appearance of several new problems. In my framework this complex transition can be mirrored by a double movement along the CV and the RC axis, i.e. by the combination of a conceptual generalization and an environmental specification. The movement from the Newtonian-Leibnizian calculus population to the conceptual population reconstructing 18th century analysis is then the movement from a conceptual population exhibiting high RC and low CV to a conceptual population exhibiting low RC and high CV.

- Kitcher's 3rd transition: the third transition described by Kitcher is the one from

the 18th century analysis to the early 19th century analysis of Cauchy and his contemporaries. Kitcher (Kitcher, 1984, p. 247,248,270) depicts this transition as an example of systematization of a body of mathematics. Specifically, he stresses that this systematization is what drives Cauchy's foundational worries about making analysis more rigorous. In my framework, this transition is then an example of what I called environmental generalization, i.e. a step from a conceptual population with several mathematical problems to a conceptual population related to some central and general problems. The movement from the 18th century analysis conceptual population to the conceptual population representing early 19th century analysis is then a movement from a population exhibiting low RC to a population exhibiting high RC.

- Kitcher's 4th transition: the fourth and final transition of Kitcher's case study is the one from Cauchy's early 19th century analysis to the late 19th century analysis of Weierstrass and Dedekind. Kitcher (Kitcher, 1984, pp. 254-262,271) conceptualizes this transition as a typical example of rigorization by formalization. In my framework, we can also understand this transition as an example of (what I call) formalization, i.e. a step from a conceptual population related to a non-formalized body of mathematics to a population representing a formalized mathematical theory. The movement from early 19th century analysis conceptual population to the conceptual population representing late 19th century analysis is then a movement from a population exhibiting $Cont$, i.e. having a continuous distribution of heuristic power amongst its variants, to a population lacking $Cont$, i.e. exhibiting a discrete distribution of heuristic power.

Analogously with Kitcher's inter-practice transitions, my framework is able to model much of Ferreirós' recent proposal of an interplay of practices (at least the part concerning his 'theoretical' framework) and its systematic links between one practice and another one. An evolutionary history analogous to the one I just sketched for Kitcher's case study could arguably rationally reconstruct one of Ferreirós' case studies such as the evolution of real numbers (Ferreirós, 2015, pp. 206-246).

Furthermore, the augmented Lakatosian space could be able to describe more coarse-grained dynamics of mathematical evolution. Major historical transitions in philosophical views of the mathematical community could for instance be represented as major changes of population density within the Lakatosian space. Take for instance the so-called 'structuralist turn' of late 19th century mathematics. The new attention to systematization and rigor of mathematical theories could be represented as a major change of population density along the $Cont$ and the ES axes of my framework. A plethora of evolutionary histories of concepts from different parts of mathematics would have transitioned around that time from conceptual histories exhibiting $Cont$ and high ES, to conceptual histories with a lack of $Cont$ and low ES, mirroring the axiomatization and formalization of the related mathematical theories. In the same way, one could also think about foundational enterprises in the history of mathematics as changes in the population-density inside the

Lakatosian space. A given foundation would then 'drag' more and more evolutionary histories to the specific part of Lakatosian space inhabited by the main concepts belonging to the foundational theory, with a sort of quicksand or black-hole effect.

4.3 Applications of the Framework: Three Cases of Mathematical Selection

In the last section, I presented my evolutionary framework for conceptual change in mathematics. We have seen how my framework is centered around the notion of a conceptual population, understood as a set of conceptual variants coping with a set of mathematical problems and partially ordered by a given heuristic power ordering. I defined two different kinds of conceptual populations, Lakatosian and Euclidean populations, corresponding to two opposite ideals of evolutionary dynamics that a given conceptual population may exhibit. I augmented my framework with a four-dimensional space, i.e. the Lakatosian space, the dimensions of which correspond to parameters tracking four different aspects of the evolutionary dynamic of a given conceptual population. Thanks to these spatial tools, Lakatosian and Euclidean populations can be understood as different regions of the Lakatosian space and important concepts in philosophy of mathematics such as axiomatization and formalization can be given a novel characterization as certain parts of the space. Moreover, we saw how the Lakatosian space, when augmented with a time-dimension, is able to reconstruct the whole evolutionary history of a mathematical concept as a succession of different conceptual population. I showed also how this extension to my basic frameork allow us to mirror also fine-grained inter-population kinds of conceptual changes, such as Kitcher's inter-practice transitions, as specific types of movement along the axis of the augmented Lakatosian space.

In this section, I will show how my framework can be applied to historical cases of conceptual change in mathematics. More specifically, I will show how a mathematical conceptual history can be represented in my framework as a conceptual population and how its evolutionary dynamic can be judged to be a case of mathematical selection or evolutionary drift. The Lakatosian space then becomes a conceptual space for classifying episodes of the history of mathematics in terms of the evolutionary features exhibited by their rationally reconstructed conceptual population. I will then demonstrate how, thanks to the four dimensions of the Lakatosian space, my framework is able to provide a rich understanding of the evolutionary dynamic of a given episode of conceptual change. In order to achieve this, I will analyze three case studies: Lakatos' own example of Euler's conjecture and the concept of polyhedron (Lakatos, 1976), Hamilton's invention of the quaternions (Hamilton, 1843a,b, 1853; Pickering, 1995), and the pre-abstract group concepts (Wussing, 1984).

In Section 3.1, I will present the distinction between mathematical selection and evolutionary drift, in the light of which episodes of conceptual change from the history of mathematics can be judged to be more or less rational. In Section 3.2, I will present my first case study, namely the polyhedron population and the related history of Euler's

conjecture. In Section 3.3, I will reconstruct in my framework the conceptual history of the quaternion population. In Section 3.4, I will analyze my third case study, namely the pre-abstract group population. I will then draw some general conclusions about what my proposal achieves and sketch some possible directions for future work. Finally, in Section 3.5 I will present two formal applications of my framework. Specifically, in section 3.5.1, I will present a toy-model of my framework in which I will construct an actual heuristic power function in order to elucidate how my selection mechanism works. In Section 3.5.2, I will formalize my three case studies in the language of my framework in order to show how the four parameters of the Lakatosian space can be measured.

4.3.1 Mathematical selection and evolutionary drift

Applying my framework to a mathematical conceptual history involves, as its first step, to rationally reconstruct the actual history of the mathematical concept under focus as a conceptual population. After this step, the rationality of the case of conceptual change under focus can be assessed by checking whether the preference induced by the heuristic power ordering of the conceptual population is consistent with the actual choices of the mathematical community.

The rationality postulates (cf. Section 2.3.1) constraining any heuristic power ordering are fully normative and therefore my framework has a purely normative selection mechanism. As such, actual history of mathematics does not always have to follow the preference order given by these postulates. There may be some historical cases in which mathematicians selected concepts with an equal or even lower heuristic power than their competitors. In these cases, there could have been sociological or psychological factors that caused the heuristic power selection to be overturned. The heuristic power of a given concept can be easily overshadowed by a lack of familiarity with the specific mathematics connected to it, metaphysical prejudices over what a concept should or should not be, several psychological biases towards who supports a given concept, and many other similar factors. We should not forget that mathematics is a human activity. Just like in natural history, even in the history of mathematics cases of evolutionary drift are indeed present.

In evolutionary biology, the concept of evolutionary drift (also known as genetic drift or random drift, cf. Millstein, 2002; Plutynski, 2007) can be traced back to Darwin's talk of non-useful variations in his presentation of natural selection. Evolutionary drift is the chance element within evolutionary biology, usually negatively defined as the variations complementary to the ones selected by natural selection. That is, evolutionary drift is the process behind all the side-products of evolution the survival of which cannot be explained by fitness differences. Given the vagueness of this negative definition I just presented, it should not come as a surprise that the exact nature and significance of evolutionary drift has been a highly debated topic in philosophy of biology. Philosophers have in fact discussed at length what exactly evolutionary drift is (Plutynski, 2007; Millstein, 2017), how big a role does it play in respect to natural selection (Kimura, 1983; Mayr, 1983), which evolutionary factors are most correlated to it (Godfrey-Smith, 2009), and which kind of interpretation of evolutionary theory explains it best (Sober, 1984; Walsh, Lewens,

and Ariew, 2002). What is accepted by everyone about evolutionary drift is its constant presence in evolutionary theory practice and its usefulness as a conceptual complement to natural selection in evolutionary models (Binmore and Samuelson, 1999; Brandon, 2006).

In my use of evolutionary drift, I will steer as clear as possible from these philosophical debates, taking this concept to somehow denote the chance element, complementary to the workings of natural selection, responsible for all cases of non-fitness-related selections in evolution. Thus, I will talk of evolutionary drift in the context of my framework to denote all the cases of conceptual change in mathematics the selection process of which significantly involved non-heuristic-power-related considerations. In this way, I take evolutionary drift to constitute the chance factor in the history of mathematics complementary to the workings of mathematical selection.

More precisely, if a certain case of conceptual change of mathematics is consistent with the preference(s) induced by the heuristic power ordering of the conceptual population that reconstructs it, i.e. the actual concept(s) selected is the one(s) having the higher heuristic power, I will classify it as an example of mathematical selection. More specifically, I will distinguish two kinds of mathematical selection, global and local one. When this selection happens with respect to the whole environment, that is when a given concept gets selected amongst all other variants of a given conceptual population, I will talk of *global mathematical selection* or mathematical selection *tout court*. When, instead, different mathematical problems in a given environment select different conceptual variants, that is when there is no general selection of a variant but only specific selections relative to (a group of) problem(s), I will talk of *local mathematical selection*. If, instead, a given episode of conceptual change is inconsistent with the preference(s) induced by the heuristic power ordering of the related conceptual population, i.e. the actual concept(s) selected is not the one(s) having the higher heuristic power, both in relation to local and global selection, I will say that it is a case of *evolutionary drift*.

My framework gives then a novel perspective on whether mathematical conceptual change is a rational process. In my proposal, in fact, general normative rationality postulates (i.e. the postulates constraining any *hp* ordering) on conceptual selection relative to a given (set of) mathematical problem(s) are able to retrospectively assess the rationality of mathematical conceptual histories without imposing any specific constrain on their evolutionary dynamics. In this way, the emergence of mathematical concepts and problems is entirely left open to the abilities and the values of the mathematical community, but once the pool of concepts and related problems is fixed a normative rational ordering of which concept ought to have been selected in relation to a specific mathematical problems can be given. Based on this ordering, conceptual histories can be judged to be examples of mathematical selection (global and local) or evolutionary drift. This relationship between rationality and freedom in mathematical conceptual change, i.e. a selection mechanism related to fixed variants and environment that coexists with the absence of any absolute norm on the evolution of these variants and environments, is typical of the EET program view of scientific evolution, especially in its instantiations that are closer to Darwin's population thinking (cf. Mayr, 1975; Sober, 1980) such as Toulmin's one (cf. Section 1.1).

In the next subsections, I will show how thank to these notions the rationality of

mathematical conceptual histories can be assessed. In the three case studies that I will present in this work I chose, for simplicity, to treat only cases of conceptual changes that are examples of mathematical selection (both local and global), leaving the study of cases of evolutionary drift in mathematics for future work.

4.3.2 Lakatos' polyhedron example

Lakatos' own master example of the dynamic of proofs and refutations lends itself very naturally to being transformed into a conceptual population. In Lakatos' book (Lakatos, 1976), one can find no less than fifteen different definitions of a polyhedron, all facing the same proof-problem of Euler's conjecture. I am going to follow Lakatos' fictional classroom-discussion of this example, pointing the reader to Lakatos' text for historical references.

As the first conceptual variant (p_1) of our polyhedron population we can take the naive concept of polyhedron which is used in the proof-experiment of Euler's conjecture (Lakatos, 1976, pp. 6-10).

Seven other definitions of a polyhedron are obtained via the method of monster-barring, i.e. the ad hoc redefinition of a concept for excluding counterexamples: (p_2) "a solid whose surface consists of polygonal faces" (Lakatos, 1976, p. 15), (p_3) "a surface consisting of a systems of polygons" (Lakatos, 1976, p. 16), (p_4) "a system of polygons arranged in such a way that (1) exactly two polygons meet at every edge and (2) it is possible to get from the inside of any polygon to the inside of any other polygon by a route which never crosses any edge at a vertex" (Lakatos, 1976, p. 17), (p_5) "a system of edges arranged in such a way that exactly two edges meet at every vertex" (Lakatos, 1976, p. 19), (p_6) definition of p_5 plus the further condition that "the edges have no points in common except the vertices" (Lakatos, 1976, p. 19), (p_7) "in the case of a genuine polyhedron, through any arbitrary point in space there will be at least one plane whose cross-section with the polyhedron will consist of one single polygon" (Lakatos, 1976, p. 23), (p_8) "edges have two vertices" (Lakatos, 1976, p. 24)[24].

Two other variants of a polyhedron are obtained via the method of exception-barring, i.e. the ad hoc redefinition of a concept with the explicit exclusion of parts of the original extension: (p_9) "polyhedra that have no cavities, tunnels, or 'multiple structure'" (Lakatos, 1976, p. 29), (p_{10}) "convex polyhedra" (Lakatos, 1976, p. 30). Then, we have two variants of a polyhedron generated by the method of lemma-incorporation: (p_{11}) "simple polyhedra, i.e. those which, after having had a face removed, can be stretched onto a plane" (Lakatos, 1976, p. 36), (p_{12}) "simple polyhedron with all its faces simply-connected" (Lakatos, 1976, p. 38). Finally, content-increasing methods give us other three variants of a polyhedron, i.e. (p_{13}) "Georgonne-polyhedra" (Lakatos, 1976, p. 63), (p_{14}) "Legendre-polyhedra" (Lakatos, 1976, p. 63), and (p_{15}) "closed normal polyhedra" (Lakatos, 1976, p. 81). In some sense,

[24]Note that technically the last definition is not, if taken alone, a definition of a polyhedron, but in the context of Lakatos' fictional classroom discussion it is used as an additional defining feature for resisting 'counterexample 5' (Lakatos, 1976, p. 24) and as such it is a (part of a) new polyhedron variant meant to cope with Euler's conjecture.

Lakatos gives us another definition of a polyhedron, formalized in terms of vector algebra (Lakatos, 1976, pp. 112-126), but this definition is of a completely different kind than previous ones and thus should not be considered as a variant of the same conceptual population. It is a tentative formalization of both the concept of a polyhedron and Euler's conjecture in terms of vector algebra and thus pertains to a different conceptual population than the polyhedron one.

We can now reconstruct this conceptual history in my framework as a conceptual population. All the aforementioned fifteen variants of a polyhedron form the set of conceptual variants $C_p = \{p_1, p_2, \ldots, p_{15}\}$ of the polyhedron population. The environmental set is composed by the singleton of Euler's conjecture $E_p = \{ec\}$. The heuristic power ordering amongst the variants corresponds to their order of appearance in Lakatos' discussion, because each one of them is introduced as a way of dealing with a given counterexample affecting the (conjecture-pairs related to the) previous variants or via a content-increasing method. Since in my framework the number of counterexample (COUNT) and the size of the intended domain (DOM) are the two main constraints on any *hp* ordering, we can assume that $hp(p_1, ec) < hp(p_2, ec) < \ldots < hp(p_{15}, ec)$. The heuristic power ordering of the polyhedron population agrees then with Lakatos' narration of the history of the polyhedron concept. All steps from one variant to the next one are justified by the purely normative selection mechanism of my framework In the terminology of Section 3.1, this case study is then an example of mathematical selection.

The evolutionary dynamic of the polyhedron population is then clearly the one typical of Lakatosian populations, i.e. high environmental stability, high conceptual variation, high reproductive competition and a continuous distribution of heuristic power. We have in fact seen in Lakatos' reconstruction of this case study a remarkable number of different definitions of a polyhedron and only one, stable proof-problem (Euler's conjecture). In my terminology, then, the polyhedron population exhibits high environmental stability and high conceptual variation. Moreover, we saw that all the different variants of a polyhedron compete against each other in the context of proving Euler's conjecture in a truly Malthusian 'struggle for life'. The evolutionary dynamic of the polyhedron population exhibits thus high reproductive competition. Finally, the polyhedron population arguably shows a continuous distribution of heuristic power amongst its conceptual variants, i.e. similar definitions of a polyhedron have similar heuristic power. This can be seen looking at pairs of very similar definitions of a polyhedron such as (p_2, p_3) or (p_5, p_6). These pairs of variants that differ only for a minor tweak in their definition cope similarly with Euler's conjecture, i.e. their related conjecture-pairs face (almost) the same counterexamples and have the same intended domain.

We can now appreciate how the specific evolutionary dynamic of the polyhedron population allows Lakatos to describe its conceptual history as a paradigmatic example of concept-stretching. In order for a concept to be stretched via Lakatos' succession of proofs and refutations, there is a need of a stable mathematical problem and a plethora of tentative definitions of a mathematical concept. All these tentative definitions have to compete against each other for solving the same problem and there cannot be significant discrepancies of heuristic power amongst similar definitions. In other words, Lakatos' concept-stretching

model needs a certain kind of evolutionary dynamic in which conceptual populations have high environmental stability, high conceptual variation, high reproductive competition, and a continuous distribution of heuristic power. Lakatos' concept-stretching is thus perfect to describe the evolution of the kind of conceptual populations that I called Lakatosian populations. My framework is then able to answer the many critics (Feferman, 1978; Fine, 1978; Corfield, 2003; Werndl, 2009) that pointed out how the evolution of many cases of mathematical conceptual change is hardly an example of Lakatos' concept-stretching. The right domain of application of Lakatos' concept-stretching are mathematical conceptual histories that can be reconstructed as Lakatosian populations. In the next two subsection, we will see how the quaternion and the pre-abstract group populations, i.e. two examples that have been claimed to defy Lakatos' concept-stretching model (Feferman, 1978; Mormann, 2002), exhibit different evolutionary features than Lakatosian populations such as the polyhedron one.

4.3.3 Hamilton's invention of the quaternions

As my second case study I will reconstruct the conceptual history behind Hamilton's invention of the quaternions. Mormann discusses it as an example of axiomatic variation that defies Lakatos' model of concept-stretching (Mormann, 2002). I will show how, when reconstructed as a conceptual population, the quaternion population exhibits indeed a different evolutionary dynamics than the one typical of Lakatosian populations. However, thanks to the rich structure of my framework we will see that what is different in the quaternion population is not the instability of the proof-problem (as Mormann argues), but the distribution of heuristic power amongst the conceptual variants. As an historical basis for my reconstruction I will follow Hamilton's own memoirs (Hamilton, 1843a,b, 1853), together with Pickering's detailed analysis of Hamilton's practice (Pickering, 1995).

Hamilton's search for quaternions started with the idea of generalizing complex numbers to triplets. Before going into Hamilton's repeated tries into developing systems of triplets, I need to stress some basic facts about algebraic and geometric properties of complex numbers needed to understand Hamilton's generalization attempts. Central to Hamilton's research is the geometrical understanding of complex numbers, where the real and the ideal component of a number are not seen as quantities, but as coordinates of the end-point of a line segment starting from the origin in a two-dimensional plane. In this interpretation, the x-axis of the plane represents the real component of a given number, while the y-axis the imaginary one.

This correspondence between algebraic entities and line segments extends also to the operations between complex numbers, so that algebraic operations can be given a meaningful geometrical reading. Multiplication between complex numbers, an operation that constituted the core of the mathematical problem that Hamilton's higher complex numbers had to face, can be thus defined equivalently algebraically as

$$(a + ib)(c + id) = (ac - bd) + i(ad + bc)$$

or geometrically as the conjunction of two rules: "the product of two line segments is another line segment that (1) has the length given by the product of the lengths of the two segments to be multiplied, and that (2) makes an angle with the x-axis equal to the sum of the angles made by the two segments" (Pickering, 1995, p. 123).

Hamilton's search for higher complex numbers started by generalizing this geometrical reading of complex numbers to the three-dimensional case. He started thinking about another imaginary component j, geometrically represented as a line perpendicular to the two-dimensional complex plane (Hamilton, 1843b, p. 107). He also naturally assumed that $j^2 = -1$. We can take this first vague idea of a triplet as constituting the first conceptual variant (q_1) of the quaternion population.

Hamilton then focused on the algebraic operations performable on this new conceptual variant. Addition and subtraction were easily extended to the triplet case. Multiplication, instead, provided the newborn quaternion population with a stable mathematical problem. Hamilton started from the restricted case:

$$(x + iy + jz)^2 = x^2 - y^2 - z^2 + 2ixy + 2jxz + 2ijxz$$

The problem was how to understand the last term of the equation, $2ijxz$ and the product ij there contained. Hamilton's first natural choices, giving rise to two new conceptual variants of the quaternion populations, were (q_2) $ij = 1$ and (q_3) $ij = -1$. These two variants were both equally understandable from a purely algebraic point of view, but they both failed to have a reasonable geometric interpretation. Both variants were in fact still understood from the geometrical perspective of a line perpendicular to the complex plane and thus Hamilton's geometrical understanding of the multiplication operation was that "its real part ought to be $x^2 - y^2 - z^2$ and its two imaginary parts ought to have for coefficients $2xy$ and $2xz$" (Hamilton, 1843a, p. 103). The term $2ijxz$ contained in the algebraic understanding of the multiplication needed to vanish.

A new conceptual variant (q_4) arises exhibiting $ij = 0$, thus making the algebraic understanding of the multiplication in superficial agreement with Hamilton's geometrical intuitions. Only superficially, though, because to make the product of two arbitrary segments equal to zero violates the geometrical rule that wants the length of the segment product equal to the lengths of the segments multiplied (Pickering, 1995, p. 132).

Hamilton then let the commutativity assumption common to all the aforementioned conceptual variants go and assumed the more general (q_5) $ij = k$ and $ji = -k$, leaving undefined the value of k. This new conceptual variant coped with the multiplication environment far more successfully than its predecessors, achieving for the first time complete agreement between the algebraic and the geometric interpretation of multiplication for the aforementioned restricted case of the mathematical problem. Hamilton was then led to the general case of the multiplication of two arbitrary triplets and there the new conceptual variant was of little use. How one should understand the orientation of the product triplet for non-coplanar triplets?

Hamilton thus dropped the perpendicularity to the complex plane assumption, together with the orientation-part of the geometrical understanding of multiplication for complex

numbers. He first returned to the idea of $ij = 0$, this time not restricted by these two assumptions (q_6) and started working the general case only in terms of the length-part of the geometrical understanding of multiplication. Again, the algebraic and the geometrical understanding of the multiplication operation did not agree with each other, forcing Hamilton to a more radical departure from his original intuitions.

Hence, Hamilton started considering k not only as the undefined product of i and j, like it was in the variants q_5 and q_6, but as a whole new imaginary, thus obtaining the first conceptual variant in the quaternion populations with three different imaginary components (q_7). This new conceptual variant was still too unspecified to cope successfully with the multiplication problem in its general setting, since k^2 was still undefined. Three different choice for specifying this quantity naturally presented themselves to Hamilton, namely (q_8) $k^2 = 0$, (q_9) $k^2 = 1$, (q_{10}) $k^2 = -1$. The first variant, i.e. q_8, was quickly discharged for breaking again the geometrical reading of multiplication. Finally, Hamilton saw that q_{10} was the only choice that coped successfully with the multiplication environment:

> And since the order of these imaginaries is not indifferent, we cannot infer that $k^2 = ijij$ is $+1$, because $i^2 \times j^2 = -1 \times -1 = +1$. It is more likely that $k^2 = ijij = -iijj = -1$. And in fact this last assumption is necessary, if we would conform the multiplication to the law of multiplication of moduli. (Hamilton, 1843b, p. 108)

We can now model this conceptual history in terms of a conceptual population. The set of variants of the quaternion population is made of all the conceptual variants we have identified, i.e. $C_q = \{q_1, q_2, \ldots, q_{10}\}$, and the environment consists of the multiplication problem (m), $E_q = \{m\}$, with the restricted domain of coplanar triplets $restm$. Like in the case of the polyhedron population, we can assume that the hp ordering corresponds to the order in which conceptual variants appear in my recollection of Hamilton's search for quaternions. Any quaternion variant is in fact superior to the precedent ones in terms of counterexamples (COUNT) or success with the restricted domain of the problem (REST). We can therefore assume that $hp(q_1, m) < \ldots < hp(q_{10}, m)$. Like in the polyhedron population, all steps from one variant to the next one are consistent with the heuristic power ordering and thus justified by the selection mechanism of my framework. This second case study is therefore another example of mathematical selection.

If we take a look at the evolutionary dynamic of this conceptual population, at first it may seem similar to the one of Lakatosian populations such as the polyhedron population. The reconstruction of Hamiliton's invention of the quaternions provides us with many different conceptual variants and a single stable mathematical problem (the multiplication problem). The quaternion population exhibits thus high conceptual variation and high evolutionary stability. Furthermore, all the quaternion variants compete against each other in coping with the multiplication problem, making the quaternion population a conceptual population with high reproductive competition.

With respect to the three parameters of conceptual variation, environmental stability, and reproductive competition, the quaternion population exhibits the same environmental dynamic of the polyhedron population. Intuitively, however, the reconstruction of Hamilton's invention of the quaternions tells us a different story than Lakatos' fictional classroom.

Quaternion variants do not exhibit the same kind of variation that polyhedron variants have. The appearance of quaternion variants is somehow constrained by the possible ways in which their axioms can be manipulated. The story of Hamilton's research is a story of axiomatic tinkering, a story of a painstaking succession of small modifications to the definition of hyper-complex numbers needed to produce a suitable multiplication operation for this extended number domain. In this story, we saw that small modifications to the definition of a quaternion, such as the steps from q_3 to q_4 and from q_9 to q_{10}, produced huge discrepancies of effectiveness in coping with the multiplication problem. The latter case is particularly striking, since the last two variants considered by Hamilton differ only in the polarity of their specification of k^2, which is $+1$ in q_9 and -1 in q_{10}. This small difference is enough to cause a very significant hiatus in terms of heuristic power between the two variants, making q_{10} the only quaternion variant to cope with the general multiplication problem in a successful way. In the terminology of my framework, the quaternion population clearly exhibits a discrete distribution of heuristic power. With respect to this specific aspect of its evolutionary dynamic, the quaternion population is thus an Euclidean population.

In this way, the quaternion population shows us a different evolutionary dynamics than Lakatos' own example of the polyhedron concepts. The quaternion population is not a Lakatosian population, due to its lack of continuity in the distribution of heuristic power. This lack of continuity causes a discreteness in the distribution of heuristic power amongst the quaternion variants. In Section 2.3.3, I stressed how this discreteness is a symptom of an highly axiomatized body of mathematics. The present case study offers a paradigmatic example of this connection, showing how an axiomatized conceptual history such as Hamilton's invention of the quaternions presents a discrete distribution of heuristic power amongst its variants. This discreteness is caused by the underlying mechanism at work in Hamilton's research, namely the manipulation of axiomatic systems (Pickering, 1995; Schlimm, 2013) that constrains the possible conceptual variants.

Thanks to the fine-grained structure of my framework, we can then reappraise the reason why Hamilton's invention of the quaternions is a different case of conceptual change than Lakatos' polyhedron example. Mormann (Mormann, 2002), in fact, discusses this case study as an example of axiomatic variation that defies Lakatos' concept-stretching. He states that the reason why the conceptual history of quaternions defies Lakatos' model of conceptual change is the fact that also the environment changes together with the conceptual variants. According to Mormann, the conceptual history of Hamilton's quaternions "does not leave intact the theorems or laws which were originally considered to be the touchstone of its respectability" (Mormann, 2002, p. 144). Mormann is indeed right in claiming that cases of axiomatic conceptual change such as the invention of the quaternions exhibits a different environmental dynamics than Lakatos' own examples, but his diagnoses only partially captures what is specific to the variation exhibited by axiomatized bodies of mathematics. We have in fact seen that the quaternion population exhibits high environmental stability in my framework. It is indeed true that Hamilton, in the conceptual variation from triplets to quaternions, dropped more than one assumption that he had on how multiplication with higher complex numbers should work, but this does not imply a

change in the abstract mathematical environment. The problem of having a multiplication operation meaningful both from an algebraic and a geometric point of view was the only, stable mathematical problem faced by all the quaternion variants. Mormann's focus on the token-like tentative proofs causes him to miss the stability of the more abstract type-like mathematical problem. Moreover, in his description of axiomatic variation he seems to conflate the lack of environmental stability typical of formalized bodies of mathematics and the lack of a continuous distribution of heuristic power typical of highly axiomatized bodies of mathematics. This conflation makes him not realize that it is the discreteness of heuristic power distribution amongst the quaternion variants the reason why this conceptual history differs from Lakatos' own examples. In sum, the difference between Lakatosian populations such as the polyhedron population and conceptual populations like the quaternion one is not the changing environment, it is the lack of continuity in the heuristic power distribution. This reappraisal of this example of conceptual change shows how the four dimensions of the Lakatosian space allow more fine-grained analyses of the evolutionary dynamics of mathematical conceptual histories than previous proposals.

4.3.4 Pre-abstract group theory

As my third case study, I will focus on the history of pre-abstract group concepts, using as historical reference the detailed reconstruction of Wussing (Wussing, 1984). I will show how this conceptual history exhibits an evolutionary dynamic different from both the polyhedron and the quaternion population.

Pre-abstract group concepts were developed between 1770 and 1880 in relation to three connected but independent fields of mathematical inquiry: number-theory, algebra, and geometry. Specifically, they arose in the context of three-specific mathematical problems: the classification of number forms in number theory, the general solvability of algebraic equations in algebra, and the search for ordering principles in geometry.

In reconstructing this conceptual population, I will proceed in an unhistorical way, treating every mathematical problem and the related conceptual variants independently. This is just for reconstructing more clearly the conceptual evolution of the population and it should not be interpreted as assuming the historical independence of problems and related groups of conceptual variants from one another. I will denote the three mathematical problems composing the environment of this population with e_1, e_2, e_3 and the thirteen conceptual variants with $g_1, \ldots, g_{13}$.

The first mathematical problem where some group variant implicitly appeared was the problem of developing a general theory of forms (such as binary quadratic forms) in number theory (e_1). Namely, Wussing shows how Euler's theory of power residues involved in its partitioning of reminders "a clear example of group-theoretic thinking" (Wussing, 1984, p. 49). We can use the implicit, vague and heavily underdefined group-theoretic notion at work in Euler's paper as our first conceptual variant of the group population (g_1). Gauss' work gives us the next two group variants that emerged from this problem. The first one is his notion of 'congruence' (g_2) that he used to structure and extend Euler's theory of power residue (Wussing, 1984, pp. 52-54). The second one is the notion of 'composition

of forms' (g_3), which constituted the center of Gauss' general theory of forms (Wussing, 1984, pp. 55-61). The final conceptual variants within the pre-abstract group population that emerged in the context of number theory was Kronecker's axiomatization of a finite abelian group (g_4) and his proof of the related basis theorem (Wussing, 1984, pp. 61-67).

The second set of group variants I am going to focus on is the one related to the problem of solving algebraic equations of higher degree (e_2). The first group variant can be traced to Lagrange's seminal 'reflections' on the solvability of algebraic equations. Lagrange was the first to undertake a structural study of algebraic equations (Wussing, 1984, pp. 71-79). The central offspring of his study was the connection between the solvability of algebraic equations and the concept of permutation. Specifically, Lagrange realized that the degree of the resolvent of a given equation is the number of different values that the roots of the original equation take when permuted in all the possible ways. This implicit notion of permutation constitutes another group variant (g_5), the first that emerged in the context of algebra. The next steps in the theory of permutations give us two other conceptual variants. Ruffini built on Lagrange's theory, asserting for the first time the unsolvability by radicals of equations with degree higher than four. In his work one can find a general classification of permutations, simple and various kinds of complex ones, where he used the notion of permutation with implicit group-theoretic character (g_6) (Wussing, 1984, pp. 80-84). Cauchy improved further the theory of permutation with his concept of 'system of conjugate substitution' (g_7), with which he implicitly defined (a version of) the permutation-theoretic concept of group in terms of its generator. Finally, Galois was the first to define explicitly the permutation-theoretic concept of a group (g_8), understood as necessarily closed under multiplication (Wussing, 1984, pp. 111-117). He used this notion for defining the 'Galois group of an equation', which together with the pivotal property of the normality of a subgroup, allowed him to assign at every equation a permutation group whose structure reveals all the essential properties of the equation, including whether it is solvable by radicals.

The last set of pre-abstract group variants is the one coping with the search for ordering principles in geometry (e_3). As our first variant that emerged in this context we can take Möbius' notion of 'affinity' (g_9) used in his intuitive classification of geometric relations (Wussing, 1984, pp. 35-42). The next conceptual step in ordering geometries is Cayley's notion of 'invariant' (g_{10}), which he used in his abstract classificatory efforts. These steps in the search for ordering principles led famously to the Erlangen Program and its group-theoretic classification of geometries. In regards to new pre-abstract group variants, we owe to the Erlangen Program a new explicit definition of group (Wussing, 1984, pp. 187-193). Klein defined a group not in terms of permutations (like Galois did), but he spelled out his variant of the group concept in terms of transformations (g_{11}). After the Erlangen Program, Klein and Lie respectively developed two other variants of the group concept, obtained by extending and sharpening the still quite under-defined notion of transformation group (Wussing, 1984, pp. 205-223). We owe to Klein the notion of an infinite discrete group of transformations (g_{12}) and to Lie the notion of a continuous group of transformation (g_{13}).

We can now model this conceptual history as a conceptual population. The set of variants of the pre-abstract group population consists of all the aforementioned conceptual

variants, i.e. $C_g = \{g_1, g_2, \ldots, g_{13}\}$. The environment consists of three mathematical problems, i.e. the classification of number forms (e_1), the solvability of algebraic equations (e_2), the search for ordering principles in geometry (e_3): $E_g = \{e_1, e_2, e_3\}$. As we have seen, the task of solving these three problems gave rise to three different sets of group variants, each one of them with its own selected definition of group, i.e. finite abelian groups, permutation groups, transformation groups. All these three notions were selected as the culmination of a series of implicit and explicit group-theoretic notions, each one with a more general intended domain or more successfully adapted to a restricted version of the problem than the precedent one. Guided by the rationality principles of intended domain size (DOM) and restricted case applications (REST), we can assume the *hp* ordering of this conceptual population is composed by the different chains: $hp(g_1, e_1) < hp(g_2, e_1) < hp(g_3, e_1) < hp(g_4, e_1)$; $hp(g_5, e_2) < hp(g_6, e_2) < hp(g_7, e_2) < hp(g_8, e_2)$; and $hp(g_9, e_3) < hp(g_{10}, e_3) < hp(g_{11}, e_3) < hp(g_{12}, e_3), hp(g_{13}, e_3)$. With regards to mathematical selection, in this case study we have to carefully distinguish between local and global selection. The three different heuristic power chains give in fact rise to three different local selection mechanisms specific to the mathematical problem under focus. With respect to one of the three mathematical problems composing the environment of the population, conceptual variants can be ordered in terms of heuristic power. At the general level of the whole population, however, the pre-abstract group population does not select any conceptual variant as the one to be preferred. There is, for instance, no reason to generally prefer the transformational group concept to the permutational one (and vice-versa) for its heuristic power. No general selection is induced by the heuristic power ordering of this population. The actual history of pre-abstract group theory is consistent with the combination of locally selected conceptual variants and the absence of a generally preferred one. For each one of the three mathematical problems, a conceptual variant got selected and all of them kept being used successfully in their respective areas for the whole conceptual history. This third case study is thus an example of local mathematical selection.

We can then assess which kind of evolutionary dynamic the pre-abstract group population exhibits. With respect to conceptual variation and environmental stability, the evolution of this conceptual population is similar to the one of the other two case studies we discussed. The reconstruction of the pre-abstract group concept history gives us many different conceptual variants and three stable mathematical problems. Just like the polyhedron and the quaternion population, the pre-abstract group population exhibits high conceptual variation and high environmental stability. With respect to the distribution of heuristic power, like the polyhedron population and unlike the quaternion population, the pre-abstract group population exhibits a continuous distribution of heuristic power. Similar pre-abstract group variants, such as g_6 and g_7, have very similar heuristic power. This continuity is due to the lack of axiomatization of the pre-abstract group concepts and it was about to change in a few years with the development of the abstract group concept (and the related conceptual population) (Wussing, 1984, pp. 230-254). Is therefore the pre-abstract group population another example of a Lakatosian population? Just a quick look at the degree of reproductive competition in this population provides a negative answer to this question. The pre-abstract group population, in fact, dramatically lacks

(almost) any reproductive competition whatsoever. A striking symptom of this lack of reproductive competition is the complete absence of any counterexample in the history of the pre-abstract group concepts. The evolution of these concepts proceeded as a series of generalizations and further applications, without any significant dialectic between proofs and refutations. A further factor that contributed to the lack of reproductive competition is the plurality of mathematical problems in the environment of this population. This plurality of problems made possible the coexistence of different pre-abstract group variants, each of them very successful in coping with its own related problem. The three aforementioned locally preferred pre-abstract group variants, i.e. finite abelian groups, permutations groups, transformation groups, evolved collectively, each one of them improving their respective predecessors within the context of their specific problem.

We can now appreciate the specific kind of evolutionary dynamic that the pre-abstract group population exhibits. In comparison to the quaternion population, this conceptual population is more Lakatosian with respect to the distribution of heuristic power, but it is more Euclidean with respect to reproductive competition. This kind of evolutionary dynamic is typical of mathematical conceptual histories driven by "internal organization" (cf. Feferman 1978, p. 174) or "systematization" (Kitcher 1984, pp. 217-225), where the cooperative and collective aspect of mathematical evolution is more prominent than the proofs and refutation aspect. My framework is able to adequately represent this kind of evolutionary dynamics thanks to the separation of the reproductive competition aspect from the other three parameters.

More generally, the three case studies I presented demonstrate how my framework allows a very fine-grained classification of the evolutionary dynamics of mathematical conceptual histories. Examples of conceptual change can be rationally reconstructed as conceptual populations and judged to be more or less consistent with the heuristic power ordering of the related population. Conceptual populations can then be classified with respect to the four parameters of the Lakatosian spaces, tracking specific aspects of their environmental dynamics. These four parameters break down the opposition between Lakatosian and Euclidean populations into a plurality of evolutionary features with respect to which a mathematical conceptual history can be judged to be more Lakatosian or more Euclidean (or none of the above, in the case of populations lacking evolutionary stability). Different combinations of these four parameters give rise to different evolutionary dynamics, each one of them occupying a different part of the Lakatosian space (Figure 2).

Let me recall the main steps of the last two sections. Building upon Mormann's evolutionary reading of Lakatos and Godfrey-Smith's population-based Darwinism, I proposed a general evolutionary framework for conceptual change in mathematics. The framework is made of three main ingredients: the notion of a conceptual population, the opposition between Lakatosian and Euclidean populations, and the spatial tools of the Lakatosian space.

I showed how my framework achieves a general evolutionary account of conceptual change in mathematics compatible with the diversity of evolutionary dynamics that the history of mathematics exhibits. I demonstrate how different mathematical conceptual histories can be reconstructed in my framework as conceptual populations. Thanks to the

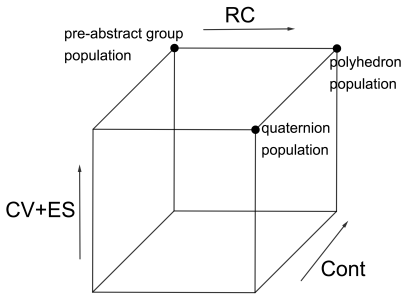


Figure 4.2: A three-dimensional representation of the Lakatosian space showing the parts of the space corresponding to the kinds of evolutionary dynamics exhibited by the three case studies.

normative selection mechanism of my framework, the rationality of these episodes of conceptual change can then be assessed by judging whether these conceptual histories are cases of mathematical selection or evolutionary drift. Moreover, thanks to the rich dimensional structure of the Lakatosian space, the specific evolutionary dynamic exemplified by these conceptual histories can be classified with respect to four different aspects: conceptual variation, environmental stability, reproductive competition, and distribution of heuristic power. Different combination of these parameters are connected to different kinds of evolutionary dynamics, making a giving conceptual population more similar to Lakatosian populations or more closer to Euclidean ones (or different from both). Furthermore, certain parts of the Lakatosian space corresponding to specific combination of the four parameters give a novel characterization of important concepts in philosophy of mathematics such as formalization or axiomatization.

As I stressed at the beginning of Section 2, with this framework that I just presented I tried to approach the EET program in a bottom-up way, focusing on a specific selection mechanism (i.e. the heuristic power ordering) involving only few conceptual elements (i.e. conceptual variants and mathematical problems) within a single scientific discipline (i.e. mathematics). Despite the simplicity and the narrow scope of the framework, we saw the many insights that it is able to give in assessing different case studies from the history of mathematics. My proposal is furthermore just the beginning of a research program and as such it is open to extensions in multiple parts. As we saw in Section 2.3.4, augmenting my basic framework with a time-dimension allows one to model entire evolutionary histories of mathematical concepts as a succession of conceptual populations and the related inter-population kinds of changes as specific movements along the axis of the augmented Lakatosian space. Similarly, other extensions of my framework, via the addition of further dimensions to the Lakatosian space or of additional elements to conceptual populations, can be envisaged in order to treat more aspects of the evolution of mathematical concepts, such as the emergence of mathematical problems, the inheritance mechanisms of selected conceptual variants, the similarity relations between conceptual variants, the relationship between different mathematical fields, and many others significant factors. This bottom-up way of a simple framework and several modular extensions appear

a promising way of gradually building a precise and historically testable evolutionary model of scientific conceptual change.

4.3.5 Formal addenda: a toy heuristic power function and the three case studies formalized

In this last part of this section, I will present two formal addenda to the presentation of my evolutionary framework for mathematical selection. The first addendum consists of a toy-example (Section 3.5.1) in which I construct an actual heuristic power function. This simple exercise in model-building will demonstrate the satisfiability of the rationality postulates that constrain any heuristic power ordering in my framework. The construction of a simple heuristic power function will also show how the selection mechanism of my framework works in practice. In the second addendum I will instead formalize the three case studies presented in this work. Specifically, I will show how these conceptual histories can be reconstructed as formal conceptual populations and how the four parameters of the Lakatosian space can be measured.

A toy example

Let *pe* be the proof-problem consisting of the conjecture $\forall x(B(x) \rightarrow S(x))$ for a finite domain of twelve individuals $\{a, \ldots, l\}$. Let $restpe_1, restpe_2$ be two restricted cases of the proof-problem *pe* for a sub-domain respectively of three $\{a, b, d\}$ and eight $\{a, b, c, d, e, f, g, h\}$ individuals (Figure 2).

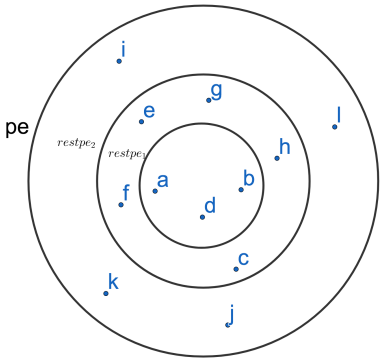


Figure 4.3: A representation of the toy-case setting. Points represent the twelve single instances from a to l, while the three circles represent the domain of the proof-problem *pe*, together with its restricted cases $restpe_1$ and $restpe_2$.

The fifteen individuals have the following properties:

- (a) : $A(a), S(a), C(a), \neg D(a), L(a), T(a), Z(a)$;

- $(b):\ A(b), S(b), C(b), \neg D(b), L(b), T(b), Z(b)$;
- $(c):\ A(c), \neg S(c), C(c), D(c), \neg L(c), \neg T(c), \neg Z(c)$;
- $(d):\ \neg A(d), \neg S(d), \neg C(d), D(d), L(d), T(d), \neg Z(d)$;
- $(e):\ A(e), \neg S(e), \neg C(e), D(e), L(e), T(e), \neg Z(e)$;
- $(f):\ A(f), \neg S(f), C(f), \neg D(f), L(f), T(f), Z(f)$;
- $(g):\ \neg A(g), S(g), C(g), \neg D(g), L(g), T(g), Z(g)$;
- $(h):\ A(h), \neg S(h), C(h), \neg D(h), L(h), \neg T(h), Z(h)$;
- $(i):\ A(i), \neg S(i), C(i), \neg D(i), \neg L(i), T(i), \neg Z(i)$;
- $(j):\ A(j), S(j), C(j), D(j), L(j), T(j), Z(j)$;
- $(k):\ A(k), S(k), C(k), \neg D(k), L(k), T(k), Z(k)$;
- $(l):\ A(l), \neg S(l), C(l), \neg D(l), L(l), T(l), \neg Z(l)$.

Let $c_1, \ldots, c_6$ be six different conceptual variants of the concept B defined as follows:

- $c_1\ B(x) := A(x)$;
- $c_2\ B(x) := A(x) \wedge C(x)$;
- $c_3\ B(x) := A(x) \wedge C(x) \wedge \neg D(x)$;
- $c_4\ B(x) := C(x) \wedge L(x)$;
- $c_5\ B(x) := C(x) \wedge T(x)$;
- $c_6\ B(x) := Z(x)$.

This toy-setting may appear a little bit unrealistic at first, but (modulo several simplicity assumptions) it represents a typical Lakatosian dynamics of proofs and refutations such as the famous polyhedron case study. Consider the following fictional narrative. A mathematician wants to find a property $B(x)$ such that whenever an individual object has it, the object possesses also the property $S(x)$. She observes the restricted domain $restpe_1$, noting that if $A(x)$ holds, then $S(x)$ holds too. She thus chooses $A(x)$ (c_1) as a definition of $B(x)$ and formulates the related conjecture $\forall x(A(x) \rightarrow S(x))$. Focusing on the broader domain $restpe_2$, this first conjecture is soon falsified by the individual e for which $A(e)$ and $\neg S(e)$ hold. Our mathematician notes that in this case (differently from the cases in the first restricted domain) $\neg C(e)$ holds. She therefore refines her conjecture redefining $B(x)$ as $A(x) \wedge C(x)$ (c_2), thus shielding her conjecture from counterexample e. The positive instance f seems to corroborate her refinement, but c presents itself as

a counterexample (also) for the new conjecture, since $A(c) \wedge C(c) \wedge \neg S(c)$ holds. Our fictional mathematician refines again her conjecture adding a third condition $\neg D(x)$ to her definition of $B(x)$ (c_3), protecting the new conjecture from c. This series of restrictions has successfully shielded the conjectures from counterexample but it has also restricted the domain of the conjecture, as it is exemplified by cases like g that the conjecture does not cover. In Lakatosian terms, the content of the theorem has decreased. In order to counter that, our mathematician "inspects her proof searching for a deeper theorem"The result of her inspection is a new definition of $B(x)$ as $C(x) \wedge L(x)$ (c_4), the related conjecture of which is as shielded from counterexamples as the precedent one but with a bigger domain. The case h provides a new counterexample to the new conjecture (and to all the old ones), though. Our mathematician refines another time her conjecture with a new conceptual variant, defining $B(x)$ as $C(x) \wedge T(x)$ (c_5). The related refined conjecture is shielded from counterexample h and thus successfully copes with all cases of the restricted domain $restpe_2$. The bigger domain pe holds another unpleasant surprise, though. The case i presents a counterexample for the new conjecture, but not for the precedent definition of $B(x)$. However, the latest conceptual variant is still the best definition at our mathematician disposal, coping successfully with all cases of $restpe_2$ and with both j and k in the bigger domain. Another counterexample to all the conjectures, l, is found. Can our mathematician inspect her proof and find another definition of $B(x)$ capable finally to cope successfully with the overall domain of the proof-problem pe? After a pain-staking inspection of the proof, our mathematician finds a new definition of $B(x)$, $Z(x)$ (c_6), the related conjecture of which copes successfully with the general proof-problem pe. At last, a 'natural', 'intuitive', 'self-evident', 'fruitful', 'simple' definition of $B(x)$ has been found.

