The 'Theoretical Virtues' Theory: Towards a Joint Solution to the Meta-Descriptive and Meta-Normative Problems

Panagiotis Saranteas

University of
St Andrews

This thesis is submitted in partial fulfilment for the degree

of Master of Philosophy (MPhil)

at the University of St Andrews

November 2020

Candidate's declaration

I, Panagiotis Saranteas, do hereby certify that this thesis, submitted for the degree of MPhil, which is approximately 40,000 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for any degree.

I was admitted as a research student at the University of St Andrews in September 2017.

I confirm that no funding was received for this work.

Date 07/11/2020 Signature of candidate

Supervisor's declaration

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of MPhil in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

Date 07/11/2020 Signature of supervisor

Permission for publication

In submitting this thesis to the University of St Andrews we understand that we are giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. We also understand, unless exempt by an award of an embargo as requested below, that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that this thesis will be electronically accessible for personal or research use and that the library has the right to migrate this thesis into new electronic forms as required to ensure continued access to the thesis.

I, Panagiotis Saranteas, confirm that my thesis does not contain any third-party material that requires copyright clearance.

The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

Printed copy

No embargo on print copy.

Electronic copy

No embargo on electronic copy.

Date 07/11/2020 Signature of candidate

11/2020 Signature of supervisor

Underpinning Research Data or Digital Outputs

Candidate's declaration

I, Panagiotis Saranteas, hereby certify that no requirements to deposit original research data or digital outputs apply to this thesis and that, where appropriate, secondary data used have been referenced in the full text of my thesis.

Date 07/11/2020 Signature of candidate

TITLE

The 'Theoretical Virtues' Theory: Towards a Joint Solution to the Meta-Descriptive and Meta-Normative Problems

ABSTRACT

There are at least four main problems that the 'theoretical virtues' theory needs to solve. Those are the descriptive, meta-descriptive, normative, and meta-normative problems. The first two problems are about virtue discovery and its methods, and the latter two are about virtue justification and its methods. The principles constituting a solution to the descriptive problem need to adequately describe the order of epistemic priority between the theoretical-explanatory virtues, and their combinations in various degrees, based on the considered judgments of expert natural and social scientists, and philosophers of science, on actual cases of explanatory-theory choice. The principles constituting a solution to the normative problem need to adequately describe the order of epistemic priority that should guide the judgments of both experts and non-experts in rational explanatory-theory choice in science, philosophy, and everyday reasoning. In this study it is argued that the two meta-level problems are prior to the two lower-level problems, and also, that they are interdependent and can be solved jointly. Such a joint solution is explored with the aim of eventually solving the two lower-level problems. After evaluating the most recent solutions that have been proposed to the descriptive and normative problems, the historical philosophy of science method and experimental philosophy of science method are also evaluated for their meta-descriptive and meta-normative potential. After their limitations are examined, it is argued that as part of any adequate solution to the two meta-level problems, the concept of a 'theoretical-explanatory virtue' would need to be conceptually re-engineered into, one, combinations of features of explanatory theories, and, two, a set of principles that describe each feature's and each combination's weight in an order of epistemic priority. Finally, the proposed solution to the meta-level problems is further developed by utilizing techniques from the methods of philosophical artificial intelligence and reflective equilibrium.

TABLE OF CONTENTS

CHAPTER 1
Introduction

1. Reasoning up from First Principles and Navigating the Conceptual Terrain

Assume there are two explanatory theories T1 and T2 of a certain phenomenon or set of phenomena. Any one of the following four things can be the case: T1 is preferable to T2, T2 is preferable to T1, T1 and T2 are equally preferable; or, for whatever reason, we cannot form a clear judgment as to whether T1 is preferable, T2 is preferable, or T1 and T2 are equally preferable. There are at least two types of questions that one can ask about the two explanatory theories. First, are they equal or is one of them higher in the order of preference, and if one of them is preferable which one is it? The second more important question is, why? That is, if one of them is preferable then why is it preferable, or what makes it preferable, and if they are equally preferable then why are they so? The answer to the first question is a judgment of preference either in favour of T1, in favour of T2, or in favour of their equality; with a fourth option being 'no judgment', that is, a statement to the inability to make a judgment of preference for whatever reason in this particular case. The answer to the second question is itself an explanatory theory T3, this time of the phenomena of our judgment of explanatory-theory preference. T3 would partly consist of a set of proposed descriptive principles from which the particular judgment that was actually made can be derived, and partly consist of a set of proposed normative principles of epistemic priority that aimed to justify that judgment and also guide our future explanatory-theory preference judgments. Again, an alternative option being a statement to the inability to justify why we made the judgment of explanatory-theory preference that we did. Lastly, being an explanatory theory itself, T3 would also have to self-apply and so would need to be shown to be the most preferable among the competing meta-level explanatory theories of the phenomena of our judgment of explanatory-theory preference at least given its own proposed normative principles; which would in no way be incoherent, viciously regressive, or viciously circular. One such meta-level explanatory theory is the 'theoretical virtues' theory. So now that the basics are in place we need to build our way up by strategically selecting the most useful among the conceptual tools available to us to pursue the above questions.

Concerning the conceptual apparatus that has been used in the philosophical literature up to the current research, the most common form that an answer to the second question takes is that, for example, T1 is preferable to T2 because T1 provides a "better", "lovelier", or otherwise more 'choice-worthy' explanation than T2.[1] Specifically, under the 'theoretical virtues' theory a 'better explanation' is commonly taken to be an explanation which displays more "theoretical-explanatory virtues"[2] or the 'virtues of a good theory or explanation', and so is more 'theoretically or explanatorily virtuous', than the alternatives.[3] Although I will be considering these concepts and their limitations in more detail further on, the reader may already be becoming suspicious of the depth and even the validity of a description of such a form, that is, of the form: X is preferable to Y because X is 'better' than Y, and it is 'better' because it has more 'good things' about it than Y. Furthermore, it should also be noted that the theory seems to presuppose a commitment to the position that scientific theories for example are in the business of explanation to begin with, which has been disputed.[4] However, for the purposes of this study I will focus on theories that are explanatory, as opposed to, say, merely instrumentally useful for making predictions, and the virtues that apply to them. I will leave these issues aside for now and pick some of them up in later sections. Moving on, I would like to first focus on the

[1] See Lipton, Peter, 2004, *Inference to the Best Explanation*; Mackonis, Adolfas, 2013, *Inference to the Best Explanation, Coherence and Other Explanatory Virtues*; Harman, Gilbert, 1965, *The Inference to the Best Explanation*; also for 'lovelier' see Barnes, Eric C., 1995, *Inference to the Loveliest Explanation*.

[2] For an example of the use of the concept of a 'theoretical-explanatory virtue' or 'theoretical-explanatory value' in the philosophical literature see, Kramer, Matthew H., 2018, *H. L. A. Hart: The Nature of Law*, s. 6.1. 'Theoretical-Explanatory Virtues'.

[3] For examples of philosophers who use the concept of 'explanatory hypothesis' see Thagard, Paul, 1989, *Explanatory Coherence*, p. 435ff; for examples of philosophers who use the concepts of 'theory preference' or 'theory choice' see Kuhn, Thomas, 1977, *The Essential Tension*, ch. 13, 'Objectivity, Value Judgment, and Theory Choice'; also Schindler, Samuel, 2018, *Theoretical Virtues in Science*.

[4] Duhem, Pierre, 1914/1991, *The Aim and Structure of Physical Theory*, ch. 1, s. 1 'Physical Theory Considered as Explanation', p. 7ff; Russell, Bertrand, 1912/1992, *On the Notion of Cause*.

concepts that are used in formulating the questions posed for this study and motivate them over alternatives.

To begin, instead of the concept of 'theoretical-explanatory virtues' researchers have used the concepts of "theoretical virtues"[5] or "virtues of good theories"[6], "epistemic virtues" or "cognitive virtues"[7], and even "epistemic values", "cognitive values"[8], "aesthetic", "practical," and "pragmatic" virtues and values[9]. An issue with some of these other concepts is that they are 'ideologically loaded', that is they come with a number of connotations and assumptions that need not necessarily be accepted when trying to research the questions set for this study. To be more specific let me take these alternative concepts in turn and give some reasons as to why they should not be preferred. First, the use of 'values' instead of 'virtues' does not add anything significant, both 'virtues' and 'values' basically analyse into 'good things' that something has, in our case, an explanatory theory of a set of phenomena. The task, however, is to discover those 'good things' and explain why they are 'good', or 'virtuous', or 'valuable'. Therefore, while the concept of 'virtue' will temporarily be preferred over the concept of 'value', later on I will argue that the 'theoretical-explanatory virtues' conceptualization should be conceptually re-engineered into one that is more conducive to rational inquiry and problem-solving, and which will be referred to as the 'features and principles' conceptualization.

Now the concepts 'epistemic', 'cognitive', 'practical', 'aesthetic', and 'pragmatic' come with semantic baggage and they mean something over and above that there is

---

[5] Keas, Michael N., 2018, *Systematizing the Theoretical Virtues*; Schindler, Samuel, 2018, *Theoretical Virtues in Science*.

[6] McMullin, Ernan, 2014, *The Virtues of a Good Theory*.

[7] Mackonis, 2013.

[8] Douglas, Heather, 2013, *The Value of Cognitive Values*; also Laudan, Larry, 1984, *Science and Values*, ch. 3 'Closing the Evaluative Circle: Resolving Disagreements about Cognitive Values'.

[9] See Keas, 2018, p. 2762, "aesthetic virtues"; for use of "aesthetic", "practical", and "pragmatic" see Popper, Karl, 1959, *The Logic of Scientific Discover*, ch. 7 'Simplicity'; for use of "epistemic virtues" and "pragmatic virtues" see Schindler, 2018, ch. 1, s 1.3. 'Theoretical Virtues: Epistemic or Pragmatic?'.

something ‘good’ about an explanatory theory. For example, some may mean that these virtues have something to do with the theory itself, assuming that the latter is a separate ‘object’ which may stand in a certain relation to ‘the world’, or that, on the contrary, they have primarily to do with our cognition, and so on. The same applies to the use of those terms in a more indirect way, as in the case of the ‘aesthetic concept of simplicity’, the ‘pragmatic concept of simplicity’, or the ‘epistemic concept of simplicity’[10]; all these formulations come preloaded with assumptions and connotations about the particular virtue in question. These are things to be shown, not to be assumed in the analysis of the very concept itself. That is why it would be best to avoid their use in formulating the research questions. In sum, these are some of the main reasons why the use of the concept of ‘theoretical-explanatory virtue’ is preferable to the above alternatives for now. Having navigated the conceptual terrain and strategically selected the most useful conceptual tools, we can now proceed to the four main research problems of the inquiry and the goals of this particular study.

1.1. The Main Research Problems of the Inquiry and the Goals of This Study

This study is partly an inquiry into the nature of what so far has been referred to as the ‘theoretical-explanatory virtues’. Specifically, an inquiry into the formative role these virtues play in the judgments of natural and social scientists, and philosophers of science, and into the methods that have been utilized for their discovery and justification with the purpose of constructing and justifying principles of epistemic priority. As to the main research problems of this study, it can be safely assumed that the 'theoretical virtues' theory is but one of a number of potential explanatory theories of the phenomena of our explanatory-theory preference, and any such theory needs to adequately solve, or at least be shown to be able to solve, the descriptive, meta-descriptive, normative, and meta-normative problems. That is in order for it to minimally be deemed adequate and so be admissible into the theoretical competition for bestness, where the explanatory theory that solves at least those problems most

[10] See Popper, 1959, p. 122ff.

plausibly, and, importantly, most virtuously, is to be considered the best or most preferable relative to all other competitors. The first two problems are about virtue discovery, and the latter two problems are about virtue justification. The descriptive problem is about discovering the descriptive principles of epistemic priority that adequately describe the order of epistemic priority between the theoretical-explanatory virtues and their combinations, based on the considered judgments made by expert natural and social scientists and philosophers of science on actual cases of explanatory-theory choice.

The meta-descriptive problem is about which method and theoretical framework is the best for discovering those descriptive principles, and about the distinction between sufficiently and insufficiently descriptive principles. These sub-problems, and also the related one of rational choice between sufficiently descriptive, yet competing, sets of principles, are only meaningful within the framework of the best among the admissible meta-descriptive solutions. Therefore, for at least these reasons, the meta-descriptive is prior to the descriptive problem. The normative problem is about justifying the normative principles of epistemic priority that describe the order of epistemic priority between the theoretical-explanatory virtues and their combinations at various degrees that imply judgments of explanatory-theory preference that expert and non-expert scientists and philosophers of science, on the one hand, and non-scientists and non-philosophers-of-science, on the other hand, should adopt in their scientific, philosophical, and everyday reasoning. The meta-normative problem is about which method and theoretical framework is the best for justifying those normative principles. In this case also, the distinction between sufficiently and insufficiently normative principles, and the subsequent rational choice between competing yet sufficiently normative principles, are only meaningful within the framework of the best among the admissible meta-normative solutions. Therefore, the meta-normative is also prior to the normative problem.

In closing this introductory section, it would be useful to provide a statement of the main goals that are aimed to be achieved with regards to the four main research problems. Before I do that, however, I would like to clearly distinguish between the goals that would be achieved by a successful inquiry into these research questions, and the goals that can practically be achieved with regards to the sub-problems that are the main focus of this particular study. Aspirationally, the aim would be to fully solve the descriptive problem by, discovering every feature of explanatory theories that plays a formative role in, or otherwise significantly impacts or influences, our judgments of explanatory-theory preference. That is, from the perspective of the 'theoretical virtues' theory, discover every theoretical-explanatory virtue. It would furthermore aim to discover whether there is any significant level of variation in the order of epistemic priority within and between each of these features and their combinations. Then if no significant variation in the order of epistemic priority is observed such a theory would aim to provide a complete list of descriptive principles that adequately describe the judgments of theoretical-explanatory preference made by the natural and social scientists and philosophers of science involved. If, on the contrary, significant variation is observed beyond reasonable doubt, then it would aim to provide different lists of principles each of which adequately described each of the variant sets of judgments. It should be noted that if a random pattern is observed in the judgments of these natural epistemic agents, then the 'theoretical virtues' theory would be in disagreement with the phenomena, and therefore would be wrong. However, such randomness would then itself be a puzzling yet interesting phenomenon in need of further theoretical explanation.

Aspirationally, again, another goal would be to fully solve the normative problem by justifying the normative principles that adequately describe the order of epistemic priority within and between every feature and between every combination of features, or ‘virtues’ in this case, that scientists, philosophers of science, and everyday reasoners should adopt and adhere to when engaging in rational choice between competing explanatory theories of any given set of phenomena. Importantly, in the case of an

existence of significant variation, one would need to either (a) justify the choice of one set of normative principles of epistemic priority over the others, or (b) show that all variant sets of normative principles are somehow justified, or (c) show that none are justified, say, because of the very existence of said variation, or some other plausible reason. Of course given the scope of those problems it must be conceded that these are not achievable goals within the context of the present inquiry. Therefore, I would now also like to state the two main goals that are practically achievable within the scope and limits of this study.

First, evaluating current proposed solutions to the descriptive problem, and making progress in solving the meta-descriptive problem for the 'theoretical virtues' theory, by further improving the framework for, and method of, virtue discovery that has been developed by philosophers and philosophically-minded scientists so far, so that it can eventually achieve the first of the aspirational goals above, that is, to solve the descriptive problem. Analogously, the second goal would be evaluating current proposed solutions to the normative problem, and making progress in solving the meta-normative problem by further improving the framework for, and method of, virtue justification in way that builds on the methods that have been proposed so far, and that allows us to effectively achieve the second of the aspirational goals above, namely, solving the normative problem. On a side note, in this study I will also show that these two goals are in fact interrelated, and that the two frameworks and methods are actually interdependent and when brought together can form a unified solution to the two meta-level problems. Having stated the two main goals that this study aims to achieve in the interest of establishing a common framework of understanding and a common frame of reference with the reader, I will now proceed to a brief overall outline of the study before moving on to the philosophical literature evaluation with these goals in mind.

1.2. Overall Outline of the Study by Chapter

Following the above quadruple distinction between the main research problems outlined above, the study is divided into four main chapters, in addition to the introductory and concluding ones, two on virtue discovery and two on virtue justification, and in each of the chapters the most recent proposed solutions to each of the four problems will respectively be explained and evaluated. Chapter two, the first on virtue discovery, will evaluate the philosophical progress and latest proposals made towards solving the descriptive problem, with the underlying aim of better understanding and solving the meta-descriptive problem later on. After considering the relation between the descriptive and the meta-descriptive problems, and providing a short outline (§2), I will give a general picture of the philosophical progress that has been made towards solving the descriptive problem so far (§2.1). Then I will briefly examine the origins from which the 'theoretical virtues' theory progressively developed, and explain in short the relevant foundational work done by philosophers like Aristotle, William of Ockham, and Thomas Kuhn (§2.2). After that, I will evaluate the most recent proposal to the descriptive problem to be found in the set of theoretical-explanatory virtues proposed by Keas, who in his work has taken into consideration, and also adopts elements from, most of the major solutions proposed before him (§2.3). By examining and evaluating his proposal, and going through each one of the twelve virtues he discusses in his proposal (§2.3.1-7), I will also be doing the necessary preparatory work and establishing the theoretical framework for studying and evaluating the proposals made to the normative, meta-descriptive, and meta-normative problems in the chapters that follow.