The rationality of the narrative just presented is shown in my framework by finding a suitable hp ordering that warrants the choices made by our fictional mathematician. For illustration only, in this toy-example we will induce this ordering from a simple function (conveniently called also hp). The hp function from conjecture-pairs to natural number is then defined as follows:

$$hp(c_i, pe) = \frac{10}{count(c_i, pe) + 1} + \frac{dom(c_i, pe)}{10} + \frac{restr(c_i, pe)}{20}$$

where $count(c_i, pe)$ is the function counting the number of counterexamples for a given conjecture-pair, i.e. the number of individuals in the domain that have the property $B(x)$ (as defined by c_i) but not the property $S(x)$, $dom(c_i, pe)$ counts the possible instances of the conjecture-pair, i.e. the number of individuals in the domain having the property $B(x)$ (as defined by c_i), and $rest(c_i, pe)$ is the function counting the number of domains $(restpe_1, restpe_2, pe)$ of the proof-problem with which the given conceptual variants copes successfully (i.e. there are no counterexamples for the related conjecture-pair within the domain).

These functions applied to the six different conjecture-pairs give the following result:

- *count*: $count(c_1, pe) = 5$, $count(c_2, pe) = 4$, $count(c_3, pe) = 2$, $count(c_4, pe) = 2$, $count(c_5, pe) = 2$, $count(c_6, pe) = 0$;

- *dom*: $dom(c_1, pe) = 9$, $dom(c_2, pe) = 8$, $dom(c_3, pe) = 7$, $dom(c_4, pe) = 8$, $dom(c_5, pe) = 8$, $dom(c_6, pe) = 6$;
- *rest*: $dom(c_1, pe) = 1$, $dom(c_2, pe) = 1$, $dom(c_3, pe) = 1$, $dom(c_4, pe) = 1$, $dom(c_5, pe) = 2$, $dom(c_6, pe) = 3$;
- Therefore we have $hp(c_1, pe) = \frac{10}{6+1} + \frac{9}{10} + \frac{1}{20}$, $dom(c_2, pe) = \frac{10}{5+1} + \frac{8}{10} + \frac{1}{20}$, $dom(c_3, pe) = \frac{10}{3+1} + \frac{7}{10} + \frac{1}{20}$, $dom(c_4, pe) = \frac{10}{3+1} + \frac{8}{10} + \frac{1}{20}$, $dom(c_5, pe) = \frac{10}{3+1} + \frac{8}{10} + \frac{2}{20}$, $dom(c_6, pe) = \frac{10}{1} + \frac{6}{10} + \frac{3}{20}$.

It is easy to see that our *hp* function respects all the three rationality postulates of my framework (cf. Section 2.3.1). The selection ordering induced by our *hp* function is then $hp(c_1, pe) < hp(c_2, pe) < hp(c_3, pe) < hp(c_4, pe) < hp(c_5, pe) < hp(c_6, pe)$ and therefore c_6 is the conceptual variant that ought to be preferred.

Formalization of the three case studies

- **Polyhedron Population** (PP)

  $PP = \langle C_p, E_p, hp \rangle$ where $C_p = \{p_1, p_2, \ldots, p_{15}\}$, $E_p = \{ec\}$ and $hp = hp(p_1, ec) < hp(p_2, ec) < \ldots < hp(p_{15}, ec)$. The Environmental history consists of the same environmental set $E_{H_p} = \{E_p\}$.

  Conceptual Variation: $CV(PP) = |C_p| = 15$

  Environmental Stability: $ES(PP) = \frac{|\cap E_{H_p}|}{|\cup E_{H_p}|} = 1$

  Reproductive Competition: $RC(PP) = \frac{|C_p|}{|\cup E_{H_p}|} = 15$

  Continuity: $\forall p_x, p_y \in C_{PP}(p_x \approx p_y \rightarrow hp(p_x, ec) \approx hp(p_y, ec))$ is satisfied.

- **Quaternion Population** (QP)

  $PP = \langle C_q, E_q, hp \rangle$ where $C_q = \{q_1, q_2, \ldots, q_{10}\}$, $E_q = \{m\}$ and $hp(q_1, m) < \ldots < hp(q_{10}, m)$. The environmental history consists of the same environmental set $E_{H_q} = \{E_q\}$.

  Conceptual Variation: $CV(QP) = |C_q| = 10$

  Environmental Stability: $RC(QP) = \frac{|C_q|}{|\cup E_{H_q}|} = 10$

  Reproductive Competition: $ES(QP) = \frac{|\cap E_{H_q}|}{|\cup E_{H_q}|} = 1$

  Continuity: $\forall q_x, q_y \in C_{QP}(q_x \approx q_y \rightarrow hp(q_x, m) \approx hp(q_y, m))$ is not satisfied.

- **Pre-abstract Group Population** (GP)

  $PP = \langle C_g, E_g, hp \rangle$ where $C_g = \{g_1, g_2, \ldots, g_{13}\}$, $E_g = \{e_1, e_2, e_3\}$ and $hp(g_1, e_1) < hp(g_2, e_1) < hp(g_3, e_1) < hp(g_4, e_1)$; $hp(g_5, e_2) < hp(g_6, e_2) < hp(g_7, e_2) < hp(g_8, e_2)$;

and $hp(g_9, e_3) < hp(g_{10}, e_3) < hp(g_{11}, e_3) < hp(g_{12}, e_3), hp(g_{13}, e_3)$. The Environmental history consists of the same environmental set $E_{H_g} = \{E_g\}$.

Conceptual Variation: $CV(GP) = |C_g| = 13$

Environmental Stability: $ES(GP) = \frac{|\cap E_{H_g}|}{|\cup E_{H_g}|} = 1$

Reproductive Competition: $RC(GP) = \frac{|C_g|}{|\cup E_{H_g}|} = \approx 4$

Continuity: $\forall g_x, g_y \in C_{GP}, \forall e_i \in E_g(g_x \approx g_y \rightarrow hp(g_x, e_i) \approx hp(g_y, e_i))$ is satisfied.

4.4 Assessing Evolutionary Models in the Toolbox Framework

In this final section, I will analyze how evolutionary models of conceptual change in mathematics can be classified within the Toolbox framework, i.e. the meta-framework for assessing models of conceptual change that I presented in Chapter 2. More specifically, we will see how the Darwinian models of conceptual evolution such as my framework for mathematical selection can be assessed along the nine evaluative dimensions of the Toolbox framework: units of selection, concept ontology, concept structure, kinds and degrees of conceptual change, degree of normativity, effectiveness of normative judgment, assumptions and consequences for conceptual change in science, assumptions and consequence for conceptual change in philosophy, metaphilosophical assumptions and implications. Let us survey how Darwinian models of conceptual evolution performs in these dimensions, one by one, then.

Units of selection This dimension judges models of conceptual change according to the level of abstraction at which they identify conceptual entities as meaningful units of change. In the case of Darwinian models of conceptual change, the meaningful unit of conceptual change is considered to be a conceptual population. The level of abstraction at which evolutionary models of conceptual change understand conceptual change is thus a set of conceptual variants, i.e. concepts designed to solve similar scientific problems. The need of taking into account a whole set of conceptual variants, and not a single concept like other models of conceptual change, is crucially connected with the Darwinian populational thinking (Mayr, 1975) and the related concepts of variation, fitness, and adaptation within the context of a whole population of entities. Despite the main actors of evolution are always individuals, evolutionary change has to be understood in a collective context.

Concept ontology This dimension focuses on the compatibility of a given model of conceptual change with the different philosophical positions on the ontology of concepts. As we saw in Sections 1 and 2, evolutionary models of conceptual change such as my framework for conceptual selection in mathematics are not tied to any specific ontological position about concepts. However, as it was heavily stressed by Toulmin's insistence on the

concept of scientific possibility (cf. Section 1.1), evolutionary models of conceptual change necessarily engage only with the public aspect of scientific concepts as understood by the related scientific community. As such, evolutionary models are particularly compatible with linguistic and abstract views about concept ontology. Nevertheless, enthusiasts of a psychological or a worldly view can easily have a deflationary understanding of evolutionary models use of concepts by seeing them as working to the linguistic or abstract public correlate of a given concept. For instance, Gärdenfors (Gärdenfors, 2014) seems to have a similar reading of evolutionary accounts of language and concepts in mind in his account of how meanings can be shared amongst different individuals.

Concept structure This dimension focuses instead on how a given model of conceptual change assumes the structure of concepts to be constituted. Evolutionary models of conceptual change are quite neutral on matters of conceptual structure, since it does not usually directly play a role in the evaluation of the fitness of a given concept or in the related selection mechanisms. What matters for evolutionary models of conceptual change is how a given concept faces the related scientific problems in comparison to the performances of its competitors. Depending on the specific scientific problems at issue, evaluating a concept performance may involve different parts of a concept role or function. This multiple structures possibly involved in evaluating a concept performance are the reason why Toulmin championed a hybrid view of concepts as micro-institutions (cf. Section 1.1) and they could be put forward as evidence for a strong compatibility of evolutionary models of conceptual change with a hybrid view of conceptual structures (cf. Chapter 2, Section 1.2). That said, it should be clear that a deflationary reading of the evaluation of these multiple structures is of course a natural possibility and as such evolutionary models seem compatible with most theories of conceptual structure.

Kinds and Degrees of conceptual change This dimension focuses on the kinds and degrees of conceptual change that a given model of conceptual change identifies. Evolutionary models of conceptual change identify two kinds of changes, intra-population and inter-population changes. The first kind of change corresponds to changes happening within a given conceptual population, such as the appearance of new conceptual variants and new problems (or their disappearance). The second kind of changes underlies the transitions from one conceptual population to another one, corresponding to more radical changes in the scientific problems and the related conceptual variants. As it was stressed by several authors (e.g. Toulmin 1972; Kitcher 1984; Hull 1988a), the difference between intra-population and inter-population changes is definitely not a sharp one, since it crucially depends on the model building activity of identifying a given conceptual population. As we saw in applying my conceptual selection framework to actual case studies from the history of mathematics, there is always a certain degree of arbitrariness in the identification of both a set of conceptual variants and a set of scientific problems. Thus the difference between intra-population and inter-population changes should be taken as a lightweight pragmatic difference strongly dependent on the specific model of conceptual change and

the specific reconstruction of the case study under focus.

Degree of normativity This dimension tracks the extent to which a given model of conceptual change is more or less normative in judging episodes of conceptual change. As we saw in Section 3, evolutionary models of conceptual change such as the one presented in this chapter are able to somehow assess, in a quasi-normative way, the rationality of a given historical episode of conceptual change. They can in fact judge a given conceptual history to be a case of selection or drift, mirroring the standard quasi-normative distinction between "intentional"and random changes in the evolution of a given biological population. If a case of scientific selection corresponds to an intentional, rationally driven, choice of the fittest variant(s) in the conceptual population, cases of drift correspond to episodes where an equally fit (or sometimes even a less fit) variant gets selected due to some external influences or circumstances. It should be of course noted that, as Toulmin lucidly stressed (cf. Section 1.1), purely rational selections and purely external scientific drifts should be understood as ideal extremes between which most episodes of conceptual change in science lie.

Effectiveness of normative judgment This dimension focuses on how effective the normative judgment of a model of conceptual change is. In the case of evolutionary models of conceptual change, the question is how sharp and trustworthy can we be when judging a given conceptual history as an episode of scientific selection or scientific drift. Not surprisingly, the judgment is highly dependent on the specific rational reconstruction of the historical case under focus. In fact, as we saw in the case studies presented in Section 3, judging a case study to be an example of selection or drift is highly dependent on the conceptual variants and the scientific problems identified, as well as on the heuristic power function chosen and on the assessment of the heuristic power of a given variant in relation to a given problem. All these crucial elements in the evaluation of the rationality of a given case study are of course dependent on how the case study is rationally reconstructed. As such, the normative judgment of evolutionary models of conceptual change depends on the historical and philosophical faithfulness of the underlying rational reconstruction and should be judged by the usual philosophical and historical pragmatic tools available.

Assumptions/consequences for conceptual change in science This dimension focuses on the assumptions and the consequences of a given model of conceptual change in relation to the problems that scientific conceptual change poses in philosophy of science. Evolutionary models of conceptual change are of course connected with a vision of scientific progress and rationality consistent with the ideals of evolutionary epistemology and the EET program. As we saw in Section 1, the central tenet of the EET program is the belief in a significant analogy between biological and scientific evolution. Scientific progress is then seen by evolutionary epistemologists as akin to the progressiveness of biological evolution, i.e. as not goal-directed and always strongly dependent on the environment and the variation at a given time. Evolutionary models of conceptual change are thus consis-

tent with a belief in scientific progress, albeit of the non-teleological and fallibilist kind paradigmatically exemplified by Popper's ideals (cf. Popper 1974a). Similarly, the view of scientific rationality depicted by evolutionary models of conceptual change is an instrumental rationality dependent on the shared values and goals of the scientific community and on the specific problems that they face. Moreover it is a purely backward-looking kind of rationality that, just like the workings of natural selection, can only be properly assessed long after the selection process has taken place.

Assumptions/consequences for conceptual change in philosophy This dimension focuses on the assumptions and the consequences of a given model of conceptual change in relation to the problems that philosophical conceptual change poses in philosophy. As in the previous case of scientific conceptual change, evolutionary models of conceptual change bring to the problem of conceptual change the idea(l)s of evolutionary epistemology. Philosophical concepts then, just like every other kind of intellectual product, are also subject to intellectual selection and as such the problem of conceptual change is as real for philosophical concepts as it is for scientific ones. The interesting question that this evolutionary perspective on philosophical conceptual change brings forwards is the following: what are the selection mechanisms behind conceptual change in philosophy? Of a special interest is the subquestion of whether these mechanisms are similar to the ones at play in scientific selection and if not, what is different between them. To my knowledge, not much effort has been put by philosophers in applying evolutionary models of conceptual change to philosophical conceptual histories. This interesting yet largely unexplored research question could perhaps shed a new light on the old question whether philosophy progresses in the same way in which science does.

Metaphilosophical assumptions and implications This dimension focuses on the metaphilosophical background that a given model of conceptual change has. The metaphilosophical background of evolutionary models of conceptual change is of course the one of evolutionary epistemology in all its scope and depth. According to evolutionary epistemology, philosophical activity and its products should take into serious consideration their role and status of evolutionary products. What a truly evolutionary epistemology, and more generally an evolutionary philosophy, amounts to is (as we saw in Section 1) debatable, but it seems safe to assume that a truly evolutionary philosophy would involve a radical reform of many philosophical fields. The ideal of an evolutionary epistemology is also usually connected with the naturalization of epistemology and related normative philosophical enterprises such as metaphysics and ethics. The debate on whether such philosophical disciplines can or should be naturalized is even longer and more controversial than the discussion on evolutionary epistemology and as such I will not dare to enter into it.

Chapter 5

Indeterminate Models of Conceptual Change

The focus of this chapter will be on what I will call indeterminate models of conceptual change, i.e. frameworks for understanding conceptual change in which semantic indeterminacies are modeled as central features of this phenomenon. This is not to say that all the other models of conceptual change do not take semantic indeterminacy into consideration. As we saw in the previous chapters, in fact, conceptual change is a phenomenon remarkably prone to various kinds of semantic indeterminacy such as vagueness and ambiguity. Moreover, models of conceptual change necessarily have to wrestle with the indeterminacy of their subject-matter, allowing a good dose of open-endedness and pluralism in their pictures of conceptual change. Nevertheless, all the models of conceptual change that we saw so far treat semantic indeterminacies as contingent factors in the evolution of concepts, not particularly central to the mechanisms by virtue of which concept change and therefore not central also to their modeling strategies. Indeterminate models of conceptual change, instead, conceptualize semantic indeterminacies as one of the central aspects, indeed as one of the central engines, of conceptual change[1].

In this chapter, I will focus specifically on two indeterminate models of conceptual change, i.e. Waismann's (Waismann, 1945) open texture model and Mark Wilson's (Wilson, 2006) framework of patches and facades. We will see how both these models understand conceptual change as a phenomenon centered around the openness and the plasticity of our concepts and theories about the world. Consistently with the central role assigned to semantic indeterminacies by these models, we will see how the usual concepts and tools of models of conceptual change are reshaped by these models in order to make room for the wanderings of scientific and philosophical concepts and theories. Moreover, both

[1]This characterization of indeterminate models of conceptual change is evidently not extremely precise nor exhaustive and, as such, it should be taken as a pragmatic one. Using my classification of models of scientific conceptual change presented in Chapter 2, indeterminate models could be grouped together with semantics models or with pragmatic models. Nevertheless, I chose to treat indeterminate models of conceptual change as a standalone kind of model because, as it will be clear later, they share a common general conception of the phenomenon of conceptual change and its philosophical implications.

Waismann and Wilson use the case of conceptual change as a basis for proposing a more general revisionary approach to philosophers' received view of semantic indeterminacies. If, in fact, the significance of semantic indeterminacies for philosophical problems is hardly a revolutionary proposal, both Waismann and Wilson spent a lot of ink in arguing for the positive role of semantic indeterminacies in the connections between our language and the world. Thus, Waismann's and Wilson's indeterminate models of conceptual change are not just two paradigmatic examples of a certain way of modeling conceptual change, but they underlie also a radical, original approach to philosophy of language and philosophy of science *tout court*. A paradigmatic example of this reshaping is Wilson's revisionary proposal of substituting our usual understandings of scientific concepts and theories with his notions of patches and facades. In order to understand what such revisionary proposals precisely consist of, I will rationally reconstruct (much of) Wilson's theory of conceptual change within a modified structuralist framework for reconstructing scientific theories.

In Section 1, I will present Waismann's open texture model, describing the philosophical background of Waismann's seminal work, together with some recent attempted reconstruction of Waismann's notion of open texture. I will argue that Waismann's open texture and his related understanding of conceptual change are best analyzed together with the other central notion of Waismann's philosophy of language, namely the notion of language strata. In Section 2, I will present Mark Wilson's account of conceptual behavior, describing his diagnosis of why analytic philosophy has often neglected the central role of semantic indeterminacies in conceptual affairs and presenting his revisionary framework of patches and facades. In Section 3, I will then show how Wilson's framework can be rationally reconstructed within a modified version of the structuralist view of scientific theories. More specifically, I will show how my modified structuralist framework that I will call Wilson-Structuralism is able to give a precise semantic reconstruction of many of the conceptual wanderings and indeterminacies described by Wilson as specific set-theoretic relations between parts of a scientific theory. Finally, in Section 4, I will analyze indeterminate models of conceptual change such as Waismann's and Wilson's ones through the lenses of the Toolbox framework.

5.1 Waismann's Open Texture and Language Strata

In this section, I will present Waismann's model of conceptual change as it can be reconstructed from Waismann's philosophical work. In order to do that, I will briefly analyze Waismann's philosophy of language from a historical and an abstract point of view, focusing particularly on the related, central notions of open texture and language strata. We will see that a correct understanding of these two notions and their relations, together with a general appreciation of their place in Waismann's philosophical thought, gives us a indeterminate model of conceptual change.

The originality of Waismann's philosophy has been for many years not appreciated by mainstream analytic philosophy. Waismann has in fact often been considered just a minor character in Wittgenstein's philosophical evolution. In recent years, the original-

ity of Waismann's philosophical contributions has been instead re-appraised, both from a contemporary (Makovec and Shapiro, 2019) and a historical (McGuinness, 2011) point of view. As it has been stressed by several scholars (Waismann, 1965; McGuinness, 2011; Lavers, 2019), Waismann did not just offer several original contributions in his late years in Oxford, but even in his Wittgensteinian phase he clarified and developed many Wittgensteinian themes and ideas to such an extent that his works cannot be considered mere interpretations of Wittgenstein's philosophy. In what follows, I will not offer a full account of Waismann's philosophy and its originality, but I will instead concentrate on his mature contributions devoted to problems in philosophy of language with a particular focus on the aforementioned notions of open texture and language strata.

5.1.1 Open texture

The notion of open texture is the most famous contribution of Waismann to philosophical terminology and analytic philosophy. Moreover, thanks to the work of Hart (Hart, 2012), open texture has transcended the disciplinary boundaries of philosophy and it has become an important concept in legal theory and practice[2]. In recent years, open texture has been the subject of both historical and analytical analyses that seek to clarify its meaning and its scope, in Waismann's thought as well as in contemporary philosophy (cf. the papers collected in Makovec and Shapiro 2019).

Waismann introduced open texture in his paper "Verifiability" (Waismann, 1945), in the context of his defense of a broadly empiricist attitude towards language and epistemology against accuses of reductionism[3]. The main goal of this paper is to distinguish healthy forms of empiricism from crude forms of radical reductionism, such as the project of translating all material statements into sense data ones. According to Waismann, the former positions consist in emphasizing the confirmation and disconfirmation of scientific statements as a central task of any reasonable epistemological project and can be successfully defended by critiques. Radical reductionist projects, instead, are for Waismann ill-conceived projects, due to the open-texture of most of our empirical concepts. Roughly speaking, open texture denotes for Waismann the essential incompleteness and openness of many of our empirical concepts. In contrast to some completely formalized and precise concepts, Waismann stresses the fact that is often unclear how to apply empirical concepts in unexpected situations. This essential plasticity of our empirical concepts is what causes the impossibility of a complete verification of any statements about the material world. Consequently, this impossibility determines the failure of any attempt to fully translate our material objects statements into phenomenalist language. Such a translation would, according to Waismann, require in fact to know in advance the conditions of verification a material statement, a requirement made impossible by the open texture of most of our empirical concepts. In Waismann's own words:

[2]For a full account of open texture history and significance in legal theory, see (Bix, 1991, 2019).

[3]It should be noted that the concept of open texture can be traced back to some remarks made by Waismann regarding the epistemological status of empirical hypotheses in his (Waismann, 193?).

> "Open texture is a very fundamental characteristics of most, though not of all, empirical concepts, and it is this texture which prevents us from verifying conclusively most of our empirical statements. Take any material object statements. The terms which occur in it are non-exhaustive; that means that we cannot foresee completely all possible conditions in which they are to be used; there will always remain a possibility, however faint, that we have not taken into account something or other that may be relevant to their usage; and that means that we cannot foresee completely all the possible circumstances in which the statement is true or in which it is false. There will always remain a margin of uncertainty. Thus the absence of a conclusive verification is directly due to the open texture of the terms concerned" (Waismann, 1945, p. 43).

Leaving aside Waismann's use of open texture as an impossibility argument against any form of reductionism, the assessment of which would require lots of historical context and would be hardly relevant to our present topic, let us focus on open texture as a purely semantic phenomenon[4].

Waismann warns us that spelling out the usages of most of our empirical concepts is a never-ending task. For how much established the semantic understanding of a certain empirical concept can be, we can always encounter new surprising conditions in which we do not know how to apply a given term. So that, even for what may appear perfectly stable empirical concepts such as cat, friend, and gold, the possibility of uncertainty given by their open texture presents itself in the form of gigantic cats, disappearing friends, and radioactive gold (cf. Waismann 1945, pp. 41-42). In contrast to (what he takes to be) the essential completeness of definitions in formal mathematics, Waismann takes our empirical concepts and statements to be bound to be amendable and revisable in light of surprising experiences. Note that Waismann aptly stresses that there can be open-texture without the terms involved exhibiting any vagueness at all. Natural kind terms, such as gold, for which no borderline cases usually arise and that thus no one would arguably consider vague, are in fact paradigmatic examples of open texture, as Waismann's example of a possible future discovery of a radioactive gold-like substance shows. As Waismann puts it, open texture is not vagueness, but "something like the possibility of vagueness" (Waismann, 1945, p. 42).

For what concerns our present topic, i.e. conceptual change, we can now see how Waismann's open texture implies the inherent revisability of most of our empirical concepts, thereby forbidding any conceptual analysis to claim complete success. Most of our empirical concepts, according to Waismann, are not stable and fixed entities, but they are fundamentally inexhaustive descriptions of the external world. As such, we cannot know or fix the meaning of many of our empirical terms once and for all, but we are forced to keep discovering and changing their semantic in the light of experience and usage. Any wannabe model of concepts and conceptual change, then, has according to Waismann to take into account the phenomenon of open texture.

[4]I will just note here that Waismann seems to be building a kind of a straw-man of reductionism, by claiming that reductionism equates with the impossible task of a complete once-and-for-all perfect translation of the vocabulary of the to-be-reduced parts of language.

We saw then that Waismann introduces open texture as the inherent open-endedness and revisability of many empirical concepts. Despite its intuitive strength, Waismann's presentation and the examples that he chose do not give us a completely clear picture of the semantic import of open texture. What exactly does the possibility of vagueness mean? Is there something like a test for ascertaining whether a given concept exhibits open texture? Are there semantic or cognitive limits to the revisability of empirical concepts? In other words, it is not clear what exactly a full theory of open texture would amount to. In the aforementioned recent surge of interest in Waismann and open texure, some contemporary philosophers have tried to clarify the notion of open texture, proposing renewed definitions and understandings of it.

The most known and more detailed proposal of what open texture consists of is indeed Stewart Shapiro's one (Shapiro, 2006a,a, 2013; Shapiro and Roberts, 2019). Shapiro retrieved Waismann's notion as a pivotal part of his contextualist account of vagueness (Shapiro, 2006a). According to Shapiro (Shapiro, 2006a, p. 10), open texture amounts to the possibility for competent speaker to decide either way in different contexts whether a certain term can be applied to a certain object. Defined in this way, open texture assures the existence of borderline cases in the application of certain terms, understanding these cases as unsettled by linguistic and pragmatic rules. In contrast to Waismann's original definition of open texture, Shapiro's open texture is then a mainly linguistic phenomenon inherently intertwined with the existence of vagueness and borderline cases. Shapiro (Shapiro, 2006a, p. 211) acknowledges this difference between his version of open texture and Waismann's one, but he (Shapiro, 2006a, pp. 212-215) argues that other works of Waismann, such as Waismann's (Waismann, 1949-1953) series of papers on analyticity, seem consistent with his open-texture-cum-vagueness unsettledness of borderline cases.

In more recent years, Shapiro used and defined the notion of open texture in slightly different ways. In a couple of papers (Shapiro, 2006b, 2013) on the intuitive concept of effective calculability (cf. Chapter 3, Section 3), he in fact used a more standard Waismannian definition of open texture as openness of a given concept. More interestingly, in a recent co-authored paper with Craige Roberts (Shapiro and Roberts, 2019), Shapiro gave a new definition of open texture, different from Waismann's one and even from the one I just presented, as a specific kind of linguistic and factual indeterminacy. According to this more recent definition (Shapiro and Roberts, 2019, p. 190), a predicate exhibits open texture if and only if it is possible for there to be an object such that nothing concerning the predicate established use nor concerning the non-linguistic facts, determines whether the predicate applies to the object. This definition, that within Shapiro's account of vagueness can be arguably shown to be equivalent to the previous one that he gave, does not stress anymore the possibility of going either way in the application of a term, but it focuses instead on the absence of linguistic and factual determination in the open texture cases.

Shapiro and Roberts discuss also how this indetermination dovetails nicely with Waismann's contextual notion of analyticity. Many example of open texture seem in fact, according to Shapiro and Craige (Shapiro and Roberts, 2019, p. 196), to display what linguists consider presupposition failures, i.e. cases in which standard presuppositions of our language do not apply, making the truth value of the related statement indeterminate.

These presuppositions are exactly the kind of revisable definitions that Waismann equates with analytic statements. Thus, open texture can be understood as the essential linguistic and factual underdetermination of most of our concepts that allow us to revise them. This revision prompted by the appearing of new cases forces us to change our established usages and presuppositions, thereby making our language evolve. Seen in this light, then, open texture functions as an engine of conceptual change in both philosophy and science (Shapiro and Roberts, 2019, pp. 205-206). Our language and our theories about the world are flexible and plastic in their evolution because most of our empirical terms are not fully determined in their usages and presuppositions. Semantic indeterminacies such as vagueness and open texture are thus not a problem to be solved, but they are engines of linguistic and scientific growth.

It should be noted that in recent years other reconstructions of what Waismann's open texture consists of have been proposed in the philosophical literature. Despite none of these reconstructions is as developed as Shapiro's one, they offer some interesting different ways of rationally reconstructing Waismann's notion. I will briefly mention their main insights in what follows. Waismann's open texture has been alternatively reconstructed in terms of expansion outside the standard domain of application of a term (Tanswell, 2018), assertory definitions (Vecht, 2020), or as a specific kind of prototype view of concepts (Zeifert, 2020). Despite the specific differences in framing the semantic implications of open texture between the different proposals, all these definitions, as well as Shapiro's ones, agree on stressing the positive role of open texture for conceptual change in science and in philosophy. I will not take a specific stance on which one of these definitions is a more accurate reconstruction of Waismann's notion. What I will do instead is stress an aspect of Waismann's open texture that has not been at the center of any of these reconstructions. Despite its negligence, we will see that this aspect of Waismann's open texture is pivotal to understand his conception of conceptual change and linguistic evolution. This neglected aspect of Waismann's open texture is the fact that open texture is just a specific aspect of the more general phenomenon that Waismann calls the stratification of language. In order to understand this connection, I need to briefly present the other central notion in Waismann's philosophy of language, namely the notion of language strata.

5.1.2 Language strata

If, as we saw, the notion of open texture has been recently at the center of a renewed scholarly interest in the philosophical literature, Waismann's notion of language strata has not received analogous attention. This lack of attention is unfortunate because, as we will see, behind the term language strata lies a very original theory of linguistic and philosophical practice. Moreover, I will argue that a proper understanding of the notion of language strata and its place in Waismann's philosophy of language allows us to better understand also the extent and the scope of Waismann's notion of open texture.

Even though a similar conception of language as a stratified entity can be found in various remarks in his previous works (cf. Waismann 1936, 1940, 1946a), Waismann explicitly introduces the notion of language strata in a two-part homonymous paper (Waismann,

1946b, 1953). In this work, Waismann introduces the idea that language is divided into layers, i.e. what he calls strata, that have different semantic, pragmatic, and epistemological properties. Different language strata might, according to Waismann (Waismann, 1946b, pp. 94-99), differ for what concerns their inner logic, their completeness, the texture of their concepts (i.e. whether their concepts exhibits open texture or not), the standard of verification that are valid within them, and even the notion of truth to which statements of a stratum are subject. So, language strata are parts of language that have a somewhat homogeneous mixture of semantic, pragmatic, and epistemological properties different from the one at work in other parts of language. Examples of kinds of statements that seem to belong to different language strata are for Waismann (Waismann, 1946b, p. 93) scientific laws, statements about the external world, phenomenological statements, statements about dreams, memories, and fictional statements.

The main insight of Waismann's language strata and their underlying picture of language is a localized holism in semantic and epistemological matters intertwined with an explicit pluralism and anti-reductionism. Against (what Waismann takes to be) the beliefs of both Wittgenstein and the ordinary language philosophy movement, our language has very few universal properties that can be discovered (cf. Waismann 1946b, p. 101). There is no universal logic nor universal verification procedure that all our statements exhibit (Waismann, 1945, 1946a,b). Different logics and different verification procedures are instead exhibited by different parts of our language. This recognition of the existence of different language strata forces us, according to Waismann, to abandon any heavy reductionism or monistic theory about language nature and functions. Instead of searching for universal properties, then, philosophers should pay more attention to the different parts of our language and their subtle interconnections (Waismann, 1953, pp. 118-121). By paying more attention to language strata, argues Waismann, philosophers will realize that many traditional philosophical problems such as the search for a correct logic or a correct notion of truth are just pseudo-problems caused by conflating two or more strata:

> "Thus language seems to be separated into strata by gaps over which one may jump but which cannot be bridged by logical processes. This fact accounts for many of the traditional problems in philosophy. The core of such a problem often lies in the difficulty of passing from one stratum to another. To give examples: If we start from sense datum statements and ask how we can arrive at material object statements, we are faced with the problem of perception; if we start from material object statements and ask how we can arrive at physical laws, we are studying the problem of induction; if we pursue the relations in the reverse order, i.e. if we travel from physical laws to material object statements and from the latter to sense datum statements, we are embarking on the problem of verification; and so on" (Waismann, 1946b, p. 100).

Gaps between language strata cannot then be resolved or explained away by philosophical theories, but they have instead to be recognized and accepted in their looseness and indeterminateness. Our linguistic practices, including science, naturally organize themselves according to Waismann in localized holistic domains characterized by distinctive semantic

and epistemological properties. The boundaries of these domains and their connections are often indeterminate and not easily knowable by us. An adequate philosophical depiction of language has then to recognize the plurality of language strata and their indeterminacy.

Now that I have briefly described Waismann's picture of language strata, I can explain my previous comments on the significance of language strata for a correct understanding of open texture and Waismann's picture of conceptual change. A neglected aspect of open texture is that it is just a specific aspect of the more general semantic and epistemological indeterminacy caused by the stratification of language. The porosity of most of our concepts, as Waismann's original term (*Porosität der Begriffe*) for open texture is more accurately translated, is just one aspect of the indeterminacy that language strata exhibit. Other kinds of conceptual porosities can in fact be found in Waismann's texts. Waismann speaks explicitly also about the porosity of our inferences (*poröse schlüsse*, Waismann 1945, p. 50), stressing the looseness and indeterminacy of the interconnections between different language strata. Thus, as concepts are not fully determined in advance in their usages and presuppositions, logical connections between parts of our languages are also for Waismann often not fully specified and fixed in advance. Similarly, from Waismann's (Waismann 1946b, pp. 94-99, Waismann 1953, pp. 112-117) descriptions of the other semantic and epistemological components of a language stratum, we can assume that also properties like truth, verifiability, completeness exhibit some kind of porosity, together with the language strata themselves.

Language is then for Waismann an inherently plastic and adaptive entity, in which different parts structure themselves through different local rules that are often not fixed and fully determined, but they change accordingly to practical needs of the related linguistic practices. As in the specific case of open texture, also other kinds of semantic indeterminacies are for Waismann positive engine of conceptual and theoretical change. The looseness of many of our concepts, our inferences, and many other of our linguistic tools gives us the possibility of revising and changing our conceptual tools according to practical needs and certain regulative principles (Waismann, 1945, pp. 63-65). Conceptual change is then for Waismann an ubiquituous necessary phenomenon by virtue of which language adapts to the unexpected situations that the world creates to us. Any adequate account of how this change works must take into central account the porosity and stratification of our language. Seen in this way, Waismann's open texture is then a devastating attack to any form of essentialism about concepts and kinds. Our knowledge of the world and of our linguistic practices can never be considered final and fixed, but we always have to pay attention to the subtle ways in which our language organize itself in order to successfully refer to the world.

5.2 Wilson's Conceptual Wanderings

After having seen Waismann's seminal notions of open texture and language strata, together with his underlying conception of language and conceptual change, the focus of this subsection will be on another indeterminate theory of conceptual change: Mark Wilson's

(Wilson, 2006) framework of patches and facades. As we will see, even though Wilson does not refer to Waismann in any of his works and there is no direct genetic link between Waismann's and Wilson's philosophical work, Wilson's account of linguistic and conceptual change shares many assumptions of Waismann's view of language. In contrast to Waismann's scattered remarks, however, Wilson's framework offers us a complete and fine-grained account of conceptual behavior that will allow us to understand better the characteristics of indeterminate model of conceptual changes.

Wilson expressed his views on concepts in many different places throughout his career. In my exposition, I will use as my main reference "Wandering Significance" (Wilson, 2006), building on the tools and the terminology he used there, making connections with other related works of his when needed.

The main theme of Wilson's work is that concepts are primarily adaptive localized tools for evaluating worldly activities. Analytic philosophy, according to Wilson, has mostly analyzed concepts and other 'terms of evaluations' wrongly, not understanding their intrinsically practical and contextual nature. This misunderstanding is at the heart of what Wilson calls 'the classical picture of concepts' (Wilson, 2006, pp. 139-146), i.e. a list of beliefs about concepts that constitutes the received view of conceptual behavior in analytic philosophy[5]. Fathers of analytic philosophy like Frege and Russell may have disagreed on specific epistemological views about how we acquire concepts or how they are structured, but they shared a core of semantical presuppositions about what concepts are. Amongst the many presuppositions forming the classical picture of concepts, the most important one (and dangerous one for Wilson) are the existence of fixed and stable conceptual contents, the possibility of successfully refining unclear concepts via conceptual analysis, and the fixity of truth-values of our claims involving concepts. In other words, concepts for the classical picture are stable entities, gluing together our language with the world. Mismatches between the outside world and our representations of it arise due to our failure of correctly grasping the true essence of a given concept.

Opposing this classical picture of concepts is another general set of semantical presuppositions about concepts that Wilson calls "anti-classical thinking" (Wilson, 2006, pp. 236-242). Anti-classicists thinkers such as Quine and (the later) Wittgenstein refused completely the classical picture of concepts, replacing it with a use-based conception of concepts as directives for performing actions[6]. In this view, the meaning of a concept cannot be grasped in isolation but only holistically in connection with other semantical and pragmatical expressions of our language.

Wilson sees the dialectic between classical and anti-classical thinking about concepts

[5]What Wilson calls the 'classical picture' of concepts should not be confused with what psychologists and philosophers of mind usually call the 'classical theory', i.e. the definitional view of conceptual structure that we saw in Chapter 2.

[6]It should be noted that certain parts of Quine's philosophy may appear quite entrenched in the classical picture of concepts, such as his satisfaction with Tarskian semantics (if understood in a disquotational manner). I will not consider here these matters of Quinean scholarship, but it may be argued that Quine was not as monolithically anti-classical as Wilson depicts him.

as one of the main driving forces of the birth and the development of analytic philosophy[7]. Despite their opposition, Wilson stresses that both these pictures of conceptual behavior share some unhealthy philosophical attitudes. Both classical and anti-classical thinkers suffer of what Wilson calls the "Theory T syndrome" (Wilson, 2006, p. 126), namely the bias of forcing all kinds of conceptual phenomena to fit into a neat single philosophical theory. The Theory T syndrome suppresses the individuality of individual concepts and the complexity of real world phenomena due to its obsession with monistic explanations. Thus, for instance, anti-classical thinkers like Quine, who correctly criticized the classical myth of straightforward coordination between our predicates and physical attributes, overshoot dramatically in banning any direct reference whatsoever from their philosophical views, suffering from what Wilson calls the "fear of attribute naming" (Wilson, 2006, pp. 262-273) and being thus condemned to a "hazy holism" (Wilson, 2006, pp. 280-286) for all kinds of concepts.

Wilson's remedy for the Theory T syndrome and more generally for analytic philosophy is to make a synthesis of classical and anti-classical thinking, which he calls (following Hume) "mitigated skepticism" (Wilson, 2006, pp. 599-605) about concepts. Certain concepts behave more like classical predicates, others play instead more practical roles and can therefore be understood only in a broader pragmatic context. Moreover, even the same concept can play a more classical role at a certain point of its history and a more anti-classical one at a later time. A mitigated skeptic ought not only to accept inter-conceptual pluralism in the way in which different concepts refer to the world, but also the intra-conceptual semantical "seasonalities" in the use of a given concept at different times or in different contexts. In order to have an adequate philosophical theory of how our language refers to the world, then, we have to give up entirely the idea of concepts as fixed semantical cores of our predicate usages. Wilson argues that even for the case of very simple predicates such as color ones, beneath their apparent simplicity lies a very complex web of contextual usages which cannot be grouped in a single semantical entity. So that the real crime of the Theory T syndrome is to neglect the complex 'personality' of our concepts, treating them all in the same simplistic way.

Wilson's mitigated skepticism covers both ordinary language and science. A great part of his work is devoted to show that even in physics our apparently straightforward ways of describing the world often hide quite convoluted referential architectures. In a painstakingly detailed analysis of several macroscopic predicates of classical mechanics such as 'force' or 'hardness', Wilson shows how their usages in scientific practice exhibit a complex patchwork of localized usages tailored for the specific contexts in which they operate. The complexity of how many of our scientific terms refer is not a contingent product of a human activity, it is often the only way in which we can say something meaningful about the world. The key to understand the necessity of this semantical complexity is to acknowledge that describing the macroscopic world is a very difficult affair. The number of

[7]Wilson traces back the roots of the classical/anti-classical picture of concepts, as well as the related birth of analytic philosophy, to debates about the foundation of classical mechanics in the nineteenth-century. For a critical survey of Wilson's historical analysis, see (Friedman, 2010).

variables that we must take into account even in a simple macroscopic scenario of classical mechanics such as, say, the trajectory of a cannon ball is so high that describing all of them accurately requires a computational power that exceeds by far our limited intellectual capacities and measurement abilities. This is why the practice of science requires what Wilson calls *physics avoidance* or variable reduction (Wilson, 2017, Ch. 2), i.e. a patchwork description of heavily simplified local domains connected in a way that efficiently reduces the complexity of the computations needed to describe a phenomenon.

5.2.1 Patches and facades

Wilson's view of conceptual behavior, both in ordinary language and in science, requires a patchwork structure that allows concepts to adapt contextually to localized usages. Wilson (Wilson, 2006, pp. 377-390) describes this patchwork semantical structure in terms of patches and facades.

Patches constitute the basic unit of Wilson' reconstruction of conceptual behavior. A *patch* is a localized mini-theory about (a specific part of) the world. A patch is composed by five different types of elements: vocabulary, domain, local reasoning tools, boundaries, and translation principles. The vocabulary of a patch is made of different linguistic entities such as predicates, names, relation symbols, and some limited logical and mathematical resources. The domain of a patch is a subset of a basic domain of physical facts, to which elements of vocabulary refer. Predicates, for instance, refer to one (or more) physical attribute(s) in the domain, while names denote constant elements. Local reasoning tools contain inferences and constraints on elements of the vocabulary that are valid within the domain of the patch. The boundaries of a patch constrain the contexts to which the patch can be applied. Finally, the translation principles are rules regulating how information can be exported and imported between a given patch and other patches connected to it.

A *facade* is a set of patches over a given domain of physical facts. Facades play the role of scientific theories, being collections of interconnected localized parts of our languages describing different aspects of a given phenomenon. Wilson thinks of facades as atlases of specific maps (Wilson, 2006, pp. 289-296), the patches, in which every map is useful for a given purpose, but no map has the foundational, privileged role that philosophers of science often assign to a certain 'constitutive' part of a theory. Every map has its own partially distorted way of representing the world and there is no neutral epistemic perspective from which one can judge a given map to be more truthful than another one. The technical difference between facades and traditional understandings of theories lies in the inherent multi-valuedness of how elements of a facade are connected. Wilson requires only partial connections between patches of a given facade, allowing a great deal of indeterminacy in the way in which localized linguistic usages work together to achieve a global description of a phenomenon. Different patches are allowed to assign different physical referents to a common predicate, forming what Wilson calls an *uneven facade* (Wilson, 2006, p. 324), a behavior exemplified by the concept of force in classical mechanics (Wilson, 2006, pp. 158-165, 175-182). Patches can also block the export of certain inferences or reasoning tools at the boundaries, so that two patches can share a common predicate that refers to the same

attribute(s) in every patch, but the inferences connected with this predicate may change from one patch to the other. Wilson calls this inference-blocking behavior a *Stokes facade* (Wilson, 2006, p. 324), since a paradigmatic example of it is the Stokes phenomenon in optics (Wilson, 2006, pp. 319-327). Patches can furthermore have partially or fully overlapping domains and they are allowed to give radically different descriptions of the same subset of the physical domain. Thanks to adequate translation principles, constraining the exchange of information between the patches, these different descriptions do not produce an inconsistency. This is how contemporary multi-scalar models in engineering science manage to describe the complex behavior of some materials (Wilson, 2017, Ch. 1). Patches can also be connected in such a way that a given predicate figures in the vocabulary of several interconnected patches without any common inter-patch reference, i.e. referring to different attributes in every patch it figures. This phenomenon is called by Wilson quasi-attributes or *ghost properties* (Wilson, 2006, p. 273) and it is strikingly exemplified by the concept of hardness (Wilson, 2006, pp. 335-355). More generally, facades must be thought as very dynamical entities, whose structure of interconnected patches changes on the basis of the practical needs of scientific research, creating new patches that often force the overall structure of a facade to adjust itself.

Facades come equipped with what Wilson calls a *semantical picture* (Wilson, 2006, p. 307), i.e. a description of how the vocabulary of a given patch matches the part of the world it is designed to cover. Semantical pictures are then, in Wilson's cartographic metaphor, like prefaces to atlases, explaining the peculiarities and the distortions that a given map exhibits. Since the interconnections between patches of a facade are multivalued and dynamical, semantical pictures need to be periodically revised. In particular, Wilson describes a very common phenomenon that causes such a revision, i.e. what he calls the *canonical developmental history* of a predicate (Wilson, 2006, pp. 534-535). This phenomenon starts with the extension of a given predicate to a new patch the semantical underpinning of which are unclear. If this new context of usage of the predicate proves practically successful, this new patch becomes sufficiently established to cause (in due time) the development of an adequate semantical picture for it. This semantical picture for the new patch can force, modulo conceptual inconsistencies, scientists to replace the old semantical pictures of the predicate under focus with the new one also in its older contexts of usages, concluding a canonical developmental history. This last step of the history, i.e. the replacement of old semantical pictures with a new one is dubbed by Wilson *semantic detoxification* (Wilson, 2006, pp. 545-552). Canonical histories of predicates exemplify for Wilson why the classical and the anti-classical picture of concepts are just unhealthy philosophical hypostatizations of different moments of a predicate life. In extending a predicate to new usages, people behave anti-classically, following practical successes of languages and not paying much attention to how this success is semantically justified. Later, when this extension has proven to be sufficiently successful and stable, semantical worries enter the picture and successful usages of a predicate need to be classically justified in their worldly correlations.

The bestiary of conceptual wanderings described by Wilson can then be seen, despite the aforementioned lack of any explicit genetic connections between the two philosophies,

as a more fine-grained description of the porosity of our conceptual and inferential tools that Waismann stressed with his notions of open texture and language strata. In the structure of uneven facades and the related wanderings of our scientific terms we can see the open texture of most of our empirical terms, while stokes facades with their loose inferential connections give us a model of the looseness of inferences between different language strata. In Wilson's incessant reminder of how linguistic individualities and the specific pragmatic contexts affect the semantic of our scientific predicates we can see a reminiscence of Waismann's anti-reductionism stance.

Wilson offers then a deflationary understanding of scientific conceptual behavior centered around the notions of facades and patches. The partial and plural connections between patches of a given facade make the overall structure dynamically revisable according to the practical needs of science, leaving space for the inevitable conceptual wanderings necessary for this very human activity.

5.3 Taming Conceptual Wanderings: Wilson-Structuralism

We saw in the last section how Mark Wilson (Wilson, 2006, 2017) present a highly original account of conceptual behavior that challenges many received views about concepts, reference, and conceptual change in analytic philosophy. Despite the vast praise of Wilson's work (Brandom, 2010; Friedman, 2010; Pincock, 2010; Carus, 2012a), few attempts have been made to give a precise semantic reconstruction of his framework.