Chapter three, the first on virtue justification, will evaluate the most recent efforts that have been made on solving the normative problem. Analogously to what happened in chapter two, it will begin by explaining in more detail the difference and the relation between the normative and meta-normative problem for the 'theoretical virtues' theory (§3). Then, after a brief examination of Keas' claims on the justification of individual theoretical-explanatory virtues (§3.1, 3.1.1), a sub-problem which has been studied in quite great detail, attention will be shifted to proposals for solving the sub-

problem of the justification of the order of epistemic priority between these virtues, which unlike the former sub-problem has remained comparatively neglected (§3.2). After outlining and examining Keas' main positions on the epistemic priority between what he refers to as 'virtue classes', I evaluate his overall theoretical framework and normative proposals (§3.2.1). Afterwards, I will move on to the second of the two most developed proposals on solving the normative sub-problem on the order of epistemic priority between the virtues, namely, Douglas' normative proposal for resolving or 'dissolving' what she refers to as epistemic 'tensions' between the virtues, more accurately, between the four groups of virtues which she distinguishes (§3.3). First, I am going to discuss and evaluate her proposals with regards to tensions to be found between these four groups of virtues, the first two of which she characterizes as 'minimal criteria' and the other two of which she characterises as 'ideal desiderata' that explanatory theories can exhibit (§3.3.1). Then I will move on to Douglas' proposal concerning the tensions she observes within each of the two 'minimal criteria' groups of virtues, the tensions within the first of the 'ideal desiderata' groups, and, lastly, the tensions within the second of the 'ideal desiderata groups' (§3.3.2-4). Finally, I will end this chapter with an evaluation of Douglas' overall theoretical framework and normative proposals.

Chapter four, which will be the second on virtue discovery, will focus on proposals for solving the meta-descriptive problem. After outlining the sections that this chapter will consist in and the main aims for each section (§4), I will move directly into examining and evaluating the historical philosophy of science method and as a method of virtue discovery (§4.1); at which point I will go through what that evaluation shows to be its main limitations in adequately fulfilling this function (§4.1.1). Then I will shift attention to what is the latest methodological proposal for solving the descriptive problem, which is currently referred to as the 'experimental philosophy of science' method (§4.2), and give some examples of current philosophical research utilizing this method and briefly present some of the results (§4.2.1-2). Afterwards, I will go on to discussing what its main limitations are as a solution to the descriptive problem (§4.2.3), before

proposing some ways to further develop it in ways that would allow it to successfully overcome those limitations (§4.3). At that point I will begin by further examining the first limitation (§4.3.1), which mainly has to do with conceptual imprecision and even in some cases incoherence, before moving on to a proposal for overcoming this limitation by conceptually re-engineering the theoretical-explanatory virtues (§4.3.2). After that I will consider how conceptually re-engineering the 'virtues' into features, combinations of features, and principles, would also allow for overcoming the second limitation by allowing the method to effectively and efficiently discover the order of epistemic priority between them (§4.3.3). Lastly, I will examine some objections to the 'features and principles' version of the experimental philosophy of science method, namely, such that it would lead to very low rates of progress (§4.4), to which I will reply with a proposal for further methodological improvement that can potentially be achieved through adopting techniques from what is currently referred to as 'philosophical artificial intelligence' (§4.4.1). Against this approach, I will consider a further objection in response, namely, that this philosophical-AI version of XPhiSci would be unexplanatory and uninformative (§4.4.2). Finally, I will attempt to overcome this objection through proposing a further refinement to the method of philosophical artificial intelligence, drawing on an approach that is currently referred to as 'Inverse Reinforcement Learning'.

Chapter five, which will be the second on virtue justification, will mainly concentrate on examining and evaluating proposals to the meta-normative problem. First the chapter outlines the sections in which it is divided and the main purpose of each of these sections (§5). After which it begins with the main part of the analysis by an evaluation of HPS and XPhiSci this time not as methods of virtue discovery, but as methods of virtue justification (§5.1-2). Then, after considering the limitations of these two methods at the meta-normative level, the discussion moves on to the evaluation of the philosophical AI version of XPhiSci or 'φAI-XPhiSci', specifically the one utilizing IRL (§5.3). At that point an objection is raised that φAI-XPhiSci is in fact insufficiently normative as a method of virtue justification, to which I consider a reply that the

objection would be overcome through utilizing the reflective equilibrium method (§5.3.1). In response, a further objection is raised that aims to show that even then the method would be insufficiently normative because of concerns that due to its design it would then be bound to a narrow version of reflective equilibrium which has general normative limitations. To which a reply is offered showing that wide reflective equilibrium under certain conditions would overcome these issues and so the objection would be unsuccessful (§5.3.2-4). Chapter six will be the final chapter and will consist of some concluding remarks and also ideas for further research (§6-6.1).

CHAPTER 2

Virtue Discovery - The Descriptive Problem

2. The Descriptive and Meta-Descriptive Problems

In order for the 'theoretical virtues' theory to be descriptively adequate, one would need to first satisfactorily answer the prior question as to what the best method is for discovering the virtues, or the principles of epistemic priority that are actually used in rational explanatory-theory choice. For that reason, in a later chapter, I am going to examine the meta-descriptive proposals that have been forward so far, namely, the methods and theoretical frameworks of historical philosophy of science (HPS) and more recently of experimental philosophy of science (XPhiSci), and then consider potential ways for improving on them. However, in order to be able to evaluate and further develop solutions to the meta-descriptive problem, it would be useful to examine the descriptive results, the proposed virtues that is, that form the current presumptive solution to the descriptive problem. Thus in this chapter I will begin with a brief outline of the philosophical progress that has been made so far in solving the descriptive problem, and will mainly focus on examining the most current solutions that have been put forward in the form of a set of 'discovered' theoretical-explanatory virtues. I will then examine each of the virtues that have been proposed as part of this solution, and also evaluate them and the reasons for including them in the set of discovered virtues.

2.1. The Philosophical Progress So Far on Solving the Descriptive Problem

When one considers the philosophical development of the inquiry into the nature and function of the theoretical-explanatory virtues, one can differentiate three periods that can be demarcated by two significant leaps in philosophical methodology, and accompanying theoretical framework, which allowed for an increase in the rate of progress in this inquiry. The first period is characterised by a series of philosophical examinations of individual virtues such as simplicity or parsimony, but there was no explicit goal of discovering or justifying the full list of virtues of explanatory theories or a set of principles from which judgments on cases could be derived. The second

period, begins with the advent of the first of the two methodological and theoretical advances, or what is commonly referred to as the 'Historical Philosophy of Science' method or HPS, where we now have an explicit program of discovery and justification of the major virtues starting with philosophers like Thomas Kuhn. It is during this period that different virtue sets were proposed and there is a continuous examination of their role in philosophical and scientific practise, but also there are major debates of their normative status which I will examine in a later chapter. The third and final period can be demarcated by the advent of what is currently called the 'experimental philosophy of science' method or XPhiSci, which could be characterised as a more developed version of experimental philosophy applied to the philosophy of science and explanation. This method has not come to replace HPS but to complement it and add an empirical element to the search for virtues and their justification. As I will argue, neither HPS nor XPhiSci are fully developed in terms of method and theoretical framework, and after examining their achievements I will also consider their limitations in later chapters. One of the goals of this paper is to explore some ways to improve on these methods with the aim of hopefully increasing the rate of philosophical progress in this inquiry. As I will argue later, this can be made possible by certain advances that have been made in philosophical methodology but will require a conceptual re-engineering intervention with regards to how we formulate the theoretical-explanatory virtues themselves.

But before going on with the literature evaluation let us be clear once again as to what specifically we are looking for, and so choose our historical approach accordingly. To restate, the first question to be considered concerned which explanatory theories we find to be preferable to which others, and, as I mentioned before, the various answers to this question are judgments not principles. So one historical approach would be to outline the history of which major explanatory theories have been judged preferable and chosen over which other major explanatory theories at certain moments in time in the past in different fields of inquiry. This is largely a data-accumulation task, which is important, but is not the task I will engage in in this study. I will only refer to certain

evidence of historical judgments in the process of evaluating the various historically significant approaches to explaining those judgments. Another historical approach would be one that goes through the philosophical record with the aim to find answers that have been proposed with regards to the second question, namely, as to why an explanatory theory is preferable when it is preferable, or what makes it preferable; or what makes two explanatory theories equally preferable if they are judged to be so. This is the approach I will adopt for the rest of this chapter, and so I will go through the record of one particular viewpoint, namely, the one that explains explanatory-theory preference in terms of theoretical-explanatory virtues. That is roughly the position that certain explanatory theories are preferable to others because they exhibit certain features called virtues that their competitors do not. After clarifying this let me begin with briefly examining the origins from which the 'theoretical virtues' theory progressively developed.

2.2. Origins - a Brief Note on Aristotle, William of Ockham, and Thomas Kuhn

The main task of this study is not to do history of philosophy, that is, to provide a detailed picture of the development of the 'theoretical virtues' theory and engage in hermeneutic debates about textual evidence. However, no study on the function of theoretical-explanatory virtues would be complete without a brief mention of the philosophical origin of the concepts to be discussed, before leaping forward to the current debates. In the period before the development of HPS there were individual philosophers who as part of their inquiry explicitly adopted or proposed what we now call theoretical-explanatory virtues and many times also explicitly named them. An example of a name that lasts until today is ‘parsimony’ which is either to be considered as another name for ‘simplicity’ or a special case of simplicity depending on the context. It is important to note before delving into this part of philosophical history that most of this period is pre-scientific, set before Galileo, and even mostly set before 1833 around when the term ‘scientist’ started to be adopted into common usage. Thus it should be made clear that these questions are widely understood to be traditionally

philosophical problems and concerns, and are strongly grounded to this day within the scope of philosophical inquiry. However, what is noticeable about this period is that there was no discernible method that aimed specifically at discovering 'novel' virtues or examining the nature of our judgment of explanatory-theory preference; in other words, there was no 'theoretical virtues' theory as it were, such as the one we will see being formulated during the next period. Thus these philosophers did not consider and study the various theoretical-explanatory virtues as a whole, nor did they propose a complete set of such virtues for the purpose of studying them as a system. However, they did develop some kinds of proposals as to why for example theoretical explanations that displayed simplicity or parsimony were better explanations, whether that was the proposed simplicity of nature, or God's way of designing things, or something else. But this was done mostly for individual virtues, not for the totality of virtues. In other words, there was no philosopher who claimed that all virtues in their totality are virtues because of the way 'nature-out-there is' or because of the way 'God designs things'.

Taking the virtue of 'parsimony' as a case in focus, I will generally follow Sober in discussing the two major philosophers that comprise the originators of the important basic ideas about this virtue that survive to this day; Aristotle and William of Ockham.[11] According to Sober, Aristotle in his *Physics* states that when we try to explain a chain of motion "we ought [...] to suppose that there is one rather than many" unmoved first mover, that is to "assume only one mover, the first of unmoved things, which being eternal will be the principle of motion to everything else."[12] In other words, all other things being equal, when considering competing explanatory theories of a particular phenomenon of motion, prefer the one that has the feature of fewer causes. And why, because "we assume, basing our assumptions on what we see, that nature does

[11] Sober, Elliott, 2015, *Ockham's Razors: A User's Manual*, ch. 1 'A history of parsimony in thin slices (from Aristotle to Morgan)'.
[12] Sober, 2015, pp. 8-9; originally in Aristotle, 1995, *Physics*, 8.6.259a (following Sober's citation notation).

nothing in vain in so far as is possible in each case".[13] After Aristotle, the discussion was moved forward by the second philosopher of interest, William of Ockham. According to philosophers that came after him he held a version of what is referred to as the 'razor' principle, that "entities should not be multiplied beyond necessity."[14] As Sober observes, however, "Ockham didn't say much about the principle" and "relied on the fact that he and other philosophers found it sensible"[15]. It is worth pointing out that an important question remains, namely, why do we perceive nature as doing nothing in vain? In other words, why do we 'see' parsimony when attempting to explain nature? Now, the focus of this study is the 'theoretical virtues' theory which did not come to be until much later during the next period and, as I said before, the contribution of a complete historical evaluation between the times of Aristotle and Ockham and the period that followed would be minimal. So instead of pursuing this now, I will fast-forward to the next significant methodological leap that allowed the philosophical discussion to progress further.

What happened during the next period of philosophical interest was the development of a method and theoretical framework which laid the foundations for a new era of systematic study of theoretical-explanatory virtues. A method commonly referred to as "Historical Philosophy of Science" or "Philosophy of Science by Historical Means" (HPS).[16] An influential philosopher in that period was Thomas Kuhn who was one of the first to put forward a specific list of virtues, a proposed solution, in other words, to the descriptive problem of virtue discovery. The virtues he identified were "accuracy,

[13] Sober, 2015, pp. 6-8; originally in Aristotle, 1995, *Generation of Animals*, V.8.788b21; or alternatively, ibid., *Movement of Animals*, 2, 704b11-17, "nature does nothing in vain, but always does what is best, from among the possibilities".
[14] Sober, 2015, p. 5; originally in Ockham, William of, "it is futile to do with more what can be done with fewer" and "Plurality should not be posited without necessity".
[15] Sober, 2015, p. 5.
[16] Schindler, 2018, ch. 7, s. 7.2.2 'Kuhn on HPS and the Kuhnian Mode of HPS', p. 196-201; See also Thagard, Paul, 1988, *Computational Philosophy of Science*, ch. 7, s. 7.3 'Historical Philosophy of Science', pp. 115-119.

consistency, scope (unification), simplicity, and fruitfulness".[17] According to Kuhn an explanatory theory exhibits the virtue of accuracy, if "within its domain [...] consequences deducible from a theory should be in demonstrated agreement with the results of existing experiments and observations." By an explanatory theory having consistency he means both being consistent "internally or with itself" and also being consistent "with other currently accepted theories applicable to related aspects of nature." By having scope Kuhn means that an explanatory theory's consequences "should extend far beyond the particular observations, laws, or sub-theories it was initially designed to explain." A simple theory is one that brings "order to phenomena that in its absence would be individually isolated and, as a set, confused." Lastly, a fruitful theory is one that discloses "new research findings" or, in other words, "new phenomena or previously unnoted relationships among those already known." In their totality according to Kuhn these virtues comprise "*the* shared basis for theory choice" between "an established theory and an upstart competitor".[18] Since richer and more developed proposals have been put forward since Kuhn, instead of examining the original formulation and analysis of these virtues, it would be more conducive to achieving the goals of the study to proceed directly to evaluating the most current descriptive proposals which are nonetheless informed by it.

2.3. The Most Recent Major Proposals for Solving the Descriptive Problem

Some of the most recent proposals developed after Kuhn that endeavour to systematize the theoretical-explanatory virtues have been put forward by Keas, Douglas, Mackonis, and McMullin.[19] For the first few chapters the main works of reference will be *Systematizing the Theoretical Virtues* by Keas, in which he has drawn

[17] Keas, 2018, p. 2763; originally in Kuhn, Thomas, 1977, *The Essential Tension*, ch. 13 'Objectivity, Value judgment, and Theory Choice', p. 321ff.
[18] Kuhn, Thomas, 1977, p. 322.
[19] Keas, Michael N., 2018, *Systematizing the Theoretical Virtues*; Douglas, Heather, 2013, *The Value of Cognitive Values*; Mackonis, Adolfas, 2013, *Inference to the Best Explanation, Coherence and Other Explanatory Virtues*; McMullin, Ernan, 2014, *The Virtues of a Good Theory*.

on these previous works and attempted to bring them together and improve upon them, and *The Value of Cognitive Values* by Douglas, which I will evaluate mostly on the next chapter on the normative problem. To begin, in his paper Keas identifies twelve virtues and classifies them into four categories. The categories are: "Evidential virtues", "Coherential virtues", "Aesthetic virtues", and "Diachronic virtues".[20] In the first category, of evidential virtues, he lists: "Evidential accuracy", "Causal adequacy", and "Explanatory depth". In the second category, of coherential virtues, he includes: "Internal consistency", "Internal coherence", and "Universal coherence". In the third category of aesthetic virtues he lists, "Beauty", "Simplicity", and "Unification". Finally, in the fourth and last category, of diachronic virtues, he includes: "Durability", "Fruitfulness", and "Applicability". I will now examine them and evaluate each of them as to their potential for being part of a solution to the descriptive problem.

2.3.1. Evidential accuracy

Let us begin with the virtue of evidential accuracy which Keas sums up as the quality of "fit[ting] the empirical evidence well."[21] In more detail, let us assume there is a set of phenomena to be explained and two competing explanatory theories which both "fit" that set of phenomena equally well. Then we could say that, all other things being equal, those two theories are equally virtuous and we have no reason to prefer one over the other. Keas gives a "historical case study" as an example, namely the two competing astronomical systems of geocentrism and heliocentrism which were equal, or "roughly equal," with regards to this virtue and bar other virtues they were equally preferable.[22] In his words "if evidential accuracy were the only recognized criterion for theory choice at this time, then astronomers would have had insufficient reason to accept the heliocentric theory."[23] However, one may wonder whether evidential accuracy is a virtue at all. Can there be an explanatory theory that does not fit the

[20] Keas, 2018, pp. 2762-2763ff.
[21] Ibid., p. 2765.
[22] Ibid.
[23] Ibid.

empirical evidence? One could object that such a theory would be inadmissible, and so that evidential accuracy would not be a virtue to begin with. On the other hand, the way Keas conceived of it, evidential accuracy is not an all-or-nothing matter. So an explanatory theory that is not evidentially accurate is not necessarily evidentially completely inaccurate. There are gradations. So to clear Keas' point a bit further, we should be clear that a theory can be evidentially completely inaccurate which would actually make such a theory, not just not preferable, but unexplanatory and so inadmissible. Moreover, a theory can be evidentially somewhat inaccurate but still fitting some of the empirical evidence. And then a theory can be evidentially totally accurate, that is, if fits the empirical evidence perfectly well.

So theories that fit the empirical evidence less than perfectly well can be more or be less accurate than competing theories for the virtue of empirical accuracy to be applicable, but not completely inaccurate. Correspondingly, the explanatory theory that fits the evidence better is the more explanatorily virtuous one. However, one could reasonably ask whether there is more than one way for an explanatory theory to fit the evidence. In other words, are there different varieties of evidential fit? Keas does not say much about this point. He only mentions it as part of the example he gives that evidential accuracy has been "conceived since antiquity mainly as the match between mathematical astronomical theory (chiefly combinations of circular motions) and the observed apparent motions of planets."[24] That is, a theory fits the evidence in this sense if the mathematical or numerical descriptions or predictions derived from it concerning the object or phenomenon at hand, match the observations about that object or phenomenon. This is not the place to go into this in more detail, but if it were shown that there were other major ways that an explanatory theory could fit the empirical evidence, then more would need to be said as to whether one variety of evidential fit is prior to, or more otherwise important than, the others. If one kind was prior, then the theory that featured it would be more virtuous with regards to this virtue, if all kinds were equal, then the theories would be equally preferable or choice-

[24] Ibid.

worthy. Therefore, the important thing to note here on top of Keas' description is that there is a gradation and an order of epistemic priority within the virtue of empirical accuracy.

2.3.2. Causal Adequacy and Explanatory Depth

According to Keas' proposal, within virtue classes the virtues "sequentially follow a repeating pattern of progressive disclosure and expansion".[25] So the next virtue in the category of evidential virtues is disclosing something more about the explanatory theory than just how well it fits the evidence. Specifically, it says something about the causes that produce the effects or phenomena or empirical evidence that the theory aims to explain. That virtue is causal adequacy and a theory demonstrates this virtue if "it specifies causal factors that plausibly produce the effects in need of explanation."[26] In other words, it tells more of the causal story of how these phenomena came to be in the first place. Well, how is it different from evidential accuracy? A theory can fit the empirical evidence well but provide no picture of the causal mechanism behind the formation of that very same empirical phenomena. So two explanatory theories T1 and T2 can be equally evidentially accurate, but if T2 specifies at least one cause from which the set of phenomena to be explained can be derived, and T1 does not, then T2 is more casually adequate and so more virtuous overall.

The next virtue in this class, explanatory depth, takes it one step by incorporating into the analysis the number of levels of causal explanation that an explanatory theory displays. An explanatory theory has this virtue if, "it excels in [a] causal history depth *or* [b] in other depth measures such as the range of counterfactual questions that its law-like generalizations answer regarding the item being explained".[27] Let us unpack this quite dense analysis, there needs to be clarity about at least three things here: (1) the

[25] Ibid., p. 2762.
[26] Ibid., p. 2765; As clarified in the introduction, the explanatoriness of theories in science is assumed, but has been disputed, see Duhem, Pierre, 1914/1991; Russell, Bertrand, 1912/1992.
[27] Keas, 2018, p. 2766, emphasis and notation added.

concept of 'causal history depth', (2) what 'other depth measures' amounts to, and (3) what the role of counterfactuals is in this analysis. Causal history depth seems to be the most straightforward. Take a certain phenomenon to be explained: a person died. There are two competing explanatory theories concerning that phenomenon. Explanatory theory T1 says that the person died because a tree fell on them. Explanatory theory T2 says that the person died because a tree fell on them, the tree fell because of an explosion, the explosion happened because of a reaction between two substances which when combined tend to react in an explosive manner, and the substances were accidentally mixed together by a technician in a nearby lab.

Now, for the purpose of continuity with the previous discussion let us acknowledge that both T1 and T2 display the previous two virtues of empirical accuracy and causal adequacy. That is because they both "fit the evidence equally well", and they both "specify" some "causal factors that plausibly produce the effects in need of explanation".[28] However, T2, on top of everything else, displays the virtue of explanatory depth given that it "excels in causal history depth", and given this example, we can now begin elucidating the first of the three terms above. So the concept of 'causal history depth' means that each level of causal explanation is itself causally explained; in the case of T2, going up to four levels of explanation. Level 1: The death of the person is causally explained by the fall of the tree; level 2: the fall of the tree is causally explained by the explosion; level 3: the explosion is causally explained by a reaction between two substances that tend to react when mixed; and level 4: the mixing of the substances that tend to react is causally explained by the actions of the technician in the nearby lab. However, Keas notes that causal history depth, or explanatory depth, is further analysed in terms of "how far back in a *linear* or *branching* causal chain one is able to go".[29] So this virtue becomes even more complicated when thought of in this manner.

[28] Ibid., p. 2765.
[29] Ibid., p. 2766, emphasis added.

Let me try to unpack this a bit more. In the example, T2 was more virtuous with regards to explanatory depth than T1, but specifically, it was so because its "causal chain" went "further back" in a "linear" manner; there was no "branching". If a third explanatory theory T3 were introduced in the competition, which had all the virtues that T2 had, but whose causal chain were branching multiple times, then, then a number of questions could be raised. Questions as to what its status would be with regards to the virtue of explanatory depth, and what its competitive position would be in relation to the other theories. Keas does not go into more depth on this matter, but either one of three things can be the case. First, an explanatory theory whose causal chain is equally deep in linear terms but displays more branching than its competitor could be considered, all other things being equal, more virtuous. One could argue for this position by claiming that branches add to the explanatory depth of the theory or somehow offer a richer causal history for the phenomena to be explained. Unless, that is, the first position is nuanced by a condition that says that branching adds to explanatory depth only if the competing theories are equally causally deep in the linear sense first, and so branching becomes a secondary consideration. Second, an explanatory theory whose causal chain is equally deep in linear terms but displays more branching than its competitor could be considered, all other things being equal, less virtuous. Here one would find it difficult to argue for such a position within the scope and limits of the virtue of explanatory depth itself, that is, while holding all other virtues equal.

However, one could make a case for this position on grounds of another virtue, such as the one that Keas refers to as 'simplicity', which I will examine later on. Briefly, one could say that since both theories have the same number of levels of explanation but one of the theories branches much more, then the first one would be more virtuous because it is equally explanatorily deep but simpler. Unless, that is, the virtue of explanatory depth, both in linear and branching terms, takes total priority over simplicity in the order of epistemic priority between the virtues, and so considerations of simplicity become secondary. I will examine such tensions between virtues in a later

chapter when evaluating the work of Douglas.[30] The third and last position one could hold, is that causal chain branching makes no difference to the explanatory depth or simplicity of the theory. That is, the only thing that matters is the number of levels of explanation in a linear manner. Such a position would entail that the two explanatory theories would be equally virtuous, and so equally preferable. To recap, assume you have two explanatory theories T1 and T2. In T1 the phenomenon P is explained by cause C1, which is in turn explained by cause C2, which is in turn explained by cause C3. In T2, the phenomenon P is explained by cause C1, which is in turn explained by cause C2, which is in turn explained by cause C3 and also by cause C4 in a branching manner. So according to the first of the three positions above, the one where branching adds to explanatory depth, T2 is more explanatorily deep and so more virtuous than T1. According to the second position where branching renders the theory less simple, then a judgment in favor of T1 or T2 will depend on the position of simplicity in the order of priority between virtues. According to the third position, where branching does not affect a theory in terms of explanatory depth or otherwise, T1 and T2 would be equally virtuous and so equally preferable.

Now to make things even more complicated Keas notes that "explanatory depth comes in at least two varieties" one about "events" and one about "laws".[31] These relate to the three things we need to get some more clarity on, to restate, (1) the concept of 'causal history depth', (2) what 'other' depth measures amounts to, (3) what the role of counterfactuals is in this analysis. The first variety of explanatory depth concerning events Keas associates directly with the concept of causal history depth, which is the one that was discussed at length above. The second major variety of explanatory depth, the one concerning laws, is related to the other two things to be clarified above; Keas refers to it as the "Hitchcock–Woodward explanatory depth" variety.[32] With regards to that variety a theory has explanatory depth if it exhibits a

[30] Douglas, Heather, 2013.
[31] Keas, 2018, p. 2766.
[32] Ibid., p. 2766; originally in, Hitchcock, Christopher, and Woodward, James, 2003, *Explanatory Generalizations, part II: Plumbing Explanatory Depth*.

high level of generality or range "with respect to *other possible properties of the very object or system that is the focus of explanation*",[33] or otherwise "handles a larger range of counterfactual questions about the same kind of phenomena".[34]

Let us unpack this a bit further. So a theory has explanatory depth in this sense if it can tell a 'rich' causal story about the object or system that it is supposed to explain not just regarding its actual properties, but also with regards to variant possible properties which that object or system may have had. And here is where the counterfactuals come in. Keas gives the example of Newton's and Galileo's rival accounts of free fall.[35] Both Newton's and Galileo's theory accounted for objects free falling very near the earth's surface. But Newton's theory of free fall could also account for objects free falling very near the earth's surface, even had the earth counterfactually been different with regards to its mass and radius. In other worlds, in all possible worlds close to the actual one where the Earth is slightly more or less massive and its radius changed correspondingly, then, in that world, Newton's theory would be able to account for objects free falling near the earth's surface and Galileo's theory would not. In that sense, therefore, Newton's theory is more explanatorily deep than Galileo's.

Now that we have the two "varieties" of explanatory depth one may understandably wonder, what happens when one theory is more explanatorily deep in the 'causal history depth' sense and its rival is more explanatorily deep in the 'Hitchcock–Woodward explanatory depth' sense? Which one is more virtuous? Keas does not have a position on this in his work, but there are at least three options here. First, one could claim that 'causal historical depth' is higher in the order of epistemic priority than 'Hitchcock–Woodward explanatory depth' and so the first theory is more virtuous. One could argue a position like that for example by claiming that depth measures with regard to the actual world are more important than depth measures with regard to

[33] Keas, 2018, p. 2767; originally in Hitchcock, Christopher, and Woodward, James, 2003, *Explanatory Generalizations, Part II*, p. 182, emphasis original.
[34] Keas, 2018, p. 2767.
[35] Ibid.

close possible worlds. It is a potential way to argue which I only give for the sake of example, so I will not explore this further. Second, one could claim the reverse and argue that 'Hitchcock–Woodward explanatory depth' is higher in the order of epistemic priority than 'causal historical depth' and so the second theory in the above example is more virtuous. A position like this could potentially be supported by showing that being able to explain a phenomenon not only in the actual world but on close possible worlds is superior to just explaining that same phenomenon only in the actual world, even if a rival theory gives a richer causal history of the phenomenon in the actual world. Third and last, the two theories could be said to be equally virtuous. However, for this to be informative there would need to be some 'bridge principles' that show that the two depth measures can be usefully compared.

2.3.3. Internal Consistency

Let us now move to the next class of virtues, namely the "coherential theoretical virtues", and start with 'internal consistency'.[36] Keas sums it up as a virtue that an explanatory theory has when its "components are not contradictory"; specifically, "in the sense of formal logical coherence", not externally, but "within the theory itself".[37] The addition of this virtue initially seems pretty straightforward but upon closer examination one may find it difficult to comprehend how an explanatory theory that did not have this feature would even be up for consideration. In other words, one would reasonably expect that virtues are something that a theory can have, but also can be without. If an explanatory theory cannot but have something, then that something cannot plausibly be claimed to be a virtue, every explanatory theory would have to have it in order for it to be an explanatory theory. Thus one could argue that a proposed explanatory theory that is internally inconsistent is not admissible as an explanatory theory in the first place. Explanatory theory admissibility is prior to explanatory-theory preference or choice, that is, it is prior to any consideration of

[36] Ibid., p. 2769.
[37] Ibid., p. 2769-70; cf. Douglas, 2013.

virtue or goodness. If it is not good enough to be an explanatory theory, then there is no point in discussing whether it is a good theory or not, since a basic condition of admissibility has not been met.

Similarly, nobody has proposed a theoretical virtue of 'explanatoriness'. It is just assumed that it is a feature of every admissible theory.[38] A theory that did not explain the phenomenon that it is meant to explain would not just be a less virtuous theory, but an inadmissible explanatory theory with regard to that phenomenon. Let me give an example. Take the following phenomenon to be explained: a tree fell in the forest. An explanatory theory of this phenomenon can be: 'because somebody chopped it down with an axe'. On the other hand, 'because elephants are grey' is not an admissible explanatory theory with regards to the given phenomenon, because it is not explanatory and so not admissible in the first place. Neither are the explanatory theories 'because an elephant tore it down and elephants do not exist' and 'because elephants are grey all-over and green all-over'. The first one is contradictory or 'internally inconsistent' and so is unexplanatory, even though the first conjunct on its own would. The second one would also be unexplanatory for at least two individually sufficient reasons, one, it is internally inconsistent and, two, had it not been so, it would be unexplanatory because it explains nothing about the fallen tree. Therefore both would be inadmissible.

But here is another way of seeing the function of internal consistency altogether, one which is compatible with Keas' claim that it is a virtue of theories. Explanatory theories are mostly larger in scale and more complex than the ones discussed in the above examples. They have many different parts and sometimes some of those parts do not work well together; they produce inconsistency. We do not, normally, abandon a theory at the first sign of a minor inconsistency, dismiss it as inadmissible, and run to the next available competitor. That is because the theory may be rather virtuous in

[38] Once more, and as clarified above, although this is presupposed for the purposes this study, it has been disputed at least in the context of theories in science; cf. Duhem, Pierre, 1914/1991; Russell, Bertrand, 1912/1992.

many other respects and the minor internal inconsistency may be able to be worked out in time. Also, theory choice is about rationality as much as it is about formal logic. The theoretical-explanatory virtues, internal consistency included, are supposed to be, among other things, guides to rationality, and sometimes what is rationally required may not be cohering perfectly well with the principles of formal logic. For example, if a theory is internally logically inconsistent, then it is so in all possible worlds including the actual world. It is also the case that in those possible worlds where we have 'discovered' the inconsistency we may be rationally required under certain circumstances to dismiss the theory.

However, let us say that in the actual world none of the scientists or philosophers of science working on that theory have 'discovered' or even suspected that the theory is internally inconsistent; which is a common scenario. In this case the principles of formal logic would imply that we should abandon the theory, but the principles of rationality would imply that we should abandon the theory only if we discover an inconsistency. But even the latter principle has been questioned by philosophers like Gilbert Harman. In his work *Change in View* he considers whether what he calls the "Logical Inconsistency Principle" is a good normative principle of reasoning or rationality, or in his words, a good normative principle of belief "revision".[39] The principle implies that one should always avoid logical inconsistency. The version of the principle that would apply to rationality would be that one is always rationally required to avoid logical inconsistency in one's beliefs, views, or, we could add, explanatory theories.

The question is whether that principle is normatively plausible as a principle of rationality. Harman gives the following counterexamples:

> "To see that the Logical Inconsistency Principle has its exceptions observe that sometimes one discovers one's views are inconsistent and does not know how

[39] Harman, Gilbert, 1986, *Change in View*, ch. 2, p. 11.

> to revise them in order to avoid inconsistency without great cost. In that case, the best response may be to keep the inconsistency and try to avoid inferences that exploit it. This happens in everyday life whenever one simply does not have time to figure out what to do about a discovered inconsistency. It can also happen on more reflective occasions. For example, there is the sort of inconsistency that arises when one believes that not all one's beliefs could be true. One might well be justified in continuing to believe that and each of one's other beliefs as well."[40]

Thus there are at least two occasions that put stress on the normative plausibility of the principle specifically with regards to explanatory theories. First, an explanatory theory could have some components that are inconsistent with each other, and we may know it, but we do not know how to overcome that inconsistency without great cost. We could avoid building on, or drawing inferences from, the inconsistent components but otherwise keep holding the explanatory theory and using it to explain phenomena.

Second, we could believe that the theory is internally inconsistent, but we may have not figured out which of the components are mutually inconsistent. In both cases we may still keep the explanatory theory and therefore reject the version of the Logical Inconsistency Principle that is constructed for rationality. I will not analyze Harman's view any further, since the point of the discussion is the virtue of internal consistency and not the relationship of rationality and logic in general. The last thing to say on the matter is that thinking of this virtue more in the context of rationality and theory choice rather than simply formal logic, one could further claim that an explanatory theory can have more or less of it. Meaning that a theory which displays only minimal internal inconsistency between only a few of its components is to be considered more virtuous than a theory that displays extensive inconsistency between many of its components.

[40] Ibid., ch. 2, pp. 15-16.

2.3.4. Internal Coherence and Universal Coherence

Moving on, at the next level of progressive generality within the coherential virtues class Keas places the virtue of internal coherence. He analyses it as the virtue that an explanatory theory has if its "components are coordinated into an intuitively plausible whole."[41] In his exposition Keas quotes Schindler as claiming that this virtue "evades definition by necessary and sufficient conditions"[42] at least in its positive formulation. For that reason, he gives a negative, or "vice", formulation of the virtue as follows: "a theory lacks internal coherence to the extent that it incorporates ad hoc hypotheses."[43] In consequence, an explanatory theory that aims to be maximally virtuous with regards to this virtue must eliminate any ad hoc hypotheses it may incorporate. On the other hand, the more ad hoc hypotheses it has as part of its theoretical components the less internally coherent, and so more 'vicious', it is. A theoretical component such as a hypothesis is ad hoc according to Keas if it is "attached to a theory in order to solve an isolated problem", but it also falls under either of three criteria that render it "illegitimate".[44] The first criterion is that of being "insufficiently testable", and one way for a hypothesis to be this way is for it to be imprecise. Let me add that if it is sufficiently vague and the predictions you can derive from it are also vague enough, then it could be made to fit any kind of evidence and so it cannot be shown to be false; which renders it unfalsifiable. The second criterion is that of not explaining any other significant facts "beyond the data that prompted its construction", and the third criterion is that of not fitting with the other theoretical components of the theory coherently or properly, but only in an "awkward, arbitrary, or superficial" manner.[45]

---

[41] Keas, 2018, p. 2770.
[42] Ibid., p. 2770; originally in Schindler, Samuel, 2014, *Novelty, coherence, and Mendeleev's periodic table*.
[43] Keas, 2018, p. 2770.
[44] Ibid., p. 2771.
[45] Ibid.

Moving on, the coherential virtues progress expansively from the virtue of internal consistency, where explanatory theories are expected to not have contradictory theoretical components internal to the theory, to the virtue of internal coherence, where explanatory theories' theoretical components are supposed to cohere well with one another and not display ad hocness, to, finally, the virtue of universal coherence. Explanatory theories that exhibit this virtue are expected to cohere well with other theories or, in Keas' words, other "warranted beliefs", or at least to not be contrary to them in any obvious way.[46] In yet another formulation such a theory would be one that "sits well within one's total knowledge, especially the knowledge most firmly justified and most comparable to the theory in question."[47] To illustrate this virtue, Keas gives the example of two competing theories, namely, big bang cosmology and steady state theory. The latter did not cohere well universally with certain other theories and warranted beliefs, and was less virtuous, all other things being equal, than the former one. Specifically, steady state theory, in trying to be able to achieve one of its theoretical goals, held that new matter was being continually created. Which did not cohere with the theory of conservation of matter and energy, and generally violated the intuitions and arguments that stood behind it. The main thing to say in evaluating this virtue, is that, in its current formulation, it seems imprecise and "insufficiently testable", to use one of Keas' own criteria, and the principles for its application in deriving judgments of explanatory-theory preference are underdeveloped.