In this section, I will show how a modified version of the structuralist view of scientific theories (Sneed, 1979; Stegmüller, 1976; Balzer et al., 1987; Balzer and Moulines, 1996) is able to rationally reconstruct Wilson's framework of patches and facades. At first sight, the choice of reconstructing Wilson's ideas within a structuralist framework may appear quite surprising. Few frameworks in philosophy of science appear as distant as the structuralist one from Wilson's work. Wilson's reconstruction of scientific theories is in fact informal, committed to a realist understanding of scientific terms, and oriented towards scientific practice. In comparison, structuralism in philosophy of science is characterized by a more formal approach to the subject matter, it is often combined with a somewhat anti-realist understanding of theoretical terms, and it lacks an analogous heavy focus on the practice of science. Despite these differences, I will show that there are some interesting connections between the reconstruction of scientific theories offered by Wilson and the one championed by the structuralists. Moreover, we will see that the Structuralist framework, when adequately modified to eliminate its hierarchical understanding of scientific theories, is able to offer a precise semantic reconstruction of Wilson's ideas.

More specifically, I will show how my modified structuralist framework, i.e. what I will call *Wilson-Structuralism*, offers a semantic reconstruction of scientific theories capable of modeling Wilson's account of conceptual behavior. Specifically, I will argue that Theory-Elements and Wilson-Theory-Nets explicate respectively Wilson's patches and facades,

thanks to the relaxed inter-elements constraints and the weak-specialization relationship of Wilson-Structuralism. In order to support my claim, I will demonstrate how several wandering phenomena described by Wilson can be adequately understood in a more abstract way within my framework. I will also further strengthen my case by showing how one of Wilson's main case studies of the wandering behavior of scientific terms, i.e. viscous fluids forces in classical mechanics, can be adequately reconstructed within Wilson-Structuralism.

This work has then a three-fold aim. The first aim is to offer a rational reconstruction of Wilson's framework of patches and facades, thanks to which many of the wandering phenomena described by Wilson can be given a precise semantical understanding within a formal framework. The second aim is to offer a modified version of the structuralist framework that eliminates its original hierarchical understanding of scientific theories, offering an alternative way of reconstructing scientific theories that allows the semantic indeterminacy prescribed by Wilson. The third, more general, aim of this work is to show a surprising connection between two *prima facie* very different ways of reconstructing scientific theories, namely Structuralism and Wilson's framework of patches and facades.

First, in the next subsection, I will present Structuralism in the philosophy of science and its model-theoretic way of reconstructing scientific theories. Then, I will present Wilson-Structuralism, i.e. my modified structuralist framework that eliminates the hierarchical aspect of the structuralist reconstruction of scientific theories. I will show how this modified structuralism is able to explicate Wilson's patches and facades, adequately representing several wandering-phenomena described by Wilson. I will also present a rational reconstruction in my framework of one of Wilson's main case studies, i.e. the case of viscous fluids forces.

5.3.1 Structuralism in philosophy of science

The research program of Structuralism, understood as the model-theoretical way of reconstructing scientific theories that started with the work of Sneed, has been presented in different forms throughout the years (Sneed, 1979; Stegmüller, 1976; Balzer et al., 1987; Balzer and Moulines, 1996). In my presentation I will use as my main reference (Balzer et al., 1987), which is arguably the most mature and complete presentation of the structuralist program.

As the name suggests and as we briefly saw in Chapter 2 in our discussion of conceptual change in science, the key idea of the structuralist program is that scientific theories and their conceptual dynamics are best reconstructed in terms of structures. This structure-centered view was originally supposed to constitute a non-statement view of scientific theories, contrasting with (what allegedly was) the logical empiricists' statement view orthodoxy[8]. Scientific theories are best reconstructed not as a bundle of statements, but instead as a collection of set-theoretic structures.

[8]It should be noted that recent scholarship has convincingly argued that the opposition between statement and non-statement view of scientific theories has been largely overstated by the proponents of non-statement views. For recent perspectives bridging this alleged opposition see (Schurz, 2014a,b; Andreas, 2014; Lutz, 2014).

Specifically, the smallest unit of scientific theories reconstruction in the structuralist framework is a *theory-element*, i.e. a ordered pair $T = \langle K, I \rangle$ where K is a theory-core and I is the set of intended applications of the theory element[9]. Theory-elements represent law-like scientific statements and are the building blocks of the structuralist reconstruction of scientific theories. A *theory-core* is a quintuple $K = \langle M_p(T), M(T), M_{pp}(T), GC(T), GL(T) \rangle$ that represents the theoretical framework of a given theory element, including its conceptual framework, its models, its possible empirical applications, the connections between its different applications, and the relations between the theory-element and other theory-elements of the same scientific theory.

Formally, the first component of a theory core $M_p(T)$ is the class of *potential models* of the theory-element, i.e. the set of structures of the type $\langle D_1, \ldots, D_k, A_1, \ldots, A_k, n_1, \ldots, n_p, t_1, \ldots, t_q \rangle$ that satisfy the non-law-like axioms (i.e. the statements expressible via typifications and characterizations, Balzer et al., 1987, pp. 14-17) of the theory, where $D_1, \ldots, D_k$ are sets of empirical objects, $A_1, \ldots, A_k$ sets of mathematical objects, $n_1, \ldots, n_p, t_1, \ldots, t_q$ (respectively) non-theoretical and theoretical relations between members of the two kinds of sets (usually functions from empirical objects to mathematical ones)[10]. Potential models are thus the conceptual framework of a given theory-element, representing the kind of entities for which it is meaningful to ask whether they satisfy the law-like statement represented by the theory-element. The second component of the theory core $M(T)$ is the class of *models* of the theory element, i.e. the subset of potential models including all and only the ones satisfying the law-like axioms (i.e. the statements that are neither characterizations or typifications) of the theory-element. Intuitively, models are the entities for which the law-like statement represented by the theory-element is true. Then we have the class of *partial potential models* of the theory-element $M_{pp}(T)$, i.e. the set of structures of type $\langle D_1, \ldots, D_k, A_1, \ldots, A_k, n_1, \ldots, n_p \rangle$ for which $\langle D_1, \ldots, D_k, A_1, \ldots, A_k, n_1, \ldots, n_p, t_1, \ldots, t_q \rangle$ is a potential model of T. Partial potential models are thus truncated potential models, in which the only relations remaining are the non-theoretical ones. Intuitively, they represent the empirical entities to which the law-like statement of the theory-element can be applied. The *global constraint* of the theory-element $GC(T) = \bigcap\{C_1(T), \ldots, C_n(T)\}$ is the combination of all constraints $C_1, \ldots, C_n$ between different applications of the theory-element. A constraint C is the set of admissible combinations of potential models of the theory element, i.e. $C \subseteq \wp(M_p)$ such that $C \neq \emptyset$, $\emptyset \notin C$, and $\forall x \in M_p : \{x\} \in C$. Intuitively, constraints represent physical and conceptual requirements between different applications of the law-like statement represented by the theory-element. Examples of

[9]Here I use the first of several simplifications of the structural framework that are needed to keep the presentation of this complex framework contained, namely, I identify theory-elements with what the structuralist actually call idealized theory-elements (Balzer et al., 1987, pp. 89-92). In what follows I will consistently use the idealized version of empirical claims, theory-nets, and all the other related notions.

[10]For simplicity, I do not present here the structuralist way of distinguishing theoretical and non-theoretical relations/predicates of a theory. The distinction is centered around the notion of a T-admissible method of determination (Balzer et al., 1987, pp. 47-78) and its exposition would require the introduction of several auxiliary notions. I chose thus not to inflate the presentation of the structuralist framework, assuming the possibility of making this distinction.

constraints are the equality constraint for mass in classical particle mechanics or the extensivity of energy in simple equilibrium thermodynamics. Finally, the last component of a theory-element is its *global link* $GL(T) = \bigcap\{L_1(T), \ldots, L_n(T)\}$, i.e. the intersection of all abstract and concrete links $L_1(T), \ldots, L_n(T)$ between (components of potential models or potential models of) the theory-element T and (components of potential models or potential models of) other theory-elements $T_1, \ldots, T_n$ relevant for the theory-element. Links $L \subseteq M_p \times M'_p$ are admissible combinations of potential models of different theory elements, i.e. constraints between different theory-elements[11]. Intuitively, they represent constraints that certain law-like scientific statements of a given theory impose on their law-like neighbors for maintaining the consistency of the overall theory, such as the kinematical axioms which determine position and the chronological conditions for the measurement of time in classical particle mechanics.

If the theory-core is the theoretical framework of a theory-element, then the set of *intended applications* I is its empirical part. The set of intended applications represents in fact the empirical circumstances to which a given theory-element is intended to apply. Formally, the set I is a subset of the class of partial potential models of the theory-element, but it is meant to be only pragmatically defined as a list of paradigmatic examples that a theory has to cover in order to be successful. The success of a given theory-element can be formally expressed by the related *empirical claim* of a theory-element, defined as $I \in Cn(K)$ (the set of intended applications is included in the content of the theory-element), where $Cn(K) := r(K)(\wp(M(T)) \cap GC(T) \cap \wp(GL(T)))$ is the content of the theory-element. The content of the theory element is the set of sets of partial potential models, obtained cutting out the theoretical terms of potential models via the operation $r(K) : M_p \rightarrow M_{pp}$, that can be augmented by theoretical terms such that the resulting potential models are models $M(T)$ that satisfy all constraints $GC(T)$ and all inter-theoretical links $GL(T)$, i.e. the members of the set $\wp(M(T)) \cap GC(T) \cap \wp(GL(T)$.

In order to clarify all these model-theoretic definitions, I will present a simple example of a theory-element, namely the one reconstructing classical particle mechanics (CPM) (Balzer et al., 1987, pp. 103-107). This theory-element is the foundational unit of the structuralist reconstruction of classical mechanics, representing the most general presentation of Newton's second-law of motion. For brevity, I will not discuss the intertheoretical links of CPM, focusing only on the self-contained components of the theory-element. The paradigmatic intended applications of the classical particle mechanics theory-element are the solar system, the pendulum, the projectile, and the harmonic oscillator (all members of the set $I(CPM)$). A potential model of classical particle mechanics is constructed as follows: x is a $M_p(CPM)$ iff:

- $x = \langle P, T, S, \mathbb{N}, \mathbb{R}, c_1, c_2, s, m, f\rangle$

[11]Here I am again simplifying the presentation of the structuralist framework, omitting the definition of concrete links (i.e. links between single components of partial potential models) and cutting a degree of complexity in the definition of a global link. For a full presentation of intertheoretical links see (Balzer et al., 1987, pp. 57-62, 78-79)

- P is a finite non-empty set, T, S are non-empty set, $\mathbb{N}, \mathbb{R}$ denote respectively the set of natural and real numbers.
- $c_1 : T \to \mathbb{R}$ and $c_2 : S \to \mathbb{R}^3$ are bijective and they denote respectively the coordination function for time and space.
- $s : P \times T \to S$ denotes the position function, $c_2 \circ s_p \circ \breve{c_1}$ is smooth for all $p \in P$.
- $m : P \to \mathbb{R}^+$ denotes the mass function and $f : P \times T \times \mathbb{N} \to \mathbb{R}^3$ denotes the force function (made of $\mathbb{N}$ force components).

This is then the conceptual framework of the theory-element representing Newton's second law of motion in its general form, making precise the kind of entities that may satisfy Newton's law. A model of classical particle mechanics is then a potential model that actually satisfies Newton's second law. More precisely, x is a $M(CPM)$ iff:

- $x = \langle P, T, S, \mathbb{N}, \mathbb{R}, c_1, c_2, s, m, f \rangle$ is a $M_p(CPM)$
- $\forall p \in P$ and $a \in \mathbb{R} : m(p)D^2r(p, a) = \Sigma_{i \in \mathbb{N}} f(p, \breve{c}_1(a), i)$ (i.e. Newton's second law of motion).

An example of a constraint for CPM is the equality constraint for force. This constraint requires the i-th component force acting on a given particle at a given time to be the same independently from the system to which the particle belongs:

$$C(CPM) = \{X | \emptyset \neq X \subseteq M_p(CPM) \wedge \forall x, y \in X, \forall p, t, i$$

$$(p \in P_x \cap P_y \wedge t \in T_x \cap T_y \wedge i \in \mathbb{N} \to f_x(p, t, i) = f_y(p, t, i))\}$$

Theory-elements are then organized in larger units of scientific theories reconstruction called *theory-nets*. Theory-nets are collections of theory-elements sharing a significant part of their structure, organized via a relation of specialization. Intuitively theory-nets are meant to reconstruct large scientific theories, such as classical particle mechanics *tout court*. These large scientific theories are thought as hierarchies of more and more specialized theory-elements, representing specific sub-fields of applications of the more general theory-element in which more restrictive conditions hold (e.g. Hooke's law for elastic forces). Formally, a theory-net is a poset (i.e. a partially ordered set) $N = \langle \bar{T}, \sigma \rangle$, where $\bar{T}$ is a non-empty, finite set of theory-elements and $\sigma : \bar{T} \times \bar{T}$ is a specialization relation such that $T' \sigma T$ (T' is a specialization of T) iff $M_p(T') = M_p(T)$, $M_{pp}(T') = M_{pp}(T)$, $M(T') \subseteq M(T)$, $GC(T') \subseteq GC(T)$, $GL(T') \subseteq GL(T)$, and $I(T') \subseteq I(T)$. Thus a theory-net is a collection of theory-elements sharing the same classes of potential and partially potential models and partially ordered in terms of set-theoretic inclusion of their classes of models, admissible combinations of potential models (both intra- and inter-theory element ones), and intended applications. Structuralists stress that further restrictions should be imposed on the specialization-ordering of the theory-elements of a given

theory-net for harmonizing the empirical claims of theory-elements into a substantial global claim sharing a common empirical ground. Two important restrictions that contribute to this harmonization of empirical claims are connectedness, i.e. $\forall T_i, T_j \in \bar{T}\ \exists T_{k_1}, \ldots, T_{k_n}$ such that $(T_i \sigma T_{k_1} \vee T_{k_1} \sigma T_i) \wedge \ldots \wedge (T_j \sigma T_{k_n} \vee T_{k_n} \sigma T_j)$, and having a singleton-basis, i.e. $B(N) = \{T | T \in \bar{T} \wedge \forall T' \in \bar{T}(T \neq T' \rightarrow \neg T \sigma T')\}$ contains exactly one element. A connected theory-net having a singleton-basis is called a *theory-tree*. The most detailed example of scientific theory reconstruction in the structuralist program, i.e. classical particle mechanics, is a theory-tree. Theory-trees can thus be considered the paradigmatic structuralist unit of reconstruction of big scientific theories. It is important to note that the two harmonizing restrictions that theory-nets must satisfy in order to classify as theory-trees, i.e. connectedness and singleton-basis, strengthen the hierarchical aspect of the structuralist reconstruction of scientific theories. Scientific theories that can be reconstructed as theory-trees have in fact a single, most general and most fundamental theory-element (i.e. the singleton-basis) of which every other theory-element is a specialization. Every other law-like statement of the scientific theory so reconstructed is then understood as merely specifying more restrictive conditions under which the fundamental law holds.

As in the case of theory-elements, I will present a simple example of a theory-net, namely a (very small) part of the theory-tree reconstructing the whole of classical mechanics. The CPM theory-element that we have seen before is the singleton-basis of the classical particle mechanics theory-tree and thus all the other theory-elements are specializations of CPM, imposing further constraints on CPM-models, constraints, and links. The very detailed structuralist reconstruction of the CPM theory-tree (Balzer et al., 1987, pp. 180-191) singles out four different main lines of specializations of the CPM theory-element, namely symmetry forces, position-dependent forces, velocity-dependent forces, and time-dependent forces. In what follows, I will briefly present only two theory-elements belonging to the velocity-dependent forces part of the theory-tree, for both their simplicity and their relevance for Wilson's case studies. A simple example of a specialization of CPM is given by the theory-element of velocity-dependent classical particle mechanics ($VCPM$). The models of $VCPM$ are models of CPM in which at least one component force of a given particle depends on the particle velocity. Formally, x is a $M(VCPM)$ iff:

- $x = \langle P, T, S, \mathbb{N}, \mathbb{R}, c_1, c_2, s, m, f \rangle$ is a $M(CPM)$;
- $\exists p \in P, i \in \mathbb{N}$ such that for all $a \in \mathbb{R}$: $f(p, \check{c}_1(a), i) = F(Dr(p, a), a)$;
- $\exists p, a$ and $j \in \{1, 2, 3\}$ such that $D_j F(Dr(p, a), a) \neq 0$.

Another specialization of CPM, which is also a specialization of $VCPM$, is the theory-element of simple frictional classical particle mechanics ($SFCPM$). The models of $SFCPM$ are models of $VCPM$ (and thus of CPM) in which at least one component of frictional force is determined by a power of the velocity alone. Formally, x is a $M(SFCPM)$ iff:

- $x = \langle P, T, S, \mathbb{N}, \mathbb{R}, c_1, c_2, s, m, f \rangle$ is a $M(VCPM)$;

- $\exists p \in P, i \in N$ such that for all $a \in \mathbb{N}$: $f(p, \check{c}_1(a), i) = b(p, i)(Dr(p, a))^l$ with $b : P \times \mathbb{N} \to \mathbb{R}$ and $l \in \mathbb{N}, l \geq 1$.

These are thus the central notions of the structuralist framework for reconstructing scientific theories, centered around the concepts of theory-elements and theory-nets. These two notions denote the specific kinds of structures representing (respectively) localized and general scientific theories[12]. As it will be clear in the next section, these two notions, when adequately modified, correspond quite naturally to Wilson's notions of patches and facades.

5.3.2 Wilson-Structuralism

After having presented the structuralist framework in its original form, I will show how by eliminating its hierarchical aspect one obtains a semantic view of scientific theories capable of reconstructing many of the wandering phenomena described by Wilson. Thus, I will not use Wilson's theory as a target phenomenon for a structuralist reconstruction of the kind through which structuralists analyzed various kinds of scientific theories. What I will do, instead, is to show how to change two central structuralist notions in order to make the structuralist framework able to reconstruct scientific theories in a way that allows the kind of semantic indeterminacy prescribed by Wilson. More specifically, I will show how the structuralist notions of theory-elements and theory-nets, when the latter is adequately modified, provide a formal equivalent of Wilson's notions of patches and facades.

I will first present my modified structuralist framework in its general form and then I will show how it can rationally reconstruct one of Wilson's main examples of conceptual wandering in classical mechanics, i.e. the case of viscous fluids forces.

Wilson-Theory-Nets

We have seen in the last section how the structuralist reconstruction of scientific theories is centered around two notions, theory-elements and theory-nets. If the former is meant to reconstruct law-like specific parts of a scientific theory, the latter organizes all these specific parts into a more coherent whole. The level of generality of these two structuralist notions exactly corresponds to the one of Wilson's patches and facades. Moreover, theory-elements, just like Wilson's patches, are micro-theories about a specific part of the world (i.e. the models of the theory-elements), made of a conceptual and linguistic part (i.e. their potential models) together with their own reasoning tools and connections with neighboring theories (i.e. their constraints and inter-theoretical links). Theory-nets, then, exactly like Wilson's facades, are collections of micro-theories (i.e. theory-elements) over a given macro-domain that organize the connections between these micro-theories by constraining their components (via their specialization relation).

[12]I must stress that in my presentation I focused only on a small (albeit central) part of the structuralist galaxy of units of scientific theory reconstructions. Important pragmatic and dynamic extensions of theory-nets and theory-elements are for instance non-idealized theory-nets (Balzer et al., 1987, pp. 357-362), theory-evolutions (Balzer et al., 1987, pp. 216-221), and theory-holons (Balzer et al., 1987, pp. 387-407).

Can we then easily reconstruct in the structuralist framework Wilson's patches as theory-elements and Wilson's facades as theory-nets? Unfortunately not. The problem with this tentative mapping is the aforementioned heavily hierarchical aspect of the structuralist reconstruction of scientific theories. Theory-nets (and a fortiori theory-trees) organize, in fact, theory-elements into a strongly hierarchical chain of specialization relation(s), where specific applications (in the intuitive sense of the word) of a theory are supposed to be always conceptually reducible to more general law-like statements. Applications of a scientific theory are then in the structuralist reconstruction of a theory just model-theoretic precisifications of the related fundamental theory-element. Theory-nets are then a perfect example of the kind of semantic finality based received view of scientific theories that Wilson repeatedly attacks in his work. The whole bestiary of semantic wanderings presented in "Wandering Significance" can be seen as a list of ways in which the uses of scientific terms defy a structuralist-like hierarchical reconstruction of a scientific theory. Wilson, in fact, repeatedly shows how applications and non-fundamental law-like statements of a scientific theory are not reducible to mere precisifications of a more general law, but they often expand, twist, and extend the uses of the scientific terms and the reasoning tools of the theory in unexpected ways. Despite the aforementioned similarities between the two notions, structuralist theory-nets cannot then (in their canonical form) adequately mimic Wilson's facades, due to the rigidity in their hierarchical organization of theory-elements.

In order to solve this problem, I will now present a modified structuralist framework, i.e. Wilson-Structuralism, in which I eliminate this hierarchical aspect of the structuralist reconstruction of scientific theories. I will keep the coarse-grained organization of Structuralism, but I will drastically change its representation of how the different law-like statements of a given scientific theory are organized. More specifically, I will change the definition of a theory-net and the related specialization relation. I will not require theory-elements of the same theory-net to be related by subset inclusion of models, constraints, and links, but only by the non-empty intersection between these components. This change will allow the modified theory-nets to enjoy the 'multi-valuedness' needed to adequately represent several wanderings phenomena described by Wilson.

Formally, I will leave completely unchanged the structuralist definition of a theory-element. Just like in classical structuralism, in Wilson-structuralism a theory-element T is a couple $T = \langle K, I \rangle$ where $K = \langle M_p, M, M, GC, GL \rangle$ is a theory core and I denotes the intended applications of the theory-element. All components of the theory-core are defined exactly like we saw in Section 3, as well as the intended applications.

Theory-elements in Wilson-Structuralism are meant to explicate Wilson's patches, i.e. self-contained micro-theories about a subset of a given domain. The components of a patch that Wilson describes can be adequately mapped to the ones of a given theory-element. The vocabulary of a patch is represented in Wilson-structuralism by the potential models of the related theory-element. Potential models are in fact the conceptual framework of a given theory-element, representing the linguistic part of the law-like statement. The domain of a Wilson's patch is instead mapped to the models of the related theory-element. Models of a theory-element depict in fact in an anti-realist way the possible scenarios that satisfy the law-like statement, all the possible 'denotations' of the related scientific terms. The

boundaries of Wilson's patches are instead mapped to the intended applications of a theory-element, i.e. to the pragmatically defined subset of the class of partial potential models that represents the empirical situations to which the theory-element should apply. The reasoning tools of a Wilson's patch are mapped to the constraints of the theory-element. This mapping is justified by noting that Wilson's patches are equipped with a variety of reasoning tools that encompasses also the kind of physical and conceptual requirements on scientific terms that are framed by the structuralists as constraints on the related theory-element. Moreover, it has been recently shown how more paradigmatic reasoning tools such as deductive inferences can be represented in the structuralist framework as constraints on the acceptable combinations of potential models of a theory-element (Andreas, 2013). Finally, a patch translation-principles, i.e. the rules that norm the import and the export of information with other neighbor patches, are mapped to the intertheoretical links of the theory-element, a component that does the same exact job of Wilson's translation-principles between different theory-elements.

If theory-elements adequately represent Wilson's patches, the role of facades is played in Wilson-structuralism by an adequately modified version of theory-nets. The key difference between what I will call Wilson-theory-nets and the traditional definition that I presented in the last section is the specialization relation. We have seen that in traditional structuralism the specialization relation of theory-nets hierarchically orders theory-elements by requiring (weak) subset inclusion of their models, constraints, links, and intended applications (and equality of potential models and partial potential models). Formally, in a traditional theory-net, a theory-element T' is a specialization of another theory-element T, i.e. $T'\sigma T$, iff

$$M_p(T) = M_p(T'),\ M_{pp}(T) = M_{pp}(T'), I(T') \subseteq I(T),$$

$$M(T') \subseteq M(T), GC(T') \subseteq GC(T), GL(T') \subseteq GL(T).$$

In Wilson-structuralism, I replace this specialization-relation with a weaker version that I will call *weak specialization*. Like orthodox specialization, weak-specialization requires equality of potential models and partial potential models, as well as weak subset-inclusion of intended applications, but it only requires non-empty intersection between the models, the constraints, and the links of the two theory-elements. This means that a theory-element is a weak-specialization of another one if they share the same conceptual framework and empirical ground (aka potential models and partial potential models), its range of applications is (weakly) included in the one of the other, and they have compatible models, constraints, and inter-theoretical links.

Formally, a *Wilson-Theory-Net* (WTN) is a poset $N^W = \langle \bar{T}^W, w\sigma \rangle$, where $\bar{T}^W$ is a non-empty, finite set of RW-theory-elements and $w\sigma : \bar{T}^W \times \bar{T}^W$ is a *weak specialization* relation such that $T'^W w\sigma T^W$ (T'^W is a weak specialization of T^W) iff

$$M_p(T) = M_p(T'),\ M_{pp}(T) = M_{pp}(T'), I(T') \subseteq I(T),$$

$$GC(T') \cap GC(T) \neq \emptyset, GL(T') \cap GL(T) \neq \emptyset, M(T') \cap M(T) \neq \emptyset$$

.

These new definitions allow in a Wilson-theory-net specializations of theory-elements to have different (although compatible) models, constraints, and inter-theoretical links than the more general theory-element of which they are a specialization. Thus, the hierarchical structure of a traditional theory-net is maintained in a Wilson-theory-net only for what concerns the intended applications, while the theoretical relationships between the cores of the theory-elements are allowed far more diversity (modulo direct-neighbor compatibility).

This diversity is the key element that allows Wilson-theory-nets to adequately represent Wilson's facades. The relaxed constraints imposed on the components of the different theory-elements of a Wilson-theory-net allow several conceptual wanderings described by Wilson, properly understood as particular set-theoretic relationships between components of different theory-elements. In order to be represented in Wilson-structuralism, Wilson's wanderings have in fact to be reconstructed as differences in the mathematical structure that (parts of) the related theory-elements denote. This is because of the differences in the degree of realism and externalism in the semantics of scientific terms between Wilson's and the structuralist framework. As I hinted in the Introduction, in fact, if Wilson freely talks about an external reality in which scientific terms take their reference by aligning themselves with a (set of) attribute(s), the structuralist framework reconstruct scientific terms as certain kinds of functions or relations, i.e. as certain components of the potential models of a given theory-element. Specific reconstructions and views about the ontological status of theoretical terms vary within the structuralist camp, from arguably anti-realist reconstruction (e.g. Sneed 1979; Andreas 2014) to more neutralist approaches (e.g. Stegmüller 1976; Moulines 1991), but no specific version of the structuralist framework conceptualizes scientific terms in a strongly realist way like Wilson. According to Structuralists, the denotation of a given scientific term is made of all the possible abstract structures referred to by the occurrences of the related function term in the actual models of the related theory-elements. Thus, Wilson's inter-patches changes in the alignment between certain predicates and the attributes they refer to have to be reconstructed, in a structuralist framework, in a more abstract way as certain kinds of differences in the structures denoted by (all the occurrences of) the function-terms in the models of the related theory-elements. As a guide to this abstract representation of the semantic indeterminacies described by Wilson, I will use the following metaphysical assumption:

> Assumption: Two different attributes cannot be represented by the same class of mathematical entities ($\approx$ contraposition of Leibniz's identity of indiscernibles). Therefore, if a certain predicate refers to two different attributes in two different contexts, this predicate cannot be represented by the same mathematical entity in both contexts.

The idea behind this quasi-Leibnizian assumption is that difference in scientific terms reference must have a correlate in the mathematical representation of that part of a sci-

entific theory. For instance, if a given predicate refers to two different attributes in two different contexts (i.e. Wilson's uneven facades construction), then its mathematical representation has to be different as well in the two contexts. Differences in reference must have some discernibles consequences in the logical reconstruction of a theory. Formally, in order to talk about a predicate and its representation in my framework, I use the set-theoretic projection function $\Pi(T, i_1 \dots i_x)$ = set of all entities appearing in places $i_1 \dots i_x$ in theory T (cf. Balzer et al. 1987, p. 61, Moulines 1981, pp. 214-215). This function picks out from each tuple in T the entities occupying the $i_1 \dots i_x$ places, thus having as a domain the set of all ordered tuples of T and as a range the set of all entities occupying the places $i_1 \dots i_x$ in tuples of T. So, for instance, the function $\Pi(M(CPM), f)$ refers to the union of all the force functions f appearing in the models of the CPM theory-element, thus picking out from each tuple in $M(CPM)$ its force component. This projection function, together with the metaphysical assumption above, constitute the main tools for representing inter-patches changes of scientific term meaning in Wilson-structuralism. Then, I will say that the meaning of a given scientific term is different in two directly connected patches if and only if the two theory-elements representing these patches, provided that they are directly connected by a weak-specialization relation, have models with sets of functions representing the given scientific terms that are incomparable with respect to the subset relation. Informally, this incomparability constrain assures us that the meaning of the scientific term under focus is truly different in the two related patches (and not just a specification or a generalization of each other).

We can now see how Wilson-theory-nets are able to explicate several of the wandering phenomena described by Wilson. For instance, uneven facades, i.e. facades in which some patches assign a different referent to a common predicate (such as the case of the force predicate in classical mechanics), are represented in Wilson-structuralism as Wilson-theory-nets in which (at least) two theory-elements, directly related by weak-specialization, have models with incomparable (with respect to the subset relation) sets of functions representing a given scientific term, i.e. $\exists T, T' \in \bar{T}^W$ such that $Tw\sigma T'$ and $\Pi(M(T), t) \not\supseteq \Pi(M(T'), t)$ and $\Pi(M(T), t) \not\subseteq \Pi(M(T'), t)$. This incomparability condition implies that the union of all functions t appearing in the models of T is incomparable with respect to the subset relation to the union of all functions t appearing in the models of T'. Thus, some functions t in the models of T have to be different from any function appearing in the models of T' and some functions t in the models of T' have to be different from any function appearing in the models of T. However, since the two theory-element T and T' are in a weak-specialization relation, at least one function t appearing in their models has to be the same, because of the non-empty intersection of the models required by the weak-specialization relation. This formal condition mirrors in an abstract way the fact that patches in an uneven facade drag a given predicate into matching attributes that are, relative to a common application, incompatible with its original meaning. Note that the incomparability of sets of functions appearing in the models of two theory-elements implies the incomparability of the models and thus it is in stark contrast with the structuralist orthodox specialization relation (that requires weak subset inclusion of the models).

Stokes facades, i.e. facades in which different patches validate different inferences involv-

ing a common predicate (such as the Stokes phenomenon in optics), are instead represented in Wilson-structuralism as Wilson-theory-nets in which at least two theory-elements, directly related by weak-specialization, have the same models but incomparable (with respect to the subset relation) constraints, i.e. $\exists T, T' \in \bar{T}^W$ such that $Tw\sigma T'$, $M(T) = M(T')$ and $GC(T) \not\supseteq GC(T')$ and $GC(T) \not\subseteq GC(T')$. This formal condition represents the fact that in a Stokes facade patches share the same meaning of their scientific terms (i.e. they refer to the same attributes) at the cost of limiting the validity of certain reasoning tools at the patch boundaries.

Wilson-theory-nets are also compatible with the existence of ghost properties, i.e. an extreme case of uneven facades in which a given term has a different alignment in every patch and there is no common attribute that the predicate denotes (such as the case of hardness). We can represent this phenomenon as a Wilson-theory-net where every two theory-elements have incomparable (with respect to the subset relation) sets of function representing a given theoretical term and in which the intersection between all the models of all the theory-element is empty, i.e. $\exists t$ such that $\forall T, T' \in \bar{T}^W(\Pi(M(T), t) \not\supseteq \Pi(M(T'), t)$, $(\Pi(M(T), t) \not\subseteq \Pi(M(T'), t))$, and $\bigcap\{M(T)|T \in \bar{T}^W\} = \emptyset$. Moreover, the weak specialization relation $w\sigma$ of a Wilson-theory-net is designed to make possible loop-structures of specializations where no patch is more fundamental than another one (i.e. what Wilsons calls the "lousy encyclopedia phenomenon"). These cycles are allowed by weak specialization because, in contrast to the orthodox specialization relation, it is not anti-symmetric.

More generally, the weakening of the specialization relation and the consequent less homogeneous core-net (the net of all theoretical core of theory-elements in a given theory-net) of a Wilson-theory-net allow Wilson-structuralism to represent several of the eerie internal organizations of scientific theories described by Wilson such as incompatible descriptions of the world, Escherian geometries of patches inter-connection, and horizontal and vertical multi-valuedness of patches. Wilson-theory-elements can for instance have the same intended applications but incomparable models, thereby representing Wilson's patches offering incompatible descriptions of the same part of the domain. Cycles of weak-specialization relations may occur together with incomparable (with respect to the subset relation) constraints, models, and links between connected theory-elements, creating multiple possible ways in which the fundamentality of a theory-element in a Wilson-theory-net can be assessed.

Wilson-Structuralism is then able to adequately represent several wandering phenomena described by Wilson, achieving a logical reconstruction of scientific theories free from the semantic finality of classical Structuralism. Theory-elements and Wilson-theory-nets rationally reconstruct Wilson's patches and facades in a precise formal framework, in which several wanderings described by Wilson can be understood in an abstract way as specific set-theoretic relations between components of different theory-elements. Moreover, Wilson-Structuralism can be seen also as a complementary generalization of (a part of) the structuralist framework. In fact, Wilson-theory-nets have as specific cases the traditionally defined theory-nets that we saw in Section 3.1, i.e. Wilson-theory-nets in which all the weak specialization relations are also traditional specialization relations. If none of the semantic wanderings described by Wilson occur between any of its theory-elements, in fact,

a Wilson-theory-net is just a structuralist theory-net, where all the models, constraints, and links of the specialized theory-elements are subsets of the ones of the theory-element of which it is a specialization. As an extreme case of this lack of wanderings, we can also have a structuralist theory-tree as a specific case of a Wilson-theory-net in which there is a single theory-element of which all other theory-elements are specializations. This specific case can mirror what Wilson calls a flat structure facade, i.e. a facade that presents no wandering between its patches and thus can be said to consist "essentially one patch, that covers its whole domain adequately" (Wilson, 2006, p. 379).

This modified structuralism joins other recent attempts of renewing classical Structuralism by simplifying and improving its behemothic framework. Andreas' "Carnapian Structuralism" (Andreas, 2010, 2014) is an example of a kind of structuralism more compatible with other contemporary philosophical views about scientific theories reconstruction. Carnapian Structuralism restyled the structuralist framework through a reader-friendly system of postulates built around the notion of a theoretical expansion of a partial potential model. Wilson-structuralism takes a more radical departure from classical Structuralism than Carnapian Structuralism, radically weakening several semantic presupposition of the orthodox framework, but they both try to bring the structuralist way of reconstructing scientific theories closer to contemporary philosophy of science.

Taming conceptual wanderings: the case of viscous fluids forces

In order to make clearer how Wilson-Structuralism represents Wilson's framework of patches and facades, I will sketch how one can reconstruct as a Wilson-Theory-Net one of Wilson's main case studies of wandering referents in science, namely viscous fluids 'forces' in classical mechanics.

I have stressed in Section 2 how a considerable part of (Wilson, 2006) is dedicated to show how the apparently neat theory-structure of classical mechanics hides a complexity of wandering patches of usage ingeniously connected in a facade-like way. One of Wilson's (Wilson, 2006, pp. 157-165, 175-182) favorite examples of this semantic phenomenon is the concept of force. Through a detailed analysis of several subfields of classical mechanics, Wilson shows that this central concept of Newtonian physics is remarkably prone to change physical referents from one application to another one. In particular, the efforts of nineteenth-century physicists in extending Newtonian mechanics to more and more macroscopic phenomena pushed the predicate 'force' to be attached to attributes radically differing from any true force. Wilson (Wilson, 2006, pp. 158-159) stresses for instance the case of viscous "forces" in which the predicate force denotes net losses and gains of momentum caused by molecules leaving or entering the fluid "particle"İt is only thanks to what Wilson (Wilson, 2017, p. 368) calls a computational opportunity, then, that the behavior of fluids can be described with mathematical tools analogous to the ones used in more traditional parts of classical mechanics (cf. Wilson 2006, pp. 175-176). This computational opportunity, together with the aforementioned pivotal change of reference of the force predicate, allowed then physicists to claim that the Navier-Stokes equations for viscous fluids are just a specialization of Newton second law of motion, thereby annexing the underlying behavior

of viscous fluid to the phenomena adequately described by classical mechanics. Analogous strategies of theory expansion via 'property dragging' are behind the case of frictional and elastic forces (Wilson, 2006, pp. 175-176), parts of classical mechanics where the predicate force gets attached respectively to a net effect cause by the strength of the substratum and to a measure of internal stress.

In Wilson's terminology, the changes of referents for the predicate force in different parts of classical mechanics are a paradigmatic case of uneven facade, in which 'force' refers to different physical attributes in the viscous fluid, the frictional, and the elastic patches of classical mechanics. Given the metaphysical assumption in the last subsection, this change of referent implies different mathematical representations of the force predicate in the logical reconstruction of these parts of classical mechanics. So that structuralist theory-elements representing these 'forces' must have incomparable sets of force functions (and thus incomparable models). This incomparability is explicitly forbidden in the classical structuralist reconstruction of classical particle mechanics, where the models of every theory element have to be a subset of the ones satisfying Newton's second law of motion (i.e. the fundamental CPM theory-element). This limitation in the models makes every force function in the models of any theory-element in CPM just a specialization of the force function in the foundational theory-element corresponding to Newton's second law of motion. So that structuralists are forced to either reconstruct the viscous fluids and the frictional force theory-elements within CPM, characterizing them as simple additions of further constraints on a given component of the force function (just like we saw for the velocity-dependent forces theory-element in Section 3) or to reconstruct these parts of classical mechanics as belonging to different, albeit related by suitable links, theory-nets. The latter option is exemplified by Moulines' (Moulines, 1981, 2013) reconstruction of thermodynamics as (what he calls) a theory-frame. Both options, from Wilson's perspective, are not adequate reconstructions, since they either hide the change in meaning of the force predicate as a simple specification (the former reconstruction within the same theory-net) or they unnaturally divide the connected usages of force in classical mechanics into several theory-nets (the latter, multi-theory-nets type of reconstruction). Both kind of reconstruction are thus examples of the kind of semantic finality in traditional logical reconstructions of classical mechanics that Wilson (Wilson, 1998, 2014) argues against. The concept of force used in applying classical mechanics to viscous fluids cannot be a mere specialization of the one employed in Newton's second law of motion since the physical attribute denoted by the predicate is radically different from the one to which force aligns itself in CPM. At the same time, viscous fluids forces, according to Wilson, should be reconstructed together with all the other forces in classical mechanics, since they are part of the same facade obtained by gradual extensions of the force predicate in new domains (i.e. the phenomenon that Wilson calls "property-dragging nucleation", cf. Wilson 2006, p. 194).

Wilson-Structuralism allows this difference in the interpretation of the force predicate to be adequately reconstructed within a single Wilson-theory-net. We have in fact seen that Wilson-Theory-Nets can be uneven facades, i.e. Wilson-theory-nets in which (at least) two theory-elements, directly connected by a weak-specialization relation, have models with incomparable (with respect to the subset relation) sets of functions for a given term (and thus

incomparable models). So that, in reconstructing classical particle mechanics in Wilson-structuralism, the viscous fluid theory-element $ViscCPM$ can belong to the same theory-net of the CPM theory-element, the former being a weak-specialization of the latter that assigns a different meaning to the force function. More formally, since the force predicate denotes different physical attributes in the fundamental CPM patch and in the viscous fluids patch, we can assume that the CPM theory-element and the $ViscCPM$ theory-element (respectively representing the two patches in Wilson-structuralism) have incomparable sets of force functions in their models: $\Pi(M(CPM), f) \nsubseteq \Pi(M(ViscCPM), f)$ and $\Pi(M(CPM), f) \nsupseteq \Pi(M(ViscCPM), f)$. Then, the $ViscCPM$ theory-element can still be a weak-specialization of the CPM theory-element (or of another theory-element connected to CPM such as the velocity-dependent force theory-element), since in Wilson-structuralism a theory-element can be a weak-specialization of another one despite having incomparable models (a condition implied by having incomparable sets of functions in the models). So, that, assuming that the CPM and the $ViscCPM$ theory-element have at least one model jointly satisfying them and that viscous fluids forces can be mathematically reconstructed with the same typification of the force occurring in the CPM potential models (both conditions seem intuitively justified by the aforementioned existence of the computational opportunity described by Wilson), we can say that the viscous fluid theory-element is a weak-specialization of the classical particle mechanics theory-element: $ViscCPM\ w\sigma\ CPM$.

This sketch of a case study shows how Wilson-structuralism is able to reconstruct Wilson's framework of patches and facades, making precise in which sense in certain parts of classical mechanics the force predicate is attached to deviant attributes. The case study focused on the case of viscous fluids, but analogous theory-elements forming uneven facades can be built for the aforementioned cases of frictional and elastic forces. Furthermore, Wilson-Structuralism can adequately reconstruct in the same way other kinds of wandering phenomena described by Wilson. For instance, as already mentioned in the last subsection, Stokes facades, i.e. facades where different patches validate different inferences involving a common predicate, can be represented by Wilson-Theory-Nets having at least two theory-elements that have the same models but incomparable constraints. Thus, one could reconstruct Wilson's paradigmatic example of a Stokes facade, namely the Stokes phenomenon (Wilson, 2006, pp. 319-327), through three different theory-elements, sharing the same models but having as incomparable constraints the tree dominant behaviors of the light intensity predicate described by Wilson. Extreme uneven facades that include ghost properties, i.e. predicates with a different interpretation in every patch and no core meaning, such as the one describing the behavior of the hardness predicate (Wilson, 2006, pp. 335-355) can be similarly reconstructed by Wilson-theory-nets in which for a given predicate such as 'hard' every possible pair of theory-elements have incomparable sets of functions in their models and the overall intersection of the models of all theory-elements of the Wilson-theory-net is empty.

Let me recap the main steps of the present section. Starting from Wilson's analysis of the complex ways in which language refers to the world, I presented his framework of patches and facades. I then pointed to the surprising connections between two central

notions of the structuralist reconstruction of scientific theories, i.e. theory-elements and theory nets, and Wilson's patches and facades. We have seen however that the heavily hierarchical aspect of the structuralist framework poses a problem for any tentative structuralist reconstruction of Wilson's notions. I then presented my modified version of the structuralist framework, i.e. Wilson-Structuralism. I showed how this modified structuralist framework, relaxing the definition of a theory-net and the related specialization relation, is able to eliminate the hierarchical aspect of original Structuralism. We have then seen how in Wilson-Structuralism theory-elements and Wilson-theory-nets adequately represent Wilson's notions of patches and facades, allowing a precise reconstruction of several wandering phenomena described by Wilson such as the case of viscous fluids forces in classical mechanics.

Wilson-Structuralism achieves then a precise semantic reconstruction of many conceptual wanderings described by Wilson and a structuralist reconstruction of scientific theories more compatible with the nuances and the dynamics of scientific practice. Wilson-theory-nets provide in fact a more general alternative to the orthodox notion of theory-net, by allowing one to reconstruct within the same theory-net radical change of meanings and other conceptual wanderings that might be present in the target phenomena. Wilson-theory-nets and Wilson-Structuralism *tout court* should then be thought as a further tool in the structuralist toolbox that allows a precise semantic reconstruction of several semantic indeterminacies that can be found in many scientific theories.

In connection to these achievements, various directions for future work present themselves. A natural extension is to expand the scope of Wilson-Structuralism, taking into account also more pragmatic structuralist notions of scientific theory reconstruction such as theory-evolution, paradigm-driven theory-nets, and crystallizations (Moulines, 2011, 2013, 2014). From these extensions, I expect Wilson-Structuralism to achieve interesting complementary alternatives of these structuralist notions, arguably more suitable for allowing a vast range of semantic indeterminacies within their logical reconstruction of scientific theories. For instance, I expect this expanded Wilson-structuralism to be able to model dynamic wanderings such as semantic detoxification (Wilson, 2006, pp. 545-552), asymptotic connections between patches (Batterman, 2001; Wilson, 2017), and Machian explications (Wilson, 2012a; Carus, 2012a). Other promising ways of extending this framework would be to add linguistic and pragmatic contexts in order to model also Wilson's context-adjusting models of predicate extension (Wilson, 1982, 2012b) and to merge Wilson-Structuralism with Carnapian Structuralism (Andreas, 2014, 2020) and with accounts of deductive reasoning in structuralist frameworks (Andreas, 2013) in order to model Wilson's contextual notion of inference validity and logical inconsistencies (Wilson, 1994, 2000a). It would also be interesting to merge Wilson-Structuralism with other reconstructions of scientific theories devised to capture the indeterminacy of theoretical terms such as Carnap's ϵ-term methodology (Carnap, 1956; Schiemer and Gratzl, 2016; Leitgeb and Carus, 2020; Leitgeb, MS).

5.4 Assessing Indeterminate Models in the Toolbox Framework

In this final section, I will analyze how indeterminate models of conceptual change can be classified within the Toolbox framework, i.e. the meta-framework for assessing models of conceptual change that I presented in Chapter 2. More specifically, we will see how models such as Wilson's and Waismann's one can be assessed along the nine evaluative dimensions of the Toolbox framework: units of selection, concept ontology, concept structure, kinds and degrees of conceptual change, degree of normativity, effectiveness of normative judgment, assumptions and consequences for conceptual change in science, assumptions and consequence for conceptual change in philosophy, metaphilosophical assumptions and implications. Let us survey how indeterminate models of conceptual change perform in these dimensions, one by one, then.

Units of selection This dimension judges models of conceptual change according to the level of abstraction at which they identify conceptual entities as meaningful units of change. Both Waismann and Wilson share a certain localized holism in conceptual affairs that makes them recognize small parts of language as the smallest meaningful units of conceptual change. In order to understand how and why a concept change we have to look at the broader chunk of linguistic practice in which the concept is used, i.e. to the related language stratum, for Waismann, and to the related patch, for Wilson. This necessity to assess conceptual change in a (localized) holistic way is of course determined by the ubiquity of semantic indeterminacies that indeterminate models of conceptual change recognize. In order to understand how a concept change we have to ascertain the specific semantic, pragmatic, and epistemological properties of the local parts of language involved.

Concept ontology This dimension focuses on the compatibility of a given model of conceptual change with the different philosophical positions on the ontology of concepts. We saw how indeterminate models of conceptual change stress how change in linguistic practices is ubiquitous and never-ending. Consequently, both Wilson and Waismann strongly oppose any conception of concepts that see them as stable entities. Concepts are not something given to us once and for all, but they are adaptive tools that constantly change consistently with the linguistic practices of which they belong. In such models of conceptual change, then, there is no place for the fixity of the abstract view of concept ontology. Indeterminate models of conceptual change seem instead compatible with all the other three main views on concept ontology, i.e. the psychological, the linguistic, and the worldly view. Due to their attention to linguistic practices, It seems in fact most natural to couple indeterminate models of conceptual change with a linguistic view of concept ontology. That said, it seems possible to combine these models with the psychological or the worldly view of concept ontology thorough a somewhat deflationary reading of the emphasis on language that indeterminate models have. For instance, I mentioned in Section 2 that Waismann's open texture model of conceptual change has been recently reconstructed

within a psychological view of concepts based on prototype theory (Zeifert, 2020). For what concerns the worldly view of concepts, instead, Wilson (Wilson, 2017, Ch. 9) sometimes seems to argue for a worldly conception of scientific concepts.