2.3.5. Beauty, Simplicity, and Unification

The third class virtues, called the aesthetic theoretical virtues, are also about things fitting together well. However, this time the emphasis on fittingness is not with regards to the theory's theoretical components fitting well together and with other theories

[46] Ibid.
[47] Ibid.

and warranted beliefs in a "logical-conceptual" way, but in an "aesthetic" manner by having a certain "shape", or "aesthetic fittingness".[48] The first of those virtues is beauty and an explanatory theory possesses it when it "evokes aesthetic pleasure in properly functioning and sufficiently informed persons", while allowing for some "degree of cultural and individual variation of aesthetic experience."[49] There is a lot to unpack here. The following components of the analysis should be clarified: (1) what is Keas' preferred account of aesthetic pleasure and aesthetic experience is; (2) what a properly functioning and sufficiently informed person amount to; and (3) what the impact and the role of a certain degree of cultural and individual variation in those qualities would be on the formulation and applicability of this virtue. Before doing that, let us note the three major examples of qualities that trigger the evocation of beauty. Keas names "symmetry", "aptness", and "surprising inevitability" as some of the "properties of theories and mathematical proofs" that have this effect.[50] He does not go into more details on those cases of beauty.

On the other hand, Keas raises the status of a further two "special cases of beauty" to being separate virtues themselves, namely, simplicity and unification.[51] He analyses them respectively as explaining "the *same facts* as rival theories, but with *less* theoretical content" and explaining "*more kinds of facts* than rival theories with the *same* amount of theoretical content."[52] He notes that these two virtues are complementary and are both about the "style of informativeness" of an explanatory theory, where the first one increases informativeness by reducing theoretical content relative to competitors, and the second one by increasing the number of kinds of phenomena that are explained, relative to competitors.[53] But what counts as less theoretical content? Keas mentions quite a few ways to analyse this concept. It can be

[48] Ibid., p. 2772.
[49] Ibid.
[50] Ibid., p. 2773.
[51] Ibid., p. 2775.
[52] Ibid., emphasis original.
[53] Ibid.

analysed as postulating “fewer entities,” fewer “kinds of entities”, “fewer laws”, or laws “relating fewer variables”; or as positing more “concise basic theoretical principles” or “fewer primitive explanatory ideas”; or as raising “fewer additional explanatory questions”.[54] I will examine some aspects of simplicity and unification in later chapters and evaluate some of their various formulations in more detail.

2.3.6. Durability

The last class of theoretical-explanatory virtues Keas calls ‘diachronic virtues’, and the reason for the name is that this class has a “distinctive temporal dimension” that the other three previous classes lack.[55] The first of these virtues is durability which a theory displays if “it has survived testing by successful prediction or by plausible accommodation of new unanticipated data (or both)”.[56] For a theory to be durable in this sense it has to be testable. So Keas notes two things about that, first, that testability is a “prerequisite” of durability or even potentially a “constituent of” it, and, second, that a theory’s durability ‘score’ after successful testing is proportional to its testability. On the other hand, an explanatory theory whose “predictions are disconfirmed”, or which introduces “ad hoc hypotheses” on the face of such disconfirming evidence, scores lower on durability.[57] By definition, at the moment of introduction, and so before any testing, it is impossible for an explanatory theory to have any durability score. That does not mean that the theory has a negative score with regards to the virtue, it just means that this virtue does not yet apply.

However, it seems natural to assume that it stands to lose when pitted against a competing theory that, all other things being equal, has a positive score with regards to this virtue. If later on the theory scores higher on durability, then the tables may be

[54] Ibid., pp. 2775-2776; cf. Beebe, James R., 2009, *The Abductivist Reply to Skepticism*, and Swinburne, Richard, 1997, *Simplicity as Evidence of Truth*.
[55] Keas, 2018, p. 2780; ch. McMullin, Ernan, 2014, *The Virtues of a Good Theory*, and Axtell, Guy, 2014, *Bridging a Fault Line*.
[56] Keas, 2018, p. 2781.
[57] Ibid.

turned. At this point, because of the temporal element, the question could be raised as to whether the theory will be able to 'catch up' with, or to supersede, an older theory. The answer to which would have to be positive, otherwise the formulation of this virtue would be rather implausible given that older theories would always have an unbeatable advantage, which we know is descriptively not the case. The reason why it is possible for this to happen lies in the analysis of durability. Things are not absolutely linear, and that is evident for at least two reasons. First, the competing theory may face disconfirming evidence in the future, or it may introduce ad hoc hypotheses, and so start scoring comparatively lower in durability. Second, a theory that is more testable than its competitor increases its durability score at a much faster rate that its less testable competitors even if they have not faced disconfirming evidence. Therefore, this virtue is built so as to withstand such potential criticism, and so retains its plausibility.

There is another concern about the analysis of this virtue that Keas provides, namely, something important may be missing in the description, specifically about testability. How do you compare two competing explanatory theories that have been successfully subjected to testing, but each theory has been subjected to qualitatively different tests? A quick answer would be that one should wait before both theories have been tested given the same tests before assigning each a score based on this virtue. However, this may not be possible, and I do not mean not possible in the sense that we do not have the resources to test both theories, but in the sense that there may be different standards of testing for different types of theories. That is, a case where, in principle, neither theory can be tested by the test that the other one can. In this case the virtue of durability may just be inapplicable. One could respond to this concern by pointing out that, if the theories cannot in principle be tested by the same tests then they are not actually in competition. This may actually imply that the theories are in fact explaining different and disparate phenomena, or there are otherwise incommensurable. Which is fair enough, but what happens when both theories can be subjected to the same kinds of tests but there are more than one applicable tests, and

the theories perform differently with regard to these different tests. Which explanatory theory would be more virtuous? What is missing in this case, which would allow us to satisfactorily answer this question, is a theory about the order of epistemic priority between the tests. If the tests are plausibly ordered according to their significance then the theories that do better on the tests that are higher in the order of epistemic priority would be, all other things being equal, more virtuous than competing theories which may have done better on lower ranking tests. In the end, the analysis suggests that virtue of durability can plausibly be defended as being part of a presumptive solution to the descriptive problem for the 'theoretical virtues' theory, however the conditions for its attainment are not fully described.

2.3.7. Fruitfulness and Applicability

At the next level of progressive generality within the diachronic virtues we find the virtue of fruitfulness, which a theory exhibits when "it generates additional discovery by means such as successful novel prediction, unification, and non ad hoc theoretical elaboration."[58] The difference between this particular virtue and the previous one has a lot to do with two varieties of prediction; which we could call 'normal prediction' and 'novel prediction' respectively. When an explanatory theory exhibits normal predictive success, that is when the predictions "formulated in the context of a theory's construction" are successful, or are "verified", then that theory gains in durability; in Keas' words the theory achieves "conservation" by successfully "passing tests to survive."[59] On the other hand, when an explanatory theory exhibits novel predictive success, that is, when the predictions "not conceived in conjunction with [the] theory's construction" are verified, or "confirmed by observation", then that theory gains in fruitfulness; in other words, it exhibits "innovation" by "stimulating further discovery".[60] In a more detailed analysis, McMullin conceptually analyses 'novel

[58] Ibid.
[59] Ibid.
[60] Ibid.

prediction' as the variety of prediction that, when successful, "would count as unexpected", and by that he means that "the novel result lies to some degree outside the scope of the data originally accommodated by the theory". He further notes that there are gradations in unexpectedness and therefore, we can assume, in the level of novelty.[61]

Finally, at the highest level of generality within the diachronic class we find the final 'discovered' virtue, namely applicability. An explanatory theory displays this virtue if it is "used to guide successful action" or used to "enhance technological control", and that action or control "provides more effective outcomes than what is possible in the absence of the theory."[62] This virtue belongs in the diachronic virtue class because there is a necessary temporal element that constrains applicability, since obtaining scientific knowledge is prior to applying it. When the explanatory theory is successfully applied "as the basis for a new or improved technology", and by 'successfully' it is meant that this technology 'works', only then can it be said to exhibit the virtue of applicability. Keas also notes that applicability is less often exhibited by explanatory theories of "how things originated", and more often by explanatory theories of "how things work". The reason for that is related to the fact that "experimentally controlled prediction" plays a lesser role in the theories of the first kind.[63] What Keas does not consider is whether applicability in certain ways, or certain areas, is more important with regards to these virtues than others.

For example, assume there are two explanatory theories T1 and T2 of a certain set of phenomena. Explanatory theory T1 is more applicable in the sense of guiding successful action, such us helping us craft and adopt more efficient and effective energy production and consumption policies, but explanatory theory T2 is more applicable in the sense of enhancing our technological control, as in helping us make

[61] Ibid., p. 2781; originally in McMullin, Ernan, 2014, *The Virtues of a Good Theory*, p. 505.
[62] Keas, 2018, p. 2785.
[63] Ibid., p. 2787.

more efficient batteries and energy storage units. It is unclear whether T1 or T2 are equally preferable in this case, or one of them is to be considered more virtuous or otherwise more preferable. As with a number of the other virtues evaluated above further elaboration on this matter is not provided to a satisfactory degree. Therefore, although fruitfulness and applicability can be defended as part of a presumptive solution to the descriptive problem of 'virtue discovery' given that there is no major flaw or objection undermining them, the fact that the principles for their application in rational explanatory-theory choice have remained underdeveloped is a significant point of concern. In the end, having examined the most recent literature on the theoretical-explanatory virtues discovered since the adoption of HPS, and evaluated them for their descriptive potential, I will now move on to examining the most recent normative proposals in the area of virtue justification.

CHAPTER 3

Virtue Justification - The Normative Problem

3. The Normative and the Meta-Normative Problems

Merely discovering all the theoretical-explanatory virtues, and even settling on the best method of virtue discovery, would not be immediately useful to scientific and philosophical inquiry. That is, it would not be enough to have a descriptively and meta-descriptively adequate theory describing all the features of explanatory theories that play a formative role in our judgments of explanatory-theory preference, and the principles that determine the order of epistemic priority within and between those features. Such a theory would be insufficiently normative unless the discovery aided in, or was accompanied by, the justification of a set of principles of epistemic priority that we could use in rational explanatory-theory choice, specifically, in determining which principles we should use when choosing between competing explanatory theories. The normative challenge here is two-part. The prior problem is meta-normative in nature, and is partly characterised by the question: how do we justify the virtues and the order of epistemic priority between them?

As explained in the introduction, this problem has priority because it has to be answered satisfactorily before one would be in a position to answer the subsequent normative problem. In a later chapter, I will consider the meta-normative status of the most significant of the recent methods of justification, some of which were used to derive the following proposed solutions to the normative problem discussed below. However, in order to do that it would be useful to first evaluate certain attempts that have been made at solving parts of the normative problem, and learn from their achievements and limitations. In question form, the normative problem can be formulated as follows: which virtues, and which principles describing the order of epistemic priority between them, are justified and should be used in rational explanatory-theory choice? So in what follows, I will briefly outline Keas' justification positions for the twelve major virtues, then move on to recent work by Douglas on the sub-problem of justifying the order of epistemic priority between some of the virtues, while evaluating their normative proposals.

3.1. Keas on the Justification of Individual Virtues

Given that Keas' work is heavier on the descriptive side, it is no surprise that many times he takes established normative positions on some virtues for granted, often slightly refining them or choosing between competing positions. For example, there is no explicit motivation specifically arguing for the normativity or justification of any of the individual virtues in the group of evidential virtues, which includes evidential accuracy, causal adequacy, and explanatory depth.[64] That is, apart from historical case studies and what Keas claims to be the "prima facie epistemic priority of evidence in theory choice".[65] The question could immediately be raised as to whether HPS, which relies on such historical case studies, is sufficiently normative to justify these and other virtues, but I will return to this issue later on when I discuss the meta-normative problem. With regards to the group of coherential virtues, which, to restate, consists of the virtues of internal consistency, internal coherence, and universal coherence, there is also no explicit argumentative support for the normativity of these individual virtues.[66] Again, except for an appeal to certain historical case studies and Keas' claim that, along with the evidential virtues, they are "widely understood to be of intrinsic epistemic value" and each of these virtues is, individually, "either a truth requirement or indicates the likely attainment of approximate truth."[67] The same goes for the group of diachronic virtues, that is the virtues of durability, fruitfulness, and applicability,[68] which Keas characterises as, "likewise", having "intrinsic epistemic value", since they "involve predictions that later are shown to have been approximately true beliefs about the future",[69] and thus constituting an "epistemically intensified means of theory development".[70] To which, however, unlike with regards to the evidential and

[64] Keas, Michael N., 2018, *Systematizing the Theoretical Virtues*, pp. 2765-2769.
[65] Ibid., p. 2769.
[66] Ibid., pp. 2769-2772.
[67] Ibid., p. 2772.
[68] Ibid., pp. 2781-2787.
[69] Ibid., p. 2772.
[70] Ibid., p. 2789.

coherential virtues, he adds the slight elaboration that they are also characterised by a temporal dimension which he considers to be of "considerable epistemic importance."[71] Keas does not explicitly argue why that is the case, but presumably it is because, among other things, they are at the same time of intrinsic epistemic value, so are truth-indicative, but cannot be possessed by newly created theories. Somethings that can play an important role later on when considering the order of epistemic priority between the diachronic virtues, on the one hand, and the evidential and coherential virtues, on the other.

3.1.1. The Normativity of the Aesthetic Class of Virtues

Things become a bit more complicated when Keas discusses the group of aesthetic virtues, comprised of the virtues of beauty, simplicity, and unification,[72] to which he ascribes "zero to modest" epistemic value.[73] I will examine his positions concerning this group in more detail since it is important for the discussion concerning normative issues between virtues. Because if it were the case that all, or some, of these virtues have zero epistemic value, then this would imply that they would be at the bottom of the order of epistemic priority in relation to other virtues, but if they have modest epistemic value then their position in the order would be less obvious. Also if they had zero epistemic value, then there would be little point in discussing the order of epistemic priority within the group itself. However, if they had modest epistemic value, then the order within the group would be an open question worth answering. Keas begins his discussion of this group of virtues by rejecting aesthetic relativism with regards to the virtue of beauty. He understands that position as implying that "no judgments about beauty or ugliness [...] are more correct than others", which, if correct, would render it "difficult to see how any aesthetic theoretical virtues could be of rational importance in theory choice".[74] What Keas seems to imply here is that there

[71] Ibid., p. 2780.
[72] Ibid., pp. 2773-2780.
[73] Ibid., p. 2788.
[74] Ibid., p. 2773.

would be no question of whether a theory that exhibited beauty, simplicity, and unification were more probably true, or more worth accepting, than a competing theory, since the normativity of these virtues would be undermined by epistemic relativism. However, Keas does also not hold the reverse position that since beauty is not subject to aesthetic relativism, then the virtue of beauty is thus normatively justified to be used in rational explanatory-theory choice.

On the contrary, he holds that none of the virtues in this group have any "intrinsic epistemic value", as do the evidential and coherential virtues, or, as I will discuss later, the diachronic virtues do. He does, on the other hand, endeavour to ascribe "extrinsic" epistemic value to them, or at least some of them. By which he means that each of them "promotes, *without indicating*, truth attainment", even though they are not a requirement for truth.[75] At this point, one could object that the way one discovers a virtue should imply its level justification. In other words, if beauty had neither intrinsic nor extrinsic epistemic value, then why would it be a virtue to begin with? A reply could be that a virtue can be discovered, say, by its being observed to be the reason, or part of the reason, that expert scientists and philosophers of science have chosen an explanatory theory over its competitors, but that the normative work is a separate, even though still related, task. So beauty features in many cases of explanatory-theory choice, but its epistemic value, whether intrinsic or extrinsic, can still be an open question. That question, in turn, may need to be answered through a meta-normative approach that is very different from the meta-descriptive approach that was used to discover it. Meaning that even though we know that this virtue is being appealed to by the experts in attempting to justify their judgments in actual cases of explanatory-theory choice, we may not yet be justified in claiming that it should be adopted as a principle of rational explanatory-theory choice. I discuss this point in more detail in the chapter on the meta-normative problem.

[75] Ibid., p. 2772, emphasis added.

But then the question still remains as to how beauty can promote truth without indicating it. Keas answers this question by reasoning as follows: first he claims at different points in his work that simplicity and unification are at least "epistemically relevant in theory choice",[76] and, according to some accounts, of "extrinsic epistemic value";[77] and even goes as far as allowing for them to be "(possibly) intrinsically *epistemic*".[78] Then he adopts the "contention"[79] that these two virtues are "special cases"[80] of beauty, or, in a different formulation, that "general aesthetic experience inclines researchers toward recognizing and cultivating simplicity and unification as special kinds of beauty"[81]. Lastly, from all this he concludes that the virtue of beauty therefore "likely possesses extrinsic epistemic value", which, if it follows at all, would not be very informative. Generally, what is important to keep in mind, especially for the meta-normative issues that will be discussed later on, is Keas' appeal to scientists' judgments and actual scientific practice, for both the epistemic value of simplicity and unification and the extrinsic epistemic value of beauty. Much can be said about this, however, since a lot of work has been done on normative issues with regards to the justification of individual virtues, but there has not been significant progress made in the justification of the order of epistemic priority between them, it would be more fruitful to move on to examining the latter part of the normative problem. Therefore, in sum, this is the brief outline of the main normative positions Keas adopts in his proposal concerning the twelve virtues individually. It will be useful to have in mind when evaluating his and others' normative positions concerning the order of epistemic priority between them, which I will now begin.

3.2. Keas on Justifying the Order of Epistemic Priority Between Virtue Classes

[76] Ibid., p. 2775.
[77] Ibid., p. 2777.
[78] Ibid., p. 2775, emphasis original.
[79] Ibid., p. 2772.
[80] Ibid., p. 2772.
[81] Ibid., p. 2775.

Keas claims that "virtues are the traits of a theory that show it is *probably true* or *worth accepting*",[82] but also goes a step further in claiming that the first three out of the four theoretical-explanatory virtues classes he demarcates, respectively, the evidential, coherential, and aesthetic classes, are "arranged in decreasing order of epistemic weight."[83] On a sidenote, the fourth class, namely the group of diachronic virtues is a special case due to its temporal dimension and he finds its position in the order a "thorny issue" which his systematization "does not settle".[84] Now the first part of the claim above, also implies that he is noncommittal with regards to the relation between normativity, or justification, and truth, since an explanatory theory can presumably 'be worth accepting' without it being 'probably true'. Which of course does not mean that if it were 'probably false' it would be worth accepting. Moreover, in the second part he states that the order of epistemic priority is *between classes*, which further implies that he has no explicit positions with regards to the order of epistemic priority *within classes*, the order *within virtues* or *between virtues* in general, or the order between *combinations of virtues*. The only thing he says about the relation of virtues within each class is that they "sequentially follow a repeating pattern of progressive disclosure and expansion", and unless he implicitly assumes that a virtue that discloses more, and is more expansive, is thereby higher in the order of epistemic priority than competing virtues that are less so, then we can assume that he is noncommittal on this matter too.[85] Also, to restate, another distinction Keas makes is between "intrinsic" and "extrinsic" epistemic value, where virtues with intrinsic epistemic value either "indicate the likely attainment of approximate truth" or are "a requirement for truth".[86] On the contrary, a virtue with extrinsic epistemic value "promotes the attainment of truth without itself being an indicator or requirement of

[82] Ibid., p. 2761, emphasis added.
[83] Ibid., p. 2772.
[84] Ibid., p. 2783.
[85] Ibid., p. 2762.
[86] Ibid., p. 2772.

truth."[87] It would be useful to keep these distinctions in mind during the discussion that follows.

3.2.1. Evaluation of Keas' Theoretical Framework and Normative Proposals

Let us take a closer look at Keas' claims about the order of epistemic priority between his virtue classes and evaluate them in turn. The first claim of interest is that all of the evidential virtues are higher in the order of epistemic priority than all of the coherential virtues, and that all of the coherential virtues are in turn higher in the order of epistemic priority than all of the aesthetic virtues. This means that an explanatory theory T1 that scores higher on either of the theoretical-explanatory virtues of evidential accuracy, causal adequacy, or explanatory depth is higher in the order of epistemic priority than any competing explanatory theory T2 that scores lower on these virtues. Importantly, not just all other things being equal, but even if T2 scores higher than T1 on either or all of the virtues of internal consistency, internal coherence, universal coherence, beauty, simplicity, or unification. It also means that if any explanatory theory T3 scores higher on either of the virtues of internal consistency, internal coherence, or universal coherence, it is higher in the order of epistemic priority than any explanatory theory T4 that scores lower on those, even if it scores higher on either or all of the virtues of beauty, simplicity, or unification. Therefore, T1 and T3 would be more probably true, or more worth accepting, than T2 and T4 respectively given the virtues each displays. However, if this were accepted to be the case, then the following would also have to be accepted to be case: If explanatory theory T5 scores higher than its competitor on evidential accuracy but lacked internal consistency, lacked internal and universal coherence, lacked beauty, simplicity, and unification, then it would more probably be true or more worth accepting than a competing explanatory theory T6 that scored lower on evidential

[87] Ibid.; originally in Steel, Daniel, 2010, *Epistemic Values and the Argument from Inductive Risk*.

accuracy, but excelled in all those other virtues that T5 lacked; which seems counterintuitive.

One philosopher who would have to support such a position would be Ernan McMullin who claims that evidential accuracy is "primary" among the theoretical-explanatory virtues, and one of a kind, since, in his words, "account[ing] for the data already in hand" is the "first requirement" of an explanatory theory.[88] On a sidenote, Laudan and Douglas, in contrast to McMullin, would also add internal consistency in this category.[89] Specifically, Douglas claims that these two virtues are "minimal criteria" and that they "come first, and both must be met."[90] But let us consider McMullin's stronger position first to see if it holds. Part of the point McMullin is making can be captured in the following statement by Richard Feynman: "It does not make any difference how beautiful your guess is. It does not make any difference how smart you are, who made the guess, or what his name is - if it disagrees with experiment it is wrong."[91] Of course McMullin is making an even stronger claim than that. His position implies that the virtue of evidential accuracy has the highest epistemic weight, and so is ranked highest in the order of epistemic priority. The explanatory theory that displays it would be, according to Keas' normative view, more probably true, or more worth accepting, than all competitors who do not. Of course one could object that this is not how things are normally done, that is, we do not abandon an otherwise highly virtuous explanatory theory at the first sign of trouble. There are potentialities that we have to take into consideration, such as that there may be things wrong with the design of the experiment and that may be the reason the explanatory theory lacks evidential accuracy. However, such criticism would be beside the point. As Feynman clarifies, the theory is to be deemed wrong if it disagrees with experiment, "after the experiment has been checked, the calculations have been checked, and the thing has

[88] Keas, 2018, p. 2764; originally in McMullin, 2014, pp. 563-564.
[89] This according to Douglas, Heather, 2013, p. 798; originally in Laudan, Larry, 2004, *The Epistemic, the Cognitive, and the Social*.
[90] Douglas, 2013, p. 801.
[91] Feynman, Richard, 1965/1985, *The Character of Physical Law*, p. 156.

been rubbed back and forth a few times to make sure that the consequences are logical consequences from the guess, and that in fact it disagrees with a very carefully checked experiment."[92] That is, with regards to the example above, the claim that the explanatory theory that is evidentially accurate is preferable to the one that is not, but has all these other virtues at the highest degree, already assumes this. Namely, that the evidential accuracy of T5 and T6 have been determined *after* checking for experimental or other kinds of errors. Therefore, although the position may initially seem counterintuitive under closer examination it seems more plausible. Moreover, if one followed Laudan and added internal consistency to T5, but not T6, on top of evidential accuracy, then the resultant 'weaker' claim in favour of T5 would be even more plausible.

Things become more difficult to discern when one considers what may be characterised as a more controversial implication of Keas' position on the order of epistemic priority between the virtues. If the virtue classes are in an absolute order of epistemic priority, then that means that each of the virtues within a particular class is higher in the order of epistemic priority than all of the virtues within the class beneath it. Thus the position also implies the following. Assume that two competing explanatory theories T7 and T8 equally excel in evidential accuracy, causal adequacy, and explanatory depth from the evidential virtues class, and also in internal consistency from the coherential virtues class. Further assume that T7, on top of that, excels in internal and universal coherence while exhibiting none of the aesthetic virtues, and that the reverse is the case for T8; that is, T8 excels in all of the aesthetic virtues while exhibiting neither internal nor universal coherence. Then according to Keas' normative position T7 would be more probably true, or otherwise more worth accepting, than T8. This is more controversial given the value scientists and philosophers of science have placed on simplicity and unification. One could object that the above described implication is not actually counterintuitive, and also that it has not explicitly been argued in the literature that considerations of ad hocness are

[92] Feynman, 1965/1985, pp. 156-157.

less significant in epistemic value matters than considerations of simplicity and unification. Which is an important observation but does not resolve the issue, since one could slightly alter the example by adding excellence in the virtue of universal coherence to T8 too. In that case the claim implied by Keas' position would be that, all other things being equal, an explanatory theory that "sits well with (or is not obviously contrary to) other warranted beliefs", is more likely to be true or otherwise more worth accepting than an explanatory theory that does not sit well with other warranted beliefs or explanatory theories. Even given the presumption that the latter theory "evokes aesthetic pleasure in property functioning and sufficiently informed persons", explains "the same facts" than the former theory "with less theoretical content", and, on top of that, "explains more kinds of facts" without adding theoretical content.[93] That is in fact counterintuitive, since it seems more likely that the fact that the latter theory is not sitting well with other warranted beliefs we may hold, shows that those other beliefs need to be re-examined. Which would count against Keas' position. In conclusion, the analysis so far indicates that although Keas' work has elucidated many parts of the normative problem, his proposals and presumptive positions on virtue justification cannot be defended effectively and need to be further developed.

3.3. Douglas on Justifying the Order of Epistemic Priority Between the Virtues

Another philosopher who has endeavoured to resolve parts of the problem of justifying the order of epistemic priority between the virtues is Heather Douglas. In her paper *The Value of Cognitive Values*, which is how she refers to part of what in this study is referred to as 'theoretical-explanatory virtues', she recognizes that there are "clear tensions and trade-offs among the various values", as for example when one "might gain in simplicity but lose scope".[94] She sees this task as one of "reducing the tensions among the values", which is a less ambitious goal than discovering and

[93] Keas, 2018, p. 2762.
[94] Douglas, Heather, 2013, p. 797.

justifying the full order of epistemic priority between them.[95] Douglas distinguishes the set of virtues she considers into four distinct groups each with a different epistemic status, and her set does not exactly overlap with Keas'. However throughout the exposition and evaluation there will be notes on which of Douglas' virtues match Keas' virtues, and which groups are similar to Keas' groups. In what follows I will first discuss Douglas' claims about the tensions between the four groups of virtuous she differentiates. Then I will consider her positions on the tension within the first two groups of virtues, and afterwards, her positions on the latter two groups of virtues. Lastly, I will evaluate Douglas' overall theoretical framework and normative proposals.

3.3.1. Tensions Between Douglas' Four Groups of Virtues

Douglas locates the first instance of tension as being between evidential accuracy, which she refers to as 'empirical adequacy', and internal consistency, which, in contrast to Keas, she actually groups together with empirical adequacy, and considers both as minimal criteria applied respectively to "the theory in relation to evidence"; specifically to "existing evidence, not all possible evidence," and to "the theory per se".[96] As I mentioned above she 'resolves' this tension by placing those two virtues jointly at the highest point in the order of epistemic priority between the virtues, by claiming that they "come first, and both must be met."[97] Keas would not be in agreement with that since his position is that epistemic adequacy, or 'evidential accuracy' as he calls it, belongs to the class of evidential virtues which he claims to be higher in the order of epistemic priority than the class of coherential virtues, which is where he places internal consistency. Douglas recognizes that in scientific practice "scientists may still choose to pursue the development of a theory [...] even in the face of failings in [...] minimal criteria", if the theory nevertheless exhibits what Keas refers to as 'simplicity' and 'unification'. However, she adds that in that case they must do it

[95] Ibid., p. 801.
[96] Ibid., pp. 799-800.
[97] Ibid., p. 801.

"with full acknowledgment that the theory is inadequate as it stands and must be corrected [...] as quickly as possible."[98] One of the reasons for this claim is that Douglas holds that these virtues "give no assurance as to whether the claims that instantiate them are true", but only that "we are more likely to hone in on the truth".[99] However, I will reconsider this position in the evaluative analysis of the claims made concerning questions of epistemic value of individual virtues later on. Thus the solution to the first kind of tension that Douglas considers between empirical adequacy and internal consistency against other virtues is similar to the one offered by Laudan, namely, joint epistemic priority of both virtues over all other virtues.[100] Which is a different position than Keas, since Douglas recognizes no tension between the two virtues themselves while Keas claims that evidential accuracy is epistemically prior to internal consistency.[101]

The second instance of tension Douglas locates between a further two groups of virtues. Those virtues that she claims to be "instantiated by theories only", and those "instantiated by the relations between theories and evidence", as for example the virtue of fruitfulness which Douglas refers to as "novel prediction".[102] The ones in the first of those groups do not "provide a reason to accept a theory as well supported or true or reliable" and are "simply not epistemic", while the virtues in the second group "do have genuine epistemic import."[103] At least for this reason one would assume that Douglas would hold the position that the second group of virtues, the ones that are instantiated by the relations between theories and evidence, would be more epistemically weighty, or be higher in the order of epistemic priority, than the first group of virtues, the ones that are instantiated by theories only. After all, if a virtue is epistemic then it should be more heavily epistemically weighted than a virtue that is

[98] Ibid., p. 802.
[99] Ibid., p. 800.
[100] Laudan, 2004.
[101] Keas, 2018.
[102] Douglas, 2013, pp. 800-801.
[103] Ibid., pp. 802-803.

"simply not epistemic." Strangely, Douglas does not hold this position. She claims that there is no tension at all between these two groups, or in her words "any apparent conflict dissolves", since the two groups "aim at different purposes".[104] Specifically, she says that if one aims at satisfying one's need of "epistemic assurance" with regards to what is our "best available knowledge at the moment", then the virtues of the second group apply, but the virtues of the first group "have no bearing" at all, given that they are only applicable when one "wants to justify future research endeavours."[105]

However, this position does not hold very well. Presumably, when one wants to justify future research endeavours, one needs to be able to show that the explanatory theory that they suggest be further elaborated, or experimentally tested, is "more probably true" or otherwise "more worth accepting" that its competitors. That is, one needs to show that the explanatory theory under consideration is more epistemically weighty than its competitors. That could mean either one of the following two things. One, virtues belonging to the first group, instantiated by the theory only, have some epistemic weight, or, two, they are in fact not applicable in situations where one wants to justify future research endeavors. In any case, to conclude, given that the claim that Douglas bases her position on seems somewhat unclear and arguably even implausible, we are left with no good reason to hold that the conflict has dissolved. It may however be resolved. One way for it to be resolved would be to agree with Douglas' assessment of the epistimicity of the virtues involved, but hold that virtues in the second group, like novel prediction, are higher in the order of epistemic priority than virtues in the first group, since the latter have no epistemic standing. Another way would be to hold that the latter group of virtues does have epistemic standing, and then resolve the tension in favour of either one of the two groups.

3.3.2. Tensions Within the Two 'Minimal Criteria' Groups of Virtues

[104] Ibid., 2013, p. 804.
[105] Ibid., 2013, p. 804.

Douglas also examines tensions *within* these groups. Just to recap, the first two groups of virtues that were discussed, are composed of virtues that Douglas claims are "minimal criteria" for explanatory theories. These are respectively instantiated by "the theory per se", such as the virtues of internal consistency, or instantiated by "the relation between the theory and the evidence", such as the virtue of evidential accuracy or "empirical adequacy" as she calls it. Concerning those two groups, Douglas claims that there is no tension between them, which is presumably because, as she claims, they are equally mandatory for theories to have so they have equal epistemic value, and also, because they are jointly higher in the order of epistemic priority than all other groups of virtues. She further claims that there are no tensions within those groups, however, she provides no explicit support for that claim. One has to assume then that this must be because by her analysis the virtues in those groups are "necessary" for an explanatory theory to have, so there could not be any tension between the members of each of the groups.[106] Which is acceptable if one assumes that they are actually necessary requirements, that there are no gradations involved, and, if gradations were involved, that an explanatory theory would have to exhibit those virtues at the maximum degree or be deemed not to possess them at all. But this is not an obvious truth about those virtues and Douglas does not provide any support as to why these are the case.

As discussed before, Keas in his virtue set conceives of evidential accuracy and internal consistency in ways that allow them to have gradations. With regards to evidential accuracy, which Keas analyses as the virtue that an explanatory theory has when "it fits the empirical evidence well", Keas does not explicitly state that there are gradations.[107] However, a statement he makes in the example he gives to illustrate this virtue implicitly assumes that there are. When discussing two competing explanatory theories in astronomy, the geocentric astronomical system and the heliocentric

[106] Ibid., p. 804.
[107] Keas, 2018, p. 2765.

astronomical system, he says that "[p]rior to Galileo's telescopic discoveries" their evidential accuracy was "roughly equal."[108] So stating that their evidential accuracy was roughly equal means that for Keas there are various levels of accuracy, and also that various levels of accuracy are acceptable even though more accuracy is considered preferable to less accuracy. With regards to internal consistency Keas clarifies that he considers internal consistency in relation to the explanatory theory' various theoretical components. Again it is not an all or nothing matter for him since a theory can be mildly inconsistent, say when two minor components of the theory, such as two minor hypotheses, are inconsistent with each other, which is something that could be worked out later on. However, the explanatory theory could also be maximally or severely inconsistent when, respectively, none of its components are consistent with each other, or when its major components such as a group of hypotheses that are integral to the theory are inconsistent with each other.

It should be stressed once more that in the framework within which we are discussing, these virtues are examined with regards to principles of rationality not merely with regards to principles of logic. And as I discussed in a previous chapter, philosophers like Harman have argued that these two sets of principles do not necessarily overlap in the case of logical inconsistency. Therefore, there is good reason to believe that both the virtue of evidential accuracy, and that of internal consistency, exhibit gradations. In that case Douglas' claim that there are no tensions between the group that contains evidential accuracy and the group that contains internal consistency could be undermined. At least because it is not clear what holds when two rival explanatory theories that, all other things being equal, both display the two virtues but at different degrees. For example, if explanatory theory T1 is maximally evidentially accurate but only minimally internally consistent, and explanatory theory T2 is maximally internally consistent and just minimally evidentially accurate, then which one is more virtuous? Or, given that according to Douglas the two virtues are both epistemic, which one is higher in the order of epistemic priority? The answer is not immediately clear, and so

[108] Ibid., p. 2765.

Douglas is wrong to assume that there is no tension between the two groups. Furthermore, given that Douglas allows that there are other virtues in each of the groups, it is also unfounded to hold that there is no potential tension within each of the groups. Depending on which virtues those are they may be questions as to which are epistemically weightier, so potential tension within the groups cannot be excluded. In sum, it is not the case that there are no tensions between or within the two groups of virtues that are minimal criteria for explanatory theories, and so questions as to the order of epistemic priority between these virtues remain unanswered.

3.3.3. Tensions Within the First of the 'Ideal Desiderata' Groups of Virtues

Moving on, the latter two groups Douglas characterises as "ideal desiderata", as opposed to "minimal criteria". Her position is that, within the first of those groups of virtues, those "instantiated by theories only", there are tensions, but those are "productive tensions". Within the second group, the virtues "instantiated by the relations between theories and evidence", after analysis, "there remain some tensions", as for example, between what she calls the virtues of "explanatory coherence" and "novel prediction".[109] I now will examine this position in more detail. Douglas claims that the virtues in the first group are not epistemic and serve only "pragmatic" concerns with regards to what she refers to as the "fruitfulness of the theory" which exhibits them, and has to do with "the ease with which scientists will be able to use the theory in new contexts" or "to devise new tests" in order to "refine, revise, or if need be overhaul completely the theory."[110] This is different from the virtue of fruitfulness in Keas' set. Now since Douglas ascribes no epistemic dimension to this class of virtues, there is, by implication, no epistemic tension between them, and so questions about the order of epistemic priority between them would not even arise.