Concept structure This dimension focuses instead on how a given model of conceptual change assumes the structure of concepts to be constituted. Indeterminate models of conceptual change do not dwell so much into matters of conceptual structure, so that they seem consistent with (almost) all of the theories about concept structure that we saw in Chapter 2. The only aspect of conceptual structure that plays a central role in these models is the extreme variability and context-dependency of concepts usage. As we saw in this section, both Waismann and Wilsons stress repeatedly that behind many seemingly monolithic concepts lies a complex bundle of different entities that perform different works in the various contexts in which they are employed. Whatever theory of conceptual structure one prefers, it has to allow a great degree of variability and context-sensitivity in its description of concepts in order to be adequate to the conception of conceptual change given by these indeterminate models. Pluralist theories of concepts seem then particularly apt to the task, with their explicit stress of the variability of inter-concept and intra-concept structure. Also more traditional theories of concepts such as prototype and theory theories seem to be able to allow the kind of conceptual variability stress by both Waismann and Wilson. As I mentioned in Section 1, prototype theories have been explicitly used to offer a reconstruction of Waismann's notion of open texture (Zeifert, 2020). For what concerns theory theories, as we saw briefly in Chapter 2, they have been developed for taking into account contextual effects of background knowledge on conceptual structure and are thus very much equipped for being coupled with the localized holism typical of indeterminate models of conceptual change.

Kinds and Degrees of conceptual change This dimension focuses on the kinds and degrees of conceptual change that a given model of conceptual change identifies. Indeterminate models of conceptual change do not really conceptualize conceptual change as a phenomenon that lends itself to be divided in degrees or kinds. As we saw in Section 1 and 2, Waismann and Wilson understand conceptual change as the natural result of the constant plastic adaptation of language to the world. This adaptation is a very indeterminate process, of which we often have only partial knowledge and control. As a consequence of this inherent indeterminacy, it is difficult for a model of conceptual change to rigidly classify its subject-matter in kinds or degrees.

Degree of normativity This dimension tracks the extent to which a given model of conceptual change is more or less normative in judging episodes of conceptual change. The indeterminacy at the heart of indeterminate models of conceptual change strongly reduces the space of informed choice in conceptual change. Concepts constantly change in response to external pressures and they often wander outside our control and understanding. Acknowledging this state of affairs, indeterminate models of conceptual change

inevitably downsize the normativity of their judgments on episodes of conceptual change. Nevertheless, we saw that a space for theoretical choice arises in certain occasions thanks to the underdetermination of conceptual usages by linguistic rules and non-linguistic facts, i.e. what Waismann calls open-texture and Wilson calls conceptual plasticity. In science and in ordinary language, then, we can sometime direct the linguistic evolution towards different results (cf. Wilson 2012a). This possibility of directing conceptual change allows then indeterminate models of conceptual change to judge with a certain degree of normativity some episodes of conceptual change, appraising the choices of the related linguistic communities with the benefit of hindsight.

Effectiveness of normative judgment This dimension focuses on how effective the normative judgment of a model of conceptual change is. Given what I said in the previous paragraph, it should not come as a surprise that the effectiveness of the normativity judgments within indeterminate models of conceptual change is a rather weak one. As we saw, the ubiquitous semantic and epistemological indeterminacies recognized by these models as constitutive aspects of conceptual change leaves few space for theoretical choice and even less space for normative judgments on these choices. Even in the few episodes of conceptual change that lend themselves easily to a normative judgment, indeterminate models are bound to give very weak judgments due to the inherent indeterminacy and complexity of the semantics behind our language that they postulate. If we often have a rather dubious knowledge and control of our conceptual tools, we should not expect a better knowledge and control of our models of these tools.

Assumptions/consequences for conceptual change in science This dimension focuses on the assumptions and the consequences of a given model of conceptual change in relation to the problems that scientific conceptual change poses in philosophy of science. Indeterminate models of conceptual change understand scientific conceptual change as completely analogous to ordinary conceptual change. Every linguistic practice change consistently with the requirements of the communities involved and the external pressure of the phenomena to which it refers. The fact that scientific concepts change is thus understood by indeterminate models of conceptual change as a completely natural and unproblematic phenomenon, that should not make us skeptic of the epistemic and ontological status of our best scientific theories (cf. Wilson 2000b). Seen from this perspective, indeterminate models of conceptual change are positive news for defenders of scientific progress and realism. However, accounts of conceptual change such as Waismann's and Wilson's one have also a more negative and revisionary message for philosophers of science. The traditional view of scientific concepts as stable and fixed entities, as well as the related heavily simplified philosophical models of their meaning and reference mechanisms, are deeply mistaken. Scientific concepts describe the world often with the aid of subtle and complex semantic architectures. Moreover, these architectures are remarkably sensible to the context and they are very much prone to change when it is required by the interfacial accomodations behind them. As such, indeterminate models of conceptual change

make scientific conceptual change a complex yet absolutely necessary object of study for philosophers of science that are interested in how science describes the world.

Assumptions/consequences for conceptual change in philosophy This dimension focuses on the assumptions and the consequences of a given model of conceptual change in relation to the problems that philosophical conceptual change poses in philosophy. We saw that indeterminate models of conceptual change understand their subject matter as an ubiquitous phenomenon at work indiscriminately in (almost) all linguistic practices. Thus, according to these models, philosophical concepts are subject to conceptual change as much as any other kinds of concepts. Moreover, the many kinds of semantic and epistemological indeterminacies around which the conception of conceptual change of these models is centered are of course present also in the case of philosophical concepts. The anti-essentialism typical of accounts of conceptual behavior such as Waismann's and Wilson's ones forbids any idea of philosophical concepts as stable and fixed entities, as well as any dream of a final conceptual analysis of these entities. In philosophy as well in science, philosophers should pay more attention to the non-trivial ways in which our language works and changes. Certain philosophical problems are in fact diagnosed by indeterminate models of conceptual change as stemming out from an inadequate account of how our philosophical and scientific languages work. Waismann (Waismann, 1946b; Fischer, 2019) stresses, for instance, that traditional philosophical problems seem to arise from a lack of appreciation of different language strata and their connections, while Wilson (Wilson, 2008, 2017, MS) repeatedly calls out some contemporary metaphysical problems as based on mistaken accounts of the semantics behind our best scientific theories.

Metaphilosophical assumptions and implications This dimension focuses on the metaphilosophical background that a given model of conceptual change has. As the differences between Waismann's and Wilson's articulation of their proposal exemplify, indeterminate models of conceptual change do not seem to share a common metaphilosophical background. Nevertheless, some general morals of metaphilosophical interest can be drawn from these models. First, indeterminate models of conceptual change show how philosophers should pay a great deal of attention to the linguistic practices in which a given discussion or theory is framed. Both in ordinary language and in science, language can often deceive us, hiding a surprisingly complex semantic architecture behind some apparently easy linguistic forms. Both Wilson and Waismann give us several examples of philosophical pseudo-problems that arise from not paying enough attention to linguistic practices. This call for attention to the linguistic practice has some connections with both the ordinary language and the pragmatist movements in analytic philosophy. A second metaphilosophical moral that can be drawn from models of conceptual change such as Waismann's and Wilson's is that our linguistic and conceptual tools are the byproduct of both worldly and human factors. As such, conceptual change is a phenomenon only partly under our control and foresight. Indeterminate models of conceptual change present, then, a conception of conceptual change that strikes a middle ground between the discovery-driven traditional

approach of conceptual analysis and the invention-like active ideal of conceptual engineering (cf. Chapter 2, Section 2.2). As Wilson (Wilson, 2006, p. 287) puts it, the refinement of our conceptual tools is a matter of interfacial accommodations, always prone to wander outside our control in order to take into account the unexpected situations that the world creates. Intentionally driven conceptual engineering then, in philosophy like in any other human activity, should be thought as a rare, specific case of the more general phenomenon of conceptual change (cf. Wilson 2012a; Decock 2021).

Chapter 6

Cognitive Models of Conceptual Change

The focus of this chapter will be on cognitive models of conceptual change, i.e. models that frame conceptual change as a specific kind of change in the cognitive structure underlying related theories. Specifically, I will concentrate on cognitive models of scientific conceptual change. As we briefly saw in Chapter 2, in the last fifty years tools from cognitive science have been extensively used to provide a novel understanding of scientific theories and their dynamics. As a consequence of these cognitive models of science, also the dynamics of scientific concepts have been modeled as specific kinds of change in the cognitive representation of scientific theories and concepts.

In this chapter, I will present three different cognitive models of conceptual change, differing in the cognitive architecture at the center of their model of scientific change. We will see Thagard's (Thagard, 1992) model of conceptual revolutions in science based on the notion of a conceptual system, frame-based (Andersen et al., 2006; Schurz and Votsis, 2014; Kornmesser and Schurz, 2018) models of scientific change, and finally Gärdenfors' and Zenker's (Gärdenfors and Zenker, 2011, 2013) model of scientific change based on conceptual spaces. As this list shows, cognitive models of conceptual change have relied on different cognitive structures in order to represent the dynamics of scientific concepts. Despite the difference between their specific way of representing the cognitive structure of scientific knowledge, we will see that all these cognitive frameworks share the belief that cognitive models of science are superior to logic-based models in modeling their subject-matter. The details or arguments of this superiority statement may vary depending on which specific cognitive architecture is employed, but the core of this superiority thesis is that the way in which knowledge is represented by cognitive models such as frames or conceptual spaces is procedurally different from logical reconstruction of it and it gives a better description of how human knowledge actually change[1].

A paradigmatic claim of the superiority of cognitive representations of scientific knowl-

[1]Instances of such superiority statements can be found in (Thagard, 1984, 1988; Giere, 1988; Thagard, 1992; Gärdenfors, 2000; Andersen et al., 2006; Zenker, 2014).

edge in comparison to logic-based ones is Thagard's (Thagard, 1984, 1992) argument for the autonomy of conceptual change with respect to belief changes. Thagard's autonomy argument, or Thagard's challenge (Park, 2010) as it is sometimes called in the literature, is particularly interesting because it can be seen as a challenge to logicians (and in particular to belief revision theorists) to come up with a logical model of conceptual change as faithful as the cognitive ones. In this chapter, I will present a novel model of conceptual revision that is able to mirror much of Thagard's model of conceptual change within a logical belief-revision-like system[2]. This conceptual revision system is a belief-revision-like system that works at the conceptual level of abstraction and that is therefore able to mimic cognitive model of conceptual change at their native level of abstraction.

In Section 1, I will present three different types of cognitive model of conceptual change, discussing their perks and limitations. Specifically, in Section 1.1 I will present Thagard's model of conceptual revolutions, in Section 1.2 I will talk about frame-based model of conceptual change and in Section 1.3 I will focus on a model of scientific change based on the theory of conceptual spaces. In Section 2, I will instead present a novel framework of conceptual revision, i.e. a belief-revision-like logical system that works on conceptual structures. I will present this novel framework in full generality, defining suitable expansion, revision, and contraction operations for it and showing how it satisfies several AGM-like rationality postulates for conceptual revision. In Section 3, I will show how this framework for conceptual revision is able to model almost all the kinds of conceptual change by Thagard in his model of conceptual revolutions. I will also show how a paradigmatic example of scientific revolution, i.e. the chemical revolution, can be reconstructed as a series of conceptual revisions and contractions in my framework. Finally, in Section 4, I will assess cognitive models of conceptual change such as Thagard's one using the nine dimension of my Toolbox framework.

6.1 Cognitive Models of Conceptual Change

Philosophers and cognitive scientists have used a variety of tools from cognitive science for modeling the dynamics of scientific theories and concepts. Consequently, one can find cognitive models of conceptual change using very different tools for representing the changes in conceptual knowledge that occur in science. As a showcase of some different cognitive tools that have been used as a basis for modeling scientific theories and concepts, I will present in this section three different types of cognitive model of conceptual change centered around three different cognitive representation of conceptual knowledge.

I will first present Thagard's model of conceptual revolutions, a cognitive model of conceptual change centered around the notion of a conceptual system. The second type of cognitive model of conceptual change that I am going to present in this section is made of frame-based models. As the name suggests, these models represents scientific theories and

[2]Let me stress again that the model of conceptual revision presented here was developed in full collaboration with Sena Bozdag and, as such, the material presented in Section 2 and 3 of these chapter has to be considered an entirely collaborative product.

their conceptual dynamics by the means of frames. Finally, we will see Gärdenfors' and Zenker's cognitive model of scientific change based on the theory of conceptual space (cf. Chapter 3, Section 4.2).

Before presenting and comparing these three kinds of cognitive model of conceptual change, thereby describing their differences, I must stress their similarities. In fact, as it will be clearer at the end of this section, conceptual systems, frames, and conceptual spaces give a roughly similar picture of human knowledge acquisition and change. All three types of knowledge representation, just like virtually all the other accepted models in contemporary cognitive science, share a very concept-based account of human knowledge where the acquisition and the revision of knowledge is largely the product of the interaction of many different default reasoning mechanisms that constantly contribute at different levels of voluntariness and abstraction to what we call knowledge. In the light of this large-scale common picture of human knowledge, it seems intuitively justifiable to envisage the possibility of hybrid cognitive models of conceptual change that represent the dynamics of human knowledge with the aid of several specific cognitive architectures. Indeed in the literature on cognitive models of conceptual change one can find such hybrid models, as it is exemplified by Giere's (Giere, 1988) influential cluster-of-schemata model of scientific change.

6.1.1 Thagard's model of conceptual revolutions

As I mentioned before, Thagard (Thagard, 1992) developed a fine-grained cognitivist model of scientific theory change centered around transformations in conceptual systems. Conceptual systems are complex structures roughly similar to frames (Minsky, 1975; Gamerschlag et. al., 2013), but (usually) not recursive. They are made of concepts and objects nodes connected via different kinds of links such as kind-links, instance-links, rule-links, and part-links. Changes in science then correspond to different modifications of these links. Specifically, scientific revolutions involve major transformations in part-links and in kind-links inside a conceptual system.

More specifically, there are two kinds of nodes and four kinds of links that can figure in a conceptual system. Nodes can be concept nodes or object nodes, mirroring respectively concepts and objects. Concept nodes can be connected with other concept nodes via three kinds of links (kind-links, part-links, rule-links) and with other object nodes via another kind of links, i.e. instance-links[3]:

- Kind-links (from concepts to concepts): intuitive reading 'is a kind of', example 'the canary is a kind of bird'.

[3]Note that Thagard in presenting his framework mentions also a fifth kind of link, property-links (Thagard, 1992, p. 31). This kind of links is supposed to mirror the information of a given object possessing a given property, but it does not seem to play any role into Thagard's model of conceptual change. It is in fact not mentioned in his abstract presentation of the model (Thagard, 1992, pp. 34-39) nor in any of the case studies (Thagard, 1992, pp. 131-224). I chose therefore to omit this kind of link from this presentation.

- Part-links (from concepts to concepts): intuitive reading 'a whole has a given part', examples 'the beak is a part of birds', 'fins are part of fishes'.

- Rule-links (from concepts to concepts): intuitive reading as expressing generic relations between concepts, example 'canaries are yellow'.

- Instance-links (from objects to concepts): intuitive reading 'is an instance of', example 'Tweety is a canary'.

The most important kinds of links are the ones between conceptual nodes. Kind-links and part-links specify what the constituents of (a part of) the world are according to a given conceptual system. Concepts within conceptual systems are organized in kind-hierarchies and part-hierarchies, i.e. sets of kind-links and part-links that are constrained in a tree-like form in order to give a consistent picture of (a part of) the world. Rule-links instead represent factual information and default reasoning mechanisms codified within the conceptual system. They are not organized in a hierarchy, but they can be divided between weak-rules and strong-rules depending on the strength of the information they represent.

Conceptual changes on a given conceptual system are then ordered by Thagard (Thagard, 1992, p. 35) in terms of how radical they are, from the least to the most radical:

1. Instance-addition: adding an instance relation saying that a given individual is an instance of a given concept, e.g. 'that blob in the distance is a whale'.

2. Rule-addition: adding a rule relation, e.g. 'whales can be found in the Arctic ocean' or 'whales eat sardines'[4].

3. Part-addition: adding a new part-relation, e.g. 'whales have spleens'.

4. Kind-addition: adding a new kind-relation, e.g. 'a dolphin is a kind of whale'.

5. Concept-addition: adding new concept, e.g. 'sound-wave' or 'narwhal'.

6. Kind-collapse: collapsing part of a kind-hierarchy, abandoning a previous distinction, e.g. when Darwin collapsed species and varieties within a species distinction.

7. Hierarchy-reorganization: shifting concepts or parts of the kind and part-hierarchies to another part of the hierarchies, i.e. *branch-jumping* such as Darwin's shift of humans to the animal-mammal part of the kind-hierarchy. It may also involve transformation of part-relations onto kind-relations and *vice versa*.

[4]Note that Thagard actually divides the rule-addition kind of conceptual change in two distinctive sub-types: weak-rule and strong-rule addition. Since Thagard's distinction between weak and strong rules is entirely pragmatical (Thagard, 1992, p. 35), being it based on the problem-solving power of a rule, I collapsed for simplicity these two types of changes in one type.

8. Tree-switching: changing the organizational principle of the kind-hierarchy, e.g. Darwin's switch from a morphological kind-hierarchy to an evolutionary one.

Thagard defended his concept-based model and the autonomy of conceptual change arguing that these revolutionary changes cannot be modeled by belief-revision theories. This supposed impossibility of modeling radical conceptual change within a belief-revision framework has been dubbed *Thagard's challenge* (Park, 2010). Specifically, Thagard's challenge claims that strong kinds of conceptual change are irreducible to belief-revision types of changes, because the former involves holistic recombinations of links and nodes in a given conceptual system that cannot be modeled by any piece-meal belief-revision operation. This irreducibility shows for Thagard how concept-based representations of knowledge, despite being expressively equivalent to first-order logic, are procedurally different from it (Thagard, 1984, 1988).

More specifically, Thagard's challenge consists of the claim that belief revision systems can model just the first two degrees of conceptual change in Thagard's hierarchy, i.e. instance-addition and rule-addition, but not the other six (Thagard, 1992, p. 36). Both instance-addition and rule-addition represent in fact piecemeal additions that do not involve any recombination in the part- and kind-hierarchies of a given conceptual system. These two kind of changes can then be adequately mirrored as changes at the belief-level, revising for instance the extension of a predicate and its prototypical instances (Strößner, 2020a, 2021). The other six, more radical kinds of conceptual changes are more holistic types of changes, since they involve the adjustment of the part- and kind-hierarchies (as well as rule and instance-relations) of the whole conceptual system. These changes represent in fact how scientists in revolutionary times add new concept, delete old concepts, drastically reorganize kind and part-hierarchies, and sometimes they even change the organizational principle of the hierarchical tree. Due to their holistic character, these changes cannot be easily mirrored as changes at the belief level like the first two. These revolutionary changes, then, are for Thagard (Thagard, 1992, p. 28) evidence that conceptual change is irreducible to belief-revision.

Thanks to the expressive power of conceptual systems, then, Thagard presented a very fine-grained cognitive analysis of conceptual change in science. Episodes of conceptual change can be analyzed in terms of which kinds of changes take place in the conceptual systems involved. Depending on how much the related conceptual system is modified in the process, historical episodes of conceptual change in science can be classified in terms of how radical they were and which kind of rational continuity between successive scientific theories they maintained. Thagard (Thagard, 1992, pp. 39-47, 131-223) showcased the power of his model of conceptual change by giving a detailed cognitive analysis of several alleged conceptual revolutions in the history of science in terms of specific modifications of conceptual systems. Thagard's case studies included the chemical revolution, the geological revolution, the Copernican revolution, and the Darwinian revolution.

6.1.2 Frame-based models of scientific conceptual change

The second type of cognitive models of conceptual change that I am going to present consists of frame-based models. Frames (Minsky, 1975; Barsalou, 1992; Gamerschlag et. al., 2013) are a very successful way of representing knowledge that has been employed for more than forty years in artificial intelligence and cognitive science. There are various kinds of frames, differing in structure and complexity, but all frames share a core set of representational tenets.

The two main kinds of components of a frame are attributes and values. Attributes represent certain salient properties that instances of a given concept tend to possess, while values add more information about the default way in which these properties are exemplified. For instance, the salient property of being a male is represented by the frame for the concept of a bachelor by the combination of the attribute 'sex' and its value 'male'. This attribute-value structure typically represents default knowledge that contributes to the prototype of a certain concept. Frames, in fact, can be thought as a specific, elaborate, representation of the prototype theory of concepts (cf. Ch. 2, Section 1.2). Just like other prototypical representations of concepts such as feature lists or schemata, frames represent a certain concept in terms of the most typical properties that its instance tend to possess. Specifically, these typical properties are represented by frames as specific attribute-value combinations. Central to frames is also the recursiveness of this core attribute-value structure. This recursiveness allows frames to nest this prototypical description of concepts in order to store, economically, impressive amounts of default knowledge as nets of frames.

As I mentioned before, there are different kinds of frames that differ in their specific elaboration of the basic attribute-value structure common to all frames. A common elaboration of the attribute-value structure is, for instance, the addition of constraints between different attributes of the same (or even of different) frame(s) that limits the possible combinations of values of some attributes. Such constraints are particularly effective for representing simple inductive inferences and causal dependencies between different concepts. More generally, it is safe to say that frames, in comparison to other cognitive ways of representing conceptual knowledge, are particularly apt to represent several forms of everyday reasoning (e.g. default, defeasible, non-monotonic, etc.), thanks to the plasticity of their basic attribute-value structure (especially when augmented with suitable constraints) and its aforementioned recursive structure.

Given their success in representing conceptual and inferential knowledge, it is not surprising that several philosophers and cognitive scientists have chosen frames as the background of their model of conceptual change. A paradigmatic example of frame-based models of conceptual change is Andersen's, Barker's, and Chen's (Andersen et al., 2006) neo-Kuhnian model of scientific revolution. This neo-Kuhnian model is based on Barsalou's dynamic frames (Barsalou, 1992; Barsalou and Hale, 1993), i.e. specific frames that allow attributes to have several values at once and different kinds of structural constraints on both attributes and values. Scientific theories are then represented as (what they call) conceptual structures, i.e. specific hierarchies of nested dynamic frames that have to satisfy certain consistency requirements.

The authors (Andersen et al., 2006, pp. 67-69) listed three hierarchical principles that stable conceptual structures have to respect: the No-Overlap Principle, the Exhaustion Principle, and the Inclusion Principle. The first principle prescribes the exclusivity of contrasting concepts, forbidding any object to be categorized as an instance of two or more contrasting concept (e.g. an animal cannot be both a dog and a cat). The second principle prescribes the completeness of subordinate concepts with respect to superordinate ones, saying that an instance of a given superordinate concept must be an instance also of one of its subordinate ones (e.g. an animal has to be either a mammal, or a reptile, etc. It cannot be something completely foreign to any of the subordinate animal concepts). Finally, the third hierarchical principle prescribes that all instance of a subordinate concept have to be also instances of the related superordinate concept (e.g. all cats are mammals). These three hierarchical principles together force a conceptual structure to have a general internal consistency, making thus a hierarchy of dynamic frames apt to represent the cognitive structure of a given scientific theory.

Andersen, Barker, and Chen use then these hierarchies of dynamic frames to give a cognitive explication (in Carnap's sense of the term, see Ch. 3) of Kuhn's theory of scientific revolution. As any good neo-Kuhnian theory, the three authors (Andersen et al., 2006, pp. 66-86) start with distinguishing normal science with revolutionary science. Normal science consists of changes in a scientific theory that do not significantly change the related conceptual structure such as changing specific value constraints or introducing new subordinate concepts. These small changes are simple ways of resolving small violations of the hierarchical principles of a conceptual structure such as the observation of an uncategorizable object (e.g. a new species of a bird). Revolutionary science, instead, involves far more radical changes that do not leave the conceptual structure intact. If, in fact, changes in normal science might involve the addition of new attributes and/or values in one of more frames, revolutionary changes often involve the holistic re-thinking of the similarity and dissimilarity relations between object that are in the background of a given hierarchy of dynamic frames. This modification of the basic ways in which instances are categorized is Anderson's, Barke's, and Chen's explication of Kuhn's famous gestalt-switch picture of a scientific revolution. The three authors (Andersen et al., 2006, pp. 87-117) also give a fine-grained analysis of scientific revolutions, distinguishing different degrees of taxonomic incommensurability in terms of radical mismatches on the basic categorization structure between pre-revolution and post-revolution theories. In radical scientific revolutions, then, sever lack of inter-theoretical communication is produced by mutually inconsistent sets of attribute nodes that categorize a certain phenomenon on the basis of radically different similarity and dissimilarity relations. The three authors (Andersen et al., 2006, pp. 130-162) give various examples of scientific revolutions as competing hierarchies of dynamic frames, showing how the degree of communication between rival theories can be ascertained via checking how much the conceptual structure of a given theory violates the rival conceptual structure and its application of the hierarchical principles.

Similarly to Andersen's, Barker's, and Chen's neo-Kuhnian model, other frame-based models of scientific conceptual change use nets of nested frames to represent the cognitive structure of scientific theories, understanding different types of scientific changes as differ-

ent modifications in the related frame structures. Other types of frames, different from Barsalou's dynamic frames, have been used by different frame-based models of scientific conceptual change such as Kornmesser's and Schurz's theory-frame model (Kornmesser and Schurz, 2018).

Notwithstanding the specific type of frame structure implemented, all frame-based models of science make pivotal use of the plasticity of the recursive attribute-value structure and the related applicable constraints in order to model scientific conceptual change as a variety of modifications of different strength that gradually transform the hierarchical frame-structure related to a given scientific theory. This fine-grained analysis of scientific dynamics is, according to the supporters of frame-based models, only possible thanks to the expressive power of frames that allow us to ascertain rational continuity in the deep cognitive structure behind scientific theories even when the more superficial linguistic counterparts present radical differences.

6.1.3 Scientific change as dimensional change

The third type of cognitive model of conceptual change that we are going to see in this section is made of models based on the theory of conceptual spaces. We already encountered the theory of conceptual spaces in Chapter 3, where I used it to give a formal explication of Carnap's notion of explication[5]. As a matter of fact, the model of Carnapian explication that I presented in Chapter 3 can be indeed considered a cognitive model of conceptual change based on conceptual spaces. However, in this section I will not focus on my own model of scientific change based on conceptual spaces, but I will instead present a different model developed by Gärdenfors and Zenker (Gärdenfors and Zenker, 2011, 2013; Zenker and Gärdenfors, 2014, 2015a; Masterton, Zenker, and Gärdenfors, 2017) in a series of papers.

If I used conceptual spaces to give an explication of Carnapian explication (cf. Ch. 3, Sect. 4), Gärdenfors and Zenker used the expressive power of conceptual spaces theory to give a somewhat neo-structuralist model of scientific conceptual change. Just like the structuralist program in philosophy of science that we saw in Chapter 5 (cf. Ch. 5, Sect. 3.1), in fact, Gärdenfors' and Zenker's framework understands a scientific theory primarily as a bundle of structures. The key difference with classical Structuralism is that, according to Gärdenfors and Zenker (Zenker and Gärdenfors, 2014), the structure of a scientific theory is best reconstructed with the tools of the theory of conceptual spaces (and not in a model-theoretic way, like it is prescribed by orthodox Structuralism). More specifically, a given scientific theory can be reconstructed via what the authors call a *conceptual framework*, i.e. a (bundle of) conceptual space(s) representing the theoretical framework of the theory as a collection of appropriate (sets of) dimensions. So that, instead of a set of (sets of) model-theoretic structure like in orthodox Structuralism, the structure of a given scientific

[5]I invite the interested reader to see again Chapter 3, Section 4.1 for a presentation of the theory of conceptual spaces, together with relevant bibliography. For a fully formalized presentation of the theory, see instead (Raubal, 2004; Lewis and Lawry, 2016; Bechberger and Kühnbeger, 2017).

theory is given by a set of (sets of) cognitive dimensions suitably grouped in respective domains and adequately interconnected by certain constraints.

Gärdenfors and Zenker (Gärdenfors and Zenker, 2011, 2013) used their reconstruction of scientific theories as conceptual frameworks to give a cognitive model of scientific conceptual change based on five different types of changes in the conceptual framework of the theory. These five changes are ordered with respect to their radicality, from the least radical to the most radical:

- Addition/Deletion of special laws: the least radical conceptual change consists of the addition (or the deletion) of a special law (e.g. Hooke's law for elastic forces in Newtonian Mechanics). In a conceptual framework, the addition or deletion of a special law modifies in fact only the intra- and inter-dimensional constraints, but it does not change the dimensional structure of the scientific theory.

- Change of scale or metric: the second type of change in Gärdenfors' and Zenker's model is a change in the scale or the metric of a given (sets of) dimension(s) (e.g. the change from the Celsius to the Kelvin scale). This change implies a modification of the information given by the modified dimension(s) and thus constitutes a more radical change than the addition/deletion of special laws.

- Change in the importance of dimensions: the third type of change consists of a change in the salience or the importance of cognitive dimensions in the conceptual framework of a given scientific theory. In scientific dynamics, changes in the importance of dimensions can strongly modify the ontological import of a given scientific theory (e.g. the weakening of the importance of the color dimension in the eighteenth century chemistry).

- Change in the separability of dimensions: The fourth type of change is the change of a cognitive domain in the conceptual framework of a given scientific theory. That is, it consists of a change in the separability of two or more cognitive dimensions (e.g. the step from Newtonian space and time to Einsteinian space-time). A change in the separability of dimensions (and thus of domain) implies a substantive rethinking of the measurement methods and of the epistemological and ontological status of (some parts of) the theoretical framework of a given scientific theory.

- Addition and deletion of dimensions: The most radical form of conceptual change in Gärdenfors' and Zenker's taxonomy consists of the the addition or deletion of cognitive dimensions. When, in fact, some dimensions are added or deleted from the conceptual framework of a scientific theory, its cognitive structure is radically altered up to the point of creating a breakdown of rational continuity between successive theories (e.g. the definitive deletion of the ether dimension carried out by twentieth-century physics).

This five-step taxonomy of conceptual changes allows Gärdenfors and Zenker (Gärdenfors and Zenker, 2011, 2013; Masterton, Zenker, and Gärdenfors, 2017) to give a cognitive

analysis of the changes involved in the passage from Newtonian Mechanics to Special Relativity. More generally, their cognitive model of conceptual change allows to break-down major scientific conceptual changes into a step-by-step modification of the related conceptual frameworks and their cognitive dimensions.

We have then seen three different types of cognitive model of conceptual change, respectively centered around three different cognitive architectures: conceptual systems, frames, and conceptual spaces. Despite the differences in the basic set-up and in the specific way in which conceptual knowledge is represented and modified, all these three types of cognitive model offer a fine-grained analysis of conceptual change in science as a specific taxonomy of modifications of the cognitive structure related to a scientific theory. Specific episodes of conceptual change or specific semantic phenomena can be more or less suited to be reconstructed in a specific kind of cognitive model (cf. Strößner 2020b), but in general it is safe to stress again that all these three types of cognitive model (as well as virtually every other type of cognitive model that can be found in the literature) give a very similar picture of the phenomenon of conceptual change and its epistemological and semantic implications.

6.2 A Model of Conceptual Revision

We saw in Section 1.1 Thagard's fine-grained cognitive model of conceptual change, his argument for the autonomy of conceptual change and his related challenge to belief-revision theorists. Despite the enormous expansion of the belief-revision literature in the last thirty years (Hansson, 1999) and recent work connecting it with philosophy of science (Olsson and Enqvist, 2010), Thagard's challenge has not received so much attention. I will show in this section a way of taking up Thagard's challenge by developing a belief-revision-like framework capable of modeling the radical types of conceptual change described by Thagard. Specifically, I will present a conceptual revision framework in which one can revise and contract conceptual structures, i.e. set-theoretic representations of Thagard's conceptual systems. The change operators will be reminiscent of the ones used in base-generated belief change theories (Rott, 2001; Hansson, 1999), but working on conceptual structures instead of belief bases.

The choice of units of revision, i.e. conceptual structures, makes this conceptual revision system differ from other applications of belief-revision to the problem of scientific change. Traditionally, belief-revision theories deal with piece-meal changes in a belief set similar to the kind of changes happening in normal science (cf. Gärdenfors and Rott 1995). In applying these theories to the problem of scientific change, logicians have focused on mirroring changes in scientific theories as changes in (usually structured) belief sets (Martin and Osherson, 1998; Cresto, 2008; Hansson, 2010; Andreas, 2011; Strößner, 2020a, 2021). This belief-centered take on scientific change is exactly the reason why Thagard claims that belief revision theories are not adequate for representing conceptual change (Thagard, 1988, 1990). We instead chose to model conceptual change at its native level of abstraction, without any reference to the belief level[6]. This will be achieved by lifting the

[6]Again, let me stress that the content of this section and the next one is a product of a full collaboration

methodology of belief revision theories to the conceptual level. As a result, the aim of the change operations of my conceptual revision system will then be the preservation of the consistency of conceptual structures. This consistency is understood as the satisfaction of some structural constraints on the components of a conceptual structure that ensure the overall consistency of the knowledge represented by it. The knowledge represented via conceptual structures is similar to the content represented by description logics (Wolter and Zakharyaschev, 1999), since they also represent knowledge about concept hierarchies[7].

In this section, I will present this novel belief-revision-like model of conceptual revision. More specifically, I will first define a revision and a contraction operation that work on conceptual structures. Then, I will show how the conceptual revision model satisfies several rationality postulates analogous to the AGM ones for belief revision theories (Alchourrón, Gärdenfors, and Makinson, 1985).

6.2.1 Conceptual structures and conceptual hierarchies

The conceptual revision model will be equipped with a change mechanism similar to the one of base-generated belief revision framework, but the units of change are structure mirroring Thagard's conceptual systems rather than belief bases. In this way, Thagard's changes can be mirrored at their native level of abstraction, namely the conceptual level.

The conceptual revision framework takes as its units of changes set-theoretic entities which are called *conceptual structures*. We define two different domains, one for concepts and one for individual objects, as the primary elements of a conceptual structure. Our conceptual structures enrich these two basic domains with different relations between elements of these domains. Mirroring Thagard's system, we define three two-place relations between elements of the concept domain (kind-relation, part-relation, rule-relation) and one two-place relation between elements of the object domain and elements of the concept domain (instance-relation).

Formally, a *conceptual structure* is defined as follows: $CS = \langle \mathcal{C}, \mathcal{O}, K, P, R, I \rangle$ is a conceptual structure iff,

- $\mathcal{C}$ and $\mathcal{O}$ are (possibly empty) finite domains of (respectively) atomic concepts and individual objects.
- $K = \{\langle x, y\rangle, \ldots\}$ and $P = \{\langle x, y\rangle, \ldots\}$ are two-place irreflexive relations between elements of the concept domain, such that $x, y \in C$ and $\langle x, y\rangle$ is an ordered pair. They represent respectively Thagard's kind and part links between concept nodes. If $\langle x, y\rangle \in K$, we write $x \sqsubset_K y$ (same for $x \sqsubset_P y$, if $\langle x, y\rangle \in P$).
- $R = \{\langle x, y\rangle, \ldots\}$, with $x, y \in C$ and $\langle x, y\rangle$ is an ordered pair, is a two-place anti-symmetric relation between elements of the concept domain. It represents Thagard's rule links between concept nodes.

between me and Sena Bozdag. As such, I will use the first person plural subject in these two sections.

[7] AGM-style and base-generated revision theories in description logics are also proposed in (Ribeiro and Wassermann, 2007) and in (Ribeiro and Wassermann, 2009)

- $I = \{\langle a, x\rangle, \ldots\}$ with $a \in O$ and $x \in C$ and $\langle a, x\rangle$ is an ordered pair, is a two-place anti-symmetric relation between elements of the object domain and elements of the concept domain. It represents Thagard's instance links between object and concept nodes.

We can then single-out specific kind-relations and part-relations through a tree-like structural requirement. Relations satisfying this requirement are then called respectively kind-hierarchies and part-hierarchies. This requirement is our way of rationally reconstructing Thagard's implicit structural requirements on conceptual systems. Similarly, we introduce criteria to single out certain rule and instance relations as *consistent* rule and instance relations. With these further criteria we mirror common constraints on how knowledge is represented in a consistent way by frames (cf. Andersen et al. 2006; Gamerschlag et. al. 2013). Then, a conceptual structure is a *conceptual hierarchy* iff its kind relation is a kind-hierarchy, its part relation is a part-hierarchy, its rule relation is a consistent rule relation, its instance relation is a consistent instance relation, and all the concepts and objects occurring in its relations are members respectively of the concept domain or the object domain.

More formally, $CH = \langle \mathcal{C}, \mathcal{O}, K_h, P_h, R_{cons}, I_{cons}\rangle$ is a conceptual hierarchy iff:

- $\mathcal{C}$ and $\mathcal{O}$ are (possibly empty) finite domains of respectively concepts and objects, which include all the concepts and objects that appear in the relations.
- K_h is a kind-hierarchy, i.e. a transitive kind-relation $K = \{\langle x, y\rangle, \ldots\}$ that, if non-empty, has a top element and from any other element of the ordering there exists a unique path to this top element.
- P_h is a part-hierarchy, i.e. a transitive part-relation $P = \{\langle x, y\rangle, \ldots\}$ that, if non-empty, has a top element and from any other element of the ordering there exists a unique path to this top element.
- R_{cons} is a consistent rule-relation, i.e. a rule-relation $R = \{\langle x, y\rangle, \ldots\}$ such that $\forall x, y, z \in C$ if $\langle x, y\rangle \in R_{cons}$ and $z \sqsubset_K x$, then $\langle z, y\rangle \in R_{cons}$.
- I_{cons} is a consistent instance-relation, i.e. an instance-relation $I = \{\langle a, x\rangle, \ldots\}$ such that $\forall x, y \in C$ and $\forall a \in O$ if $\langle a, x\rangle \in I_{cons}$ and $x \sqsubset_K y$, then $\langle a, y\rangle \in I_{cons}$.

We define the *top element in a kind-relation* (*part-relation*) as follows: given a conceptual structure H and the concept domain C_H and the kind-relation K_H (part-relation P_H) of H, a concept $a \in C_H$ is a top element in K_H (P_H) iff for all concepts $t \neq a$ in C_H which occur in a pair in K_H (P_H), $\langle t, a\rangle \in K_H$ ($\in P_H$). By *a unique path to the top element from any other element* we mean that, given a is a top element in a kind-relation (or in a part-relation), for all $t \neq a$ which occur in the kind-relation (part-relation), if $\langle t, y\rangle$ and $\langle t, z\rangle$ are pairs in the kind-relation (part-relation) such that $y \neq z$, then either $\langle z, y\rangle$ or $\langle y, z\rangle$ is also a pair in the same relation. In other words, kind-hierarchies and part-hierarchies do not allow upward branchings (Figure 6.1).

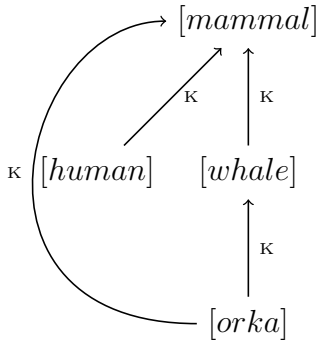


Figure 6.1: A consistent conceptual structure made of four concepts and four kind-links between them.

6.2.2 Revision on conceptual structures

In this section we will describe how a conceptual structure should be revised in our framework. Revising a conceptual structure means adding new elements (of a conceptual structure) to an existing conceptual structure, while preserving (or restoring) the consistency of the revised (new) conceptual structure. The information we want to add (or delete) can be a concept, a kind-relation, a part-relation, a rule-relation, or an instance-relation. Consistency is characterized by the idea of conceptual hierarchies, i.e. by structural restrictions on the different kinds of relations connecting concepts and objects within a conceptual structure. Therefore, the goal of our conceptual revision framework is to define change operations which preserve these structural limitations.

We start with identifying the form of the eligible arguments for a revision. Suppose we want to revise an existing conceptual structure with an instance (i.e. an instance-addition in Thagard's framework). Suppose we want to add that Bob is an orca ($I\langle Bob, orca\rangle$). If the existing structure does not already contain the concept of an orca and the object Bob, simply adding the instance link would not make sense. Hence, while formalising the arguments for conceptual revisions, we explicitly state every element included in them. That is, we express the above instance link as a (partial) conceptual structure (let us call it C) which consists of the following: $C_C = \{orca\}, O_C = \{Bob\}, I_C = \{\langle Bob, orca\rangle\}$. If we add to this structure an empty part-relation, an empty kind-relation and an empty rule-relation we obtain a full conceptual structure. Therefore our conceptual revision framework allows full or partial conceptual structures as arguments.

Next, we have to choose what kind of revision operation we want in our framework. The consistency of the revised (conceptual) structure could for instance be preserved while making the additions, or it could instead be restored after the addition process (Hansson, 2010). The former approach it is typical of the AGM belief-revision paradigm (Alchourrón, Gärdenfors, and Makinson, 1985), while the latter is common amongst base-generated revision theories (Hansson, 1999; Rott, 2001). In what follows we opt for the second approach. Given a conceptual structure and an argument of revision, we will first add them on top of each other, obtaining a possibly inconsistent new conceptual structure,

then we will retrieve the consistent parts of this structure to build the revised (consistent) conceptual structure.

In base-generated belief revision, one adds the new information to the existing body of beliefs via a set-theoretical union operation.[8] The potential inconsistency in the expanded belief set is caused by too much information. To eliminate this inconsistency, the (less entrenched) beliefs responsible for it are deleted from the new belief set.

In our framework, the inconsistency of a conceptual structure may be caused by too much information or by too little information. In fact, the pivotal requirement of transitivity for the relations of a conceptual hierarchy could be lost during revision. In order to restore the consistency of a conceptual structure, we need to eliminate the inconsistent parts and to repair the transitivity of its relations. We will deal with the transitivity issue in the first step of our revision, i.e. the addition of new information to a conceptual structure (which constitutes the operation of conceptual expansion). In the second step of our revision, we will instead deal with the problem of inconsistent parts, proposing a mechanism that retrieves consistent parts of the expanded conceptual structure[9].

Conceptual expansion We will define conceptual expansion as the process of adding new information to a conceptual structure, without necessarily preserving the consistency of the expanded structure. More specifically, conceptual expansion will be performed via the fusion models.

A tuple $CS^{\oplus} = \langle \mathcal{CS}, \oplus \rangle$ is fusion model on conceptual structures iff $\mathcal{CS}$ is a set of conceptual structures that is closed under the total conceptual fusion function $\oplus$ from $CS \times CS$ to CS, uniquely determined by the following:

- $\mathcal{C}_{A \oplus B} = \mathcal{C}_A \cup \mathcal{C}_B$
- $\mathcal{O}_{A \oplus B} = \mathcal{O}_A \cup \mathcal{O}_B$
- $K_{A \oplus B} = TC(K_A \cup K_B)$
- $P_{A \oplus B} = TC(P_A \cup P_B)$
- $R_{A \oplus B} = TC|K_{A \oplus B}(R_A \cup R_B)$

[8] We significantly simplify the mechanism of base generated revisions. In fact the new information is added on top of a structure called a *belief base* which is the foundation of an agent's beliefs. Moreover, the addition of the new information is set-theoretical only if we are dealing with *flat* belief bases which are not ordered by a preference or entrenchment relation. Full theories of belief revision usually include such orderings to account for the rationality of changing beliefs. Adding beliefs in the AGM paradigm also goes further than the union operation as it involves taking the deductive closure of the new belief set.

[9] While we could keep the expansion process simple and deal with the (possibly lost) transitive closure in the process of retrieving information, we believe it is more natural to restore the transitivity required for consistency as part of the expansion operation. One reason is that restoring transitive closure will be done by adding new links, and keeping all additions as part of the expansion and limiting the process of retrieval of information to elimination of some (less entrenched) parts of the expanded structure which contribute to the inconsistency allows simpler definitions for the two processes. Another reason is that this allows us to characterise a conceptual expansion operation which results in structures which resemble conceptual hierarchies to an extent that they are somehow useful in practice.

- $I_{A\oplus B} = TC|K_{A\oplus B}(I_A \cup I_B)$

TC stands for the transitive closure operation on our sets of pairs. For instance, the transitive closure of a kind-relation K is the smallest transitive set of pairs that contains K such that if $\langle a, b\rangle$ and $\langle b, c\rangle$ is in $TC(K)$ then $\langle a, c\rangle \in TC\{K\}$. Transitive closure on rule-relations and instance-relation are via transitivity modulo the kind-relation. Thus, an instance-relation I is transitively closed modulo the relevant kind-relation K ($TC|K$) iff given $\langle b, c\rangle \in K$ and $\langle a, b\rangle \in I$, then also $\langle a, c\rangle$ is in I.

The above model specifies how to add a full conceptual structure on top of another one. It can be generalized for fusing partial conceptual structures in the following way:

A tuple $\langle \mathcal{C}, \mathcal{O}, K, P, R, I, \bigoplus\rangle$ is a generalized fusion model on (possibly partial) conceptual structures iff $\mathcal{C}$ is a finite concept domain, $\mathcal{O}$ is a finite object domain, K is a kind-relation, P is a part-relation, R is a rule-relation, I is an instance-relation, and $\bigoplus = \{\oplus_C, \oplus_O, \oplus_K, \oplus_P, \oplus_R, \oplus_I\}$ is the family of total fusion functions from $\langle \mathcal{C}, \mathcal{O}, K, P, R, I\rangle \times \langle \mathcal{C}, \mathcal{O}, K, P, R, I\rangle$ to $\langle \mathcal{C}, \mathcal{O}, K, P, R, I\rangle$ determined uniquely by the following:

- $\oplus_C$ is a function from $\mathcal{C} \times \mathcal{C}$ to $\mathcal{C}$ such that $\mathcal{C}_A \oplus_C \mathcal{C}_B = \mathcal{C}_A \cup \mathcal{C}_B$
- $\oplus_O$ is a function from $\mathcal{O} \times \mathcal{O}$ to $\mathcal{O}$ such that $\mathcal{O}_A \oplus_O \mathcal{O}_B = \mathcal{O}_A \cup \mathcal{O}_B$
- $\oplus_K$ is a function from $K \times K$ to K such that $K_A \oplus_K K_B = TC\{K_A \cup K_B\}$
- $\oplus_P$ is a function from $P \times P$ to P such that $P_A \oplus_P P_B = TC\{P_A \cup P_B\}$
- $\oplus_R$ is a function from $R \times R$ to R such that $R_A \oplus_R R_B = TC|K(R_A \cup R_B)$
- $\oplus_I$ is a function from $I \times I$ to I such that $I_A \oplus_I I_B = TC|K(I_A \cup I_B)$

We show with an example how conceptual expansion via conceptual fusion models works. Let H be the conceptual hierarchy depicted in Figure 6.1 such that

$$\mathcal{C}_H = \{[mammal], [human], [whale], [orka]\}$$

$$\mathcal{O}_H = P_H = R_H = I_H = \emptyset.$$

$$K_H = \{\langle human, mammal\rangle, \langle whale, mammal\rangle, \langle orka, whale\rangle, \langle orka, mammal\rangle\}.$$

Let A be a (partial) conceptual structure consisting of: $\mathcal{C}_A = \{[fish], [orka]\}$, $\mathcal{O}_A = \{(Bob)\}$, $K_A = \{\langle orka, fish\rangle\}$, $I_A = \{\langle Bob, orka\rangle\}$. Then, by the (generalized) fusion model, we can obtain the conceptual structure $H \bigoplus A$ determined by the following elements:

$$\mathcal{C}_{H\bigoplus A} = \{[mammal], [human], [whale], [orka], [fish]\},$$

$$\mathcal{O}_{H\bigoplus A} = \{(Bob)\},$$

$$K_{H\bigoplus A} = \{\langle human, mammal\rangle, \langle whale, mammal\rangle, \langle orka, whale\rangle,$$

$$\langle orka, mammal\rangle, \langle orka, fish\rangle\},$$

$$I_{H \bigoplus A} = \{\langle Bob, orka\rangle, \langle Bob, fish\rangle\langle Bob, whale\rangle, \langle Bob, mammal\rangle\},$$

$$P_{H \bigoplus A} = R_{H \bigoplus A} = \emptyset.$$

The instance pairs $\langle Bob, mammal\rangle$, $\langle Bob, whale\rangle$ and $\langle Bob, fish\rangle$ in $I_{H \bigoplus A}$ are additions to the simple union of I_H and I_A via the transitive closure operation.