[109] Douglas, 2013, pp. 798, 804.
[110] Ibid., p. 802; for the virtue of fruitfulness as part of the class of diachronic virtues see Keas, 2018, pp. 2762, 2781-2783.

It is important to note that Douglas makes a peculiar move which I will now need to make explicit for the purposes of evaluation. As with the previous two groups of virtues, the ones that are minimal criteria, she makes a distinction with regards to virtue instantiation between, one, the explanatory theory itself and, two, the *relation* between the explanatory theory and the evidence. The implications of the distinction for the first two groups are not immediately apparent. That is because, the virtues are neatly separated into those virtues that are instantiated by the explanatory theory, such as internal consistency, and those that are instantiated by the above-mentioned relation. But there is no crossover. However, the implications of this distinction for the latter two groups of virtues is more clear and impactful, since it provides the basis for Douglas' position that, on the one hand, the group of virtues that are both ideal desiderata and are also instantiated by the theory per se, are non-epistemic. While, on the other hand, the group of virtues that are both ideal desiderata and are also instantiated by the relation between the theory and the evidence, are epistemic. The issue becomes even more complicated with regards to those groups since, according to Douglas' analysis, there is significant virtue crossover between these groups. She gives as examples the virtues of scope and simplicity. In brief, the virtues of simplicity and scope, when instantiated by the relation between the explanatory theory and the evidence, correspond almost exactly to Keas' virtues of simplicity and unification. When they are instantiated by the theory itself, they do not correspond to any virtues that we have seen in the virtue sets considered so far; it seems to be sui generis.

In more detail, with regards to scope Douglas supports her distinction by differentiating between actuality and potentiality. On the one hand, when the virtue of scope is instantiated merely by the explanatory theory itself, then that theory "might have the *potential* to apply to lots of different terrain or to wide swaths of the natural world (i.e., the claims it makes are of broad scope)".[111] On the other hand, when this virtue is instantiated by the relation between the explanatory theory and the evidence,

[111] Douglas, 2013, p. 799, emphasis added.

then that theory successfully applies to "actual evidence" in fact, and so actually "explains a wide range of evidence and phenomena".[112] Metaphysically it must be the case that, if with regards to actuality the virtue of scope is instantiated by the relation between the explanatory theory and the actual evidence, then with regards to potentiality the virtue of scope is instantiated by the relation between the explanatory theory and potential evidence. Which may provide reason to doubt the need to assume that this virtue can be instantiated by the explanatory theories per se, or the need to argue for such a strong distinction; and may also provide reason to hold that a distinction, if necessary at all, should instead be made between the virtue of scope being instantiated by the relation between the explanatory theory and actual evidence, and virtue of scope being instantiated by the explanatory theory and potential evidence.

With regards to simplicity Douglas does not directly discuss the distinction between potentiality and actuality and she is very brief in her description. However, she does differentiate between simplicity "describing a relation to evidence", or describing "a theory that is simple with respect to the complex and diverse evidence that it captures", on the one hand, and "a simple theory", on the other hand.[113] So if one assumes that by 'the evidence that it captures' Douglas means 'the evidence that it *actually* captures', and at the same time one takes into consideration the potential-actual distinction with regards to applicability that she relied on when analysing the virtue of scope, then one could plausibly interpret Douglas as implicitly basing her differentiation between simplicity instantiated by the theory and simplicity instantiated by the relation between the theory and the evidence on the same potential-actual distinction. The alternative would be to assume that Douglas bases this differentiation on nothing at all, which would be less hermeneutically generous. However, there are two issues with such an attempt. First, it is not clear how to do that, and second, if successful, one would also have to analyse the virtue of internal

[112] Ibid.
[113] Ibid., p. 799.

consistency in terms of the potential-actual distinction, or show that, even though the distinction does not apply, there is an alternative way to support its analysis. However, I will postpone evaluation of all these distinctions for until after the exposition of Douglas' position on the tensions within these two groups of virtues.

So what kind of tension does Douglas claim there is within the first of these groups, that are instantiated by the theory itself, and why does she consider them productive tensions? The tension at issue here is a tension relating to fruitfulness, as described above, relative to different scientists. In other words, different scientists find different virtues to be more fruitful than others in, say, designing new experiments. Some find a simpler theory more fruitful, some find a more unifying theory, or in her terminology, a theory with broader "scope", to be more fruitful, in her sense.[114] The reason Douglas considers this tension productive is that, "having diverse efforts in scientific research is a good thing for science [...] and is crucial for the eventual generation of reliable knowledge".[115] A claim which she bases on work in the "social epistemology" of science.[116] In response, one could raise the question as to whether there in fact is such a kind of non-epistemic tension, and, if there were, whether it would be productive or not. However, given that the main point of focus of this chapter is the order of epistemic priority between the theoretical-explanatory virtues and its justification, such a discussion would not be directly relevant. Douglas is basing her position on the claim that the virtues in these groups are non-epistemic, and the acceptability of this position stands and falls with this claim. A claim which is neither obvious nor universally accepted.

3.3.4. Tensions Within the Second of the 'Ideal Desiderata' Groups of Virtues

[114] Ibid., p. 802.
[115] Ibid.
[116] Ibid., referring to Solomon, Miriam, 2001, *Social Empiricism*, and Longino, Helen, 2002, *The Fate of Knowledge*.

Douglas recognizes some persistent epistemic tensions within the second of the ideal desideration groups of virtues. To restate, those are instantiated by the relation between the explanatory theory and the evidence, and, contrary to the previous group, she does ascribe "genuine epistemic import" to them.[117] In the relevant section, Douglas considers what she refers to as "Whewell's consilience", which she claims provides the "strongest epistemic assurance we have available to us", and when an explanatory theory exhibits it, then "it is hard for other theories to compete".[118] Which means that it provides the highest indication that the explanatory theory that displays it is more probably true, or otherwise more worth accepting, than any of the competing theories that lacks it. She interprets consilience not as a virtue itself but as a composite or combination of two virtues, namely "novel prediction", to which Keas refers as "fruitfulness", and "explanatory coherence" or "the successful unification of evidence", which in Keas' set would be closest to the virtue he calls "unification".[119] Between these two virtues is where the tension lies, according to Douglas.

Although according to this analysis jointly these two virtues comprising "consilience" are epistemically weightier than any competing combination of virtues, it is not clear what the case is when they compete with each other. Douglas, recognizes that there is "genuine epistemic tension" between the two virtues, and that scientists "legitimately disagree" about their relative epistemic weight, with some of them "finding greater epistemic assurance in successful novel prediction", and others in "the successful unification of evidence" or "explanatory coherence".[120] However, although she states that the scientists' disagreement is "legitimate" and the tension is "genuine" she does not explain why this is the case. Is the tension genuine because of the descriptive fact that there is legitimate disagreement between expert scientists and philosophers of science about the two virtues, or, reversely, is the disagreement between them

[117] Ibid., p. 803.
[118] Ibid.; originally in Fisch, Menachem, 1985, *Whewell's Consilience of Inductions - and Evaluation*.
[119] Douglas, 2013, p. 803; Keas, 2018, p. 2762.
[120] Douglas, 2013, p. 803.

legitimate because the tension between the virtues is genuine? These would be qualitatively different positions to hold, at least metaphysically, and Douglas analysis is not clear on this. In other words, it is not clear whether the standards of *legitimacy* and *genuineness* are related, and if they are related, then it is not clear which of those is prior. Lastly, although Douglas notes that "why these tensions arise should be clearer", she does not offer a solution to the tension, nor does she provide any guidelines that could lead one to such a solution.[121]

3.3.5. Evaluation of Douglas' Theoretical Framework and Normative Proposals

Something that has remained unchallenged throughout the discussion of Douglas' work is a certain implicit metaphysical assumption about the nature of virtues. Which is that virtues can be "instantiated by the relation" or by the "relations" between an explanatory theory and a set of evidence.[122] This is important because one of her main claims, which she repeats and emphasizes throughout her work, is that "clarifying the terrain" of the virtues, or "cognitive values" as she refers to them, through a "finer grained account" which organizes them in "four distinct groups", will help us dissolve "supposed tensions" among them, and locate any remaining actual tensions.[123] Douglas bases this quadruple distinction on her assumption that there is instantiation of virtues by both "theories per se", and by "the relation" between them and the evidence. The motivation behind this is understandable, namely clarification and finer differentiation, given that many virtues that are distinct seem to have been unjustifiably lumped together. However, it is not clear that Douglas' particular way of doing that is plausible. Many questions arise, such as whether there are two kinds of virtues at play, namely, one kind that is, or can only be, instantiated by the explanatory theory, and a second kind that is, or can only be, instantiated by the relation. Also, the question as to whether there can be 'virtuous relations', questions as to the kind of

[121] Ibid., p. 804.
[122] Ibid., pp. 802, 803.
[123] Ibid., pp. 796-797.

relation she is referring to, that is, whether there are such relations, and whether relations can in fact instantiate virtues to begin with.

Take for example simplicity. It is not clear whether the virtue of simplicity that is instantiated by the theory itself, and the virtue of simplicity that is instantiated by the relation between the theory and the evidence, is the same kind of virtue; and it is not obvious from Douglas' theory that they are not. I understand that there is a difference between the feature of having less theoretical content simpliciter, and the feature of explaining the same set of phenomena with less theoretical content. However, the object of discussion here are explanatory theories, not non-explanatory theories. So the first group of virtues is not providing us with novel information about virtues such as simplicity, but with incomplete information about the explanatory theory at hand. It merely focuses on the theory in isolation while ignoring its relation to the evidence, and so is just silent with regards to the latter. But explanatory theories do not exist in isolation, in other words, for a theory to be explanatory it has to be explaining something. Also, the main point of virtue discovery and justification is to find a way to be able to justifiably select among competing explanatory theories of a certain set of evidence. That is, in the sense that what it implies about them renders it more probably true or otherwise more worth accepting than the competition; the point is not to be able to select between theories 'out there in space'. So choosing to temporarily ignore this part of the equation in one's analysis of theoretical-explanatory virtues, and then inferring substantive epistemic conclusions about them, such as the existence of different kinds of virtues, or of different objects of instantiation that correspondingly have a lesser or greater epistemic status, would be an illegitimate move.

In the part of her work where she explicitly talks about this she claims that "[t]the object of instantiation can either be a theory per se or the theory in relation to the evidence thought to be relevant to it" and so there are "two different directions for

assessment", or in other words, "two different targets for cognitive values".[124] So immediately a certain incoherence becomes apparent, since in some places she claims that virtues are instantiated by "the relation", or "the relations", between the theory and the evidence, and in other places they are instantiated by "*the theory* in relation to the evidence". These seem to be different things, so which one is it? The relation, or the theory? When she actually argues for the group distinctions and describes the different groups, she refers to "the relation" itself, not the theory "in relation" to the evidence, so the most plausible interpretation is that she actually means to use the concept of "the relation" itself.[125] So the "object of instantiation" is the relation itself not the explanatory theory which, after all, according to Douglas, must be a different object of instantiation and related to only two of the four groups. If that is the case, then no detail is given on what this relation is, or on how it could instantiate virtues, thus the position is under-argued, and so we have no independent reason to accept it.

On the other hand, we have reasons to reject it. At least because Douglas' account of virtues is less simple than competing accounts, since it multiplies theoretical content. Specifically, adding 'relations' to her theory's ontology, and also, potentially, 'values or virtues of relations', on top of 'virtues of theories'. After all, *something* has to be instantiated in the instantiation in each case, whether or not she calls that a 'value', or 'virtue', and whether or not she attaches the characterisation 'cognitive' to that something. Douglas could reply to the second of these points, about adding a further kind of virtue, namely 'virtues of relations', that the existence of such a relation, with the feature of being able to instantiate virtues, does not imply the existence of other kinds of virtues. After all, she only explicitly claims that relations are one of "two different directions for assessment when using cognitive values", or one of "two different targets for cognitive values", the other being explanatory theories, and that

[124] Ibid., p. 799.
[125] Ibid., pp. 799-801, emphasis added.

the distinction is as "to what *the* value applies";[126] the 'the' implying that it is the same virtue which is instantiated by the two kinds of objects.

A response could be that it may be the case that this does not imply that there are two kinds of virtues in the sense of there being, for example the virtue of simplicity 1, which is instantiated by theories, and the virtue of simplicity 2, which is instantiated by the relations between theories and evidence. But it does imply that there are three kinds of virtues in the sense of there being one kind of virtue that can be instantiated only by theories, such as internal consistency, a second kind of virtue that can be instantiated only by the relation between theories and evidence, such as novel prediction, and a third kind of virtue that can be instantiated by both, such as simplicity. Therefore, Douglas' metaphysical distinction between theories and their relations to the evidence, leads to her theory's scoring lower in simplicity, since it inflates its ontology in at least two ways. First, by the introduction of these 'relations' which have the capacity to instantiate virtues, and the consequent introduction of multiple kinds of virtues, one kind that can only be instantiated by theories, a second that can only be by the 'relation', and a third that can be instantiated by both.

Furthermore, if Douglas' analysis of this kind of concept of 'relation' is that 'it is a target for cognitive values providing a direction of assessment when using them', then it still remains unclear how this very same thing is, at the same time, 'one of their objects of instantiation'. However, even if we assumed that her analysis did not imply further kinds of virtues that inflate its ontology, it would still be the case that the existence of such a relation with such characteristics is at least under-argued. After examining this analysis, it is still not clear what these relations are and how it is possible for them to instantiate virtues, even if these virtues are of the same kind that are instantiated by explanatory theories per se. Therefore, given that this distinction is a foundational component of this account of the virtues, the normative proposals built on it, including the claims about certain tensions dissolving under examination those

[126] Ibid., p. 799, emphasis added.

about the order of epistemic priority between the virtues, would be in need of further argumentation and defense. In the end, overall, Douglas' analysis further elucidated the normative problem and revealed certain important limitations of the 'theoretical virtues' theory in its current formation, which paves the way for the examination of the meta-descriptive and meta-normative problems in the chapters that follow.

CHAPTER 4

Virtue Discovery Method - The Meta-Descriptive Problem

4. Proposals for Solving the Meta-Descriptive Problem

In this chapter I am going to shift focus from proposed solutions to the two lower-level problems, that is, adequately describing and justifying the theoretical-explanatory virtues and the order of epistemic priority between them, to the first of the two meta-level problems, namely, the meta-descriptive problem. To restate the meta-descriptive problem is about the theoretical framework and the method of discovery, and the reason I will be focusing on it is that, given its priority to the descriptive problem, it has to be solved first. So in the sections that follow, I will first briefly examine HPS as a framework for, and method of, virtue discovery, and discuss some of its limitations. Then I will consider the improvements XPhiSci has brought to our capacity for virtue discovery, improvements that have rendered it the most innovative and, arguably, the most promising approach for making progress at the moment, after which I will outline its major limitations. At that point, I will propose some further improvements to the current conceptual framework and methods underlying HPS and XPhiSci, which aim to increase their capacity for virtue discovery. Finally, I will explore potential ways forward in the process of virtue discovery, which, when further developed, may provide a more satisfactory solution to the meta-descriptive problem.

4.1. Historical Philosophy of Science as a Method of Virtue Discovery

HPS's purpose is not limited to work on theoretical-explanatory virtues, it rather has the broader final goal of "accounting for scientific practices in a rational way" or, in other words, to rationally reconstruct what we consider the best practices of scientific inquiry.[127] Schindler, in his discussion of HPS, points out that this is not a unidirectional way of working, and, as Kuhn pointed out, that the process allowed for changes to be made in our theory of rationality itself when scientific best practices are taken into consideration.[128] One of the sub-goals of that multi-directional endeavour is

[127] Schindler, Samuel, 2018, *Theoretical Virtues in Science*, s. 7.2.2, p. 197.
[128] Ibid., pp. 196-197.

theoretical-explanatory virtue discovery. As a method of discovery, and in the context of the 'theoretical virtues' theory, HPS works roughly like this:[129] First, one looks into historical case studies of explanatory-theory preference, that is, you research to find which explanatory theories have been judged to be preferable over which others at different points in time in the past. Most commonly one is focusing on the judgments of expert scientists and philosophers of science. Then one tries to compile a list of virtues based on an analysis of those judgments, and these virtues are supposed to explain the phenomena of those judgments the best. These inferred virtues are also used to guide future inquiry with regards to rational theory choice in science and philosophy, however, as I will discuss later on, the element of normativity or justification in HPS needs further clarification and elaboration. This is roughly the description of the HPS proposed solution to the meta-descriptive problem of the 'theoretical virtues' theory, but I will go into more detail in the evaluation below. The HPS approach survives to this day, alongside alternative or complementary methods of solving the meta-descriptive problem, such as the experimental philosophy of science method, which I will discuss in a later section.

4.1.1. Some Limitations of HPS as a Method of Discovery

Let us examine HPS more closely and evaluate its performance as a theoretical framework for, and method of, discovery. The first part of the method is a data-accumulation task. So an important question to ask is, what is the quality of this data? Upon closer analysis it turns out that the quality may not actually be that good. That is, firstly, because under HPS in most cases the explanatory theories that are analysed are compared in their totality, as in the comparison between the phlogiston and oxygen theories of combustion. As a result, the features of those theories that are of interest for virtue discovery are not properly isolated, and working backwards from them to the features that played a formative role in the epistemic agent's judgments of

[129] Cf. Thagard, Paul, 1988, *Computational Philosophy of Science*, ch. 7, s. 7.3 'Historical Philosophy of Science', p. 118.

preference is a lossy process to say the least. This process leads to many different features being lumped together under umbrella terms such as 'the virtue of simplicity' or 'the virtue of parsimony' which results, as I will show later on, in conceptual imprecision and even incoherence.

The second thing that should worry us is that this is not how we normally operate in other areas of philosophy such as moral theory or epistemology, or at least we do not operate like this anymore. That is, we do not simply look at historical cases of moral decisions that have been made in the past so as to infer moral norms. Which would not be a good idea at least given our troubled common moral past, both with regards to the moral judgements of 'non-experts' or 'everyday moral reasoners', and also of those who were considered to be expert ethicists. Nor do we consider what, say, medieval epistemologists have considered to be 'knowledge' in the past, and then build a list of relevant 'virtues' from those judgments of what constitutes knowledge in order to derive epistemic norms. On the contrary, we try to isolate considered moral and epistemic judgments on carefully designed trolley problems and Gettier-type thought experiments purportedly exhibiting cases of knowledge, and we isolate the features we want to study.[130] Although this on its own is not sufficient to show that its results are somehow invalid, this methodological dualism with regards to HPS should at least make one suspicious of them, especially given that we have moved on from this approach in solving other philosophical problems and were arguably able to make more progress.

A third problem with HPS is that it does not provide a satisfying answer to one of the main questions that need to be answered in the virtue discovery process. Namely, to restate: if one explanatory theory is preferable to another, then why is it preferable or what makes it preferable, and if they are equally preferable what makes them so? First, let us be charitable in our interpretation and assume that somehow HPS did

[130] For the 'Trolley Problem' see, Thomson, Judith Jarvis, 1985, *The Trolley Problem*; originally in, Foot, Philippa, 1967, *The Problem of Abortion and the Doctrine of Double Effect*; for 'Gettier-type' problems see, Gettier, Edmund, 1963, *Is Justified True Belief Knowledge?*.

provide an answer to this question. So what would that answer be? It could be that these expert scientists and philosophers were able to perceive these 'virtues' and that this is why they made the judgments they made. But if that is the explanation, then we end up with a Euthyphro Dilemma, because it is also the case according to HPS that the selected features of the preferred explanatory theories become 'virtues' because they were preferred by these major scientists and philosophers.[131] This seems viciously circular and so unacceptable. There is no plausible explanatory theory offered as to why they perceived those features as virtues, instead of other features such as the language that the theory happened to have originated in, or the date of publication. Features which may be considered irrelevant, but an explanatory theory using HPS needs to be in a position to explain why they are irrelevant. In the end, even though a theoretical framework and a method HPS displays these limitations, the move to systematic study of the 'virtues' to which it led has enabled us to change perspective and see virtues in their totality, and to discover more of them and understand the relations between them.

4.2. Experimental Philosophy of Science as a Method of Virtue Discovery

Moving closer towards the current state of affairs, the method that was developed next is currently referred to as "experimental philosophy of science" or 'XPhiSci', and its related subfield focusing on scientific and general forms of theoretical explanation is referred to as the "experimental philosophy of explanation".[132] Following Samuels and Wilkenfeld I will refrain from referring to the relevant experiments examined here as 'surveys', which is how they are commonly referred to as in the philosophical literature, since, as they point out, this is a rather disparaging, unfair, and, it could be added, inaccurate description. In their words, although it is the case that "they are

[131] The original dilemma was posed by Socrates, in Plato's *Euthyphro* (10a), as to whether something is pious because it was loved by the Gods or whether it was loved by the Gods because it was pious; cf. Plato, 1963, *Collected Dialogues*.

[132] See Machery, Edouard, 2016, 'Experimental Philosophy of Science'; and also for the term 'experimental philosophy of explanation' see Lombrozo, Tania, 2016, 'Explanation' in the same volume.

survey-like in that both stimuli and probes are presented linguistically, and explicit judgments are elicited", and also that "there are interesting methodological issues regarding the scope and limits of such techniques", it should nevertheless be noted that "in contrast to surveys - which merely seek to record people's views - research in experimental philosophy almost invariably involves the control and manipulation of different variables".[133] The and the fact that many of the philosophers already mentioned before the development of this method made significant use of empirical findings, is why this method and theoretical framework comes to add value and capabilities to philosophical inquiry, not to delegitimize it. Thus XPhiSci takes its place alongside HPS, conceptual logico-linguistic analysis, thought experiments, conceptual engineering, and the rest of the panoply the philosopher has at their disposal for solving problems and making progress. So after clarifying these points let us proceed with outlining the theoretical and conceptual framework and philosophical motivation that led to these experiments being conducted.

To begin, a literature analysis indicates that the foundation that enabled the move to experimentation in the study of theoretical-explanatory virtues, was partly laid by Thagard in his work *Computational Philosophy of Science*, and his proposed model of a "computational theory of explanatory coherence."[134] Although it is important in this area, it is not engaging in data-gathering and hypothesis testing like the XPhiSci work that followed. That is, Thagard did not conduct experiments leading to empirical findings about our judgments of explanatory-theory preference, but proposed a framework or model that allowed us to be able to do this later on, and philosophers building on his work did set out to collect such empirical findings. That is why the details of this model are not what is of importance here, but the work of those who developed this new approach further, which amounts to a significant leap forward. Philosophers and philosophically-minded scientists moved from looking at patterns of

[133] Wilkenfeld, Daniel, and Samuels, Richard, 2019, *Advances in Experimental Philosophy of Science*, ch. 1 'Introduction', p. 4.
[134] Thagard, Paul, 1988, *Computational Philosophy of Science*; Thagard, Paul, 1989, *Explanatory Coherence*.

judgments of explanatory-theory preference made in the past and then trying to make generalizations that can guide our future judgments, to prediction and experimentation. The mantle was taken up by researchers like Read and Marcus-Newhall and Lombrozo, whose empirical work was partly responsible for the rise of XPhiSci. Before evaluating this method and major results in XPhiSci with regards to the theoretical-explanatory virtues, it would be useful to be familiar with their most relevant empirical findings.

4.2.1. A Brief Note on the XPhiSci Work of Read and Marcus-Newhall

In this section I will discuss a representative part of the kind of XPhiSci work that has been conducted recently. For illustration, I will be going into more detail on the work of Read and Marcus-Newhall, who conducted experiments in order to study the virtues of simplicity, of what they called 'breadth', and of what we can refer to as 'depth', and though which they found that people show stable preferences for explanatory theories that exhibit these features.[135] In a similar manner, Lombrozo, once again building on the work of Thagard, but also on the work of Read and Marcus-Newhall, conducted further XPhiSci work in this area and on what she refers to as 'simplicity as an explanatory virtue',[136] in order to make progress in our understanding of this virtue in adults but also, in other work, in children.[137] The basic XPhiSci method used in these works is similar, and pretty straightforward. Participants are presented with a variety of stimuli cases in the form of short written descriptions of a number of phenomena, and of a number of explanatory theories of some or all of those phenomena. They are also presented with a probe question about those phenomena, in order for the participants' judgments of explanatory-theory preference to be elicited, and their responses are recorded, and later analysed, to determine their statistical significance.

---

[135] Read, Stephen J., and Marcus-Newhall, Amy, 1993, *Explanatory Coherence in Social Explanations*.

[136] Lombrozo, Tania, 2007, *Simplicity and Probability in Causal Explanation*, p. 235.

[137] Lombrozo, 2007; also Bonawitz, Elizabeth B. and Lombrozo, Tania, 2012, *Occam's Rattle: Children's Use of Simplicity and Probability to Constrain Inference*.

In what follows, I will describe the main details of the experimental design used, and then go through the design of a number of the most relevant experiments in more detail while outlining the results associated with them.

In their work *Explanatory Coherence in Social Explanations* Read and Marcus-Newhall,[138] building on the work of Thagard, set out to do what is currently in the literature referred to as 'XPhiSci' work on a number of proposed principles of epistemic priority. In their formulation, three of those principles tested were:[139]

> **Principle 1:** That, "all other things being equal, people should prefer the explanation that can account for more of the evidence (breadth)."
> **Principle 2:** That, all other things being equal, "people should prefer the simplest or most parsimonious explanation, the one requiring the fewest assumptions (simplicity)."
> **Principle 3:** That, all other things being equal, "people should prefer an explanation that can be explained", (for completeness we can call this 'depth').

They conducted three experimental studies two of which were related to the three principles above, and are the most relevant to this study, namely, "Study 1", examining 'breadth' and 'simplicity', and, "Study 2", examining what has been referred to as 'depth'.[140] It is also important to note that Read and Marcus-Newhall did not examine the order of preference between those principles, in their words, "the aim of this article is to study the operation of each of several principles *independently*".[141] Thus they did not aim to discover, what Kuhn referred to as, the "weight function" for the "joint application" of those principles as will be discussed in a later section of this

[138] Read, Stephen J., and Marcus-Newhall, Amy, 1993; there is a fourth principle, that of "competitiveness" which will not be examined since it does not directly relate to the main focus of this study.
[139] Ibid., pp. 429-430.
[140] For "Study 1" see ibid., pp. 432-437, and for "Study 2" see ibid., pp. 437-438.
[141] Ibid., p. 433, emphasis added.

chapter.[142] So before moving on to the evaluation, in what follows I will first outline and discuss 'Study 1' and then 'Study 2'. Study 1 consisted of three vignettes, and I will outline the first two as an example. The first vignette contained the following phenomena to be explained:[143]

> **Phenomenon 1:** First thing in the morning, Mark was called by the district to say that he should report to the courtroom that very day.
> **Phenomenon 2:** Next, Mark was asked to respond to a call. Apparently, there was an electrical problem and some live wires were down on 11th street. He left immediately and handled the situation.
> **Phenomenon 3:** Then, Mark quickly returned to his office so that he would be ready in case of future demands. After that, there was an overflow of incidents at the office and Mark had to do an excess of paperwork. He had to spend hours filing reports and taking down complaints.

The explanatory theories made available to the participants where the following:

> **Explanatory Theory 1:** Mark is a lawyer.
> **Explanatory Theory 2:** Mark is an electrical technician.
> **Explanatory Theory 3:** Mark is an office manager.
> **Explanatory Theory 4:** Mark is a police officer.
> **Explanatory Theory 5:** Mark is a lawyer, an electrical technician, and an office manager.

The second vignette consisted of the following phenomena to be explained:

---

[142] For the original utilization of these concepts see: Kuhn, Thomas, 1977, p. 326.

[143] For clarity, the experiments are presented using the terminology of this study, for example, using the terms 'phenomenon' and 'explanatory theory' over 'fact' and 'explanation', but without alteration on the main aspects of the experiments; for more details on the vignettes see Read and Marcus-Newhall, 1993, Appendix, 'Stimulus Materials', p. 447.

Phenomenon 1: Since yesterday, Cheryl has been feeling under the weather with stomach problems. She wakes up in the morning with nausea that continues throughout the day.
Phenomenon 2: Also, Cheryl has suddenly started gaining weight.
Phenomenon 3: Moreover, Cheryl has been feeling overly tired. Even when she gets enough sleep at night, she seems to be depleted of strength and energy.

The explanatory theories made available to the participants where the following:

Explanatory Theory 1: Cheryl has stopped exercising.
Explanatory Theory 2: Cheryl has mononucleosis.
Explanatory Theory 3: Cheryl has a stomach virus.
Explanatory Theory 4: Cheryl is pregnant.
Explanatory Theory 5: Cheryl has stopped exercising, has mononucleosis, and has a stomach virus.

In sum, the main results of study 1 on the vignettes can be summarized as follows: In all cases the participants preferred explanatory theory 4. It explained all phenomena 1-3, and so it displayed the virtue of breadth at a higher degree than explanatory theories 1-3, and did so in a way that displayed the virtue of simplicity at a higher degree than explanatory theory 5. Explanatory theory 4 was judged to be preferable to the competition even though explanatory theories 1-3 were equal to it with regards to the virtue of simplicity, and explanatory theory 5 was equally to it with regards to the virtue of breadth. Thus only exploratory theory 4 displayed, at the same time, the highest simplicity and breadth. Study 2 consisted of three vignettes similar to the ones in the first study, with the difference being that, in this case, more information was given when describing the phenomena, even though the possible explanatory theories remained the same. In the first two vignettes that we are focusing on, the added information was as follows:

Phenomenon 4 (for 'Mark'): He tried to focus on the fact that he chose this profession to make the world a safer place.

Phenomenon 4 (for 'Cheryl'): Recently, her husband mentioned that it was about time they started a family.

In sum, the main results of study 2 on the vignettes can be summarized as follows:[144] In all cases the participants again preferred explanatory theory 4. It explained all phenomena 1-4, and so it displayed the virtue of breadth at a higher degree than explanatory theories 1-3, and did so in a way that displayed the virtue of simplicity at a higher degree than explanatory theory 5. The difference this time was that it also displayed the virtue of 'depth' in a higher degree than explanatory theories 1-3, but at the same degree as explanatory theory 5. Thus only exploratory theory 4 displayed, at the same time, the highest simplicity, breadth, and depth. Having outlined an indicative part of more recent XPhiSci work for the purposes of establishing a common frame of reference, I will now move on to the evaluation and to discussing some of the limitations of XPhiSci as a method of discovery.

4.2.2. Some Limitations of XPhiSci as a Method of Discovery

Although XPhiSci moved towards the right direction with regards to isolating the features of the explanatory theories that they study, which marks an improvement relative to HPS, it still does not go far enough. For example, Read and Marcus-Newhall, in the first study outlined above, did not isolate simplicity from unification, which they refer to as 'breadth'. So in reality they did not show that people prefer explanatory theories that display more simplicity, and also those that display more breadth; which in way would corroborate their proposed principles. Instead they showed that people prefer explanatory theories that display both more simplicity and more breadth than their competitors, at the same time. In other words, they showed that the combination

[144] Read and Marcus-Newhall, 1993, p. 438.

of these two virtues are preferable than the virtue of simplicity on its own and the virtue of unification on its own; which would amount to corroborating a very different principle. In their second study, something similar happened. They showed that the explanatory theory that exhibits the combination of the virtues of simplicity, breadth, and explanatory depth, to which I referred to simply as 'depth', are preferred over alternatives that only show one of these virtues. What they did not show is that people prefer an explanatory theory than alternatives that do not have this virtue, all other virtues being equal. Also they did not examine judgments of preference with regards to the epistemic priority between virtues, or between different combinations of virtues, which is an issue I will examine in more detail in a later section. Lombrozo on the other hand, did manage to isolate the virtue of simplicity and get some results, but did not examine different kinds of simplicity; and so a potential generalization of the results regarding one kind of simplicity to every kind of simplicity would be illegitimate. She also did not test the order of epistemic priority within the virtue of simplicity itself, that is, she did not endeavour to elicit judgments of preference with regards to competing explanatory theories exhibiting different kinds of simplicity. Again, this is an issue I will examine in a later section. Lastly, the number of virtues examined in XPhiSci has been quite limited, that is, given that the most current proposed sets of virtue are composed of up to twelve major virtues.[145]

Another potential limitation for XPhiSci as practiced so far is that the judgments that are taken into consideration may not be considered judgments, and are certainly not considered judgments of experts. Taking into consideration a set of potentially unconsidered judgments of non-experts in our efforts at virtue discovery may not be the best way forward. It is definitely a step backwards relative to HPS, which by default takes into consideration the judgments of expert scientists and philosophers of science. However, this position is itself amenable to XPhiSci work. That is, XPhiSci practitioners can test the considered judgments of non-experts against the considered judgments of experts, on the same set of constructed experimental cases of

[145] Cf. Keas, 2018.

explanatory-theory choice. In this way they could take advantage of all the benefits of experimental idealization and feature isolation, whilst still taking into consideration the judgments of experts. Of course, that raises the question of expert selection, which is multifaceted. Who counts as an expert in this case? The answer to this question would be a list of 'demarcation criteria' aimed at distinguishing experts from non-experts. Moreover, an expert in one area of the problem may not be an expert in another area, and given that time and energy are finite, there can be no natural epistemic agent, be they scientist or philosopher, that is an expert in every relevant area of inquiry. So whatever those demarcation criteria may be they must be relative to a particular area of inquiry. Therefore, XPhiSci philosophers can either show that there is no significant variation between the judgments of explanatory-theory preference of experts and those of non-experts. If there is significant variation, they would need to argue in favour of either one of the groups, or show that somehow both sets of judgments provide legitimate evidence in virtue discovery, and so both must be taken into consideration. In the end, these are some of the limitations that XPhiSci would need to overcome, but it nevertheless should be acknowledged that it has opened the way to further virtue discovery, and is the most promising general approach for solving the meta-descriptive problem.

4.3. Ways to Further Develop the HPS and XPhiSci Framework

Having examined the major virtues featured in the most recent proposals, and outlined the achievements and limitations of the historical philosophy of science and the experimental philosophy of science methods, it is only natural to seek ways to make improvements. A major overarching goal to which the HPS and XPhiSci methods are meant to be of service, is what Kuhn referred to as the "search for algorithmic decision procedures", in other words, the discovery of an "algorithm", that would be "able to dictate rational, unanimous choice"; drawing on, among other things, the "theoretical

virtues".[146] He saw at least two limitations with developing such a method, first, that it presupposes that "individual criteria of choice can be unambiguously stated" and, second, that "an appropriate weight function is at hand for their joint application".[147] The corresponding two reasons for these limitations he gave were that, one, "the criteria are imprecise" and, two, "they repeatedly prove to conflict with one another".[148] His judgment at the time of publication was that "unfortunately [...] little progress has been made toward the first of these desiderata and none toward the second", and so he "entirely agreed" that "the sort of algorithm which has traditionally been sought" was "a not quite attainable ideal"[149]. Neither HPS nor XPhiSci have successfully provided the scientific and philosophical community with such an algorithm so far. However, there are ways we could make progress towards that goal.

There are at least two components in the problem that Kuhn notices, that is two limitations, namely 'imprecision' *within*, and a lack of an 'order of epistemic priority' *between*, the theoretical-explanatory virtues. In more detail, the first limitation is that there is a general lack of precision and clarity with regards to the criteria of rational explanatory-theory choice, which, I will argue, is partly due to internal imprecision and unclarity in the conceptualization of the virtues themselves. This, I will show, also limits the rate of progress that can be achieved by the current version of XPhiSci as it stands. The second limitation is that the method has not been able to discover the order of epistemic priority between the theoretical-explanatory virtues, which, I will further argue, is partly a result of the first limitation. In other words, the lack of precision and clarity with regards to the theoretical-explanatory virtues, is an obstacle to making progress on discovering the "weighting function" for their "joint application". Thus solving the first problem would unlock the potential for making progress in solving the second problem. So part of Kuhn's judgment remains correct, namely, that neither of

[146] Kuhn, Thomas, 1977, pt. 2 'Metahistorical Studies', ch. 13 'Objectivity, Value Judgment, and Theory Choice', p. 326.
[147] Ibid., p. 326.
[148] Ibid., p. 322.
[149] Ibid., p. 326.

the two limitations have yet been fully overcome. This should not be very surprising since, even though we have recently moved from HPS to XPhiSci, we have not updated our conceptual apparatus accordingly to fit this framework, and have not taken full advantage of advances in adjacent fields. The second part of Kuhn's judgment, about the potential of overcoming those limitations and so achieving the two related "desiderata", is overly pessimistic. On the contrary, as I will argue below, there are ways to make progress in overcoming these limitations, and so undermine the assumptions on which Kuhn bases his judgment that the discovery of the algorithm is unattainable. In what follows, I will begin by closely examining the first limitation that is holding XPhiSci back, and then argue for a particular way to overcome it.