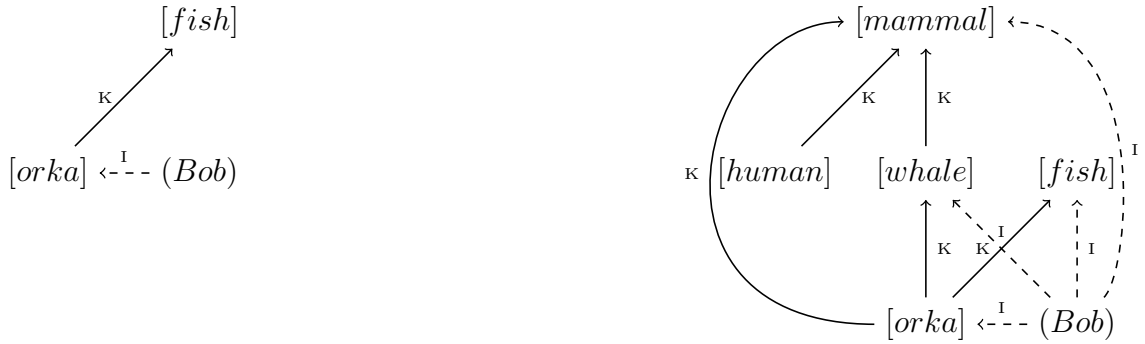


Figure 6.2: The partial conceptual structure A (on the left) and the conceptual structure $H \bigoplus A$ (on the right).

Conceptual revision Conceptual expansion may not always produce a conceptual hierarchy. Since our aim is to obtain conceptual hierarchies as the result of revisions, we propose a consistency check mechanism for restoring the consistency of conceptual structures. This mechanism is a modified version of the consolidation operation described in relation to the base generated revisions (Rott, 2001, p. 40).

In what follows, we use the notions of a *substructure* and a *maximal hierarchy within a conceptual structure*. A conceptual structure H is a substructure of a conceptual structure H' ($H \subseteq H'$) iff all the components of H are subsets of the respective components of H'. H is a maximal hierarchy within H' iff H is a substructure of H' and any expansion of H within H' is not a conceptual hierarchy.

While revising a conceptual structure, we first expand it with the argument of the revision. Since it might be the case that the expansion operation fails to produce a conceptual hierarchy, we resort to a selection mechanism which marks off the *best* maximal hierarchies within the expanded conceptual structure. The maximality principle is assumed in order to preserve as much information as possible while revising. In order to meet some rationality criteria, such selection mechanisms usually rely on an ordering of the alternatives based on the preferences of the selecting agents. In the context of scientific theory change the preferences of the agents may be shaped by principles such as conservativity, simplicity, generality, etc. It can be argued that conservativity (i.e., the principle of minimal change) should not be a preference criteria for revolutionary change, since scientific revolutions involve radical shifts in the conceptual structure of scientific theories. The identification of these preference criteria, together with their measurement and their weight in choosing

the best scientific theories is out of the scope of this paper. Hence, we will assume an arbitrary preference ordering for each set of conceptual structures prior to revisions, and we will determine how this preference ordering reacts to revisions. Since we are dealing with revolutionary change in particular, an important assumption we will hold is giving priority to new data, that is, the argument of a revision. This preference ordering is a novel addition to Thagard's framework, which does not include any way of comparing conceptual structures during the revision process[10]. We introduce this preference ordering in order to have an adequate conceptual revision mechanism.

A preference ordering may rate multiple conceptual hierarchies as the best ones. In the belief revision literature, these cases are commonly solved by taking the intersection of the selected alternatives, following the *partial meet contraction and revision* operations introduced within the AGM paradigm. However, as we will show with an example, intersecting multiple conceptual hierarchies may generate inconsistent conceptual structures. As a solution to this problem, we propose a repetitive revision operation, where the intersection mechanism is repeated until a conceptual hierarchy is obtained.

A conceptual revision model is a tuple $CS^{\oplus\leq} = \langle CS, \oplus, \leq \rangle$, such that

- $\langle CS, \oplus \rangle$ is a fusion model on conceptual structures, and
- $\leq$ is a connected preorder on a set of conceptual structures.

As determined for the fusion models and conceptual expansions, our revision models can be generalized using the generalized fusion operator $\bigoplus$ in place of the fusion operator $\oplus$. However, we require the object of the revision to be a fully formed conceptual structure, while the argument of the revision can be a partial one. Consequently, the result of a conceptual revision via our described expansion operations (particularly via the generalized version) is always a conceptual structure.

A generalized conceptual revision model is a tuple $\langle \mathcal{C}, \mathcal{O}, K, P, R, I, \bigoplus, \leq \rangle$, such that

- $\langle \mathcal{C}, \mathcal{O}, K, P, R, I, \bigoplus \rangle$ is a generalized fusion model on conceptual structures, and
- $\leq$ is a connected preorder on a set of conceptual structures.

The preorder between conceptual structures is a preference ordering. If X, Y are conceptual structures in a set, and $X \leq_{CS^{\oplus\leq}} Y$, we say the conceptual structure X is at least as preferred as the conceptual structure Y given the model $CS^{\oplus\leq}$. The best conceptual structures in a set are the ones that are minimal under $\leq$, i.e., X is minimal under $\leq$ in a set S of conceptual structures iff for all Y in S, it holds that $X \leq_{CS^{\oplus\leq}} Y$.

We propose that after the revision, the preference ordering on a set of conceptual structures changes as follows: let $\mathcal{CS}$ be a non-empty set of conceptual structures, a (possibly partial) conceptual structure H as the argument of revision, and let $\leq$ be the pre-revision preference ordering on $\mathcal{CS}$ and $\leq'$ be the revised preference ordering, then,

[10] We should note that Thagard has an evaluation mechanism between different conceptual systems based on the notion of explanatory coherence (Thagard, 2000). However, his mechanism is used only to compare fully finished conceptual structures after the changes have taken place. We leave the study of the relations between Thagard's evaluation mechanism and our preference ordering for future work.

- for all $A, B \in \mathcal{CS}$, if $H \subseteq A$ and $H \not\subseteq B$ then $A \leq' B$ and $B \not\leq' A$, and if $H \subseteq B$ and $H \not\subseteq A$ then $B \leq' A$ and $A \not\leq' B$,
- otherwise, $A \leq' B$ iff $A \leq B$.

Since our revision operation involves changing the preference order in a revision model, it is essentially a model-changing operation. Therefore, even if the expansion of the initial conceptual structure with the argument of the revision is a conceptual hierarchy, the revision operation does not reduce to expansion, since changes on the preference ordering are significant for iterating any change operation on the new conceptual structures.

To see how the revision operation is applied to conceptual structures, consider the conceptual hierarchy H and the partial conceptual structure A in the above example. Suppose we want to revise H with A instead of simply expanding the former with the latter. The first step of the conceptual revision process consists of the aforementioned expansion $H \bigoplus A$ (depicted in Figure 2). Since $H \bigoplus A$ is not a conceptual hierarchy, we continue the revision operation by identifying and intersecting the best conceptual hierarchies within $H \bigoplus A$ (determined by $\leq$ on $\mathcal{CS}$). As the new data in this revision is the partial conceptual structure A, after the revision, the conceptual hierarchies of which A is a substructure are strictly more preferred over the ones which exclude some part of A. Then, an easy way to identify the best maximal hierarchies within $H \bigoplus A$ is to start with A as the base structure and expand it within $H \bigoplus A$ until reaching a maximal conceptual hierarchy. However, it might be the case that there are no maximal hierarchies within the expanded structure that include the new data. Then, one identifies all maximal hierarchies within the expanded structure; their preference ordering is based on the 'otherwise' clause in our definition above.

The following is the unique conceptual hierarchy M (Figure 6.3) within $H \bigoplus A$ which fits the description: $\mathcal{C}_M = \{[fish], [orka]\}$, $\mathcal{O}_M = \{(Bob)\}$, $K_M = \{\langle orka, fish\rangle\}$, $I_M = \{\langle Bob, orka\rangle, \langle Bob, fish\rangle\}$, $P_M = R_M = \emptyset$. The revision of H with A finalises here.

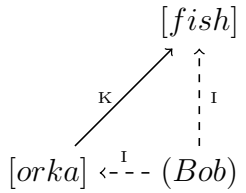


Figure 6.3: The conceptual hierarchy M.

Our next example shows how the revision operation is applied repetitively. Suppose we want to revise a conceptual structure X such that $\mathcal{C}_X = \{a, b, A, B\}$, $K_X = \{\langle a, A\rangle, \langle b, A\rangle, \langle a, B\rangle, \langle b, B\rangle, \langle B, A\rangle, \langle A, B\rangle, \}$ and $\mathcal{O}_X = P_X = R_X = I_X = \emptyset$ with the empty conceptual structure $\{\emptyset\}$. There are two maximal hierarchies within the expanded conceptual structure $X \oplus \{\emptyset\}$ whose kind relations are the following, $K_{X1} = \{\langle a, A\rangle, \langle b, A\rangle, \langle a, B\rangle, \langle b, B\rangle, \langle B, A\rangle\}$ and $K_{X2} = \{\langle a, A\rangle, \langle b, A\rangle, \langle a, B\rangle, \langle b, B\rangle, \langle A, B\rangle\}$. Suppose they are preferred equally.

Then, their intersection includes a kind-relation which does not have top element, i.e. $K_{X1\cap X2} = \{\langle a, A\rangle, \langle b, A\rangle, \langle a, B\rangle, \langle b, B\rangle\}$, that cannot be the kind-relation of a conceptual hierarchy. We therefore repeat the revision operation, first determining the best maximal hierarchies within the conceptual structure $X1 \cap X2$. These are the hierarchies with the kind-relations $K_Y = \{\langle a, A\rangle, \langle b, A\rangle\}$ and $K_Z = \{\langle a, B\rangle, \langle b, B\rangle\}$. If they are preferred equally, then the revised conceptual hierarchy has in its concept domain only the concepts a and b, together with an empty kind-relation[11].

We can then define our conceptual revision operation as follows[12]:

Given a generalized conceptual revision model $\langle \mathcal{C}, \mathcal{O}, K, P, R, I, \bigoplus, \leq\rangle$, and given H is a conceptual structure and A is a (partial) conceptual structure, H revised with A (let us denote it with $H * A$) is determined by the following cases:

1. $\cap(B : B$ is a maximal hierarchy within $H \bigoplus A$ and for all maximal hierarchies $C \subseteq H \bigoplus A$ it holds that B is at least as preferred as C based on the revised preference ordering $\leq')$ if $\cap B$ constitutes a conceptual hierarchy,

2. $\cap(C : C$ is a maximal hierarchy within $\cap B$ and for all maximal hierarchies $D \subseteq \cap B$ it holds that $C \leq' D)$ if $\cap B$ does not constitute a conceptual hierarchy and $\cap C$ constitutes a conceptual hierarchy,

3. repeat case 2 substituting $\cap B$ with $\cap C$ until reaching a conceptual hierarchy as the result of the intersection if otherwise.

6.2.3 Contraction on conceptual structures

Contracting a conceptual structure means eliminating a part of it. While our contraction operation is defined based on conceptual revision, it differs from revision significantly in terms of how the argument of a contraction should be formulated or expressed. Suppose we want to contract an instance-pair $\langle x, y\rangle$ from a conceptual structure. In regards to the arguments of revision, we required that each element that constructs a relation or a link is explicitly stated as part of the argument. An analogous way of formulating the argument of contraction would be the following $C = \{y\}, O = \{x\}, I = \{\langle x, y\rangle\}$. However, it is not (always) necessary to eliminate the concept and the object in order to eliminate the instance-link. Hence, a well-formed argument for our contraction operation does not have the limitation we proposed for revisions. Thus, an argument of contraction can be any element of a conceptual structure. As we did for revision, we require the object

[11]Both examples of conceptual revision reveal a significant amount of information loss as a result. This is connected to the revolutionary aspect of the scientific changes we want to represent. As it was famously stressed by Kuhn (Kuhn, 1970), scientific revolutions involve often the loss of information in the transition from one scientific theory to its successor, a phenomenon commonly known in philosophy of science as Kuhnian loss.

[12]We state the generalized conceptual revision models in the definition. Given that both H and A are fully formed conceptual structures, the same definition can be applied via the conceptual revision models with the fusion operator $\oplus$ in place of $\bigoplus$.

of the contraction to be a fully formed conceptual structure, while the argument of the contraction can be a partial one. Therefore, the result of a conceptual contraction is always a conceptual structure.

In order to formalize conceptual contraction, we introduce in our revision models a set-theoretical elimination operation $\ominus$:

Given that A and B are conceptual structures, $A \ominus B = A/B$, such that

- $\mathcal{C}_{A\ominus B} = \mathcal{C}_A/\mathcal{C}_B$
- $\mathcal{O}_{A\ominus B} = \mathcal{O}_A/\mathcal{O}_B$
- $K_{A\ominus B} = K_A/K_B$
- $P_{A\ominus B} = P_A/P_B$
- $R_{A\ominus B} = R_A/R_B$
- $I_{A\ominus B} = I_A/I_B$

That is, we simply eliminate B from A. It is easy to generalise the $\ominus$ operation as we did for the fusion operation:

A generalized elimination operation is a family of elimination functions $\ominus = \{\ominus_C, \ominus_O, \ominus_K, \ominus_P, \ominus_R, \ominus_I\}$ from $\langle \mathcal{C}, \mathcal{O}, K, P, R, I\rangle \times \langle \mathcal{C}, \mathcal{O}, K, P, R, I\rangle$ to $\langle \mathcal{C}, \mathcal{O}, K, P, R, I\rangle$, such that:

- $\ominus_C$ is a function from $\mathcal{C} \times \mathcal{C}$ to $\mathcal{C}$ such that $\mathcal{C}_A \ominus_C \mathcal{C}_B = \mathcal{C}_A/\mathcal{C}_B$
- $\ominus_O$ is a function from $\mathcal{O} \times \mathcal{O}$ to $\mathcal{O}$ such that $\mathcal{O}_A \ominus_O \mathcal{O}_B = \mathcal{O}_A/\mathcal{O}_B$
- $\ominus_K$ is a function from $K \times K$ to K such that $K_A \ominus_K K_B = K_A/K_B$
- $\ominus_P$ is a function from $P \times P$ to P such that $P_A \ominus_P P_B = P_A/P_B$
- $\ominus_R$ is a function from $R \times R$ to R such that $R_A \ominus_R R_B = R_A/R_B$
- $\ominus_I$ is a function from $I \times I$ to I such that $I_A \ominus_I I_B = I_A/I_B$

Note that the elimination operation does not include taking the transitive closures of the resulting relations.

A conceptual revision and contraction model is a tuple $CS^{\oplus\ominus\leq} = \langle \mathcal{CS}, \oplus, \ominus, \leq\rangle$, such that

- $\langle \mathcal{CS}, \oplus\rangle$ is a fusion model on conceptual structures,
- $\ominus$ is the set-theoretical elimination operation on conceptual structures, and
- $\leq$ is a connected preorder on a set of conceptual structures.

The revision and contraction models can be generalized using the generalized fusion operator $\bigoplus$ in place of the fusion operator $\oplus$ and the generalized contraction operator $\ominus$ in place of the $\ominus$ operator.

A generalized conceptual revision and contraction model is a tuple $CS^{\bigoplus \ominus \leq} = \langle \mathcal{CS}, \bigoplus, \ominus, \leq \rangle$, such that

- $\langle \mathcal{CS}, \bigoplus \rangle$ is a generalized fusion model on conceptual structures,
- $\ominus$ is the generalized elimination operation, and
- $\leq$ is a connected preorder on a set of conceptual structures.

As in the case of revision, in our framework the outcome of a contraction operation on a conceptual structure ought to be a conceptual hierarchy. It should also be the case that contraction operations do not expand the contracted structures with novel relations, concepts or objects[13]. As we will see, even if nothing is added to a conceptual structure through contraction, the hierarchical structure may be lost. For instance, contracting a structure with respect to a kind-link may affect the transitivity of the kind-relation hence breaking the hierarchical structure. We restore the consistency of contracted conceptual structures as we did for revised structures.

Since our contraction operation is an elimination operation, the preference ordering is not affected while contracting conceptual structures. The ordering plays the same role it did in restoring consistency, however it does not change in the process. As an example of a contraction, consider the conceptual hierarchy H', such that

$$\mathcal{C}_{H'} = \{[mammal], [whale], [orka], [narhwale]\},$$

$$\mathcal{O}_{H'} = I_{H'} = P_{H'} = R_{H'} = \emptyset$$

$$K_{H'} = \{\langle whale, mammal\rangle, \langle orka, mammal\rangle, \langle orka, whale\rangle, \langle narwhale, whale\rangle, \langle narwhale, mammal\rangle\}.$$

Consider also the partial conceptual structure A', such that $\mathcal{C}_{A'} = \{[orka], [narwhale]\}$ and $K_{A'} = \{\langle narhwale, whale\rangle, \langle orka, whale\rangle, \langle orka, mammal\rangle, \langle narwhale, mammal\rangle\}$. Suppose we want to contract H' with respect to A'. We start with the simple elimination of A' from H', obtaining $H' \ominus A'$, such that

$$\mathcal{C}_{H' \ominus A'} = \{[mammal], [whale]\}$$

$$K_{H' \ominus A'} = \{\langle whale, mammal\rangle\}$$

$$\mathcal{O}_{H' \ominus A'} = R_{H' \ominus A'} = P_{H' \ominus A'} = I_{H' \ominus A'} = \emptyset.$$

[13]This is another reason in favour of keeping the operation of adding relations or links to recover transitivity as part of the conceptual expansion. This way, we can use the exact process defined for revisions in order to retain consistency after conceptual contraction without making any additions to the conceptual structure.

The output of this particular contraction $H' - A'$ (Figure 6.4) is equal to $H' \ominus A'$, since the latter is a conceptual hierarchy, and the conceptual contraction process does not involve the change of preference ordering. If it were the case that $H' \ominus A'$ is not a conceptual hierarchy, the consistency of the resulting conceptual structure would be recovered via the same process we described for conceptual revisions, based on the initial preference ordering of $H' \ominus A'$.

Figure 6.4: The conceptual hierarchy H' (on the left) and the conceptual hierarchy $H' - A'$ (on the right).

In the following definition we state the generalized conceptual expansion and generalized conceptual elimination operators in the definition. Given both H and A are fully formed conceptual structures, the same definition can be applied via the fusion operator $\oplus$ and the elimination operator $\ominus$.

Given a generalized conceptual revision and contraction model $CS^{\bigoplus \ominus \leq}$, and given H is a conceptual structure and A is a (partial) conceptual structure, H contracted with A (let us denote it with $H - A$) is determined by the following cases:

1. $\cap(B : B$ is a maximal hierarchy within $H \ominus A$ and for all maximal hierarchies $C \subseteq H \ominus A$ it holds that B is at least as preferred as C based on the preference ordering $\leq)$ if $\cap B$ constitutes a conceptual hierarchy,

2. $\cap(C : C$ is a maximal hierarchy within $\cap B$ and for all maximal hierarchies $D \subseteq \cap B$ it holds that $C \leq D)$ if $\cap B$ does not constitute a conceptual hierarchy and $\cap C$ constitutes a conceptual hierarchy,

3. repeat case 2 substituting $\cap B$ with $\cap C$ until reaching a conceptual hierarchy as the result of the intersection if otherwise.

6.2.4 Rationality postulates for conceptual change

In this section we will show how our conceptual revision models satisfy several rationality postulates analogous to the AGM ones for belief revision (Alchourrón, Gärdenfors, and

Makinson 1985). Since our system works at the conceptual level of abstraction, we cannot in fact straightforwardly apply the AGM postulates to it. Thus, for each AGM revision postulate we will try to develop an analogous postulate at the conceptual level. We will also discuss rationality postulates for conceptual contraction, trying to comprehend the counterparts of the conceptual revision ones.

First, we show that a conceptual counterpart of the AGM closure and consistency postulates for revision is satisfied in our framework[14]. We will cal this first conceptual revision postulate the *hierarchy postulate*. This postulate amounts to the claim that a conceptual revision operation always results in a conceptual hierarchy. Recall in fact that for a conceptual structure to be a conceptual hierarchy, the information represented by the relations of the structure should not be contradictory. A conceptual hierarchy is furthermore closed, in the sense that none of the links needed for the transitive closures of the relations is missing. Hence, in our framework the consistency of a conceptual structure is intertwined with its completeness. Our framework satisfies this postulate thanks to the conjunction of the following properties: all conceptual structures have at least one maximal conceptual hierarchy as their substructure (due to their finiteness), the preference ordering always yields some minimal (most preferred) conceptual hierarchy (due to its connectedness), and a conceptual hierarchy can always be reached in finitely many iterations of our revision operation.

Next, we show that our framework satisfies a *success postulate*, i.e. the claim that if the argument of a conceptual revision is a conceptual hierarchy, the argument becomes a substructure of the revised conceptual structure. This postulate corresponds to a weakened version of the AGM success postulate for revisions[15]. For the satisfaction of this postulate, it suffices that the argument of the revision is among the minimal conceptual structures in the (revised) preference ordering. This is achieved since our revision mechanism involves exactly this reordering of the preferences when revising a conceptual structure.

The third rationality postulate we consider is the *vacuity postulate*, i.e. the requirement that if the expansion of a conceptual structure is a conceptual hierarchy, this expansion is equal to the output of the revision process without the reordering of the preference relation. This requirement corresponds to the vacuity postulate in the AGM theory and it is satisfied by our framework because such an expanded conceptual hierarchy becomes the unique maximal conceptual hierarchy within itself.

Lastly, we consider the *inclusion postulate*, i.e. the requirement that the outcome of a conceptual revision is a substructure of the result of expanding the original conceptual structure with the argument of the revision. This postulate corresponds to the AGM

[14]Our consistency claim is stronger than what is required by the AGM consistency postulate, which includes the requirement that the new belief is not a contradiction.

[15]The AGM success postulate requires inclusion of the new belief without an antecedent that says it is a consistent belief. On the other hand, the success postulate required for base-generated beliefs by Rott (Rott, 2001) and Hansson Hansson (1999) has that antecedent. We consider the weaker version of this postulate due to the strong consistency claim we established. Otherwise we have a contradiction saying the result of a conceptual structure is always consistent and if we revise a conceptual structure with a contradiction, the contradiction is part of the revised structure.

inclusion postulate for revisions. This requirement makes sure that a conceptual structure is not expanded further than what is needed to consistently include the argument of the revision. This postulate is satisfied in our framework since all the steps of our revision operation involve only substructures of the expanded conceptual structure[16].

After we mapped and analyzed some rationality postulates for conceptual revision framework, let us briefly discuss the corresponding contraction postulates. The first conceptual contraction postulate requires the result of a conceptual contraction to be a conceptual hierarchy. Since conceptual contraction involves the same consistency-recovery mechanism of conceptual revision, this principle is satisfied for reasons analogous to the revision case. The success postulate for conceptual contraction requires the argument of the contraction (a conceptual structure or a part of one) to not be a substructure of the contracted conceptual structure. A weaker version of this principle, which limits the argument of the contraction to non-empty conceptual structures or their parts, is satisfied in our framework. This is because, once the argument of the contraction is deleted from the initial conceptual structure, nothing is added to the resulting structure while rebuilding consistency. The vacuity postulate for conceptual contraction states that, if the argument of the contraction does not occur in the initial conceptual structure, then no changes are made to this structure. In our framework, this requirement is not satisfied, since it is possible that the initial conceptual structure changes in the process of consistency-recovery. A weaker version of this requirement, assuming that the initial conceptual structure is a conceptual hierarchy, is however satisfied since the initial conceptual hierarchy is the unique maximal conceptual hierarchy within itself. Lastly, we consider the requirement that the result of the contraction operation is such that, if it is expanded with the argument of the contraction, the initial conceptual structure is recovered (this requirement corresponds to the AGM recovery postulate, which is loosely the counterpart of the inclusion postulate for revision). This requirement is not satisfied in our framework, since our contraction operation may involve deleting more than the argument of the contraction (due to consistency requirements).

6.3 Conceptual Revision in Revolutionary Times

In the last section, I presented a novel conceptual revision model, showing how its revision and contraction operations satisfy several rationality postulates for conceptual change. In this section, I will show how we can mirror the dynamics of Thagard's conceptual systems in this conceptual revision system. In Section 3.1, I will demonstrate how almost every kind of change described by Thagard can be adequately represented in the conceptual revision

[16]It should be noted that there are three other basic AGM rationality postulates we did not discuss here. One is the extensionality postulate which states that revision of a belief set with classically logically equivalent arguments lead to logically equivalent revised belief sets. Since we did not comment on identity principles concerning the conceptual structures, we cannot map this requirement to our framework for now. The other two postulates are about revisions with conjunctions. We do not consider these as relevant for our current conceptual revision framework, since we did not discuss relations between structures which would correspond to logical connectives.

framework via a suitable (combination of) change operation(s) on conceptual structures. In Section 3.2, I will instead show how the conceptual revision framework can be applied to rationally reconstruct one of Thagard's main case studies of scientific revolution, namely, the chemical revolution.

6.3.1 Mirroring Thagard's kinds of changes in our conceptual revision model

As we saw in Section 1.1, Thagard described a fine-grained hierarchy of nine degrees of changes applicable to conceptual systems, ordered by their increasing strength (i.e. from the weakest to the strongest): instance-addition, rule-addition, part-addition, kind-addition, concept-addition, kind-collapse, hierarchy-reorganization, and tree-switching.

In what follows, we will discuss each of these degrees of change one by one, from the weakest to the most radical one. With the exception of tree-switching, whose case will be completely different from all the others, the structure of our discussion will take the following form. We will first present how a given kind of change operates on one of Thagard's conceptual systems. Then, we will explain informally how this kind of change can be represented in our framework. After that, we will give a formal definition of the degree of change under focus, showing how it can be seen as a special case of (a series of applications of) our revision and/or our contraction operations. Finally, we will present a toy-example of this kind of change in our framework in order to make clearer our proposed formalization.

Instance-addition. The addition of an instance-link is the least radical kind of change described by Thagard. It consists in the addition of a single instance link between one object node and one conceptual node of a given conceptual system, representing the information that a given individual is an instance of a given concept.

In our framework, we can mirror instance-addition via our conceptual revision operation, revising a given conceptual structure with a (partial) conceptual structure that includes a non-empty instance-relation. In particular, we can define three different forms of instance-addition as three different constraints on the argument of revision. The most general form, what we will call general instance-addition, consists of requiring the argument of the revision to include a non-empty instance relation. A more specific form of instance-addition, i.e. pure instance-addition, requires the argument of the revision to have instance-relation as its only non-empty relation (concept and object domains can be non-empty as well). Finally, we have an atomic instance-addition when the argument of the revision of a pure instance-addition has a single instance-pair as its instance-relation. This last form corresponds to (our interpretation of) Thagard's understanding of instance-addition.

More formally, a conceptual revision operation $H * A$ represents a *general instance-addition* iff $I_A \neq \emptyset$. A conceptual revision operation $H * A$ represents a *pure instance-addition* iff $I_A \neq \emptyset$ and $K_A = P_A = R_A = \emptyset$. A conceptual revision operation $H * A$

represents an *atomic instance-addition* iff $|I_A| = 1$ and $K_A = P_A = R_A = \emptyset$. For an example of a general instance-addition, see the conceptual revision example presented in Section 3.1.

Rule-addition. The second kind of change described by Thagard consists in adding a rule-link between two concepts nodes of a given conceptual system. This change represents adding the information that a generic holds between two concepts.

In our framework, rule-addition is represented similarly as we treated instance-addition, i.e. by requiring the argument of our revision operation to include a non-empty rule-relation. As in the previous case, three different forms of rule-addition can be defined, differing in terms of generality: general rule-addition, pure rule-addition, and atomic rule-addition.

More formally, a conceptual revision operation $H * A$ represents a *general rule-addition* iff $R_A \neq \emptyset$. A conceptual revision operation $H * A$ represents a *pure rule-addition* iff $R_A \neq \emptyset$ and $K_A = P_A = I_A = \emptyset$. A conceptual revision operation $H * A$ represents an *atomic rule-addition* iff $|R_A| = 1$ and $K_A = P_A = I_A = \emptyset$.

As a simple example of rule-addition, let H be composed by:

$$\mathcal{C}_H = \{[mammal], [whale], [orka]\}$$

$$\mathcal{O}_H = I_H = P_H = R_H = \emptyset$$

$$K_H = \{\langle whale, mammal\rangle, \langle orka, mammal\rangle, \langle orka, whale\rangle\}.$$

and let A be composed by $\mathcal{C}_A = \{[mammal], [air]\}$, $\mathcal{O}_A = \emptyset$, $R_A = \{\langle mammal, air\rangle\}$ (intuitive interpretation: mammals breath air)[17]. The output of this revision operation expands the rule-relation of H with R_A and the pairs $\langle whale, air\rangle, \langle orka, air\rangle$. We then have $H \bigoplus A = M$ where:

$$\mathcal{C}_M = \{[mammal], [whale], [orka], [air]\},$$

$$\mathcal{O}_M = P_M = I_M = \emptyset$$

$$K_M = \{\langle whale, mammal\rangle, \langle orka, whale\rangle, \langle orka, mammal\rangle\}$$

$$R_M = \{\langle mammal, air\rangle, \langle whale, air\rangle, \langle orka, air\rangle\}.$$

Since M is a conceptual hierarchy, we have $H * A = M$.

[17]Note that it would be possible in our framework to differentiate rules in terms of their intended interpretation, so that for instance the rule *breath* is represented differently from other rules (e.g. *swim*) that may be added to a given conceptual structures. We decided to follow Thagard in leaving the interpretation of the rules outside our framework, considering all rules as uninterpreted rule-pairs.

Part-addition. The third kind of change described by Thagard is called part-addition or decomposition. It consists in adding a part-link between two concept nodes of a given conceptual system, representing the information that a relation of part-hood holds between the concepts denoted by these nodes.

In our framework, part-addition is represented similarly as we treated instance-addition and rule-addition, i.e. by requiring the argument of our revision operation to include a non-empty part-relation. As in the previous cases, three different forms of part-addition can be defined, differing in terms of generality: general part-addition, pure part-addition, and atomic part-addition.

More formally, a conceptual revision operation $H * A$ represents a *general part-addition* iff $P_A \neq \emptyset$. A conceptual revision operation $H * A$ represents a *pure part-addition* iff $P_A \neq \emptyset$ and $K_A = R_A = I_A = \emptyset$. A conceptual revision operation $H * A$ represents an *atomic part-addition* iff $|P_A| = 1$ and $K_A = R_A = I_A = \emptyset$.

As a simple example of part-addition, take H to be such that:

$$\mathcal{C}_H = \{[mammal], [whale], [orka]\}$$

$$\mathcal{O}_H = P_H = R_H = I_H = \emptyset$$

$$K_H = \{\langle whale, mammal\rangle, \langle orka, mammal\rangle, \langle orka, whale\rangle\}$$

Let A be composed by $\mathcal{C}_A = \{[mammal], [lungs]\}$, $\mathcal{O}_A = \emptyset$, $P_A = \{\langle mammal, lungs\rangle$ (intuitive interpretation: mammals have lungs), and $K_A = R_A = I_A = \emptyset$. The output of this revision operation expands the part-relation of H with P_A and the pairs $\langle whale, lungs\rangle, \langle orka, lungs\rangle$.We then have $H \bigoplus A = M$ where:

$$\mathcal{C}_M = \{[mammal], [whale], [orka], [lungs]\},$$

$$\mathcal{O}_M = R_M = I_M = \emptyset$$

$$K_H = \{\langle whale, mammal\rangle, \langle orka, mammal\rangle, \langle orka, whale\rangle\}$$

$$P_M = \{\langle mammal, lungs\rangle, \langle whale, lungs\rangle, \langle orka, lungs\rangle\}.$$

Since M is a conceptual hierarchy, we have $H * A = M$ (Figure 5).

Kind-addition. The fourth kind of change described by Thagard consists in adding a kind-link between two concept nodes of a given conceptual system, representing the information that a relation of kind-hood holds between the concepts denoted by these nodes. Furthermore, Thagard, following Carey's terminology for conceptual change in child psychology (Carey, 1985), distinguishes two special cases of (series of) kind-addition(s): *coalescence* and *differentiation*. The former type of kind-addition happens when we add a superordinate conceptual node linked via a series of kind-links with some concept nodes that had no superordinate kinds before. The latter denotes instead the addition of some subordinate conceptual nodes connected via a series of kind-links with an conceptual node that before had no subordinate kinds.

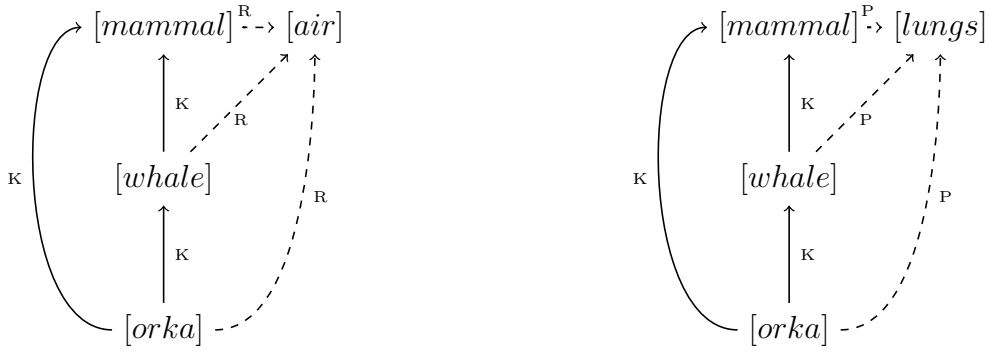


Figure 6.5: The output of the rule-addition example (on the left) and the output of the part-addition example (on the right).

In our framework, kind-addition is represented by requiring the argument of our revision operation to include a non-empty kind-relation. As in the previous case, three different forms of kind-addition can be defined, differing in terms of generality: general kind-addition, pure kind-addition, and atomic kind-addition. Coalescence and differentiation can then be represented as specific cases of general or pure kind-addition.

Formally, a conceptual revision operation $H * A$ represents a *general kind-addition* iff $K_A \neq \emptyset$. A conceptual revision operation $H * A$ represents a *pure kind-addition* iff $K_A \neq \emptyset$ and $P_A = R_A = I_A = \emptyset$. A conceptual revision operation $H * A$ represents an *atomic kind-addition* iff $|K_A| = 1$ and $P_A = R_A = I_A = \emptyset$. Furthermore, a general or pure kind-addition $H * A$ is a *coalescence* iff there exists a $x \in C_A$ such that $\langle y, x\rangle \in K_A$ and there is no w such that $\langle y, w\rangle \in K_H$. A general or pure kind-addition $H * A$ is instead a *differentiation* iff there is a $x \in C_A$ such that $\langle x, y\rangle \in K_{A_i}$ and there is no w such that $\langle w, y\rangle \in K_H$. For an example of a general kind-addition, see the conceptual revision example presented in Section 2.2.

Concept-addition. The fifth kind of change described by Thagard consists in adding a new concept node to a given conceptual system. This type of change represents the addition of a new concept to a given scientific theory[18].

In our framework, concept-addition is represented by requiring the argument of our revision operation to include a new concept. Several further restrictions can be imposed. For instance, we present here two more specific forms of concept-addition: unique concept-addition and connected concept-addition. We have a unique concept-addition when there is only one new concept in the argument of the revision (it may also include non-empty relations). We have a connected-concept addition when each new concept in the argument

[18]Thagard also stresses how concept-addition sometimes involves combining two simple concepts into a complex one (Thagard, 1992, pp. 35-36). This combination aspect of concept-addition is outside the scope of the present version of our framework, since we assumed for simplicity that the concept universe is constant.

figures in at least one relation.

Formally, a conceptual revision operation $H * A$ is a *concept-addition* iff there is an $x \in C_A$ such that $x \notin C_H$. A concept-addition $H * A$ is then a unique concept-addition iff there is only one $x \in C_A$ such that $x \notin C_H$. A concept-addition $H * A$ is then a connected concept-addition iff for all $x \in C_A$ such that $x \notin C_H$ there exists a $y \in C_A \cup O_A$ such that $\langle x, y\rangle$ or $\langle y, x\rangle$ is in $K_A \cup P_A \cup R_A \cup I_A$. For an example of a unique and connected concept-addition see the conceptual revision example in Section 2.2, the rule-addition example, or the part-addition example above.

Kind-collapse. The sixth change described by Thagard is kind-collapse, i.e. the removal of a (series of) kind-link(s) from a given conceptual system. More specifically, Thagard says that kind-collapse is the inverse change of differentiation, so that kind-collapse denotes removing all subordinate kinds of a given conceptual node.

In our framework, kind-collapse is a specific case of our contraction operation, namely, the contraction of a given conceptual structure with respect to a set of kind-pairs all of which have the same element as their second element and such that in the contracted structure this element has no subordinate kinds.

Formally, a conceptual contraction operation $H - A$ is a *kind-collapse* iff $\exists x \in C_H$ such that $K_A = \{\langle j_1, x\rangle, \ldots, \langle j_n, x\rangle\}$ and $\neg\exists y \in C_H \cup C_A$ such that $\langle y, x\rangle \in K_{H-A}$. This definition of a kind-collapse makes it the inverse process of a differentiation, just like in Thagard's system. For an example of a kind-collapse, see the contraction example in Section 2.3.

Hierarchy-reorganization. The seventh kind of change in Thagard's theory is the general process of hierarchy-reorganization or *branch-jumping*, i.e. moving a set of concept and object nodes from one part of a conceptual system to another one, thus changing (some of) their relations. This change is typical of many scientific revolutions, such as the Copernican revolution in which the earth branch-jumped from being a unique entity to a kind of planet.

In our framework branch-jumping is a specific series of our contraction and revision operations that does not involve changes to the concept-domains of the conceptual structures involved. The output of such combination is the transportation of certain parts of a given conceptual structure to a different part of it, involving some change in its relations.

Formally, we say that the sequence of contraction and revision operations $(H - A_1) * A_2$ represents a *hierarchy-reorganization* iff $\mathcal{C}_H = \mathcal{C}_{(H-A_1)*A_2}$, $\mathcal{O}_H = \mathcal{O}_{(H-A_1)*A_2}$ and either $K_H \neq K_{(H-A_1)*A_2}$ or $P_H \neq P_{(H-A_1)*A_2}$ or $R_H \neq R_{(H-A_1)*A_2}$ or $I_H \neq I_{(H-A_1)*A_2}$. Note that we leave completely open how the relations between the objects and concepts involved in the hierarchy-reorganization are transformed. Specific kinds of hierarchy-reorganization, such as part-kind transformation, can then be defined by imposing further constraints on the relations in the contraction and in the revision operation.

As an example of a hierarchy-reorganization, take H to be such that:

$$\mathcal{C}_H = \{[animal], [fish], [mammal], [whale], [orka]\},$$

$$\mathcal{O}_H = I_H = P_H = R_H = \emptyset$$

$$K_H = \{\langle whale, fish\rangle, \langle orka, fish\rangle, \langle orka, whale\rangle, \langle orka, animal\rangle, \langle whale, animal\rangle,$$
$$\langle mammal, animal\rangle, \langle fish, animal\rangle\}.$$

Let A_1 be $\{\langle whale, fish\rangle, \langle orka, whale\rangle, \langle orka, fish\rangle\}$ and A_2 be composed by $\mathcal{K}_{A_2} = \{\langle whale, mammal\rangle, \langle orka, whale\rangle\}$, $\mathcal{C}_{A_2} = \{[whale], [orka], [mammal]\}$, $\mathcal{O}_{A_2} = \{\emptyset\}$, and $P_{A_2} = R_{A_2} = I_{A_2} = \emptyset$.

The output of the hierarchy-reorganization $H - A_1 * A_2$ is then equal to the structure H' (Figure 6) where:

$$\mathcal{C}_{\mathcal{H}'} = \{[mammal], [whale], [orka], [fish], [animal]\},$$

$$\mathcal{O}'_H = R'_H = P'_H = I'_H = \emptyset$$

$$K'_H = \{\langle whale, mammal\rangle, \langle orka, mammal\rangle, \langle orka, whale\rangle, \langle orka, animal\rangle, \langle whale, animal\rangle,$$
$$\langle mammal, animal\rangle, \langle fish, animal\rangle\}.$$

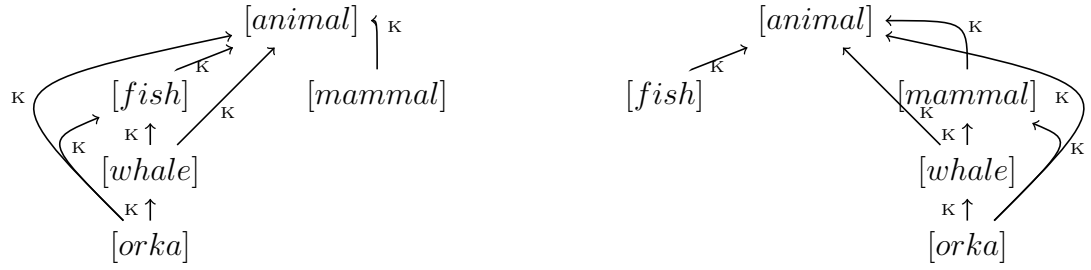


Figure 6.6: The input (on the left) and the output (on the right) of the hierarchy-reorganization example.

Tree-switching The last change described by Thagard is tree-switching, i.e. the change of the organizing principle of the whole hierarchy. This change implies thus re-interpreting any kind-relation and part-relation. An example of this kind of change is the Darwinian revolution, a revolution that involved the re-interpretation of kind-relations of biological entities as historical kinship and not as they were before as morphological similarities. This is the most radical change that can happen in science for Thagard, up to the point that it is sufficient but not necessary for having a conceptual revolution. Only certain scientific revolutions that are particularly radical exemplify tree-switching.

Since tree-switching is not really about changing the structure of a conceptual system, focusing on the external interpretation of the conceptual system, it would be at least unclear how to frame this kind of change in our framework. Using an epistemological metaphor, modeling tree-switching in our framework would be like implementing a gestalt-operation in traditional belief revision that changes the meaning of the logical consequence between

beliefs. We therefore do not treat this kind of change in the present work, focusing only on the first eight changes that affect the internal-structure of conceptual system, confident that we do not lose too much in generality, since as Thagard himself acknowledges many scientific revolutions do not even exemplify tree-switching.

6.3.2 A case study: the chemical revolution

We saw how our conceptual revision model is able to mirror eight out of nine kinds of conceptual change described by Thagard. In order to further elucidate how our revision model works, we will show how it can be applied to reconstruct one of Thagard's main case studies of radical conceptual change, i.e. the chemical revolution, as a series of applications of our change operations on conceptual structures.

The chemical revolution and the underlying battle between phlogiston and oxygen theory is one of the most famous examples of scientific revolution in the history of science. Its exact unfolding and its significance for our ideas about scientific rationality and progress have been heavily debated in philosophy, history, and sociology of science[19]. In what follows, we will steer as clear as possible of controversies about this important episode of scientific history. We will follow Thagard's (Thagard, 1990, 1992) reconstruction without committing ourselves to any specific historical or philosophical narrative.

Thagard describes the conceptual development of Laviosier's oxygen theory as a succession of four different conceptual systems, representing different historical stages of Laviosier's research: the early experiments of 1772, the developing views of 1774, the developed views of 1777, and the mature oxygen theory of 1789.

In the first half of the eighteenth century, the leading chemical theory of gases was centered around the concept of phlogiston. Phlogiston, according to Stahl who coined the name, was the inflammable principle, the basic substance responsible for the processes of combustion and 'calcination' (i.e. rusting, the production of calx). When a substance burns, it releases its phlogiston content into the air, transforming the ambient air into phlogisticated air. This saturation of the air with phlogiston was used by Priestley to explain the puzzling fact that the combustion of bodies, when it takes place in a closed vessel, often stops before the body is fully burnt. When the air is saturated by phlogiston, the burning (aka the release of phlogiston from the body) naturally stops. Phlogiston theory could thus explain many puzzling phenomena of the behavior of gases through a single principle.

The story of Laviosier's oxygen theory begins in 1772, when the young Laviosier started to focus on how exactly air combines with substances during combustion and calcination. Laviosier learned from Guyton de Morveau that metal gain weight during calcination, which in the phlogiston paradigm of the time meant that metal gain weight while losing phlogiston. Laviosier then noticed that effervescence occurs when metals are placed in

[19]A very partial primer on these literatures about the chemical revolution consists of (Toulmin, 1957; Musgrave, 1976; Thagard, 1990; Chang, 2012; Kusch, 2015; Conant, 1950; Golinski, 1992; Siegfried, 2002; Kim, 2003).

acids. He took this phenomenon to be evidence for the idea that calxes contain air of some kind. This fact could explain why metals gain weight during calcination: they gain weight because in the process of producing calx, air gets fixated into it. Here is how Thagard depicts the conceptual system related to these findings (Figure 6.7).

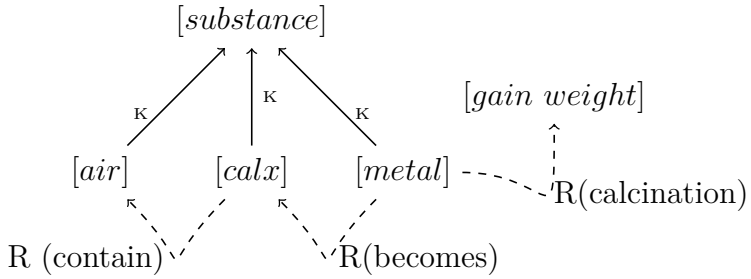


Figure 6.7: Laviosier's 1772 conceptual system. Thick lines represent kind-links, while dashed lines represent rule-links.

In this simple conceptual system, we have five conceptual nodes: substance, air, calxes, metals, gain weight. Air, calxes, and metals are three kinds of substances. This kind hierarchy is represented by the three kind links in the picture. We then have three rule links, representing the following information: metal gain weight during calcination, metals become calxes, calxes contain air (Laviosier's aforementioned hypothesis). This conceptual system can be thought as the simple conceptual correlate of the hypotheses that Laviosier made after his experiments in 1772.

We can represent in our conceptual revision framework this conceptual system with the following conceptual structure L_1:

$$\mathcal{C}_{L_1} = \{[substance], [air], [calx], [metal], [gainweight]\},$$

$$K_{L_1} = \{\langle air, substance\rangle, \langle calx, substance\rangle, \langle metal, substance\rangle\},$$

$$R_{L_1} = \{\langle calxes, air\rangle, \langle metals, calxes\rangle, \langle metals, gainweight\rangle\},$$

$$\mathcal{O}_{L_1} = I_{L_1} = P_{L_1} = \emptyset.$$

It is easy to see that this conceptual structure is a consistent one, since its kind-relation is a kind-hierarchy and its set of rules is a consistent set of rules.

The second stage of Laviosier's conceptual development, in Thagard's reconstruction, consists of his developing views in 1774. Following his hypotheses of 1772, Lavoisier conducted experiments on the combustion of phosphorous and sulfur, discovering that the products of their combustion weigh more than the originals. Thus, during combustion, just like in calcination, the weight of substances increases. Just like for calxes, Lavoisier thought that the fixation of air in the substances was responsible for this increase. More specifically, in 1774 Lavoisier proposed a rough classification of airs, specifying three kinds of air: fixed, nitrous, and common air. Common air was then the type of air that Lavoisier considered responsible for the increase of weight in the calcination of metals and in the

combustion of phosphorous and sulfur. Here is how these new conceptual developments can be represented as a conceptual system in Thagard's framework (Figure 6.8).

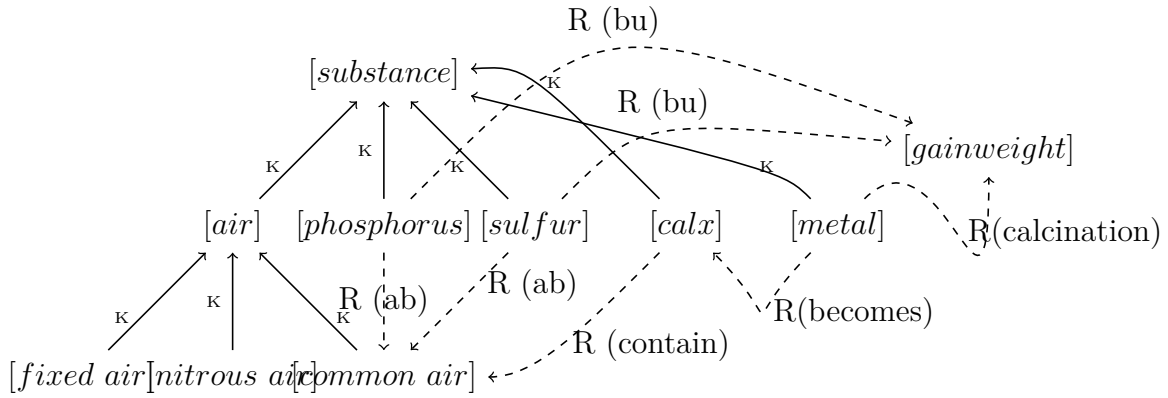


Figure 6.8: Laviosier's 1774 conceptual system. Thick lines represent kind-links, while dashed lines represent rule-links. In the rule-links, R(ab) stands for 'absorb', while R(bu) stands for 'burn'.

In respect to the 1772 conceptual system, we have here five new conceptual nodes. Two of them represent phosphorous and sulfur, the new materials with which Lavoisier experimented. The other three new conceptual nodes represent the three kinds of air that Lavoisier distinguished: fixed air, nitrous air, and common air. We have also three new kind-links, representing the fact that all these newly defined airs are kinds of air. In regards to rule links, we have four new rule-links, corresponding to the absorption of air by phosphorus and sulfur and their gain of weight while burning. Finally, one rule (calxes contain air) of the 1772 conceptual system has been contracted, replaced by the more specific rule (calxes contain common air). In Thagard's classification of degrees of conceptual change, from 1772 to 1774 Lavoisier's conceptual system has been through concept addition, kind-addition, rule-deletion, and rule-addition.

We can represent in our conceptual revision framework this conceptual system with the following conceptual structure L_2:

$$\mathcal{C}_{L_2} = \{[substance], [air], [calx], [metal], [gainweight], [phosphorus], [sulfur],$$

$$[fixedair], [nitrousair], [commonair]\},$$

$$K_{L_2} = \{\langle air, substance\rangle, \langle calx, substance\rangle, \langle metal, substance\rangle, \langle fixedair, air\rangle, \langle nitrousair, air\rangle,$$

$$\langle commonair, air\rangle, \langle commonair, substance\rangle, \langle nitrousair, substance\rangle, \langle fixedair, substance\rangle\},$$

$$R_{L_2} = \{\langle calxes, commonair\rangle, \langle metals, calxes\rangle, \langle metals, gainweight\rangle, \langle phosphorus, commonair\rangle,$$

$$\langle sulfur, commonair\rangle, \langle phosphorus, gainweight\rangle, \langle sulfur, gainweight\rangle\},$$

$$\mathcal{O}_{L_2} = I_{L_2} = P_{L_2} = \emptyset.$$

It is easy to see that this conceptual structure is a consistent one, since its kind-relation is a kind-hierarchy and its set of rules is a consistent set of rules.

We can then show how in our conceptual revision model we can mimic the transformation from the conceptual structure L_1, i.e. the one related to Lavoisier's early experiments of 1772, to the conceptual structure L_2, i.e. the one related to Lavoisier's developing views of 1774. As we mentioned earlier, this transformation involves rule-deletion, rule-addition, concept-addition, and kind-addition. As such, it can be represented via the combination of our contraction operation and our revision operation. More specifically, we contract from L_1 the rule-relation $\langle calx, air\rangle$ and then we revise the contracted structure with the conceptual structure L_{info1}:

$$\mathcal{C}_{L_{info1}} = \{[air], [calx], [gainweight], [phosphorus], [sulfur], [fixedair],$$
$$[nitrousair], [commonair]\},$$
$$K_{L_{info1}} = \{\langle fixedair, air\rangle, \langle nitrousair, air\rangle, \langle commonair, air\rangle\},$$
$$R_{L_{info1}} = \{\langle calxes, commonair\rangle, \langle phosphorus, commonair\rangle, \langle sulfur, commonair\rangle,$$
$$\langle phosphorus, gainweight\rangle, \langle sulfur, gainweight\rangle\},$$
$$\mathcal{O}_{L_{info1}} = I_{L_{info1}} = P_{L_{info1}} = \emptyset$$

This last conceptual structure consists of all and only the new information acquired by Lavoisier between the 1772 and the 1774. Formally, we get the following revision: $(L_1 - R(calx, air)) * L_{info1}$. We can see that the result of this operation is indeed equal to L_2, thanks to the pivotal addition of the kind-pairs $\langle commonair, substance\rangle$, $\langle nitrousair, substance\rangle, \langle fixedair, substance\rangle$ to ensure the transitivity of the kind-relation. This addition makes this relation a kind-hierarchy and the whole revised structure a consistent conceptual structure equal to L_2.

The third stage of Lavoisier's conceptual development according to Thagard are his developed views of 1777. Laviosier focused his efforts on refining his classification of airs, dividing common air into two components: pure air and mophette (nitrogen). He then hypothesized that pure air was the part of common air responsible for the augmentation of weight in the combustion of combustibles (i.e. substances like sulfur and phosphorus) and in the calcination of metals. Here is Lavoisier's conceptual system in 1777 (Figure 6.9).

In comparison to the 1774 conceptual system, considering the node representing mophette just a relabeling of the conceptual node representing nitrous air, we have one new conceptual node: pure air. We have the addition of two part-links (mophette is a part of common air, pure air is a part of common air), together with the addition of a new kind-link (pure air is a kind of air, not depicted in the picture), representing Lavoisier's hypothesis that common air is composed by two kinds of air: mophette and pure air. The rule-links connecting sulfur, phosphorus, and calx to common air got substituted with analogous rule-links connecting these substances with pure air, representing Lavoisier's idea that the agent responsible for combustion and calcination is this newly discovered part of common

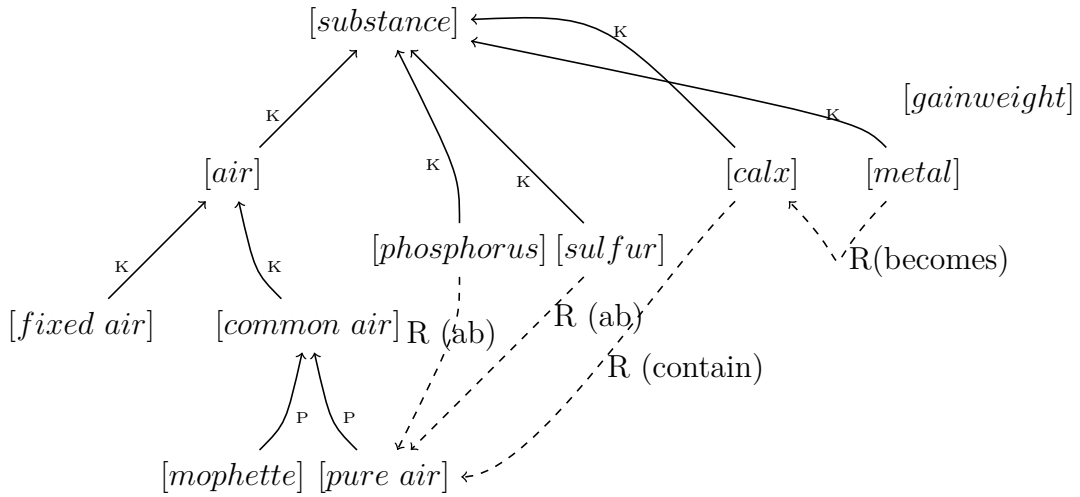


Figure 6.9: Laviosier's 1777 conceptual system. Rule-links related to the conceptual node *gainweight* and kind-links related to the conceptual nodes *mophette* and *pureair* are absent from the picture for representational clarity, but they stay the same as in the 1772 conceptual system. Thick lines represent kind-links (K) and part-links (P), while dashed lines represent rule-links.

air. From 1774 to 1777, we witness then a series of changes including concept-addition, kind-addition, rule-deletion, part-addition, and rule-addition.

We can represent the 1777 conceptual system with the following conceptual structure L_3:

$$\mathcal{C}_{L_3} = \{[substance], [air], [calx], [metal], [gainweight], [phosphorus],$$

$$[sulfur], [fixedair], [mophette], [commonair], [pureair]\},$$

$$K_{L_3} = \{\langle air, substance\rangle, \langle calx, substance\rangle, \langle metal, substance\rangle, \langle fixedair, air\rangle,$$

$$\langle commonair, air\rangle, \langle nitrousair, air\rangle, \langle pureair, air\rangle, \langle commonair, substance\rangle,$$

$$\langle nitrousair, substance\rangle, \langle pureair, substance\rangle, \langle fixedair, substance\rangle\},$$

$$R_{L_3} = \{\langle calxes, pureair\rangle, \langle metals, calxes\rangle, \langle metals, gainweight\rangle, \langle phosphorus, pureair\rangle,$$

$$\langle sulfur, pureair\rangle, \langle phosphorus, gainweight\rangle, \langle sulfur, gainweight\rangle\}$$

$$P_3 = \{\langle mophette, commonair\rangle, \langle pureair, commonair\rangle\},$$

$$\mathcal{O}_{L_3} = I_{L_3} = \emptyset.$$

It is easy to see that this conceptual structure is a consistent one, since its kind-relation and part-relation are both hierarchies and its set of rules is a consistent set of rules. We can then show how in our conceptual revision model we can model the transformation

Lavoisier's 1774 conceptual structure L_2 to Lavoisier's 1777 conceptual structure L_3. Like the step before, also this transformation can be represented via a contraction, followed by a revision. More specifically, we contract from L_2 the set of rule-relations $R_{contr2} = \{\langle calx, commonair\rangle, \langle phosphorus, commonair\rangle, \langle sulfur, commonair\rangle\}$ and then we revise the contracted structure with the conceptual structure L_{info2}:

$$\mathcal{C}_{L_{info2}} = \{[air], [calx], [phosphorus], [sulfur], [mophette], [commonair], [pureair]\},$$

$$\mathcal{O}_{L_{info2}} = I_{L_{info2}} = \emptyset,$$

$$K_{L_{info2}} = \{\langle pureair, air\rangle\},$$

$$R_{L_{info2}} = \{\langle calx, pureair\rangle, \langle phosphorus, pureair\rangle, \langle sulfur, pureair\rangle\}$$

$$P_{L_{info2}} = \{\langle pureair, commonair\rangle, \langle mophette, commonair\rangle\}.$$

This last conceptual structure represents the new information acquired by Lavoisier between the 1774 and the 1777. Formally, we get the following revision: $(L_1 - R_{contr2}) * L_{info2}$. We can see that the result of this operation is indeed equal to L_3, thanks to the pivotal addition of the kind-relation $K(pureair, substance)$ to ensure the transitivity of the kind-relation. This addition makes this relation a kind-hierarchy and the whole revised structure a consistent conceptual structure equal to L_3.

The finale stage of Lavoisier's conceptual development corresponds to his mature oxygen theory of 1789. At this point Lavoisier had completely rejected any possibility of making his theory compatible with the findings of phlogiston theorists and he has identified oxygen as the principle behind the role previously assigned to pure air in combustion and calcination. Oxygen for Lavoisier is a basic element along with light, caloric (the element then believed to be responsible for heat phenomena), hydrogen, and nitrogen. Oxygen gas is obtained when oxygen combines with the caloric element. Oxygen is the principle behind the combustion of non-metallic elements such as sulfur, phosphorus, and charcoal. Oxygen is also responsible for the oxidation of metals, understood now as the production of oxides. Lavoisier also drastically expanded the explanatory power of his oxygen theory in several directions. A crucial one was his new theory of water, understood by Lavoisier as a compound of oxygen and hydrogen, in contrast to the phlogiston theory that understood water as an element. Lavoisier in 1789 thus offered a fully worked-out, unifying chemical framework centered around the elements of oxygen and caloric. Here is how a very small part of Lavoisier's 1789 conceptual system can be rendered as a conceptual system (Figure 6.10).

In comparison to the 1777 conceptual system, considering oxygen a relabeling of pure air, we have six new conceptual nodes. Two nodes represent two new kinds of substances: elements of bodies and non-metallic ones. other three represent three kinds of elements: light, caloric. We then have one representing the new concept of oxides. The kind of airs distinguished in the 1777 conceptual system with their respective conceptual nodes, kind-links and part-links got collapsed. Instead, we have the addition of new kind-links and new part-links. We have also the addition of a rule-link (non-metallic substances

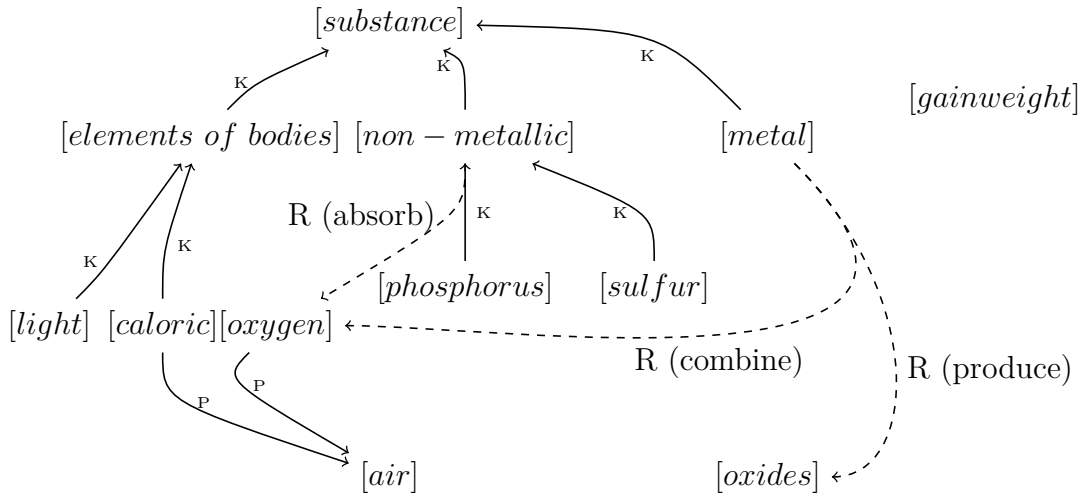


Figure 6.10: A small part of Laviosier's 1789 conceptual system. Some kind-links and rule-links are not depicted for representational clarity, but they stay the same as in the 1777 conceptual system. Thick lines represent kind-links (K) and part-links (P), while dashed lines represent rule-links.

absorb oxygen). Finally, a concept got deleted (calx), together with the related rule-links, substituted by the rule-links related to oxygen and oxides. This final step involves the deletion of concepts, kind-links, part-links, and rule-links, a hierarchy-reorganization (involving rule-relation), together with the addition of concepts, kind-links, part-links, and rule-links.

Lavoisier's 1789 conceptual system (or more actually a small part of it) can then be represented by the following consistent conceptual structure L_4:

$$\mathcal{C}_{L_4} = \{[substance], [air], [oxides], [metal], [gainweight], [phosphorus], [sulfur],$$
$$[elementsofbodies], [non-metallic], [light], [oxygen], [caloric]\},$$
$$K_{L_4} = \{\langle air, substance\rangle, \langle elementsofbodies, substance\rangle, \langle oxides, substance\rangle,$$
$$\langle metal, substance\rangle, \langle non-metallic, substance\rangle$$
$$\langle light, elementsofbodies\rangle, \langle caloric, elementsofbodies\rangle, \langle oxygen, elementsofbodies\rangle,$$
$$\langle light, substance\rangle, \langle caloric, substance\rangle, \langle oxygen, substance\rangle, \langle sulfur, substance\rangle,$$
$$\langle sulfur, non-metallic\rangle, \langle phosphorus, substance\rangle, \langle phosphorus, non-metallic\rangle\},$$
$$R_{L_4} = \{\langle metal, oxygen\rangle, \langle metal, oxides\rangle, \langle metal, gainweight\rangle, \langle phosphorus, oxygen\rangle,$$
$$\langle sulfur, oxygen\rangle, \langle phosphorus, gainweight\rangle, \langle sulfur, gainweight\rangle,$$
$$\langle non-metallic, oxygen\rangle\},$$
$$P_4 = \{\langle oxygen, air\rangle, \langle caloric, air\rangle\},$$

$$\mathcal{O}_{L_4} = I_{L_4} = \emptyset.$$

We can then show how in our conceptual revision model we can model the transformation from Lavoisier's 1777 conceptual structure L_3 to Lavoisier's 1789 conceptual structure L_4. Like the other two steps, also this transformation can be represented via a contraction, followed by a revision. More specifically, we contract from L_3 the partial conceptual structure L_{contr3}:

$$K_{L_{contr3}} = \{\langle fixedair, air\rangle, \langle commonair, air\rangle, \langle mophette, air\rangle, \langle pureair, air\rangle\}$$

$$P_{L_{contr3}} = \{\langle mophette, commonair\rangle, \langle pureair, commonair\rangle\}$$

$$R_{L_{contr3}} = \{\langle calx, pureair\rangle, \langle metal, calx\rangle\}.$$

Then, we revise the contracted structure with the conceptual structure representing (the small relevant part of) the new information acquired by Lavoisier between the 1777 and the 1789, i.e. the partial conceptual structure L_{info3}:

$$\mathcal{C}_{L_{info3}} = \{[oxygen], [light], [phosphorus], [sulfur], [caloric], [air],$$

$$[non - metallic], [elementsofbodies], [oxides]\},$$

$$K_{L_{info3}} = \{\langle elementsofbodies, substance\rangle, \langle non - metallic, substance\rangle,$$

$$\langle phosphorus, non - metallic\rangle, \langle sulfur, non - metallic\rangle, \langle light, elementsofbodies\rangle,$$

$$\langle caloric, elementsofbodies\rangle, \langle oxygen, elementsofbodies\rangle\},$$

$$R_{L_{info3}} = \{\langle non - metallic, oxygen\rangle, \langle metal, oxygen\rangle, \langle metal, oxides\rangle\}$$

$$P_{L_{info3}} = \{\langle oxygen, air\rangle, \langle caloric, air\rangle\}.$$

Formally, we get the following revision: $(L_3 - R_{contr3}) * L_{info3}$. We can easily see that the result of this operation is indeed equal to L_4.

More generally, thanks to this case study, we are able to appreciate how our conceptual revision model is able to mirror several radical degrees of conceptual change in Thagard's system with a single combination of our contraction and our revision operations.

Let us recap the main steps of the present work. Starting from Thagard's model of scientific conceptual change, we saw his taxonomy of nine degrees of conceptual change and his claim that belief revision theories can only account for the first two of them. We then presented our system of conceptual revision, i.e. a belief-revision-like system for conceptual structures. We showed how our conceptual revision and contraction operations satisfy several rationality postulates analogous to the AGM ones. We then demonstrated how our system, working at the conceptual level of abstraction, is able to mirror eight out of nine kinds of conceptual changes described by Thagard. We also showed how one of Thagard's main examples of conceptual revolution, the chemical revolution, can be rationally reconstructed in our framework as a series of applications of our conceptual revision and contraction operations.

More generally, our framework shows how belief revision theories can be mapped to the conceptual level in order to obtain a logical interpretation of radical conceptual change. The present work is only a first step towards a better understanding of the relationships between belief change and conceptual change. Several directions of future work naturally present themselves. Interesting ways of extending our framework include working with expanding domains to model conceptual combination, adding the possibility of revising conceptual structure with complex information (such as negative one, for instance) to further model logical relationships between elements of a conceptual structure, having a way of comparing differing conceptual structures in order to model Thagard's explanatory coherence notion, and also augmenting our conceptual structures in order to mimic more elaborate approaches to theory-change (e.g. Kuhn 1970; Balzer et al. 1987; Andersen et al. 2006; Masterton, Zenker, and Gärdenfors 2017; Kornmesser and Schurz 2018). These extensions would allow to model even Thagard's most radical type of conceptual change, i.e. tree-switching. It would also be interesting to merge conceptual structure with (structured) belief sets, in order to have a revision system capable of revising beliefs and concepts at the same time. Such a conceptual-plus-belief-revision system would be able to model (some of) the interesting connections between conceptual change and belief change, thereby offering a more fine-grained logical reconstruction of scientific change.

6.4 Assessing Cognitive Models in the Toolbox Framework

In this final section, I will analyze how cognitive models of conceptual change can be classified within the Toolbox framework, i.e. the meta-framework for assessing models of conceptual change that I presented in Chapter 2. More specifically, we will see how models of conceptual change based on cognitive architectures such as conceptual systems, frames, and conceptual spaces can be assessed along the nine evaluative dimensions of the Toolbox framework: units of selection, concept ontology, concept structure, kinds and degrees of conceptual change, degree of normativity, effectiveness of normative judgment, assumptions and consequences for conceptual change in science, assumptions and consequence for conceptual change in philosophy, metaphilosophical assumptions and implications. Let us survey how cognitive models of conceptual change perform in these dimensions, one by one.

Units of selection This dimension judges models of conceptual change according to the level of abstraction at which they identify conceptual entities as meaningful units of change. The three types of cognitive model of conceptual change we treated in this Chapter come equipped with different units of conceptual change, depending on the specific cognitive architecture upon which their model is based. Thus, as we saw, frame-based models might have as a meaningful unit of conceptual change a hierarchy of frames, while models based on conceptual spaces can take a whole set of (sets of) cognitive dimensions as the starting

point of their analysis. Despite the differences between these models, we can say that a common tenet of cognitive models of conceptual change is to have as meaningful units of conceptual change large sets of concepts organized in suitable hierarchies and interconnected with adequate relations and constraints. Cognitive models of conceptual change conceptualize thus their subject-matter as a large scale phenomenon, strongly focusing on inter-conceptual relations and dependencies.

Concept ontology This dimension focuses on the compatibility of a given model of conceptual change with the different philosophical positions on the ontology of concepts. Cognitive models of conceptual change are particularly apt to be coupled with a psychological view of concept ontology, due to their basic assumption of conceptualizing conceptual change as a specific kind of modification of some cognitive structure. As I stressed in Chapter 2, in fact, conceptual systems, frames, and conceptual spaces have all been used as models of conceptual knowledge acquisition and dynamic in cognitive psychology and they all can thus be given a very reasonable psychological interpretation. Moreover, contemporary cognitive science has an overwhelming preference for understanding concepts primarily as psychological entities and therefore it seems only natural to couple the heavy use of tools from cognitive science with a psychological view of concepts. Nevertheless, cognitive models of conceptual change can still be reasonably coupled with an abstract view of concept ontology, understanding the cognitive structures postulated by these models as abstract theoretical structures (cf. Gärdenfors' (Gärdenfors, 2000) scientific reading of the conceptual spaces framework). The linguistic and the worldly view of concepts seem instead *prima facie* incompatible with cognitive models of conceptual change. Cognitive architecture like frames and conceptual spaces are, in fact, often explicitly contrasted with linguistic accounts of scientific concepts and theories by their supporters, who see in their non-linguistic structure one of the reasons of the usefulness and expressiveness of these models. Similarly, cognitive structures are often coupled with internalist cognitive semantics, pushing supporters of models of conceptual change built on these structures to have often an anti-externalist background of (meta-)semantic assumptions that clashes with worldly views of concept ontology.

Concept structure This dimension focuses instead on how a given model of conceptual change assumes the structure of concepts to be constituted. The vast majority of cognitive models of conceptual change is explicitly coupled with a prototype view of concepts. As we saw in Section 1, successful cognitive architectures such as frames and conceptual spaces appeared historically as enrichment of a feature list representation of prototypes (cf. Barsalou and Hale 1993; Thagard 1984; Gärdenfors 2000). Moreover, the heavy focus on default knowledge and conceptual similarities is particularly apt to be coupled with a prototypical view of concepts. That said, enthusiasts of others cognitively-focused theories of concepts such as the exemplar view or the theory-theory can arguably find a way of interpreting cognitive models of conceptual change as building on a conceptual structure more exemplar-based or more theory-based. Moreover, as I stressed in Section 1, some cognitive

models of conceptual change (e.g. Giere 1988) have assumed a plurality of cognitive architecture in their toolkit, leaving open the possibility of coupling these kind of models with a hybrid or a pluralist view of conceptual structure. Supporters of less cognitively-oriented theories of conceptual structure that would like to employ cognitive models of conceptual spaces would have instead to resort to a more deflationary reading of these models.

Kinds and Degrees of conceptual change This dimension focuses on the kinds and degrees of conceptual change that a given model of conceptual change identifies. Cognitive models of conceptual change are particularly useful for creating very fine-grained taxonomies of kinds of conceptual change. As we saw in Section 1, in fact, all the cognitive models that I presented give rise to interesting taxonomies of changes, understanding scientific conceptual change as a series of possible modifications of specific parts of the related cognitive architectures. Whether based on conceptual systems, frames, or conceptual spaces, cognitive models of conceptual change allow an extremely subtle analysis of conceptual changes, tracing large-scale episodes of scientific change as gradual, step-by-step specific transformations of several components of a (group of) cognitive structure(s). Depending on the particular model, the taxonomy of kinds of conceptual changes recognized by cognitive models of conceptual change can vary in the number of changes isolated or in how radical certain modifications of the structure are considered to be.

Degree of normativity This dimension tracks the extent to which a given model of conceptual change is more or less normative in judging episodes of conceptual change. Cognitive models of conceptual change tend to be more descriptive than normative in their judgments of historical episodes. Discussions of what scientists should or should not have done are often eschewed by these models in favor of less value-laden discussion on the level of diachronic inter-communication between different scientific theories. That said, some philosophers have coupled their cognitive models of conceptual change with a more normative mechanism for judging historical episodes of conceptual change and their alleged rationality. An example of such normative mechanism is Thagard's (Thagard, 1992) computational model of inter-theoretical consistency that allows him to give a general measure of the empirical and theoretical coherence of competing scientific theories.

Effectiveness of normative judgment This dimension focuses on how effective the normative judgment of a model of conceptual change is. Given the aforementioned lack of normative judgments in most cognitive models of conceptual change, not a lot can be said for what regards the possible effectiveness of such judgments. Looking at the few cognitive models that are coupled with a normative mechanism, such as Thagard's (Thagard, 1992) model of conceptual revolutions, we can find that judgments of rationality or irrationality are given quite firmly. In these models, in fact, the normative judgment of (ir)rationality of a given historical episode of conceptual change is considered part of the cognitive reconstruction of a scientific history and as such a scientific endeavor itself. The normativity of these normative cognitive models of conceptual change is thus in the

eyes of their supporters a scientific one, part of what is considered to be a naturalized epistemology of science (cf. Thagard 1988; Giere 1988).

Assumptions/consequences for conceptual change in science This dimension focuses on the assumptions and the consequences of a given model of conceptual change in relation to the problems that scientific conceptual change poses in philosophy of science. Cognitive models of conceptual change depict a picture of scientific conceptual change that is highly compatible with scientific progress, scientific objectivity, and scientific realism. In fact, many cognitive models of science such as Thagard's (Thagard, 1992, pp. 103-130) one or Anderson's, Barker's, and Chen's (Andersen et al., 2006, pp. 164-179) one have as one of their main motivations the defense of a certain kind of realism and objectivity against strongly relativist and pessimist positions. Thanks to their fine-grained taxonomy of conceptual changes, cognitive models of conceptual change allow to recapture suitable notions of inter-theoretical continuity between successive scientific theories. This renewed continuity arguably diminishes the strength of pessimistic arguments for relativism and subjectivity of scientific knowledge based on the existence of scientific revolutions. Similarly, naturalized normative mechanisms for judging the rationality of episodes of scientific changes like Thagard's one give a new powerful tool of reconstruction to the supporters of scientific progress and scientific objectivity. Cognitive models of conceptual change are also usually coupled with somewhat realist position on the ontological import of scientific theories, allowing their supporters to claim new evidence in favor of positions such as constructive realism (Giere, 1988) or structural realism (Schurz and Votsis, 2014).

Assumptions/consequences for conceptual change in philosophy This dimension focuses on the assumptions and the consequences of a given model of conceptual change in relation to the problems that philosophical conceptual change poses. Cognitive models of scientific conceptual change analyze their subject-matter as completely analogous to modifications of our conceptual knowledge in non-scientific contexts. In both science and in our everyday life activities, the cognitive architectures that allow us to acquire and revise our conceptual knowledge are gradually yet constantly modified by our many theoretical and practical activities. As such, the big picture of conceptual knowledge given by cognitive models of conceptual change strongly implies the omnipresence of conceptual change in virtually all our concept-based activities. Thus, philosophical concepts, just like all the other types of concepts, should (according to these models) change together with the dynamics of philosophical practice. How do philosophical conceptual change modifies the related cognitive architectures and how different or similar this phenomenon is to its scientific counterpart has not been yet explored so much. Cognitive models of philosophical conceptual change constitute indeed a promising field of future research and, judging by the success of contemporary cognitive science and related application to scientific phenomena, have arguably the power of clarifying (and perhaps solving) some debates about conceptual change in philosophy.

Metaphilosophical assumptions and implications This dimension focuses on the metaphilosophical background that a given model of conceptual change has. The general implications that cognitive models of conceptual change may have for philosophers are all contained in the large-scale picture of human knowledge that contemporary cognitive science gives us. One of these implication is, pace Fodor (Fodor, 1998) and Machery (Machery, 2009), the everlasting centrality of conceptual knowledge for any scientifically-minded account of epistemological activities. Concepts are still the main actors of virtually all the cognitive models of human knowledge and they still provide the most meaningful units of analysis of central psychological phenomena such as categorization, abstraction, and several forms of inferential behavior. Philosophers should then focus more on concepts and conceptual knowledge when they deal with related epistemological issues. Other two general implications of cognitive models of conceptual change are the omnipresence of default reasoning mechanisms and the peripheral role of language in many cognitive processes. Contemporary cognitive science has in fact in the last forty years strongly stressed the role of default reasoning and unconscious inferences in a lot of seemingly conscious mental activities as the seemingly infinite literature on psychological biases and heuristics shows. As the growing movement of experimental philosophy argues, the recognition of the results of this kind of literature might force us to rethink traditional methods of philosophical analysis such as conceptual analysis or intuition-driven thought-experiments. At the same time, the recognition of the possibility of non-linguistic analyses of epistemological phenomena could spark some radical changes in certain deeply linguistic methodologies of analytic philosophy such as linguistic analyses of both the intuitive and the experimental kind.

Chapter 7

Conclusions

After having analyzed the four different types of model of conceptual change individually, it is now time to take a step back and look at these types of model collectively. In what follows, I will present a combined analysis of the various models of conceptual change that appeared in this work. By assessing them, comparing them, and judging them, a general conception of the phenomenon of conceptual change in science and in philosophy will appear.

In order to organize the collective analysis, I will rely on the structure of this work by comparing the four types of model of conceptual change that were analyzed (respectively) in the Chapters 3, 4, 5, and 6: pragmatic models of conceptual change, Darwinian models of conceptual change, indeterminate models of conceptual change, and cognitive models of conceptual change. Specifically, the procedure of Carnapian explication that we saw in Chapter 3 will provide a paradigmatic example of a pragmatic model of conceptual change, while the Darwinian model of conceptual evolution based on the notion of a conceptual population that I presented in Chapter 4 will be our specimen of evolutionary models of conceptual change. As instances of indeterminate models of conceptual change I will take Waismann's and Wilson's frameworks that I presented in Chapter 5, while the cognitive models of conceptual change built around (respectively) conceptual systems, frames, and conceptual spaces seen in Chapter 6 will be treated as tokens of cognitive models of conceptual change.

The comparison between these four different types of model of conceptual change will be carried out using the Toolbox framework, i.e. the meta-framework for analyzing models of conceptual change that I presented in Chapter 2. The Toolbox framework analyzes models of conceptual change along nine evaluative dimensions that check the performances of a given (type of) model of conceptual change with respect to the units of selection, the ontology of concepts, the structure of concepts, the kinds of conceptual change, the normativity, the effectiveness of the normative judgments, the assumptions and consequences for conceptual change in science, the assumptions and consequences for conceptual change in philosophy, and the metaphilosophical background that a model of conceptual change exhibits.

At the end of Chapter 3, 4, 5, and 6, I already used the Toolbox framework to analyze

the specific type of model of conceptual change that was the focus of the chapter. We can see in the following chart how the four different types of model of conceptual change performed across the nine evaluative dimensions of the Toolbox framework:

	Explication	Evolutionary	Indeterminate	Cognitive
Units	single concept	population	holism	cognitive structure
Ontology	pluralistic	public/anti-psych.	linguistic	psychology
Structure	pluralistic	plural/hybrid	plural	prototype
Kinds of CC	trans-framework	intra-pop./inter-pop.	no kinds	hierarchies
Normativity	Instrumental	selection + drift	weak	weak measures
Effectiveness	pragmatics	historical	weak	naturalized
CC in Science	value-laden	EET program	linguistic	progress
CC in Philosophy	central	evol. epist.	CC everywhere	CC everywhere
Metaphilosophy	expl. ideal	natural. epist.	language-based	cognition-based

There is a lot of to unpack in the chart above. In order to do that, I will analyze every row separately, expanding for every dimension the evaluation of a given model of conceptual change beyond the one-word summary of the above chart. In this way, we will see that beneath the surface differences in strength and weaknesses that a glimpse at the chart above shows, our analysis of these four types of model of conceptual change supports a general conception of the phenomenon of conceptual change.

The first row of the chart corresponds to the dimension of the Toolbox framework concerning the units of selection that a given type of model of conceptual change picks out. Here is the expanded row of the chart:

	Explication	Evolutionary	Indeterminate	Cognitive
Units	single concept, focus on diachronic couple ED-ET, possible extension to group of concepts	conceptual population	localized holism, focus on small linguistic practices	large-scale cognitive architecture

At first, this evaluative dimension does not present a common trend, since each type of model of conceptual change frames its subject matter through a different unit of selection. As we saw in Chapter 3, Carnapian explication focuses on a single concept or, more accurately, on the diachronic couple of concepts composed by the explicandum (ED) and the explicatum (ET). That said, we saw that Carnapian explication can be successfully used to analyze collective explications of small groups of concepts, such as the case of our phenomenal concepts of temperature (cf. Ch. 3, Sect. 4.3). The unit of selection for Darwinian models of conceptual change is instead a conceptual population, i.e. a medium-sized group of conceptual variants (cf. Ch. 4, Sect. 2). Both Waismann's and Wilson's indeterminate models of conceptual change pick out bigger units of selection such as small linguistic practices. Finally, cognitive models of conceptual change frame their subject-matter as a

phenomenon understandable by means of large-scale cognitive architectures such as (sets of) conceptual systems, frames, and conceptual spaces.

Looking at these four different units of selection a little bit closer, two common trends appear: the centrality of concepts and a localized holism. First, all four types of model have units of selection centered around concepts. Even large-scale structures such as (collections of) frames and conceptual spaces are primarily made of concepts. Secondly, all four types of model recognize that a meaningful analysis of conceptual change must take into account some degree of localized holism in its unit of analysis. This localized holism is very evident in indeterminate and cognitive models plead for the significance of context-sensitivity and background knowledge in the analysis of a given episode of conceptual change. In evolutionary and pragmatic models, instead, this localized holism is less explicit, but it can be discerned in the populational thinking of Darwinian models and in the aforementioned possibility of explicating entire groups of interrelated concepts.

The second row of the big chart corresponds instead to the dimension of the Toolbox framework dedicated to the compatibility of a given model of conceptual change with the different views on concept ontology. With respect to the big chart, I expanded this second row, splitting the concept ontology dimension in four sub-dimensions corresponding to the four main view on concept ontology (cf. Ch. 2, Sect. 1.1): the psychological, the abstract, the linguistic, and the worldly view. In order to make the table more intuitive, the compatibility of types of model with views on concept ontology is represented with five different symbols, corresponding to five different levels of agreement: strongly incompatible ($--$), incompatible ($-$), compatible ($=$), very compatible ($+$), perfectly compatible ($++$). Here is this expanded second row:

	Explication	Evolutionary	Indeterminate	Cognitive
Psychological	$=$	$-$	$+$	$++$
Abstract	$+$	$+$	$-$	$-$
Linguistic	$+$	$+$	$++$	$-$
Worldly	$-$	$+$	$=$	$--$

Looking at the four rows, we can see how compatible the four main views on concept ontology are with the four types of model of conceptual change analyzed in this work. The psychological view is of course extremely compatible with cognitive models and very compatible with indeterminate ones; it is also compatible with Carnapian explication, while instead being incompatible with evolutionary models (due to the strong preference for public/inter-subjective views of concepts of these kind of models). The abstract view is instead very compatible with both Carnapian explication and Darwinian models, but it is poorly equipped for the kind of flexibility required by indeterminate and cognitive models. The linguistic view is perfectly compatible with the indeterminate view and very compatible with both Carnapian explication and Darwinian models, while being incompatible with cognitive models. Finally, the worldly view appears very compatible with Darwinian models and compatible with indeterminate models, while being incompatible with Carnapian explication and strongly incompatible with cognitive models.

Judging simply the degree of compatibility, the linguistic view of concept ontology seems to come up as the most favored view by these four models of conceptual change. Using a simple voting rule where a '+' gives a +1 score and a '−' a −1 score (while a '=' adds 0 to the score), the linguistic view comes up with a summed preference of +3 (+1 of explication, +1 of evolutionary, +2 of indeterminate and −1 of cognitive models preferences). The linguistic view is thus more preferred than all the other three alternatives, since the psychological view scores a sum of +2, the abstract view gets 0, and the worldly view a −1.

Thus, the linguistic and the psychological views of concepts appear mostly compatible with different types of model of conceptual change. From these preferences, one can see that most models of conceptual change favor a view of concepts that allows a high degree of plasticity and flexibility, which is the main common feature of the linguistic and psychological views that both the worldly and the abstract views lack. Both the linguistic and the psychological view strike also a middle ground between strongly-subjective and strongly-objective views of concepts, allowing different kinds of inter-subjectivity in their different instantiations.

The third row of the big chart corresponds to conceptual structure, i.e. the dimension of the Toolbox framework dedicated to the compatibility of a given model of conceptual change with different accounts of how concepts are internally structured. Just like I did for the row corresponding to concept ontology, I will expand this row by splitting it into several different sub-rows. In the present case, I split it into eight different rows, corresponding to the eight main theories of conceptual structure that we saw in Chapter 2 (cf. Ch. 2, Sect. 1.2): definitional theories, functional theories, prototype theories, exemplar theories, atomic theories, ability theories, and mixed theories. Just like in the preceding chart, also in this chart I will represent the degree of compatibility of a given model with a given theory by the use of '+', '−', and '=' symbols:

	Explication	Evolutionary	Indeterminate	Cognitive
Definitional	−	=	−	−−
Functional	+	=	+	−−
Prototype	=	=	+	++
Exemplar	=	=	=	+
Atomic	−	=	−	−−
Theory	=	=	+	+
Ability	+	=	+	+
Mixed	=	+	+	+

Looking at the eight rows we can see how much each (type of) theory of conceptual structure is compatible with a given type of model of conceptual change. We can see, for instance, that definitional theories are compatible only with evolutionary models, while functional theories are instead very compatible with both Carnapian explication and indeterminate models. Prototype theories are *prima facie* compatible with all four different

theories, while exemplar and atomic theories do not fair as well. Theory theories are compatible with all models but the evolutionary ones, while ability and mixed theories seem compatible with all four types of model of conceptual change.

The simple voting rule that we used for judging the most compatible view on the ontology of concepts gives us, when applied to the case of theories of conceptual structure, the following results. The most preferred theories of conceptual structure are prototype theories, ability theories, and mixed theories, all with a summed score of +3. Theory theories, exemplar theories, and functional theories follow with a score of (respectively) +2, +1, and 0. Lastly, definitional theories and atomic theories both have a strongly negative score of (respectively) -3 and -4.

Thus, prototype, ability, and mixed theories appear the views of conceptual structure mostly compatible with different types of model of conceptual change. More generally, we can see that the various models of conceptual change are often compatible with several different theories of conceptual structures, showing therefore that they do not rely so much on specific characteristics of conceptual structure. Another general trend that can be discerned is a common preference for theories of conceptual structure that equip concepts with a rich internal structure. This is shown by noting that all such inflationist theories of conceptual structure (e.g. prototype theories, theory theories, ability theories, mixed theories) are very compatible with different models of conceptual change. Theories that give a very lightweight picture of conceptual structure, such as definitional and atomic theories, are instead incompatible with most of the models seen so far.

The fourth row of the big chart corresponds instead to the dimension of the Toolbox framework dedicated to the kinds and degrees of conceptual change that different models recognize. Here is the expanded version of this row:

	Explication	Evolutionary	Indeterminate	Cognitive
Kinds of CC	focus on trans-framework changes, external questions	intra-population vs inter-population changes (pragmatic distinction)	no kinds or degree of changes, plastic and fluid picture of CC	fine-grained hierarchies of kinds of CC, gradual modifications of cognitive structure

As the four columns show, this dimension presents a high variability of answers between the different types of model. Different models of conceptual change organize in fact their subject-matter in very different ways, making the kinds and degree of conceptual change recognized vary a lot between one model and the other. We saw in Chapter 3 that Carnapian explication focuses explicitly on one specific kind of conceptual change, i.e. trans-framework changes where one (group of) concept(s) belonging to a given linguistic framework gets (partially) substituted by another (group of) concept(s) belonging to another linguistic framework. Explication does not divide conceptual change into multiple kinds and degrees, focusing explicitly on the type of conceptual change that is philosophically more interesting, according to Carnap's framework-based metaphilosophy. Darwinian

models of conceptual change distinguish instead between changes happening within a single conceptual population and inter-population changes. Moreover, we saw in Chapter 5 that indeterminate models of conceptual change do not divide changes into degrees or kinds, claiming instead that the inherent plasticity of the phenomenon makes the changes almost continuous in structure. Finally, cognitive models give very fine-grained hierarchies of changes where several different degrees of conceptual changes are classified in terms of how radically they modify the related cognitive structures.

The number and the nature of the degrees or kinds of conceptual change recognized are then heavily dependent on the specific type of model of conceptual change that one favors. No general trend about kinds and degrees of conceptual change can thus be discerned from our analysis, since the division in kinds or degrees (or the lack of it) appears entangled with the choice of a specific type of model. The only moral that can be drawn from these results is that all four types of model treat the distinction between kinds of conceptual change as a pragmatic choice somehow artificially imposed by the model and not as a distinction essential to the phenomenon of conceptual change itself.

The fifth row of the big chart corresponds to the dimension of the Toolbox framework dedicated to the degree of normativity that a given model of conceptual change supports:

	Explication	Evolutionary	Indeterminate	Cognitive
Normativity	Normativity, but value-laden kind of rationality (instrumental)	evolutionary kind of quasi-normativity and rationality, selection vs. drift approach	weak normativity, CC as chaotic phenomenon, heavy drift	lack of focus on normativity, possible inter-subjective measures

As we saw in the preceding chapters, all the four types of model of conceptual change support some degree of normativity in their analysis of historical episodes of change, albeit of a different kind. Carnapian explication is inherently connected with a heavily value-laden kind of normativity, typical of the realm of external questions in Carnap's metaphilosophy. The normativity of Darwinian models of conceptual change is instead analogous to the quasi-normativity at work in natural selection. As we saw in Chapter 4, the important distinction between cases of selection and cases of drift allows us to give restricted and localized quasi-normative evaluations of episodes of conceptual change. A small space for normativity is also left by indeterminate models although, as Wilson repeatedly stressed in his work, the space for normative judgments is quite small in the chaotic dynamics by which concepts change. Few cognitive models of conceptual change dwell into normative judgments, although the few ones that do that (e.g. Thagard's one) present inter-subjective measures of trans-theoretical coherence that make normative judgments quite inter-subjective.

The general moral for normativity in conceptual change, according to the analysis of these four types of model, is that conceptual change is a weakly normative phenomenon,

dependent on a lot of context-dependency and value-ladenness, as well as subject to a great dose of drift and sub-optimality.

The sixth row of the big chart depicts the dimension of the Toolbox framework dedicated to the effectiveness of the normative judgments supported by a given model of conceptual change:

	Explication	Evolutionary	Indeterminate	Cognitive
Effectiveness of Normative Judgment	Pragmatic matter (instrumental rationality)	Quasi-Normativity heavily dependent on historical reconstruction, internal history reconstruction as a test for normative claims	very weak judgment, lack of knowledge and transparency	naturalized normativity, science of science

In this row, consistently with the related previous row dedicated to the normative judgments allowed by a given model, we can see a general trend towards value-laden normative judgments. The weak kind of normativity that (to a different degree) all four types of model support can only justify the kind of rationality dependent on previous agreement on shared goals and values. This instrumental kind of rationality is the kind of effectiveness that normative judgments of these models of conceptual change explicitly or implicitly favor. We saw in Chapter 3 how Carnapian explication is explicitly designed for being the vessel of instrumental rationality judgments, since Carnap's metaphilosophy allows value-free judgments only within a single linguistic framework. Darwinian models allow quasi-normative judgments only heavily dependent on very specific historical reconstructions, given the sensibility of selection and drift judgment to the specific variants and environment involved. The effectiveness of normative judgments allowed by indeterminate models is even weaker, since no such judgment can be so strongly made due to lack of knowledge and transparency in conceptual affairs that these models assume. Finally, even cognitive models that present inter-subjective normative judgments built on trans-theoreotic coherence measures, such as Thagard's one, rely explicitly on a set of shared values and goals for designing such measures, making the kind of rationality beneath these judgments approach the instrumental kind.

The seventh row of the big chart is the one corresponding to the dimension of the Toolbox framework dedicated to the assumptions and consequences for conceptual change in science that a given model has:

	Explication	Evolutionary	Indeterminate	Cognitive
CC in Science	value-laden defense of scientific progress and objectivity	EET program picture of scientific evolution (falliblist progress, no direction, pragmatic rationality)	CC in science analogous to ordinary linguistic evolution, progress exists but not easy to spot	very positive view of scientific progress, objectivity, and realism

We can see in this expanded version of the seventh row a substantial agreement between the different types of model in defending a certain degree of optimism about scientific progress and objectivity. Specifically, Carnapian explication allows, modulo previous agreements on certain scientific values and goals, a convincing defense of scientific progress and inter-subjective scientific agreement. Relative to certain sets of shared values then, the picture of science that can drawn from the ideal of explication is indeed progressive and objective. Similarly, the view of scientific progress given by Darwinian models of conceptual change is one of fallibilist inter-subjective progress relative to a given changing environment, analogous to the Darwinian picture of non-goal-directed biological evolution. Indeterminate models of conceptual change also conceptualize scientific progress as a sort of evolution, albeit analogous to the linguistic one, in which inter-subjective progress can be accomplished via the interfacial accommodations of design and drift described by Wilson. Cognitive models of conceptual change give a very strong defense of scientific progress and objectivity, although even in these models we can trace a kind of dependence of normativity on the values shared in a given scientific discipline.

Collectively, then, these four types of model of conceptual change downsize the danger that radical scientific change poses to scientific progress and objectivity, defending value-laden conceptions of these two fundamental ideals. For what concerns the debates over scientific realism, instead, we saw in the preceding chapters that the four types of model are quite neutral on the ontological import that our best scientific theories have on our picture of reality, allowing both realist and anti-realist readings of the entities that they presuppose.

The eight row of the big chart concerns instead the dimension of the Toolbox framework dedicated to the assumptions and consequences that a given model of conceptual change has with respect to the phenomenon of conceptual change in philosophy:

	Explication	Evolutionary	Indeterminate	Cognitive
CC in Philosophy	CC central to philosophy, philosophical activity as explication	evolutionary epistemology picture of philosophical activity	Ubiquity of CC and linguistic practices evolution, strong anti-essentialism	Ubiquity of CC and cognitive evolution in all human practices

Here, all the models that we have seen so far agree on the existence and the significance of philosophical conceptual change. We saw, in fact, how Carnapian explication and the related ideal of explication conceptualize the whole philosophical activity as an engineering-like task centered around the repeated explication of philosophical concepts. Darwinian models also stress the fact that, as well as in science, philosophical concepts indeed change, producing the kind of evolution of philosophical theories that evolutionary epistemology analyzes. Moreover, both indeterminate and cognitive models of conceptual change stress the ubiquity of conceptual change in every intellectual human activity, due to the inherent flexibility of (respectively) linguistic and cognitive evolution.

Thus, despite the traditional uneasiness of analytic philosophy in recognizing and conceptualizing philosophical conceptual change, the analysis of several models of concep-

tual change here presented shows the need of putting this phenomenon at the center of metaphilosophical reflections. Despite the differences in the specific extent and the specific causes of this phenomenon between different models of conceptual change, all these models agree on its significance for philosophical activity. As such, it is surprising that philosophical conceptual change has so far received very small attention from philosophers. The different models of conceptual change analyzed in this work constitute a vast array of tools for attacking the problem of philosophical conceptual change, but there is a strong lack of significant case studies. This lack of a shared repertoire of historical episodes of philosophical conceptual change constitutes the main obstacle to achieve an adequate depiction and understanding of this phenomenon.

Finally, the ninth row of the big chart corresponds to the dimension of the Toolbox framework dedicated to the metaphilosophical background of a given type of model of conceptual change:

	Explication	Evolutionary	Indeterminate	Cognitive
Meta-philosophy	Philosophical activity as a kind of engineering, explication ideal	naturalized epistemology as ideal for philosophical activity, centrality of evolutionary considerations	Centrality of linguistic drift (middle ground between analysis and engineering), centrality of linguistic considerations	Contemporary cognitive science picture of human cognition (conceptual knowledge, default/unconscious reasoning, centrality of non-linguistic phenomena)

We can see that, for what concerns the metaphilosophical background, each type of model of conceptual change has a very distinctive conception of philosophical activity that contrasts the ones behind the other models. I talked at length in Chapter 3 about the very specific metaphilosophical standpoint of Carnap's late philosophy of which the procedure of explication is the center. Philosophical activity is understood by the late Carnap as a pluralist engineering-like activity centered around the never ending adjustment of concepts and theories to the intellectual goals shared by philosophical communities. The metaphilosophical conception of evolutionary models of conceptual change is instead characterized by the ideal of truly naturalized philosophy that takes evolutionary considerations as the starting epistemological point of any philosophical reflection. Indeterminate models of conceptual change put instead at the center of their metaphilosophical conception linguistic considerations and a heavy focus on the in-depth linguistic practice at work behind our concepts and our theories, seeing as a primarily philosophical activity the diagnostics of linguistic vices. Finally, cognitive models of conceptual change depict a vision of philosophical activity centered around the large-scale image of human cognition that contemporary cognitive science gives us, i.e. a kind of naturalized philosophy centered around the dynamics of our cognitive architectures.

Despite the obvious differences between these four contrasting pictures of philosophical activity, we can see two main general morals that our analysis supports: a distrust

of traditional armchair philosophical analysis and a strong focus on extended case studies and related historical reconstructions. Traditional intuition-driven kinds of conceptual analyses are in fact not given a prominent place nor a strong justification in any of the four metaphilosophical pictures that we saw above. All these models build instead a convincing case for a move towards a more naturalized form of philosophical activity, where philosophical issues are seen in the light of their many components, such as related linguistic practice, underlying cognitive mechanism, related design imperatives, and intertwined evolutionary considerations. On the bright side, each type of model of conceptual change makes a strong use of case studies and related historical reconstruction of actual episodes of conceptual change. This emphasis on the successful reconstruction of various case studies as evidence for the goodness of a given theoretical framework can be seen as a more general plead for the development and use of more historically minded methodologies in analytic philosophy.

We have then seen a detailed comparison of the four types of model of conceptual change along the nine dimensions of the Toolbox framework. For each dimension, I stressed the differences between the models, as well as the commonalities in their approaches. The general conception of conceptual change that we got from the foregoing analysis is the following:

> Conceptual change is a multi-faceted phenomenon, centered around the dynamics of groups of concepts. Concepts seem best reconstructed as plastic and inter-subjective entities equipped with a non-trivial internal structure and subject to a certain degree of localized holism. This conceptual dynamic can be judged from a weakly normative perspective, bound to be dependent on shared values and goals. Conceptual change is then best understood as a ubiquitous phenomenon underlying all of our intellectual activities, from science to ordinary language. As such it does not pose particular problems to viable notions of scientific progress, objectivity, and realism. At the same time, the phenomenon of conceptual change must be taken into consideration by all our concept-driven intellectual activities, including philosophical and metaphilosophical reflections. An adequate understanding of the dynamics of philosophical concepts is in fact a prerequisite for analytic philosophy to develop a realistic and non-idealized picture of itself and its activities.

In connection to this conception of conceptual change, some consequences of general philosophical interest can be drawn. I will briefly mention five large-scale consequence that the present analysis arguably justifies:

- The centrality of concepts: concepts appear absolutely necessary blocks of any realistic depiction of human higher cognitive abilities and intellectual practices. In science, in ordinary language, and in philosophy, concepts play a major epistemological and semantical role. As such, concepts cannot be replaced by other kinds of entities nor they can be eliminated away.

- The ubiquity of values: apart from concepts, the other entities that have been a constant presence throughout all the present analysis are values. Theoretical and non-theoretical values are absolutely crucial for any normative judgment on how scientific concepts and theories change, as well as for any viable notion of scientific progress and objectivity.

- The inadequacy of the traditional self-image of analytic philosophy: the ubiquity of values and conceptual change in all our intellectual activities should make everyone suspicious about the traditional self-image that analytic philosophy had. The power and the finality of conceptual analysis, the value-free ideal of philosophical and scientific rationality, and the lack of historical (self-)reflections are relics of a more naive metaphilosophical age.

- The necessity of plural methodologies: the analysis of different models of conceptual change presented in this work showed the power of approaching a complex phenomenon (i.e. conceptual change) with a vast array of formal tools and philosophical methodologies. Such a methodological pluralism could be arguably equally useful in approaching other philosophical problems that have a similar multi-faceted and interdisciplinary character.

- The significance of historical reconstructions: a stable element in all the aforementioned methodologies that have been used in this work and in virtually all serious contribution to the debate over conceptual change in science and in philosophy has been the method of using historical reconstruction as case studies. This methodology, first and foremost championed by philosophers and historians of science, lends itself to be of paramount use in all parts of philosophy.

Finally, I will point to two general directions for future work that naturally present themselves in the light of this Conclusions chapter. First, it is of absolute importance for analytic philosophy to develop serious models of philosophical conceptual change in order to have a good grasp of the dynamics of philosophical concepts and their implications for philosophical activity *tout court*. Secondly, it would be great to have fine-grained formal models of values dynamics, in both science and philosophy, in order to fruitfully combine them with models of conceptual and theory change. I hope that in the future I could help filling both these gaps.

Bibliography

Alchourrón, C.E., Gärdenfors, P. and Makinson, D. (1985). "On the Logic of Theory Change: Partial Meet Contraction and Revision Functions". *The Journal of Symbolic Logic 50*(2), 510-530.

Andersen, H., Barker, P., & Chen, X. (1996): "Kuhn's mature philosophy of science and cognitive psychology". *Philosophical Psychology 9*(3), 347-363.

Andersen, H., Barker, P., & Chen, X. (2006): *The Cognitive Structure of Scientific Revolutions*. Cambridge University Press, Cambridge.

Andreas, H. (2010): "New Account of Empirical Claims in Structuralism". *Synthese 176*, 311-332.

Andreas, H. (2011): "A Structuralist Theory of Belief Revision". *Journal of Logic, Language and Information 20*(2), 205-232.

Andreas, H. (2013): "Deductive Reasoning in the Structuralist Approach". *Studia Logica 101*, 1093-1113.

Andreas, H. (2014): "Carnapian Structuralism". *Erkenntnis 79*, 1373-1391.

Andreas, H. (2020): *Dynamic Tractable Reasoning: A Modular Approach to Belief Revision*. Springer, Cham.

Awodey, S. and Carus, A.W. (2003): "Carnap versus Gödel on Syntax and Tolerance". In Parrini, P., Salmon, W.C., Salmon, M.H. (Eds.), *Logical Empiricism: Historical and Contemporary Perspectives*, University of Pittsburgh Press, Pittsburgh, 57-64.

Awodey, S. and Carus, A.W. (2007): "Carnap's Dream: Gödel, Wittgenstein, and Logical Syntax". *Synthese 159*(1), 23-45.

Awodey, S. and Carus, A.W. (2009): "From Wittgenstein's Prison to the Boundless Ocean: Carnap's Dream of Logical Syntax". In Wagner, P. (Ed.), *Carnap's Logical Syntax of Language*, Palgrave Macmillan, London, 79-108.

Balzer, W. and Moulines, C.U. (Eds.) (1996): *Structuralist Theory of Science: Focal Issues, New Results*. Walter De Gruyter, Berlin.

Balzer, W., Moulines, C.U., and Sneed, J. (1987): *An Architectonic for Science, The structuralist Program*. D. Reidel Publishing Company, Dordrecht.

Barsalou, L. (1985): "Ideals, Central Tendency, and Frequency of Instantiation as Determinants of Graded Structure". *Journal of Experimental Psychology: Learning, Memory, and Cognition 11*(4), 629-654.

Barsalou, L. (1987): "The Instability of Graded Structure: Implications for the Nature of Concepts". In Neisser, U. (Ed.), *Concepts and Conceptual Development: ecological and intellectual factors in categorization*, Cambridge University Press, Cambridge, 101-140.

Barsalou, L. (1992): "Frames, concepts, and conceptual fields". In Lehrer, A. and Feder Kittay, E. (Eds.), *Frames, Fields, and Contrasts*, Routledge, London, 21-74.

Barsalou, L. (1999): "Perceptual Symbol System". *Behavioral and Brain Sciences 22*, 577-609.

Barsalou, L. and Hale, C. (1993): "Components of conceptual representation: from feature lists to recursive frames". In Van Mechelen, I. et. al. (Eds.), *Categories and Concepts: theoretical views and inductive data analysis*, Academic Press, Cambridge (MA), 97-144.

Batterman, R. (2001): *The Devil in the Details: Asymptotic Reasoning in Explanation, Reduction, and Emergence*. Oxford University Press, Oxford.

Beaney, M. (2007): "The Analytic Turn in Early Twentieth-Century Philosophy". In Beaney, M. (Ed.), *The Analytic Turn. Essays in Early Analytic Philosophy and Phenomenology*, Routledge, New York and London, 1-30.

Beaney, M. (2013): "Analytic Philosophy and History of Philosophy: The Development of the Idea of Rational Reconstruction". In Reck, E. (Ed.), *The Historical Turn in Analytic Philosophy*, Pagrave Macmillan, London, 231-260.

Beaney, M. (2021): "Analysis". *Standford Encyclopedia of Philosophy* (Spring 2021 Edition), Edward N. Zalta (Ed.), URL=<https://plato.stanford.edu/archives/spr2021/entries/analysis/>.

Beatty, J. (1992): "Random Drift". In Keller, E.F. and Lloyd, E.A. (Eds.), *Keywords in Evolutionary Biology*, Harvard University Press, Cambridge (MA), 273-281.

Bechberger, L. and Kühnbeger, K.U. (2017): "A Thorough Formalization of Conceptual Spaces". In Kern-Isberner, G. et. al. (Eds.), *KI 2017: Advances in Artificial Intelligence*, Springer, Cham, 58-71.

Binmore, K. and Samuelson, L. (1999): "Evolutionary Drift and Equilibrium Selection". *The Review of Economic Studies 66*(2), 363-393.

Bix, B. (1991): "H.L.A. Hart and the 'Open-Texture' of Language". *Law and Philosophy 10*, 51-72.

Bix, B. (2019): "Waismann, Wittgenstein, Hart, and Beyond: The Developing Idea of 'Open-Texture' of Language and Law". In Makovec, D. and Shapiro, S. (Eds.), *Friedrich Waismann: The Open Texture of Analytic Philosophy*, Palgrave Macmillan, London, 245-260.

Bloor, D. (1976): *Knowledge and Social Imagery*. Chicago University Press, Chicago.

Borges, J.L. (1998): *Collected Fictions*. English Translation, Penguin Books, London.

Boyd, R. and Richerson, P.J. (1985): *Culture and the Evolutionary Process*. University of Chicago Press, Chicago.

Bozdag, S. and De Benedetto, M. (2022): "Taking Up Thagard's Challenge: A Formal Model of Conceptual Revision". *Journal of Philosophical Logic*, https://doi.org/10.1007/s10992-021-09650-4.

Bradie, M. (1986): "Assessing Evolutionary Epistemology". *Biology & Philosophy 1*, 401-459.

Bradie, M. (1994): "Epistemology from an Evolutionary Point of View". In Sober, E. (Ed.), *Conceptual Issues in Evolutionary Biology* (second edition), MIT Press, Cambridge (MA), 453-475.

Brandom, R. (1994): *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Harvard University Press, Cambridge (MA).

Brandom, R. (2000): *Articulating Reasons: An Introduction to Inferentialism*. Harvard University Press, Cambridge (MA).

Brandom, R. (2010): "Platforms, Patchworks, and Parking Garages: Wilson's Account of Conceptual Fine-Structure in Wandering Significance". *Philosophy and Phenomenological Research 82*(1), 183-201.

Brandon, R.N. (1978): "Adaptation and Evolutionary Theory". *Studies in History and Philosophy of Science Part A 9*(3), 181-206.

Brandon, R.N. (2006): "The Principle of Drift: Biology's First Law". *The Journal of Philosophy 103*(7), 319-335.

Bridson, M. (2008): "Geometric and Combinatorial Group Theory". In Gowers, T., Barrow-Green, J., Leader, I. (Eds.), *The Princeton Companion to Mathematics*, Princeton University Press, Princeton (NJ), 431-448.

Brown, H.I. (2000): *Conceptual Systems*. Routledge, London.

Brun, G. (2016): "Explication as a Method of Conceptual Re-Engineering". *Erkenntnis 81* (6), 1211-1241.

Brun, G. (2020): "Conceptual Re-Engineering: From Explication to Reflective Equilibrium". *Synthese 197*(3), 925-954.

Burgess, A. and Plunkett, D. (2013a): "Conceptual Ethics I". *Philosophy Compass 8*(12), 1091-1101.

Burgess, A. and Plunkett, D. (2013b): "Conceptual Ethics II". *Philosophy Compass 8*(12), 1101-1110.

Campbell, D.T. (1956): "Perception as Substitute Trial and Error". *Psychological Review 63*, 330-342.

Campbell, D.T. (1959): "Methodological Suggestions from a Comparative Psychology of Knowledge Processes". *Inquiry 2*, 152-182.

Campbell, D.T. (1960): "Blind variation and selective retentions in creative thought as in other knowledge processes". *Psychological Review 67*(6), 380-400.

Campbell, D.T. (1966): "Pattern Matching as an Essential and Distal Knowing". In Hammond, K.R. (Ed.), *The Psychology of Egon Brunswik*, Holt, Rinehart and Winston, New York, 81-106.

Campbell, D.T. (1974a): "Evolutionary Epistemology". In Schilpp, P. (Ed.), *The Philosophy of Karl R. Popper*, Open Court, LaSalle, 412-463.

Campbell, D.T. (1974b): "Unjustified Variation and Selective Retention in Scientific Discovery". In Ayala, F.J. and Dobzhansky, T. (Eds.), *Studies in the Philosophy of Biology: Reduction and Related Problems*, MacMillan, London, 139-161.

Campbell, D.T. (1987): "Selection Theory and the Sociology of Scientific Validity". In Calleabut, W. and Pinxten, R. (Eds.), *Evolutionary Epistemology: A Multiparadigm Program*, D. Reidel, Dordrecht, 139-158.

Campbell, D.T. (1988): "A General 'Selection Theory', as Implemented in Biological Evolution and in Social Belief-Transmission-with-Modification in Science". *Biology and Philosophy 3*, 171-177.

Campbell, D.T. (1997): "From Evolutionary Epistemology Via Selection Theory to a Sociology of Scientific Validity". *Evolution and Cognition 3*, 5-38.

Cappelen, H. (2018): *Fixing Language: An Essay on Conceptual Engineering*. Oxford University Press, Oxford.

Cappelen, H. (2020): "Conceptual Engineering: The Master Argument". In Cappelen, H., Plunkett, D., and Burgess, A. (Eds.), *Conceptual Engineering and Conceptual Ethics*, Oxford University Press, Oxford, 132-151.

Cappelen, H., Plunkett, D., and Burgess, A. (Eds.) (2020): *Conceptual Engineering and Conceptual Ethics*. Oxford University Press, New York.

Carey, S. (1985): *Conceptual Change in Childhood*. MIT Press, Cambridge (MA).

Carey, S. (2009): *The Origin of Concepts*. Oxford University Press, Oxford.

Carnap, R. (1928a): *Der logische Aufbau der Welt*. Weltkreis, Berlin.

Carnap, R. (1928b): *Scheinprobleme in der Philosophie*. Weltkreis, Berlin.

Carnap, R. (1934): *Logische Syntax der Sprache*. Springer, Vienna.

Carnap, R. (1947): *Meaning and Necessity*. University of Chicago Press, Chicago.

Carnap, R. (1950a): "Empiricism, Semantics, and Ontology". *Revue Internationale de Philosophie 4* (11), 20-40.

Carnap, R. (1950b): *Logical Foundations of Probability*. University of Chicago Press, Chicago.

Carnap, R. (1952): *The Continuum of Inductive Methods*. University of Chicago Press, Chicago.

Carnap, R. (1956): "The Methodological Character of Theoretical Concepts". In Feigl, H. and Scriven, M. (Eds.), *The Foundations of Science and the Concepts of Psychology and Psychoanalysis*, University of Minnesota Press, Minneapolis, 38-76.

Carnap, R. (1961): "On the use of Hilbert's ϵ-Operator in Scientific Theories". In Bar-Hillel, Y. et al. (Eds.), *Essays in the Foundation of Mathematics*, Magnes Press, Jerusalem, 154-164.

Carnap, R. (1963a): "Carnap's Intellectual Autobiography". In Schilpp, P.A. (Ed.), *The Philosophy of Rudolf Carnap*, Open Court, LaSalle, 3-84.

Carnap, R. (1963b): "The Philosopher Replies". In Schilpp, P.A. (Ed.), *The Philosophy of Rudolf Carnap*, Open Court, LaSalle, 859-1013.

Carnap, R. (1966): *Philosophical Foundations of Science*. Basic Books, New York; reprinted as *An Introduction to the Philosophy of Science*, 1972.

Carnap, R. and Jeffrey, R.C. (1971): *Studies in Inductive Logic and Probability*. Volume 1, University of California Press, Berkeley.

Carsten, K. and Awodey, S. (Eds.) (2004): *Carnap Brought Home – The View from Jena*. Open Court, LaSalle.

Cartwright, N. (1983): *How the Laws of Physics Lie*. Oxford University Press, Oxford.

Carus, A.W. (2007): *Carnap and Twentieth-Century Thought: Explication as Enlightenment*. Cambridge University Press, Cambridge.

Carus, A.W. (2012a): "Engineers and Drifters; The Ideal of Explication and Its Critics". In Wagner, P. (Ed.), *Carnap's Ideal of Explication and Naturalism*, Palgrave Macmillan, London, 225-239.

Carus, A.W. (2012b): "From Analysis to Explication". *Unpublished Manuscript*, October 2012.

Carus, A.W. (2017): "Carnapian Rationality". *Synthese 194*, 163-184.

Cavalli-Sforza, L.L. and Feldman, M.W. (1981): *Cultural Transmission and Evolution: A Quantitative Approach*. Princeton University Press, Princeton (NJ).

Chakravartty, A. (2017): "Scientific Realism". *Standford Encyclopedia of Philosophy* (Summer 2017 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2017/entries/scientific-realism/>.

Chalmers, D.J. (2020): "What is conceptual engineering and what should it be?". *Inquiry*, DOI: 10.1080/0020174X.2020.1817141.

Chang, H. (2004): *Inventing Temperature: Measurement and Scientific Progress*. Oxford University Press, New York.

Chang, H. (2012): *Is water H2O? Evidence, realism and pluralism*. Springer, New York.

Charbonneau, M. (2007): "Populations without Reproduction". *Philosophy of Science 81*, 727-740.

Church, A. (1936): "An Unsolvable Problem of Elementary Number Theory". *American journal of mathematics 58*, 345-363.

Clark, A. (1993): *Associative Engines: Connectionism, Concepts, and Representational Change*. MIT Press, Cambridge (MA).

Coffa, A. (1991): *The Semantic Tradition from Kant to Carnap: To the Vienna Station*. Cambridge University Press, Cambridge.

Cohen, L.J. (1973): "Is the Progress of Science Evolutionary?". *British Journal for the Philosophy of Science 24*, 41-61.

Conant, J.B. (1950): *The overthrow of the phlogiston theory: The chemical revolution of 1775-1789*. Harvard University Press, Cambridge (MA).

Copeland, B.J. and Shagrir, O. (2013): "Turing versus Gödel on Computability and the Mind". In Copeland B.J., Posy C., Shagrir O. (Eds.), *Computability: Gödel, Turing, Church, and Beyond*, MIT Press, Cambridge (MA), 1-33.

Corfield, D. (2002): "Argumentation and the Mathematical Process". In Kampis, G., Kvasz, L., Stölzner, M. (Eds.), *Appraising Lakatos: Mathematics, Methodology and the Man*, Kluwer Academic Publishers, Dordrecht, 115-138.

Corfield, D. (2003): *Towards a Philosophy of Real Mathematics*. Cambridge University Press, Cambridge.

Craig, E. (1990): *Knowledge and the State of Nature: An Essay in Conceptual Synthesis*. Oxford University Press, Oxford.

Crane, T. (2015): *The Mechanical Mind*. Third Edition, Routledge, London.

Creath, R. (1991): "Every Dogma has its Day". *Erkenntnis 35*(1-3), 347-389.

Creath, R. (1994): "Functionalist Theories of Meaning and the Defense of Analyticity". In Salman, W. and Wolters, G. (Eds.), *Logic, Language and the Structure of Scientific Theories*, University of Pittsburgh Press, Pittsburgh.

Creath, R. (2009): "The Gentle Strength of Tolerance: The Logical Syntax of Language and Carnap's Philosophical Programme". In Wagner, P. (Ed.), *Carnap's Logical Syntax of Language*, Palgrave Macmillan, London, 203-216.

Creath, R. (2012): "Before Explication". In Wagner, P. (Ed.), *Carnap's Ideal of Explication and Naturalism*, Palgrave Macmillan, London, 161-174.

Cresto, E. (2008): "A Model for Structural Changes of Belief ". *Studia Logica 88*, 431-451.

Cziko, G. (1995): *Without Miracles: Universal Selection Theory and the Second Darwinian Revolution*. MIT Press, Cambridge (MA).

Da Costa, N.C.A. and French, S. (2003): *Science and Partial Truth: A Unitary Approach to Models and Scientific Reasoning*. Oxford University Press, Oxford.

Dawkins, R. (1976): *The Selfish Gene*. Oxford University Press, New York.

De Benedetto, M. (2020): "Explicating 'Explication' via Conceptual Spaces". *Erkenntnis*, https://doi.org/10.1007/s10670-020-00221-8.

De Benedetto, M. (2021a): "Explication as a Three-Step Procedure: the case of the Church-Turing Thesis". *European Journal for Philosophy of Science 11*, https://doi.org/10.1007/s13194-020-00337-2.

De Benedetto, M. (2021b): "Taming conceptual wanderings: Wilson-Structuralism". *Synthese*, https://doi.org/10.1007/s11229-021-03374-3.

De Clercq, R. and Horsten, L. (2004): "Perceptual Indiscriminability: In Defence of Wright's Proof" *The Philosophical Quarterly 54*, 439-444.

Decock, L. (2021): "Conceptual Change and Conceptual Engineering: The Case of Colour Concepts". *Inquiry 64* (1-2), 168-185.

Decock, L., Dietz, R. and Douven, I. (2013): "Modelling Comparative Concepts in Conceptual Spaces". In Motomura, Y., Alastair Butler, A., and Bekki, D. (Eds.), *Lecture Notes in Computer Science 7856*, Springer, Heidelberg, 69-86.

Decock, L. and Douven, I. (2014): "What Is Graded Membership?". *Noûs* 48, 653-682.

Dennett, D.C. (2006): "Higher-Order Truths about Chmess". *Topoi 25*(1-2): 39-41.

Dershowitz, N. and Gurevich, Y. (2008): "A Natural Axiomatization of Computability and Proof of Church's Thesis". *Bulletin of Symbolic Logic 14*, 299-350.

Deutsch, M. (2020): "Speaker's Reference, Stipulation, and a Dilemma for Conceptual Engineers". *Philosophical Studies 177*(12), 3935-3957.

Dietz, R. (2013): "Comparative Concepts". *Synthese 190*(1), 139-170.

Douglas, H. (2009): *Science, Policy, and the Value-Free Ideal*. University of Pittsburgh Press, Pittsburgh.

Douven, I., Decock, L., Dietz, R., Égré, P. (2013): "Vagueness: A Conceptual Spaces Approach". *Journal of Philosophical Logic 42*(1), 137-160.

Douven, I. and Gärdenfors, P. (2019): "What Are Natural Concepts? A Design Perspective". *Mind and Language 3*, 313-334.

Dummett, M. (1973): *Frege: Philosophy of Language*. Duckworth, London.

Dummett, M. (1991): *The Logical Basis of Metaphysics*. Duckworth, London.

Dummett, M. (1993): *Seas of Language*. Oxford University Press, Oxford.

Dutilh Novaes, C. (2012): *Formal Languages in Logic: a Philosophical and Cognitive Analysis*. Cambridge University Press, Cambridge.

Dutilh Novaes, C. and Reck, E. (2017): "Carnapian Explication, Formalisms as Cognitive Tools, and the Paradox of Adequate Formalization". *Synthese 194*(1), 195-215.

Égré, P., Ripley, D. and Verheyen, S. (2019): "The Sorites Paradox in Psychology". In Oms, S. and Zardini, E. (Eds.), *The Sorites Paradox*, Cambridge University Press, Cambridge, 263-286.

Eklund, M. (2020): "Variance Theses in Ontology and Metaethics". In Cappelen, H., Plunkett, D., and Burgess, A. (Eds.), *Conceptual Engineering and Conceptual Ethics*, Oxford University Press, Oxford, 187-204.

Enç, B. (1976): "Reference of Theoretical Terms". *Noûs 10*, 261-282.

Enqvist, S. (2010): "A Structuralist Framework for the Logic of Theory Change". In Olsson, E.J. and Enqvist, S. (Eds.), *Belief Revision meets Philosophy of Science*, Vol. 21, Springer, Heidelberg, 105-136.

Etchemendy(1990): *The Concept of Logical Consequence*. Harvard University Press, Cambridge (MA).

Fadda, A. (2020): "Population Thinking in Epistemic Evolution: Bridging Cultural Evolution and the Philosophy of Science". *Journal for General Philosophy of Science 52*(2), 351-369.

Feferman, S. (1978): "The Logic of Mathematical Discovery vs. the Logical Structure of Mathematics". *PSA 1978: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 309-327.

Ferreirós, J. (2015): *Mathematical Knowledge and the Interplay of Practices*. Princeton University Press, Princeton (NJ).

Feyerabend, P. (1962): "Explanation, Reduction and Empiricism". In Feigl, H. and Maxwell, G. (Eds.), *Scientific Explanation, Space, and Time*, Minnesota Studies in the Philosophy of Science (Volume III), University of Minnesota Press, Minneapolis, 28-97.

Fine, A. (1978): "Conceptual Change in Mathematics and Science: Lakatos' Stretching Refined". *PSA 1978: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 328-341.

Fischer, E. (2019): "Linguistic Legislation and Psycholinguistic Experiments: Redeveloping Waismann's Approach". In Makovec, D. and Shapiro, S. (Eds.), *Friedrich Waismann: The Open Texture of Analytic Philosophy*, Palgrave Macmillan, London, 211-240.

Floyd, J. (2012): "Wittgenstein, Carnap, and Turing: Contrasting Notions of Analysis". In Wagner, P. (Ed.),*Carnap's Ideal of Explication and Naturalism*, Palgrave Macmillan, London, 34-46.

Fodor, J. (1975): *The Language of Thought*. Harvard University Press, Cambridge (MA).

Fodor, J. (1990): "Information and Representation". In Hanson, P. (Ed.), *Information, Language, and Cognition*, University of British Columbia Press, Vancouver, 175-190.

Fodor, J. (1998): *Concepts: Where Cognitive Science Went Wrong*. Oxford University Press, New York.

Fodor, J. (2008): *LOT 2: The Language of Thought Revisited*. Oxford University Press, New York.

Fodor, J., Garrett, M.F., Walker, E.C.T., and Parkes, C.H. (1980): "Against Definitions". *Cognition 8*(3), 263-367.

Fodor, J. and Lepore, E. (1996): "The Red Herring and the Pet Fish: Why Concepts Still Can't Be Prototypes". *Cognition 58*, 253-270.

Fracchia, J. and Lewontin, R.C. (1999): "Does Culture Evolve?". *History and Theory 8*, 52-78.

Frege, G. (1884): *Die Grundlagen der Arithmetik: Eine logisch-mathematische Untersuchung über den Begriff der Zahl*. Wilhelm Koebner, Breslau.

Frege, G. (1891): "Funktion und Begriff". Vortrag, gehalten in der Sitzung vom 9. Januar 1891 der Jenaischen Gesellschaft für Medizin und Naturwissenschaft, *Hermann Pohle*, Jena.

Frege, G. (1892a): "Über Sinn und Bedeutung". *Zeitschrift für Philosophie und philosophische Kritik 100*, 25-50.

Frege, G. (1892b): "Über Begriff und Gegenstand". *Vierteljahresschrift für wissenschaftliche Philosophie 16*, 192-205.

French, S. (2017): "Identity Conditions, Idealisations and Isomorphisms: a Defence of the Semantic Approach". *Synthese*, doi:10.1007/s11229-017-1564-z.

French, S. and Ladyman, J. (1999): "Reinflating the Semantic Approach". *International Studies in the Philosophy of Science 13*(2), 103-121.

Friedman, M. (1999): *Reconsidering Logical Positivism*. Cambridge University Press, New York.

Friedman, M. (2000): *A Parting of the Ways: Carnap, Cassirer, and Heidegger*. Open Court, New York.

Friedman, M. (2001): *Dynamics of Reason: The 1999 Kant Lectures at Standford University*. CSLI Publications, Standford (CA).

Friedman, M. (2010): "Logic, Mathematical Science, and Twentieth Century Philosophy: Mark Wilson and the Analytic Tradition". *Noûs 44*(3), 530-544.

Friedman, M. (2012): "Rational Reconstruction, Explication, and the Rejection of Metaphysics". In Wagner, P. (Ed.), *Carnap's Ideal of Explication and Naturalism*, Palgrave Macmillan, London, 190-204

Friedman, M. and Creath, R. (Eds.). (2007): *The Cambridge Companion to Carnap*. Cambridge University Press, Cambridge.

Frigg, R. and Hartmann, S. (2020): "Models in Science". *Standford Encyclopedia of Philosophy*(Spring 2020 Edition), Edward N. Zalta (Ed.), URL = <https://plato.stanford.edu/archives/spr2020/entries/models-science/>.

Galison, P. (1987): *How Experiments End*. University of Chicago Press, Chicago.

Gamerschlag, T. et. al. (Eds.) (1999): *Frames and Concept Types: Applications in Language and Philosophy*. Studies in Linguistic and Philosophy (94), Springer Science & Business Media, Berlin.

Gandy, R. O. (1980), "Church's Thesis and Principles for Mechanisms". In Barwise, J., Keisler H. K., Kunen K. (Eds.), *The Kleene Symposium*, North-Holland, Amsterdam, 123-145.

Gandy, R. O. (1988): "The Confluence of Ideas in 1936". In R. Herken (Ed.), *The Universal Turing Machine. A half-century survey*, Oxford University Press, Oxford, 55-111.

Gärdenfors, P. (2000): *Conceptual Spaces:The Geometry of Thought*. MIT Press, Cambridge (MA).

Gärdenfors, P. (2014): *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. MIT Press, Cambridge (MA).

Gärdenfors, P. (2019): "Convexity Is an Empirical Law in the Theory of Conceptual Spaces: Reply to Hernández-Conde". In Kaipainen, M. et al. (Eds.), *Conceptual Spaces: Elaborations and Applications*, Springer, Heidelberg, 77-80.

Gärdenfors, P. and Rott, H. (1995): "Belief Revision". In Gabbay, D., Hogger, C.J., and Robinson, J.A. (Eds.), *Handbook of logic in artificial intelligence and logic programming (Vol. 4): epistemic and temporal reasoning*, Oxford University Press, New York, 35-132.

Gärdenfors, P. and Zenker, F. (2011): "Using Conceptual Spaces to Model the Dynamics of Empirical Theories". In Olsson, E. and Enqvist, S. (eds.), *Belief Revision Meets Philosophy of Science*, Springer, Berlin, 137-153.

Gärdenfors, P. and Zenker, F. (2013): "Theory Change as Dimensional Change: Conceptual Spaces Applied to the Dynamics of Empirical Theories". *Synthese 190*(6), 1039-1058.

Gelfert, A. (2011): "Steps to an Ecology of Knowledge: Continuity and Change in the Genealogy of Knowledge". *Episteme 8*(1), 67-82.

Gentzen, G. (1934/35): "Untersuchungen über das logische Schließen". *Mathematische Zeitschrift 39*, 176-210, 405-431.

Gettier, E. (1963): "Is Justified True Belief Knowledge?". *Analysis 23*(6), 121-123.

Giere, R. (1988): *Explaining Science: A Cognitive Approach*. University of Chicago Press, Chicago.

Giere, R. (1999): *Science Without Laws*. University of Chicago Press, Chicago.

Gillies, D. (Ed.) (1992): *Revolutions in Mathematics*. Oxford University Press.

Gödel, K. (1934): "On Undecidable Propositions of Formal Mathematical Systems". Mimeographed lecture notes by S. C. Kleene and J. B. Rosser, reprinted with revisions in Davis, M. (Ed.), *The Undecidable: Basic papers on Undecidable Propositions, Unsolvable Problems And Computable Functions*, Raven Press, Hewlett, New York, 39-74.

Gödel, K. (1972): "A Philosophical Error in Turing's Work". Third of the three notes contained in Feferman S. et al. (Eds.), *Gödel, K.: Collected Works Volume II* (1990), Oxford University Press, Oxford, 305-306.

Godfrey-Smith, P. (2007): "Conditions for Evolution by Natural Selection". *Journal of Philosophy 104*, 489-516.

Godfrey-Smith, P. (2009): *Darwinian Populations and Natural Selection*. Oxford University Press, Oxford.

Godfrey-Smith, P. (2012): "Darwinism and Cultural Change". *Philosophical Transactions of the Royal Society B 367*, 2160-2170.

Goldfarb, W.D. (1996): "The Philosophy of Mathematics in Early Positivism". In Giere, R. N. and Richardson, A.W. (Eds.), *The Origins of Logical Empiricism* (Minnesota Studies in the Philosophy of Science, 16), University of Minnesota Press, Minneapolis, 213-230.

Goldman, A.I. and Whitcomb, D. (Eds.) (2011): *Social Epistemology: Essential Readings*. Oxford University Press, New York.

Golinski, J. (1992): *Science as Public Culture: Chemistry and Enlightenment in Britain 1760-1820*. Cambridge University Press, Cambridge.

Gontier, N., Bendegem, J.P. and Aerts, D. (Eds.) (2006): *Evolutionary Epistemology, Language and Culture: a Non-Adaptationist, System Theoretical Approach*. Springer, Dordrecht.

Gopnik, A. and Meltzoff, A. (1997): *Words, Thoughts, and Theories*. MIT Press, Cambridge (MA).

Gurevich, Y. (1991): "Evolving algebras: an Attempt to Discover Semantics". In Rozenberg G. and Salomaa A. (Eds.), *Current Trends in Theoretical Computer Science* , World Scientific, Singapore, 266-292.

Gurevich, Y. (1995): "Evolving Algebra 1993: Lipari Guide". In Börger E. (Ed.) *Specification and Validation Methods*, Oxford University Press, Oxford, 9-36.

Gurevich, Y. (1999): "The Sequential ASM Thesis". *Bulletin of European Association for Theoretical Computer Science*, February 1999, 1-32.

Gurevich, Y. (2000): "Sequential Abstract State Machines Capture Sequential Algorithms". *ACM Transactions on Computational Logic 1*, 77-111.

Gurevich, Y. (2011): "What Is an Algorithm?". In Bielikova, M. et al. (Eds.), *SOFSEM 2012: Theory and Practice of Computer Science*, Springer, Berlin, 31-42.

Gurevich, Y. (2012): "Foundational Analyses of Computation". In Cooper S.B. et al. (Eds.), *How the World Computes. Turing Centennial Conference*, Springer, Berlin, 264-275.

Gurevich, Y. (2014): "What Is an algorithm? (Revised)". In Olszewski, A. et al. (Eds.), *Church's Thesis: Logic, Mind and Nature*, Copernicus Center Press, Krakow, 215-243.

Gurevich, Y. (2015): "Semantics-to-Syntax Analyses of Algorithms". In Sommaruga G. and Strahm T. (Eds.), *Turing's revolution*, Springer, Cham, 187-206.

Gustafsson, M. (2014): "Quine's Conception of Explication - and Why It Isn't Carnap's". In Harman, G. and Lepore, E. (Eds.), *A Companion to W.V.O. Quine*, John Wiley & Sons, Hoboken (NJ), 508-525.

Haack, S. (2003): *Defending Science – Within Reason: Between Scientism and Cynism*. Prometheus Book, Amherst.

Hacking, I. (1979): "Imre Lakatos's Philosophy of Science". *The British Journal for the Philosophy of Science 30*, 381-410.

Hacking, I. (1983): *Representing and Intervening: Introductory Topics in the Philosophy of Natural Sciences*. Cambridge University Press, Cambridge.

Hacking, I. (2002): *Historical Ontology*. Harvard University Press, Cambridge (MA).

Halbach, V. (2014): *Axiomatic Theories of Truth*. Revised Edition, Cambridge University Press, Cambridge.

Hamilton, W.R. (1843a): "Quaternions. Notebook 24.5, Entry for 16 October 1843". Reprinted In *The Mathematical Papers of Sir William Rowan Hamilton*, Volume 3: *Algebra*, Cambridge University Press, Cambridge, 103-105, 1967.

Hamilton, W.R. (1843b): "Letter to Graves on Quaternions; or on a New System of Imaginaries in Algebra (17 October 1843)". *Philosophical Magazine* 25, 489-495. Reprinted In *The Mathematical Papers of Sir William Rowan Hamilton*, Volume 3: *Algebra*, Cambridge University Press, Cambridge, 106-110, 1967.

Hamilton, W.R. (1853): "Preface to Lectures on Quaternions". In *Lectures on Quaternions*, Hodges and Smith, Dublin. Reprinted In *The Mathematical Papers of Sir William Rowan Hamilton*, Volume 3: *Algebra*, Cambridge University Press, Cambridge, 117-155, 1967.

Hampton, J.A. (1979): "Polymorphous Concepts in Semantic Memory". *Journal of Verbal Learning and Verbal Behavior 18*, 441-461.

Hampton, J.A. and Gardiner, M.M. (1983): "Measures of Internal Category Structure: A Correlational Analysis of Normative Data". *British Journal of Psychology 74*, 491-516.

Hanna, J.F. (1967): "An Explication of 'Explication'". *Philosophy of Science 35*, 28-44.

Hansson, S.O. (1999): *A Textbook of Belief Dynamics: Theory Change and Database Updating*. Applied Logic Series (11), Kluwer Academic Publishers, Dordrecht.

Hansson, S.O. (2010): "Changing the Scientific Corpus". In Olsson, E.J. and Enqvist, S. (eds.), *Belief Revision meets Philosophy of Science*, Springer, Heidelberg, 43-59.

Hardin, C. and Rosenberg, A. (1982): "In Defence of Convergent Realism". *Philosophy of Science 49*, 604-615.

Harms, W. (1997): "Reliability and Novelty: Information Gain in Multi-Level Selection Systems". *Erkenntnis 46*, 335-363.

Harms, W. (2004): *Information and meaning in evolutionary processes*. Cambridge University Press, Cambridge.

Hart, H.L.A. (2012): *The Concept of Law*. 3rd Edition, Oxford University Press, Oxford (Original edition: 1961).

Haueis, P. (2021): "A generalized patchwork approach to scientific concepts". *The British Journal for the Philosophy of Science*, doi: https://www.journals.uchicago.edu/doi/10.1086/716179.

Helfman, G., Collette, B., Facey, D., Bowen, B. (2009): *The Diversity of Fishes: Biology, Evolution, and Ecology*. 2nd edition, Wiley-Blackwell, Hoboken (NJ).

Hempel, C. (1952): *Fundamentals of Concept Formation in Empirical Science*. University of Chicago Press, Chicago.

Hempel, C. (1966): *Philosophy of Natural Science*. Prentice-Hall, Englewood Cliffs.

Hernández-Conde, J.V. (2017): "A case against convexity in conceptual spaces". *Synthese 194*(10), 4011-4037.

Hesse, M.B. (1976): "Truth and Growth of Knowledge". *PSA 1976: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 261-280.

Horsten, L. (2011): *The Tarskian Turn. Deflationism and Axiomatic Truth*. MIT Press, Cambridge (MA).

Hoyningen-Huene, P. (1993): *Reconstructing Scientific Revolutions: Thomas S. Kuhn's Philosophy of Science*. University of Chicago Press, Chicago.

Hull, D. (1976): "Are Species Really Individuals?". *Systematic Zoology 25*, 174-191.

Hull, D. (1978): "A Matter of Individuality". *Philosophy of Science 45*, 335-360.

Hull, D. (1988a): *Science as a Process: An Evolutionary Account of the Social and Conceptual Development of Science*. University of Chicago Press, Chicago.

Hull, D. (1988b): "A Mechanism and Its Metaphysics: An Evolutionary Account of the Social and Conceptual Development of Science". *Biology and Philosophy 3*, 123-155.

Hull, D. (1996): "What's Wrong with Invisible-Hand Explanations?". *Philosophy of Science 64*, 117-125.

Hume, D. (1739/1978): *A Treatise of Human Nature*. Oxford University Press, Oxford.

Ichikawa, J.J. and Steup, M. (2018): "The Analysis of Knowledge". *Standford Encyclopedia of Philosophy* (Summer 2018 Edition), Edward N. Zalta (Ed.), URL = <https://plato.stanford.edu/archives/sum2018/entries/knowledge-analysis/>.

Incurvati, L. (2020): *Conceptions of Set and the Foundations of Mathematics*. Cambridge University Press, Cambridge.

Jackendoff, R. (1992): *Languages of the Mind: Essays on Mental Representation*. MIT Press, Cambridge (MA).

Jeffrey, R.C. (1994): "Carnap's Voluntarism". In Prawitz, D., Skyrms, B. and Westerståhl, D. (Eds.), *Logic, Methodology and Philosophy of Science IX* (Studies in Logic and the Foundations of Mathematics 134), Elsevier, Amsterdam, 847–866.

Justus, J. (2012): "Carnap on Concept Determination: Methodology for Philosophy of Science". *European Journal for Philosophy of Science 2*, 161-179.

Kadvany, J. (2001): *Imre Lakatos and the Guises of Reason*. Duke University Press, Durham and London.

Kamp, H. and Partee, B. (1995): "Prototype Theory and Compositionality". *Cognition 57*, 129-191.

Katz, J. (1972): *Semantic Theory*. Harper and Row, New York.

Keefe, R. (2000): *Theories of Vagueness*. Cambridge University Press, Cambridge.

Keefe, R. and Smith, P. (1996): *Vagueness: A Reader*. MIT Press, Cambridge (MA).

Keil, F. (1989): *Concepts, Kinds, and Cognitive Development*. MIT Press, Cambridge (MA).

Kim, M.G. (2003): *Affinity, that elusive dream: A genealogy of the chemical revolution*. MIT Press, Cambridge (MA)

Kimura, M. (1983): *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

Kitcher, P. (1984): *The Nature of Mathematical Knowledge*. Oxford University Press, New York.

Kitcher, P. (1995): *The Advancement of Science: Science Without Legend, Objectivity Without Illusion*. Oxford University Press, Oxford.

Kitcher, P. (2008): "Carnap and the Caterpillar". *Philosophical Topics 36*(1), 111-127.

Kleene, S.C. (1936): "General Recursive Functions of Natural Numbers". *Mathematische Annalen 112*(5), 727-742.

Kleene, S.C. (1981): "Origins of Recursive Function Theory". *Annals of the History of Computing 3*, 52-67.

Koch, S. (2020): "Engineering What? On Concepts In Conceptual Engineering". *Synthese*, https://doi.org/10.1007/s11229-020-02868-w.

Koch, S. (2021): "The Externalist Challenge to Conceptual Engineering". *Synthese 198*, 327-348.

Kolmogorov, A.N. (1953): "On the Notion of Algorithm". *Uspekhi Mat. Nauk. 8*(4), 175-176.

Kolmogorov, A.N. and V.A. Uspenski (1958): "To the Definition of an Algorithm". *Uspehi Math. Nauk. 13*(4), 3-28.

Kornmesser, S. and Schurz, G. (2018): "Analyzing Theories in the Frame Model". *Erkenntnis*, https://doi.org/10.1007/s10670-018-0078-5.

Kovac, L. (2000): "Fundamental Principles of Cognitive Biology". *Evolution and Cognition 6*, 51-69.

Knobe, J. and Nichols, S. (Eds.) (2008): *Experimental Philosophy*. Oxford University Press, Oxford.

Kripke, S. (1972): *Naming and Necessity*. Harvard University Press, Cambridge (MA).

Kroon, S. (1985): "Theoretical Terms and the Causal View of Reference". *Australasian Journal of Philosophy 63*, 142-166.

Kuhn, T.S. (1970): *The Structure of Scientific revolution*. Second Edition, University of Chicago Press, Chicago.

Kuhn, T.S. (1974): "Second Thoughts on Paradigms". In *The Structure of Scientific Theories*, Suppe, F. (Ed.), University of Illinois Press, Champaign (IL), 459-482.

Kuhn, T.S. (1976): "Theory-Change as Structure-Change". *Erkenntnis 10*, 179-199.

Kuhn, T.S. (1977): "Objectivity, Value Judgment, and Theory Choice". In Thomas Kuhn, *The Essential Tension: Selected Studies in Scientific Tradition and Change*, Chicago University Press, Chicago, 320-339.

Kuhn, T.S. (1990): "Dubbing and redubbing: the vulnerability of rigid designation". In Savage, C.W. (Ed.), *Scientific Theories*, University of Minnesota Press, Minneapolis, 298-318.

Kuhn, T.S. (1991): "The Road since Structure". *PSA 1990: Proceedings of the Biennial Meeting of the Philosophy of Science Association*(2), 3-13.

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*, Elsevier, Amsterdam, vii-xxiii.

Kusch, M. (2015): "Scientific Pluralism and the Chemical Revolution". *Studies in the History and Philosophy of Science 49*, 69-79.

Lakatos, I. (1970): "Falsification and the Methodology of Scientific Research Programmes". In Lakatos, I. and Musgrave, A. (Eds.), *Criticism and the Growth of Knowledge*, Cambridge University Press, Cambridge, 91-195.

Lakatos, I. (1971): "History of Science and its Rational Reconstructions". *PSA 1970: Proceedings of the Biennal Meeting of the Philosophy of Science Association*, 91-108.

Lakatos, I. (1976): *Proofs and Refutations: The Logic of Mathematical Discovery*. Cambridge University Press, Cambridge.

Lakatos, I. (1978): "Cauchy and the Continuum: the Significance of Non-Standard Analysis for the History and Philosophy of Mathematics". In Worrall, J. and Zahar, E. (Eds.), *Philosophical Papers Vol. 2: Mathematics, Science and Epistemology*, 43-60.

Lakatos, I. and Musgrave, A. (Eds.) (1970): *Criticism and the Growth of Knowledge*. Cambridge University Press, Cambridge.

Lakoff, G. (1987a): *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. University of Chicago Press, Chicago.

Lakoff, G. (1987b): "Cognitive Models and Prototype Theory". In Neisser, U. (Ed.), *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization*, Cambridge University Press, Cambridge, 63-100.

Lalumera, E. (2010): "Concepts Are a Functional Kind". *Behavioral and Brain Sciences 33*, 217-218.

Larvor, B. (1998): *Lakatos: an Introduction*. Routledge, London.

Larvor, B. (2001): "What is Dialectical Philosophy of Mathematics?". *Philosophia Mathematica 9*(2), 212-229.

Laudan, L. (1981): "A Confutation of Convergent Realism". *Philosophy of Science 48*(1), 19-49.

Laudan, L. (1984a): *Science and Values*. University of California Press, Berkeley.

Laudan, L. (1984b): "Discussion: Realism Without the Real". *Philosophy of Science 51*, 156-162.

Lavers, G. (2019): "Waismann: From Wittgenstein's Tafelrunde to His Wiritings on Analyticity". In Makovec, D. and Shapiro, S. (Eds.), *Friedrich Waismann: The Open-Texture of Analytic Philosophy*, Palgrave Macmillan, London, 131-158.

Leitgeb, H. (2007): "A New Analysis of Quasianalysis". *Journal of Philosophical Logic 36*(2), 181–226.

Leitgeb, H. (2009): "On Formal and Informal Provability". In Bueno O. and Linnebo Ø (Eds.), *New Waves in Philosophy of Mathematics*, Palgrave Macmillan, London, 263-299.

Leitgeb, H. (2011): "New Life for Carnap's Aufbau?". *Synthese 180*(2), 265-299.

Leitgeb, H. (2013): "Scientific Philosophy, Mathematical Philosophy, and All That". *Metaphilosophy 44*(3), 267-275.

Leitgeb, H. (MS): "Ramsification and Semantic Indeterminacy". *Manuscript*.

Leitgeb, H. and Carus, A. (2020): "Rudolf Carnap". In Edward N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy* (Summer 2020 Edition), URL = <https://plato.stanford.edu/archives/sum2020/entries/carnap/>.

Lewis, D. (1984): "Putnam's Paradox". *Australasian Journal of Philosophy 62*, 221-236.

Lewis, M. and Lawry, J. (2016): "Hierarchical Conceptual Spaces for Concept Combination". *Artificial Intelligence 237*, 204-227.

Lewontin, R.C. (1970): "The Units of Selection". *Annual Review of Ecology and Systematics 1*, 1-18.

Lewontin, R.C. (1982): "Organism and Environment". In Plotkin, H.C. (Ed.), *Learning, Development and Culture: Essays in Evolutionary Epistemology*, Wiley & Sons, Hoboken (NJ), 151-170.

Lewontin, R.C. (1985): "Adaptation". In R. Levins and R.C. Lewontin (Eds.), *The Dialectical Biologist*, Harvard University Press, Cambridge (MA), 65-84.

Locke, J. (1690/1975): *An Essay Concerning Human Understanding*. Oxford University Press, New York.

Longino, H.E. (1990): *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press, Princeton (NJ).

Lorenz, K. (1977): *Behind the Mirror: A Search for a Natural History of Human Knowledge*. Methuen, London.

Lorenz, K. (1982): "Kant's Doctrine of the a priori in the light of Contemporary Biology". In Plotkin, H.C. (Ed.), *Learning, Development, and Culture: Essays in Evolutionary Epistemology*, Wiley & Sons, Hoboken (NJ), 121-143.

Lutz, S. (2012): "On a Straw Man in the Philosophy of Science: A Defense of the Received View". *HOPOS: The Journal of the International Society for the History of Philosophy of Science 2*(1), 77-120.

Lutz, S. (2014): "What's Right with a Syntactic Approach to Theories and Models?". *Erkenntnis 79*(8 supplement), 1475-1492.

Lutz, S. (2020): "Armchair philosophy naturalized". *Synthese 197*, 1099-1125.

MacFarlane, J. (2000): *What Does It Mean to Say That Logic is Formal?*. Dissertation, University of Pittsburgh.

Machery, E. (2009): *Doing Without Concepts*. Oxford University Press, New York.

Machery, E. (2010): "Précis of *Doing Without Concepts*". *Behavioral and Brain Sciences 33*, 195-244.

Machery, E. (2017): *Philosophy Within Its Proper Bounds*. Oxford University Press, New York.

Maher, P. (2007): "Explication Defended". *Studia Logica 86*, 331-341.

Makovec, D. and Shapiro, S. (Eds.) (2019): *Friedrich Waismann: The Open Texture of Analytic Philosophy*. Palgrave Macmillan, London.

Mancosu, P. (Ed.) (1997): *From Brouwer to Hilbert: The Debate on the Foundations of Mathematics in the 1920s*. Oxford University Press, New York.

Mancosu, P. (2009): "Measuring the Size of Infinite Collections of Natural Numbers: Was Cantor's Theory of Infinite Number Inevitable?". *Review of Symbolic Logic 2*(4), 612-646.

Margolis, E. and Laurence, S. (1999): *Concepts: Core Readings*. MIT Press, Cambridge (MA).

Margolis, E. and Laurence, S. (2015): *The Conceptual Mind: New Directions in the Study of Concepts*. MIT Press, Cambridge (MA).

Margolis, E. and Laurence, S. (2019): "Concepts". *Stanford Encyclopedia of Philosophy* (Summer 2019 Edition), Zalta, E.N. (Ed.), URL = <https://plato.stanford.edu/archives/sum2019/entries/concepts/>.

Martin, E. and Osherson, D. (1998): "Belief Revision in the Service of Scientific Discovery". *Mathematical social sciences 36*(1), 57-68.

Masterton, G., Zenker, F., and Gärdenfors, P. (2017): "Using Conceptual Spaces to Exhibit Conceptual Continuity Through Scientific Theory Change". *European Journal for Philosophy of Science 7*(1), 127-150.

Mayr, E. (1975): "Typological versus Population Thinking". In *Evolution and the Diversity of Life: Selected Essays*, Harvard University Press, Cambridge (MA), 26-29.

Mayr, E. (1983): "How to Carry Out the Adaptationist Program?". *American Naturalist 121*, 324-334.

McCloskey, M.E. and Glucksberg, S. (1978): "Natural Categories: Well Defined or Fuzzy Sets?". *Memory & Cognition 6*, 462-472.

McGuinness, B.F. (Ed.) (2011): *Friedrich Waismann - Causality and Logical Positivism*. Vienna Circle Institute Yearbook, Springer, Dordrecht.

Medin, D. and Schaffer, M. (1978): "Context Theory of Classification Learning". *Psychological Review 85*, 207-238.

Menger, K. (1943): "What is Dimension?". *The American Mathematical Monthly 50*(1), 2-7.

Mesoudi, A. (2011): *Cultural Evolution: How Darwinian evolution Can Explain Human Culture and Synthesize the Social Sciences*. University of Chicago Press, Chicago.

Millikan, R. (1998): "A Common Structure for Concepts of Individuals, Stuffs, and Real Kinds: More Mama, More Milk, and More Mouse". *Behavioral and Brain Sciences 21*, 55-65.

Millikan, R. (2009): *On Clear and Confused Ideas*. Cambridge University Press, Cambridge (MA).

Millstein, R.L. (2002): "Are Random Drift and Natural Selection Conceptually Distinct?". *Biology & Philosophy 17*(1), 33-53.

Millstein, R.L. (2017): "Genetic Drift". *Standford Encyclopedia of Philosophy* (Fall 2017 Edition), Edward N. Zalta (Ed.), URL = <https://plato.stanford.edu/archives/fall2017/entries/genetic-drift/>.

Minsky, M. (1975): "A Framework for Representing Knowledge". In P. H. Winston (ed.), *The Psychology of Computer Vision*, McGraw-Hill, New York, 211-277.

Mormann, T. (1993): "Natural Predicates and the Topological Structure of Conceptual Spaces". *Synthese 95*, 219-240.

Mormann, T. (1994): "A Representational Reconstruction of Carnap's Quasianalysis". *PSA 1994: Proceedings of the Biennial Meeting of the Philosophy of Science Association* (1), 96-104.

Mormann, T. (2001): "Carnaps Philosophie als Möglichkeitwissenschaft". *Zeitschrift für philosophische Forschung 55*(1), 79-100.

Mormann, T. (2002): "Towards an Evolutionary Account of Conceptual Change in Mathematics". In Kampis, G., Kvasz, L., Stölzner, M. (Eds.), *Appraising Lakatos: Mathematics, Methodology and the Man*, Kluwer Academic Publishers, Dordrecht, 139-156.

Moulines, C.U. (1981): "An Example of a Theory-Frame: Equilibrium Thermodynamics". In Hintikka, J. et. al. (Eds.), *Probabilistic Thinking, Thermodynamics, and the Interaction of the History and Philosophy of Science*, Reidel, Dordrecht, 211-238.

Moulines, C.U. (1991): "Pragmatics in the Structuralist View of Science". In Schurz, G. and Dorn, G.J.W. (Eds.), *Advances in Scientific Philosophy*, Rodopi, Amsterdam, 313-326.

Moulines, C.U. (2011): "Cuatro Tipos de Desarrollo Teórico en las Ciencias Empiricas". *Metatheoria 1*(2), 11-27.

Moulines, C.U. (2013): "Crystallization as a Form of Scientific Semantic Change: The Case of Thermodynamics". In Küppers, B.O. et. al. (Eds.), *Evolution of Semantic Systems*, Springer, Berlin, 209-230.

Moulines, C.U. (2014): "Intertheoretical Relations and the Dynamics of Science". *Erkenntnis 79*(8), 1505-1519.

Murphy, G. (2002): *The Big Book of Concepts*. MIT Press, Cambridge (MA).

Murphy, G. and Medin, D. (1985): "The Role of Theories in Conceptual Coherence". *Psychological Review 92*(3), 289-316.

Musgrave, A. (1976): "Why Did Oxygen Supplant Phlogiston? Research Programmes in the Chemical Revolution". In C. Howson (Ed.), *Method and appraisal in the Physical Sciences*, Cambridge University Press, Cambridge, 181-209.

Nado, J. (2019): "Conceptual engineering, truth, and efficacy". *Synthese*, https://doi.org/10.1007/s11229-019-02096-x.

Nickles, T. (2017): "Scientific Revolutions". *Stanford Encyclopedia of Philosophy* (Winter 2017 Edition), Edward N. Zalta (Ed.), URL = <https://plato.stanford.edu/archives/win2017/entries/scientific-revolutions/>.

Niilinuoto, I. (2019): "Scientific Progress". *Stanford Encyclopedia of Philosophy* (Winter 2019 Edition), Edward N. Zalta (Ed.), URL = <https://plato.stanford.edu/archives/win2019/entries/scientific-progress/>.

Nosofsky, R.M. (1984): "Choice, Similarity, and the Context Theory of Classification". *Journal of Experimental Psychology: Learning, Memory, and Cognition 10*, 104-114.

Nosofsky, R.M. (1992): "Exemplars, Prototypes, and Similarity Rules". In Healy, A, kosslyn, S., and Shiffrin, R. (Eds.), *From Learning Theory to Connectionist Theory: Essays in honor of W.K. Estes*, Vol. 1, Erlbaum, Hillsdale, 149-168.

Okabe, A. et al. (2000): *Spatial tessellations*. 2nd edition, Wiley, New York.

Okasha, S. (2011): "Theory Choice and Social Choice: Kuhn versus Arrow". *Mind 120* (477), 83-115

Okasha, S. (2018): *Agents and Goals in Evolution*. Oxford University Press, Oxford.

Olsson, E.J. and Enqvist, S. (eds.) (2010): *Belief Revision meets Philosophy of Science*, Springer, Heidelberg.

Osta-Vélez, M. and Gärdenfors, P. (2020): "Category-Based Induction in Conceptual Spaces". *Journal of Mathematical Psychology 96*, 1023-1057.

Park, W. (2010): "Belief Revision vs. Conceptual Change in Mathematics". In Magnani, L., Carnielli, W., and Pizzi, C. (Eds.), *Model-Based Reasoning in Science and Technology: Abduction, Logic, and Computational Discovery*, Springer, Berlin, 121-134.

Peacocke, C. (1992): *A Study of Concepts*. MIT Press, Cambridge (MA).

Peregrin, J. (2011): *Inferentialism: Why Rules Matter*. Palgrave Macmillan, London.

Peregrin, J. (2020): "Carnap's Inferentialism". In Schuster, R. (Ed.), *Vienna Circle in Czechoslovakia*, Springer, Berlin, 97-109.

Perérez Carballo, A. (2020): "Conceptual Evaluation: Epistemic". In Cappelen, H., Plunkett, D., and Burgess, A. (Eds.), *Conceptual Engineering and Conceptual Ethics*, Oxford University Press, Oxford, 304-332.

Pickering, A. (1995): *The Mangle of Practice: Time, Agency, and Science*. University of Chicago Press, Chicago.

Pigliucci, M. and Müller, G.B. (2010): *Evolution: The Extended Synthesis*. The MIT Press, Cambridge (MA).

Pincock, C. (2010): "Exploring the Boundaries of Conceptual Evaluation". *Philosophia Mathematica 18*(3), 106-136.

Plato (c. 399-395 BC): "Euthyphro". English Translation in *Plato: Five Dialogues*, Hackett Publishing Company, Indianapolis (IN), 1981.

Plotkin, H.C. (Ed.) (1982): *Learning, Development, and Culture: Essays in Evolutionary Epistemology*. John Wiley & Sons, Hoboken (NJ).

Plutynski, A. (2007): "Drift: A Historical and Conceptual Overview". *Biological Theory 2*(2), 156-167.

Polya, G. (2004): *How to Solve It: A New Aspect of Mathematical Method*. Princeton University Press, Princeton (NJ).

Popper, K.R. (1934): *Logik der Forschung*. Julius Springer Verlag, Vienna.

Popper, K.R. (1940): "What is Dialectic?". *Mind 49*, 403-426. As reprinted in *Conjectures and Refutations: The Growth of Scientific Knowledge*, Routledge, London, 419-450.

Popper, K.R. (1957): "Philosophy of Science: A Personal Report". Reprinted as "Science: Conjectures and Refutations"in *Conjectures and Refutations: The Growth of Scientific Knowledge*, Routledge, London, 43-86.

Popper, K.R. (1963): *Conjectures and Refutations: The Growth of Scientific Knowledge*. Routledge, London.

Popper, K.R. (1972a): *Objective Knowledge: An Evolutionary Approach*. Clarendon Press, Oxford.

Popper, K.R. (1972b): "Two Faces of Common Sense: An Argument for Commonsense Realism and against Commonsense Theory of Knowledge". In *Objective Knowledge: An Evolutionary Approach*, Clarendon Press, Oxford, 32-104.

Popper, K.R. (1972c): "Of Clouds and Clocks". In *Objective Knowledge: An Evolutionary Approach*, Clarendon Press, Oxford, 206-255.

Popper, K.R. (1972d): "Evolution and the Tree of Knowledge". In *Objective Knowledge: An Evolutionary Approach*, Clarendon Press, Oxford, 256-284.

Popper, K.R. (1972e): "The Bucket and the Search Light: Two Theories of Knowledge". In *Objective Knowledge: An Evolutionary Approach*, Clarendon Press, Oxford, 341-361.

Popper, K.R. (1974a): "Autobiography of Karl Popper". In Schilpp, P.A. (Ed.), *The Philosophy of Karl Popper*, Open Court, LaSalle, 3-181.

Popper, K.R. (1974b): "Replies to my Critics". In Schilpp, P.A. (Ed.), *The Philosophy of Karl Popper*, Open Court, LaSalle, 961-1197.

Popper, K.R. (1984): "Evolutionary Epistemology". In *Evolutionary Theory: Paths into the Future*, Pollard, K.W. (Ed.), John Wiley & Sons, Hoboken (NJ), 239-255.

Post, E.L. (1936): "Finite Combinatory Processes-Formulation I". *Journal of Symbolic Logic 1*, 103-105.

Post, E.L. (1941): "Absolutely Unsolvable Problems and Relatively Undecidable Propositions: Account of an Anticipation". Unpublished draft. Reprinted in Davis, M. (Ed.), *The Undecidable: Basic papers on Undecidable Propositions, Unsolvable Problems And Computable Functions*, Raven Press, Hewlett, New York, 340-433.

Post, E.L. (1943): "Formal Reductions of the General Combinatorial Decision Problem". *Amer. J. Math. 65*, 197-215.

Prawitz, D. (1965): *Natural Deduction: A Proof-Theoretical Study*. Almqvist & Wiksell, Stockholm.

Priest, G. and Thomason, N. (2007): "60% Proof: Lakatos, Proof, And Paraconsistency". *Australasian Journal of Logic 5*, 89-100.

Prinz, J. (2002): *Furnishing the Mind: Concepts and Their Perceptual Basis*. MIT Press, Cambridge (MA).

Psillos, S. (1999): *Scientific Realism: How Science Tracks Truth*. Routledge, London.

Psillos, S. (2018): "Realism and Theory Change in Science". *Standford Encyclopedia of Philosophy* (Summer 2018 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2018/entries/realism-theory-change/>.

Putnam, H. (1970): "Is Semantics Possible?'". *Metaphilosophy 1*(3), 187-201.

Putnam, H. (1973): "Explanation and reference". Reprinted In *Philosophical Papers Vol. 2: Mind, Language and Reality*, Cambridge University Press, Cambridge (MA), 196-214, 1995.

Putnam, H. (1975): "The Meaning of 'Meaning'". Reprinted In *Philosophical Papers Vol. 2: Mind, Language and Reality*, Cambridge University Press, Cambridge (MA), 215-271, 1995.

Putnam, H. (1995): "Language and reality". In *Philosophical Papers Vol. 2: Mind, Language and Reality*, Cambridge University Press, Cambridge (MA), 272-290.

Quine, W.V.O. (1951): "Two Dogmas of Empiricism". *The Philosophical Review 60*, 20-43.

Quine, W.V.O. (1960): *Word and Object*. MIT Press, Cambridge (MA).

Quine, W.V.O. (1961): *From a Logical Point of View*. Second Edition, Harvard University Press, Cambridge (MA).

Quinon, P. (2019) "Can Church's Thesis Be Viewed as a Carnapian Explication?". *Synthese 198*(Suppl. 5), 1047-1074.

Raubal, M. (2004): "Formalizing Conceptual Spaces". In Varzi, A. and Vieu, L. (Eds.), *Formal Ontology in Information Systems: Proceedings of the Third International Conference (FOIS 2004)* vol. 114, IOS Press, Amsterdam, 153-164.

Reck, E. (2012): "Carnapian Explication: a Case Study and Critique". In Wagner, P. (Ed.), *Carnap's Ideal of Explication and Naturalism*, Palgrave Macmillan, London, 96-116.

Reiss, J. and Sprenger, J. (2020): "Scientific Objectivity". *Standford Encyclopedia of Philosophy*, (Winter 2020 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2020/entries/scientific-objectivity/>.

Renzi, B.G. (2009): "Kuhn's Evolutionary Epistemology and Its Being Undermined by Inadequate Biological Concepts". *Philosophy of Science 76*, 143-159.

Renzi, B.G. and Napolitano, G. (2011): *Evolutionary Analogies: Is the Process of Scientific Change Analogous to the Organic Change?*. Cambridge Scholars Publishing, Newcastle.

Rescher, N. (1990): *A Useful Inheritance: Evolutionary Aspects of the Theory of Knowledge*. John Wiley & Sons, Chichester.

Rey, G. (1983): "Concepts and Stereotypes". *Cognition 15*(1-3), 237-62.

Reydon, T.A.C. and Hoyningen-Huene, P. (2010): "Discussion: Kuhn's Evolutionary Analogy in *The Structure of Scientific Revolutions* and "The Road since Structure". *Philosophy of Science 77*, 468-476.

Ribeiro, M. M. and Wassermann, R. (2007): "Base revision in description logics-preliminary results". *Proceedings of the International Workshop on Ontology Dynamics (IWOD-07)*, 69-82.

Ribeiro, M. M. and Wassermann, R. (2009): "AGM revision in description logics". *Proceedings of ARCOE*, 73-85.

Richard, M. (2019): *Meaning as Species*. Oxford University Press, Oxford.

Richard, M. (2020): "The A-Project and the B-Project". In Cappelen, H., Plunkett, D., and Burgess, A. (Eds.), *Conceptual Engineering and Conceptual Ethics*, Oxford University Press, Oxford, 358-378.

Richardson, A. (2012): "Carnap's Place in Analytic Philosophy and Philosophy of Science". In Wagner, P. (Ed.), *Carnap's Ideal of Explication and Naturalism*, Pagrave Macmillan, London, 7-22.

Richardson, A. (2013): "Taking the Measure of Carnap's Philosophical Engineering: Metalogic as Metrology". In Reck, E. (Ed.), *The Historical Turn in Analytic Philosophy*, Palgrave Macmillan, London, 60-77.

Rips, L. (1995): "The Current Status of Research on Concept Combination". *Mind and Language 10*, 72-104.

Rips, L., Shoben, E.J., and Smith, E.E. (1973): "Semantic Distance and the Verification of Semantic Relations". *Journal of Verbal Learning and Verbal Behavior 12*, 1-20.

Robinson, A. (1974): *Non-Standard Analysis*. Princeton University Press, Princeton.

Rorty, R. (Ed.) (1967): *The Linguistic Turn: Recent Essays in Philosophical Method*. University of Chicago Press, Chicago.

Rosch, E. (1973): "On the Internal Structure of Perceptual and Semantic Categories". In Moore, T. (Ed.), *Cognitive Development and the Acquisition of Language*, Academic Press, New York, 111-144.

Rosch, E. (1975): "Cognitive Representation of Semantic Categories". *Journal of Experimental Psychology: General 104*, 192-233.

Rosch, E. (1978): "Principles of Categorization". In Rosch, E. and Lloyd, B. (Eds.), *Cognition and Categorization*, Lawrence Erlbaum Associates, Hillsdale, 27-48.

Rosch, E. and Mervis, C. (1975): "Family Resemblances: Studies in the Internal Structures of Categories". *Cognitive Psychology 7*, 573-605.

Rott, H. (2001): *Change, Choice and Inference: A Study of Belief Revision and Nonmonotonic Reasoning*. Clarendon Press, Oxford.

Ruse, M. (1986): *Taking Darwin Seriously: A Naturalistic Approach to Philosophy*. Blackwell, Oxford.

Russell, B. (1914): *Our Knowledge of the External World as a Field for Scientific Method in Philosophy*. Open Court, LaSalle (IL).

Sankey, H. (1997): "Taxonomic Incommensurability". In Sankey, H. (Ed.), *Rationality, Relativism, and Incommensurability*, Ashgate, London, 66-80.

Sankey, H. and Hoyningen-Huene, P. (2001): "Introduction". In Hoyningen-Huene, P. and Sankey, H. (Eds.), *Incommensurability and Related Matters*, Kluwer, Dordrecht, vii-xxxiv.

Sarto-Jackson, I. (2019): "Converging Concepts of Evolutionary Epistemology and Cognitive Biology Within a Framework of the Extended Evolutionary Synthesis". *Journal for General Philosophy of Science*, https://doi.org/10.1007/s10838-019-09479-1.

Sawyer, S. (2018): "The importance of concepts". *Proceedings of the Aristotelian Society 118* (2), 127-147.

Sawyer, S. (2020): "Talk and Thought". In Cappelen, H., Plunkett, D., and Burgess, A. (Eds.), *Conceptual Engineering and Conceptual Ethics*, Oxford University Press, Oxford, 379-395.

Scharp, K. (2013): *Replacing Truth*. Oxford University Press, Oxford.

Scharp, K. (2020): "Philosophy as the Study of Defective Concepts". In Cappelen, H., Plunkett, D., and Burgess, A. (Eds.), *Conceptual Engineering and Conceptual Ethics*, Oxford University Press, Oxford, 396-416.

Schiemer, G. (2020a): "Transfer Principles, Klein's Erlanger Program, and Methodological Structuralism". In Reck, E. and Schiemer, G. (Eds.), *The Pre-History of Mathematical Structuralism*, Oxford University Press, Oxford, 106-141.

Schiemer, G. (2020b): "Carnap's Structuralist Thesis". In Reck, E. and Schiemer, G. (Eds.), *The Pre-History of Mathematical Structuralism*, Oxford University Press, Oxford, 383-420.

Schiemer, G. and Gratzl, N. (2016): "The Epsilon-Reconstruction of Theories and Scientific Structuralism". *Erkenntnis 81*(2), 407-432.

Schlimm, D. (2012): "Mathematical Concepts and Investigative Practice". In Feest, U. and Steinle, F. (Eds.), *Scientific Concepts and Investigative Practice*, De Gruyter, Berlin, 127-147.

Schlimm, D. (2013): "Axioms in Mathematical Practice". *Philosophia Mathematica 21*(1), 37-92.

Schroeder-Heister, P. (2018): "Proof-Theoretic Semantics". *Stanford Encyclopedia of Philosophy* (Spring 2018 Edition), Edward N. Zalta (Ed.), URL = <https://plato.stanford.edu/archives/spr2018/entries/proof-theoretic-semantics/>.

Schurz, G. (2014a): "Criteria of Theoreticity: Bridging Statement and Non-Statement View". *Erkenntnis 79*, 1521-1545.

Schurz, G. (2014b): *Philosophy of Science: a unified approach*. Routledge, New York.

Schurz, G. and Votsis, I. (2014): "Reconstructing Scientific Theory Change by Means of Frames". In Gamerschlag, Th. et al. (Eds.), *Frames and Concept Types*, Springer, Berlin, 93-109.

Sellars, W. (1963): *Science, Perception and Reality*. Humanities Press, New York.

Sellars, W. (1973): "Conceptual Change". In Pearce, G. and Maynard, P. (Eds.), *Conceptual Change*, D. Reidel, Boston, 77-93.

Sen, A. (1997): "Maximization and the Act of Choice". *Econometrica 65*(4), 745-779.

Shapin, S. and Schaffer, S. (1985): *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life*. Princeton University Press, Princeton (NJ).

Shapiro, L. (2004): *The Mind Incarnate*. Oxford University Press, Oxford.

Shapiro, S. (2006a): *Vagueness in Context*. Oxford University Press, Oxford.

Shapiro, S. (2006b): "Computability, Proof, and Open Texture". In Olszewski, A, Wolenski, J. and Janusz, R. (Eds.), *Church's Thesis after 70 years*, Ontos Verlag, Frankfurt, 420-455.

Shapiro, S. (2013): "The Open-Texture of Computability". In Copeland J., Posy C., Shagrir O. (Eds.), *Computability: Gödel, Turing, Church, and Beyond*, The MIT Press, Cambridge (MA), 153-181.

Shapiro, S. and Roberts, C. (2019): "Open Texture and Analyticity". In Makovec, D. and Shapiro, S. (Eds.), *Friedrich Waismann: The Open Texture of Analytic Philosophy*, Palgrave Macmillan, London, 189-210.

Shepherd, J. and Justus, J. (2015): "X-Phi and Carnapian Explication". *Erkenntnis 80*, 381-402.

Sieg, W. (1997): "Step by Recursive Step: Chuch's Analysis of Effective Calculability". *Bulletin of Symbolic Logic 3*(2), 154-180.

Sieg, W. (2002): "Calculations by man and machine: Conceptual analysis". In W. Sieg, R. Sommer, C. Talcott (Eds.), *Reflections on the Foundations of Mathematics. Essays in honor of Solomon Feferman*, Lecture Notes in Logic 115, Assoc. for Symbolic Logic, A. K. Peters, Ltd., Natick (MA), 390-409.

Sieg, W. (2002a): "Calculations by man and machine: Mathematical presentation". In *Proceedings of the Cracow International Congress of Logic, Methodology and Philosophy of Science*, Synthese Series, Kluwer Academic Publishers, Dordrecht, 245-260.

Sieg, W. (2009): "On Computability". in A. Irvine (Ed.), *Handbook of the Philosophy of Mathematics*, Elsevier, Amsterdam, 535-630.

Sieg, W. (2013): "Axioms for Computability: Do They Allow a Proof of Church's Thesis?". In H. Zenil (Ed.), *A Computable Universe: Understanding and Exploring Nature as Computation*, World Scientific Publishing, Singapore, 99-123.

Sieg, W. (2018): "What is the Concept of Computation?". In F. Manea, R.G. Miller, D. Nowotka (Eds.), *CiE 2018: Sailing Routes in the World of Computation*, Lecture Notes in Computer Science 10936, 386-396.

Sieg, W. and Byrnes, J. (1996): "K-graph Machines: Generalizing Turing's Machines and Arguments". In P. Hajek (Ed.), *Gödel '96*, Lecture Notes in Logic 6, Springer Verlag, 98-119.

Sieg, W. and Byrnes, J. (1999): "An Abstract Model for Parallel Computations: Gandy's Thesis". *The Monist 82*(1), 150-164.

Sieg, W., Szabó, M. and McLaughlin, D. (2016): "Why Post Did [Not] Have Turing's Thesis". In Omodeo E.G. and Policriti A. (Eds.), *Martin Davis on Computability, Computational Logic, and Mathematical Foundations*, Springer, New York, 175-208.

Siegfried, R. (2002): *From Elements to Atoms: A History of Chemical Composition*. American Philosophical Society, Philadelphia.

Simon, H.A. (1981): *The Sciences of the Artificial*. MIT Press, Cambridge (MA).

Sjögren, J. (2011): "A Note on the Relation Between Formal and Informal Proof". *Acta Analytica 25*, 447-458.

Skagestadt, P. (1978): "Taking Evolution Seriously: Critical Comments on D.T. Campbell's Evolutionary Epistemology". *Monist 61* 611-621.

Skyrms, B. (2010): *Signals: Evolution, Learning, and Information*. Oxford University Press, Oxford.

Smith, E. and Medin, D. (1981): *Categories and Concepts*. Harvard University Press, Cambridge (MA).

Smith, E. and Osherson, D. (1984): "Conceptual Combination with Prototype Concepts ". *Cognitive Science 8*, 337-361.

Smith, E., Osherson, D., Rips, L., and Keane, M. (1988): "Combining prototypes: A selective modification model ". *Cognitive Science 12*, 485-527.

Smith, P. (2011): "Squeezing Arguments". *Analysis 71*(1), 22-30.

Smith, P. (2013): *An Introduction to Gödel's Theorems*. 2nd Edition, Cambridge University Press, Cambridge.

Sneed, J. (1979): *The Logical Structure of Mathematical Physics*. D. Reidel Publishing Company, Dordrecht.

Sober, E. (1980): "Evolution, Population Thinking, and Essentialism". *Philosophy of Science 47*, 350-383.

Sober, E. (1984): *The Nature of Selection*. MIT Press, Cambridge (MA).

Sober, E. (1991): "Models of Cultural Evolution". In Griffiths, P. (Ed.), *Trees of Life: Essays in the Philosophy of Biology*, Kluwer Academic Publisher, Dordrecht, 535-551.

Solomon, M. (2001): *Social Empiricism*. MIT Press, Cambridge (MA).

Stadler, F. (2015): *The Vienna Circle: Studies in the Origins, Development, and Influence of Logical Empiricism*. Revised Edition, Springer International Publishing, Berlin.

Stegmüller, W. (1976): *The Structure and Dynamics of Theories*. Springer-Verlag, Heidelberg.

Stein, H. (1992): "Was Carnap Entirely Wrong After All?". *Synthese 93*, 275-295.

Steinberger, F. (2016): "How Tolerant Can You Be? Carnap on Rationality". *Philosophy and Phenomenological Research 92*(3), 645-668.

Stevens, S.S. (1946): "On the Theory of Scales of Measurement". *Science 103*, 677-680.

Strawson, P.F. (1963): "Carnap's Views on Constructed Systems versus Natural Languages in Analytic Philosophy". In Schilpp, P. (Ed.), *The Philosophy of Rudolf Carnap*, Open Court, LaSalle, 503-518.

Strößner, C. (2020a): "Predicate Change: A Study on the Conservativity of Conceptual Change ". *Journal of Philosophical Logic 49*, 1159–1183.

Strößner, C. (2020b): "Integrating Conceptual Spaces in Frames". *IfCoLog Journal of Applied Logics 7*(5), 683-706.

Strößner, C. (2021): "Conceptual Learning and Local Incommensurability: A Dynamic Logic Approach". *Axiomathes*, https://link.springer.com/article/10.1007/s10516-021-09563-6.

Suppe, F. (1977): *The Structure of Scientific Theories*. University of Illinois Press, Urbana (IL).

Suppe, F. (1989): *The Semantic Conception of Theories and Scientific Realism*. University of Illinois Press, Chicago.

Suppes, P. (1967): "What is a Scientific Theory?". In Morgenbesser, S. (Ed.), *Philosophy of Science Today*, Basic Books, New York, 55-67.

Suppes, P. (2002): *Representation and Invariance of Scientific Structures*. CSLI Publications, Standford.

Sznajder, M. (2016): "What Conceptual Spaces Can Do for Carnap's Late Inductive Logic". *Studies in History and Philosophy of Science 56*(A), 62-71.

Sznajder, M. (2018): "Inductive Logic as Explication: The Evolution of Carnap's Notion of Logical Probability". *The Monist 101*(4), 417-440.

Tanswell, F.S. (2018): "Conceptual Engineering for Mathematical Concepts". *Inquiry 61*(8), 881-913.

Tappenden, J. (2008a): "Mathematical Concepts and Definitions". In Mancosu, P. (Ed.), *The Philosophy of Mathematical Practice*, Oxford University Press, New York, 256-275.

Tappenden, J. (2008b): "Mathematical Concepts: Fruitfulness and naturalness". In Mancosu, P. (Ed.), *The Philosophy of Mathematical Practice*, Oxford University Press, New York, 276-301.

Tarski, A. (1933): *Der Wahrheitsbegriff in den formalisierten Sprachen*. English translation, (1956) "The concept of truth in formalized languages", in *Logic, Semantics: Metamathematics*, Second Edition, Oxford University Press, Oxford. 152-278.

Ter Hark, M. (2004): *Popper, Otto Selz, and the Rise of Evolutionary Epistemology*. Cambridge University Press, Cambridge.

Thagard, P. (1980): "Against Evolutionary Epistemology". *PSA 1980: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 187-196.

Thagard, P. (1984): "Frames, Knowledge, and Inference". *Synthese 61*, 233-259.

Thagard, P. (1988): *Computational Philosophy of Science*. MIT Press, Cambridge (MA).

Thagard, P. (1990): "The Conceptual Structure of the Chemical Revolution". *Philosophy of Science 57*(2), 183-209.

Thagard, P. (1992): *Conceptual revolutions*. Princeton University Press, Princeton (NJ).

Thagard, P. (2000): *Coherence in Thought and Action*. MIT Press, Cambridge (MA).

Thomasson, A. (2020): "A Pragmatic Method for Normative Conceptual Work". In Cappelen, H., Plunkett, D., and Burgess, A. (Eds.), *Conceptual Engineering and Conceptual Ethics*, Oxford University Press, Oxford, 435-457.

Toulmin, P. (1957): "Crucial Experiments: Priestley and Lavoisier". *Journal of the History of Ideas 18*, 205-220.

Toulmin, S. (1967): "The Evolutionary Development of Natural Science". *American Scientist 55*(4), 456-471.

Toulmin, S. (1970): "From Logical Systems to Conceptual Populations". *PSA 1970: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 552-564.

Toulmin, S. (1972): *Human Understanding: The Collective Use and Evolution of Concepts*. Princeton University Press, Princeton (NJ).

Turing, A.M. (1936): "On Computable Numbers, with an Application to the Entscheidungsproblem". *Proc. London Math. Soc.* ser. 2 *42*(Parts 3 and 4) 230-265; Turing, A.M. (1937): "A correction", *ibid. 43*, 544-546.

Turing, A.M. (1954): "Solvable and Unsolvable Problems". *Science News 31*, 7-23.

Tversky, A. (1977): "Features of similarity". *Psychological Review 84*, 327-352.

Tversky, A. et al. (Eds.) (1971-1989-1990): *Foundations of Measurement*. Three volumes, Academic Press Inc., New York.

Uebel, T. (2007): *Empiricism at the Crossroads: The Vienna Circle's Protocol-Sentence Debate*. Open Courte, LaSalle (IL).

Uebel, T. (2012): "The Bipartite Conception of Metatheory and the Dialectical Conception of Explication". In Wagner, P. (Ed.), *Carnap's Ideal of Explication and Naturalism*, Palgrave Macmillan, London, 117-130.

Uebel, T. (2018): "Carnap's Transformation of Epistemology and the Development of His Metaphilosophy". In *The Monist 101*(4), 367-387.

Uspensky, A.V. (1992): "Kolmogorov and Mathematical Logic". *The Journal of Symbolic Logic 57*(2), 385-412.

Uspensky, A.V. and Semyonov, A.L. (1993): "Kolmogorov's Algorithms or Machines". In Shiryayev, A.N. (Ed.), *Selected Works of A. N. Kolmogorov*, Volume III, Mathematics and Its Applications (Soviet Series) Vol. 27, Springer, 251-260.

van Fraassen, B. (1989): *Laws and Symmetry*. Oxford University Press, Oxford.

Vecht, J.J. (2020): "Open Texture clarified". *Inquiry*, doi:10.1080/0020174X.2020.1787222.

Vicente, A. and Manrique, F.M. (2016): "The Big Concepts Paper: A Defense of Hybridism". *British Journal for the Philosophy of Science 67*, 59-88.

Wagner, P. (Ed.) (2009): *Carnap's Logical Syntax of Language*. Palgrave Macmillan, London.

Wagner, P. (Ed.) (2012): *Carnap's Ideal of Explication and Naturalism*. Palgrave Macmillan, London.

Waismann, F. (193?): "Hypotheses". Manuscript composed short before 1936. Translated in English in McGuinness, B. (Ed.), *Philosophical Papers*, 38-59, 1977.

Waismann, F. (1936): *Einführung in das mathematische Denken: Die Begriffsbildung der modernen Mathematik*. Translated in English and Reprinted as *Introduction to Mathematical Thinking: the formation of concepts in modern mathematics*, Courier, North Chelmsford (MA), 2003.

Waismann, F. (1940): "Was ist logische Analyse?". *Erkenntnis 8*, 265-289. Translated in English and reprinted in McGuinness, B. (Ed.), *Philosophical Papers*, 81-103, 1977.

Waismann, F. (1945): "Verifiability". *Proceedings of the Aristotelian Society XIX*, 119-150. As reprinted in Harré, R. (Ed.), *How I See Philosophy*, 39-66, 1968.

Waismann, F. (1946a): "Are There Alternative Logics?". *Proceedings of the Aristotelian Society XLVI*, 77-104. As reprinted in Harré, R. (Ed.), *How I See Philosophy*, 67-90, 1968.

Waismann, F. (1946b): "Language Strata: Part One". *Synthese 5*, 210-219. As reprinted in Harré, R. (Ed.), *How I See Philosophy*, 91-102, 1968.

Waismann, F. (1949-1953): "Analytic-Synthetic". Published in five parts in *Analysis 10-13*. As reprinted in Harré, R. (Ed.), *How I See Philosophy*, 122-195, 1968.

Waismann, F. (1953): "Language Strata: Part Two". *Logic and Language 2*, 11-31. As reprinted in Harré, R. (Ed.), *How I See Philosophy*, 102-121, 1968.

Waismann, F. (1965): *The Principles of Linguistic Philosophy*. Harré, R. (Ed.), Macmillan, London.

Walsh, D.M., Lewens T., and Ariew, A. (2002): "The Trials of Life: Natural Selection and Random Drift". *Philosophy of Science 69*(3), 452-473.

Weibull, J.W. (1995): *Evolutionary Game Theory*. MIT Press, Cambridge (MA).

Weiskopf, D. (2009): "The Plurality of Concepts". *Synthese 169*, 145-173.

Weiskopf, D. (2010): "The Theoretical Indispensability of Concepts". *Behavioral and Brain Sciences 33*, 228-229.

Werndl, C. (2009): "Justifying Definitions in Mathematics – Going Beyond Lakatos". *Philosophia Mathematica 17*(3), 313-340.

Whitehead, A.N. and Russell, B. (1910-1913): *Principia Mathematics*. Vol. 1-3, Cambridge University Press, Cambridge.

Wilder, R.L. (1953): "The Origin and the Growth of Mathematical Concepts". *Bulletin of the American Mathematical Society 59*, 423-448.

Williams, B. (2002): *Truth and Truthfulness: An Essay in Genealogy*. Princeton University Press, Princeton (NJ).

Williamson, T. (1990): *Identity and Discrimination*. Blackwell, Oxford.

Williamson, T. (1994): *Vagueness*. Routledge, London.

Williamson, T. (2007): *The Philosophy of Philosophy*. Blackwell, Oxford.

Wilson, E.O. (1980): *Sociobiology: The New Synthesis*. Revised Edition, Harvard University Press, Cambridge (MA).

Wilson, M. (1982): "Predicate Meets Property". *The Philosophical review 91*(4), 549-589.

Wilson, M. (1994): "Can We Trust Logical Form?". *Journal of Philosophy 91*(10), 519-544.

Wilson, M. (1998): "Classical Mechanics". In Craig, E. (Ed.), *Routledge Encyclopedia of Philosophy*, Taylor & Francis, London.

Wilson, M. (2000a): "Inference and Correlational Truth". In Gupta, M. and Chapuis, A. (Eds.), *Circularity, Definition and Truth*, Ridgeview Press, Atascadero (CA), 57-73.

Wilson, M. (2000b): "The Unreasonable Uncooperativeness of Mathematics in the Natural Science". *The Monist 83*(2), 296-314.

Wilson, M. (2006): *Wandering Significance: An Essay on Conceptual Behavior*. Clarendon Press, Oxford.

Wilson, M. (2008): "Beware of the Blob: Cautions for Would Be Metaphysicians". In Zimmerman, D.W. (Ed.), *Oxford Studies in Metaphysics: Volume 4*, Oxford University Press, Oxford, 275-318.

Wilson, M. (2012a): "The Perils of Polyanna". In Wagner, P. (Ed.), *Carnap's Ideal of Explication and Naturalism*, Palgrave-Macmillan, London, 205-224.

Wilson, M. (2012b): "Long Ago, in a Context Far Away". In Frappier M., Brown D., DiSalle R. (Eds.), *Analysis and Interpretation in the Exact Sciences*, The Western Ontario Series in Philosophy of Science vol 78, Springer, Dordrecht, 57-73.

Wilson, M. (2014): "What Is 'Classical Mechanics'Anyway?". In Batterman, R. (Ed.), *The Oxford Handbook of Philosophy of Physics*, Oxford University Press, Oxford, 43-105.

Wilson, M. (2017): *Physics Avoidance: And Other Essays in Conceptual Strategy*. Oxford University Press, Oxford.

Wilson, M. (MS): *Imitation of Rigor*. Manuscript.

Wimsatt, W.C. (1986): "Generative Entrenchment, Scientific Change, and the Analytic-Synthetic Distinction:A Developmental Model of Scientific Evolution". *Unpublished MS*, dated August 1986.

Wittgenstein, L. (1958): *Philosophical Investigations*. Third Edition, Blackwell, Oxford.

Wolter, F. and Zakharyaschev, M. (1999): "Multi-dimensional description logics". *IJCAI*, Vol. 99, 104-109.

Wright, S. (1932): "The Roles of Mutation, Inbreeding, Crossbreeding, and Selection in Evolution". *Proceedings of the Sixth International Congress of Genetics 1*, 257-266.

Wuketits, F.M. (2001): "The Philosophy of Donald T. Campbell: A Short Review and Critical Appraisal". *Biology & Philosophy 16*, 171-188.

Wussing, H. (1984): *The Genesis of the Abstract Group Concept: A Contribution to the History of the Origin of Abstract Group Theory*. MIT Press, Cambridge (MA).

Yap, A. (2010): "Feminism and Carnap's Principle of Tolerance". *Hypatia 24*, 437-454.

Zalta, E. (2001): "Fregean Senses, Modes of Presentation, and Concepts". *Philosophical Perspectives 15*, 335-359.

Zeifert, M. (2020): "Rethinking Hart: From Open Texture to Prototype Theory – Analytic Philosophy Meets Cognitive Linguistics". *International Journal for the Semiotics of Law*, https://doi.org/10.1007/s11196-020-09722-9.

Zenker, F. (2014): "From Features via Frames to Spaces: Modeling Scientific Conceptual Change without Incommensurability or Aprioricity". In Gamerschlag, T. et al. (Eds), *Frames and Concept Types: Applications in Language and Philosophy*, Springer, Berlin, 69-89.

Zenker, F. and Gärdenfors, P. (2014): "Modeling Diachronic Change in Structuralism and in Conceptual Spaces". *Erkenntnis 79*, 1547-1561.

Zenker, F. and Gärdenfors, P. (2015a): "Communication, Rationality, and Conceptual Changes in Scientific Theories". In Zenker, F. and Gärdenfors, P. (Eds.), *Applications of Conceptual Spaces: The Case for Geometric Knowledge Representation*, Springer, Berlin, 259-277.

Zenker, F. and Gärdenfors, P. (Eds.) (2015b): *Applications of Conceptual Spaces: The Case for Geometric Knowledge Representation*, Springer, Berlin.