4.3.1. Overcoming the First Limitation - Imprecision and Incoherence

As one can see from the discussion in previous chapters, despite some efforts towards systematization, there is still a general lack of unity and cohesion in the philosophical literature on, what have so far been referred to as, the 'theoretical-explanatory virtues'. As I mentioned above, there are a number of recent proposals which explicitly try to systematize them, but at the same time show significant conceptual differences with one another.[150] Also, to restate, even the main concept itself is not universally used. Variations such as 'virtues of theories', 'virtues of good theories', 'theoretical values', 'explanatory virtues', 'epistemic values', 'cognitive values', and many others mean different things to different researchers. The further thing to be pointed out and considered in this section, is that the same unclarity and imprecision is to be found with regards to the literature on the particular virtues themselves. Given these phenomena, inquiry into the matter can get very confusing. Therefore, in this section I will propose a few conceptual re-engineering interventions to the current conceptual system, which attempts to improve on the formulation of concepts that comprise the theoretical-explanatory virtues. In short, it would be better if we broke the concept of a 'virtue' into two components. Namely, into the 'features' of explanatory theories

[150] Cf. Keas, 2018, McMullin, 2014, Douglas, 2013, Adolfas, 2013.

that play a formative role in our judgments of explanatory-theory preference, along with their various combinations, and the 'principles' that relate them in an order of epistemic priority. But, before I explain more, let me motivate this shift by discussing some of the limitations that the current conceptual system faces.

To take one common example, the concept of the virtue of 'simplicity' displays many variations. For example: it sometimes gives its place to the concepts of 'parsimony', 'ontological parsimony', 'qualitative parsimony', 'causal simplicity', and many more.[151] Sometimes the variations are even accompanied by the term 'principle', as in 'the principle of simplicity', 'the principle of parsimony', and so on. Here is at least one problem with the concept of the virtue of 'simplicity' with which one is faced with when one takes into account its common usage in the philosophical literature: an explanatory theory can be 'simpler' if it is more 'parsimonious' than another. It can be more 'parsimonious' if it is more 'ontologically parsimonious' than another. And it can be more 'ontologically parsimonious' than another if it is more 'quantitatively ontologically parsimonious' than another, that is it assumes the existence of a lesser number of "individual" entities postulated.[152] But it can also be more 'ontologically parsimonious' if it is more 'qualitatively ontologically parsimonious' than another, that is if it assumes the existence of a lesser number of kinds or "types" of entities, than another explanatory theory.[153] Thus a more 'quantitatively ontologically parsimonious' explanatory theory is 'simpler' than an explanatory theory that is less 'quantitatively ontologically parsimonious'. Furthermore, a more 'qualitatively ontologically parsimonious' explanatory theory is also 'simpler' than an explanatory theory that is less 'qualitatively ontologically parsimonious'.

Now assume that there are two explanatory theories T1 and T2, such that T1 is 'quantitatively ontologically more parsimonious' than T2, while at the same time T2 is

[151] Cf. Baker, Alan, 2003, *Quantitative Parsimony and Explanatory Power*.
[152] Baker, 2003, p. 247.
[153] ibid.

'qualitatively ontologically more parsimonious' than T1. We must conclude then that T1 is, at the same time, 'simpler' than T2, and also not 'simpler' than T2; which is a contradiction so the concept of 'simplicity' displays signs of incoherence. Initially, it seems that this particular case of conceptual incoherence can be ameliorated in an ad hoc fashion by creating more virtue-type concepts for simplicity, or even by doing away with the concept of simplicity altogether and creating a number of separate, more specific, virtues that capture each of the features of a 'simpler' explanatory theory independently. However, the 'theoretical-explanatory virtues' conceptual approach is only bound to result in more incoherence in other virtues. To illustrate, another example is the concept of the virtue of 'unification' which also displays signs of incoherence in the same way as 'simplicity' does. T1 explains a larger number of facts than T2. T2 explains a larger number of kinds of facts than T1. So T1 is more 'unifying' than T2, and is also not more 'unifying' than T2; which is inconsistent so this concept too displays signs of incoherence. You can of course try to fix this problem by creating separate more specific virtues to replace 'unification' as in the case of 'simplicity'.

4.3.2. Conceptually Re-engineering 'Virtues' into 'Features and Principles'

However, if you keep doing that across all known virtues, you are bound to end up with something like the 'features and principles' conceptualization, which I will describe next.[154] That is you will need a virtue for every feature, only with double the terms and theoretical content; and so you end up with a less 'virtuous' system overall. A 'features and principles' approach would analyse the cases above into the following proposed features and sets of principles, with the first set being:

> **Proposed Set of Principles 1 'Five Features':** At least the following five elements are members of the set of all features of explanatory theories that

[154] For a similar use of the concept 'features' as in 'features of explanation' see Wilkenfeld, Daniel, and Samuels, Richard, 2019, *Advances in Experimental Philosophy of Science*, ch. 1 'Introduction', s. 'On the potential contributions of experimental philosophy of science', p. 8.

play a formative role in our judgments of explanatory-theory preference: **(a)** the number of entities (across kinds), **(b)** the number *of kinds* of entities, **(c)** the number of phenomena explained (across kinds), **(d)** the number *of kinds* of phenomena explained, and **(e)** the number *of levels of explanation* of phenomena.

The list is indicative not exhaustive. The second set of proposed principles would describe the order epistemic priority *within* these features. The second proposed set of principles would partly be constituted as follows:

Proposed Set of Principles 2 'Order of Priority Within Features': There is **(A)** a single order of priority *within* each of the features (a)-(e) across epistemic agents and across features, and also **(B)** it is adequately described by the following proposed principles of epistemic priority: **(1)** all other things being equal, explanatory theories that postulate a lesser number of entities (across kinds) are preferable to ones that postulate a greater number of entities (across kinds); **(2)** all other things being equal, explanatory theories that postulate a lesser number *of kinds* of entities are preferable to ones that postulate a greater number *of kinds* of entities; **(3)** all other things being equal, explanatory theories that explain a greater number of phenomena (across kinds) are preferable to ones that explain a lesser number of phenomena (across kinds); **(4)** all other things being equal, explanatory theories that explain a greater number *of kinds* of phenomena are preferable to ones that explain a lesser number of *kinds of* phenomena; **(5)** all other things being equal, explanatory theories that display a greater number *of levels of explanation* of phenomena are preferable to ones that display a lesser number *of levels of explanation* of phenomena; **(6)** All other things being equal, explanatory

> theories that postulate an *equal* number of entities (across kinds) are equally preferable/equal; [...][155]

Here (A) describes the presumption that there is a single order of priority *within* each of the features. This is very important since even though it is not explicitly stated in the literature, it is actually almost always presumed. However, it is a *refutable principle* and should be treated as a working assumption. With regards to (B), clearly parsing out these features allows for a more precise, and so more testable, formulation of the principles of epistemic priority. You can work with those principles without ever having to appeal to a concept of 'virtue' or 'value', neither to the particular virtue concepts of 'simplicity' or 'unification'.

Before continuing with the argument, let me explain why what has just been shown is important, and further explain the 'principles' part of the 'features and principles' conceptualization. When you call a feature or combination of features a 'virtue', say in the case of the virtue of 'quantitative ontological parsimony', you are actually doing at least two things at once. You are identifying a difference between explanatory theories with regards to at least one feature, in this case 'number of entities postulated', but you are also putting forward a *refutable* principle of epistemic priority like the ones above. Specifically, a principle that, all other things being equal, whenever an explanatory theory T1 has this feature at a certain degree, in this case a lesser 'number of entities' (across kinds), then T1 is to be preferred to a competing explanatory theory T2 which displays it at a different degree, in this case at a greater degree, accordingly. After all, one would not call something a 'virtue' if one did not assume that it is something to be preferred. At the same time you implicitly presume a *refutable* principle, such as Principle 6 above, that, all other things being equal, when two explanatory theories display this feature at the same degree, they are equal or equally

[155] Although not necessary for understanding the argument being made here, the rest of the set can be found in Appendix 1 of this study. Principles 6-10 ensure that the features and principles conceptualization achieves completeness, through accounting for something that has been neglected in the literature, namely, relations of equality within features (a)-(e).

preferable. This is one reason why the 'features and principles' conceptualization is more conducive to rational inquiry than the 'virtues' conceptualization.

Moving on with the analysis of the two example cases described above, another important distinction is between principles describing the order of epistemic priority *within features*, and principles describing the order of epistemic priority *between features*. The latter would comprise a separate set of principles, which would partially be as follows:

> **Proposed Set of Principles 3 'Order of Priority Between Features':** There is **(A)** a single order of priority *between* features of explanatory theories (a)-(e) across epistemic agents and across features, and also **(B)** it is adequately described by the following proposed principles of epistemic priority: **(11)** all other things being equal, explanatory theories that explain a greater number *of kinds* of phenomena are preferable to ones that explain a lesser number *of kinds* of phenomena but explain a greater number of phenomena (across kinds); [...] **(20)** all other things being equal, explanatory theories that postulate a lesser number *of kinds* of entities are preferable to ones that postulate a greater number *of kinds* of entities but postulate a lesser number of entities (across kinds).[156]

Again, (A) is generally presumed in the above cases and in the philosophical literature, but not explicitly stated; it too is a *refutable principle* that should be treated as a working assumption. Moving on, with regards to (B), stating the principles this way allows us to overcome incoherence and reduce imprecision, but also to argue more clearly with regards to example cases discussed above:

[156] The principles' numbers being '11' and '20' is not an accident, I have proposed further principles in between, which are listed in Appendix 1 of this study, but only these two are needed for illustration and the analysis of the example cases here.

Premise 1: The number of entities (across kinds) and the number *of kinds* of entities an explanatory theory displays, are features of explanatory theories that play a formative role in our judgments of explanatory-theory preference (from Proposed Set of Principles 1).

Premise 2: There is a single order of priority *between* the feature 'number of entities (across kinds)' and the feature 'number *of kinds* of entities', across epistemic agents and across features (from Proposed Set of Principles 3)

Premise 3: All other things being equal, explanatory theories that postulate a lesser number *of kinds* of entities, are preferable to ones that postulate a greater number *of kinds* of entities but postulate a lesser number of entities (across kinds) (Principle 20, from Proposed Set of Principles 3).

Premise 4: Explanatory theory T1 postulates a lesser number of entities (across kinds) than explanatory theory T2.

Premise 5: Explanatory theory T2 postulates a lesser number *of kinds* of entities than explanatory theory T1.

Premise 6: If P1-P5, then C.

Conclusion: T2 is preferable to T1 (from premises 1-6).

To test one's own intuitions on premise 3, 'Stimulus Case 20' below is set up to test Principle 20 of Proposed Set of Principles 3. In this stimulus case, Philons, Sophons, and Cognons are fictional entities that tend to react in an explosive manner; and it is set up as follows:

Phenomenon: There was an explosion.

Explanatory Theory 1: The explosion was the result of a reaction between 1 Philon and 1 Sophon.

Explanatory Theory 2: The explosion was the result of a reaction between 10 Cognons.

In this case T2 postulates a lesser number *of kinds* of entities than explanatory theory T1, namely, one kind of entity referred to as a Cognon, versus T2 which postulates two kinds of entities referred to as Philons and Sophons. On the other hand, T1 postulates a lesser number of entities (across kinds) than explanatory theory T2, with T1 postulating a total of 2 entities, and T2 postulating a total of 10 entities. A judgment of preference for T2 in this case, which would match the judgment derived in the conclusion of the argument, would be in accordance with Principle 20 as stated in premise 3 above, while preference for T1, or a judgment of equality between T1 and T2, would contradict Principle 20 and so count against it.

One can argue similarly for the second example case discussed before, where an explanatory theory T1 explains more phenomena (across kinds), and an explanatory theory T2 explains more *kinds of* phenomena:

> **Premise 1:** The number of phenomena explained (across kinds) and the number *of kinds* of phenomena explained, are features of explanatory theories that play a formative role in our judgments of explanatory-theory preference (from Proposed Set of Principles 1).
> **Premise 2:** There is a single order of priority *between* the feature 'number of phenomena explained (across kinds)' and the feature 'number *of kinds* of phenomena explained', across epistemic agents and across features (from Proposed Set of Principles 3).
> **Premise 3:** All other things being equal, explanatory theories that explain a greater number *of kinds* of phenomena are preferable to ones that explain a lesser number *of kinds* of phenomena but explain a greater number of phenomena (across kinds) (Principle 11, from Proposed Set of Principles 3).
> **Premise 4:** Explanatory theory T1 explains a greater number of phenomena (across kinds) than explanatory theory T2.
> **Premise 5:** Explanatory theory T2 explains a greater number *of kinds* of phenomena than explanatory theory T1.

Premise 6: If P1-P5, then C.

Conclusion: T2 is preferable to T1 (from premises 1-6).

Once more, to test one's own intuitions on premise 3 of this argument, 'Stimulus Case 11' below is set up to test Principle 11 of Proposed Set of Principles 3:

Phenomenon 1: Tree 1 fell in the forest.

Phenomenon 2: Tree 2 fell in the forest.

Phenomenon 3: Tree 3 fell in the forest.

Phenomenon 4: A person died.

Phenomenon 5: There was an explosion.

Explanatory Theory 1: Tree 1, Tree 2, and Tree 3 fell because of strong wind.

Explanatory Theory 2: The person died of old age, and the explosion happened because of an earthquake.

In this case, T2 explains a greater number *of kinds* of phenomena than explanatory theory T1, namely, it explains the death of the person and the explosion which are the only instances of their kinds of phenomena in this particular case; thus totalling 2 kinds of phenomena explained for T2. In contrast, T1 explains only 1 kind, namely, the fall of Tree 1, of Tree 2, and of Tree 3, which are three instances of the phenomenon of the kind 'felling of trees' or 'fallen trees'. At the same time, in terms of the number of phenomena explained across kinds, T1 explains a total of 3 while T2 explains a total of 2. A judgment of preference for T2 in this case, which would match the judgment derived in the conclusion of the argument, would be in accordance with Principle 11 stated in Premise 3 above, while preference for T2 or a judgment of equality between T1 and T2 would contradict and count against it. In what follows, I will show that this conceptual improvement, apart from allowing us to better account for these two example cases than the 'virtues' conceptualization, is also important for making progress with the second limitation of XPhiSci discussed above.

4.3.3 Overcoming the Second Limitation - Discovering the Order of Epistemic Priority

Moving on to tackling the second limitation, the problem with discovering the 'weighting function' for the joint application of the 'virtues' is made more salient when one realizes that principles of epistemic priority within and between *individual features* are not the only ones that need to be accommodated conceptually. Principles of epistemic priority between *combinations of features* also need to be accommodated. Now here is a problem that arises when one starts to consider such combinations, under the 'virtues' conceptualization. Assume that:

> T1 is more quantitatively ontologically parsimonious, and more qualitatively explanatory broad than T2, while T2 is more qualitatively ontologically parsimonious, and more quantitative explanatory broad than T1.

Now which explanatory theory is 'simpler' in this case, and also, given that they have an equal *number* of 'virtues', which one is more 'virtuous', T1 or T2? In a maximally developed version of the 'virtues' approach, every virtue would imply a proposed principle of priority about a feature of an explanatory theory. But every proposed principle of priority about *combinations of features* of explanatory theories would not imply a virtue. One would then have to create a virtue for every one of the potentially hundreds of different combinations of features. The formulation would then have to look something like this: T1 is combination-1-more-virtuous than T2. To contrast, under the 'features and principles' conceptualization, here is a proposed set of principles of priority between different combinations of features of explanatory theories, part of which can be used for this case:

> **Proposed Set of Principles 4 'Order of Priority Between Combinations of Features':** There is **(A)** a single order of priority *between combinations of* features of explanatory theories (a)-(e) across epistemic agents and across features, and also **(B)** it is *partly* described by the following proposed principles

> of epistemic priority: **(21)** all other things being equal, explanatory theories that explain a greater number *of kinds* of phenomena and also postulate a lesser number of entities (across kinds) are preferable to ones that explain a lesser number *of kinds* of phenomena and postulate a greater number of entities (across kinds) but explain a greater number of phenomena (across kinds) and postulate a lesser number *of kinds* of entities. [...][157]

Once more (A) is important and needs to be explicitly stated although usually it is not; it is a refutable principle and a working assumption, this time about the level of variation between *combinations of* features. In this case too, a clear argument can be formulated given these components:

> **Premise 1:** The number of entities (across kinds), the number *of kinds* of entities, the number of phenomena explained (across kinds), and the number *of kinds* of phenomena explained, are features of explanatory theories that play a formative role in our judgments of explanatory-theory preference (features (a)-(d), from Proposed Set of Principles 1).
> **Premise 2:** There is a single order of priority *between combinations of* features of explanatory theories (a)-(d) across epistemic agents and across features (from Proposed Set of Principles 4)
> **Premise 3:** All other things being equal, explanatory theories that explain a greater number *of kinds* of phenomena and also postulate a lesser number of entities (across kinds) are preferable to ones that explain a lesser number *of kinds* of phenomena and postulate a greater number of entities (across kinds) but explain a greater number of phenomena (across kinds) and postulate a lesser number *of kinds* of entities (Principle 21, from Proposed Set of Principles 4).

[157] Although not required for understanding the argument being made here, for the rest of the set, which is indicative not exhaustive, see Appendix 1, this study.

Premise 4: Explanatory theory T1 postulates a lesser number of entities (across kinds) and explains a greater number *of kinds* of phenomena than explanatory theory T2.

Premise 5: Explanatory theory T2 postulates a lesser number *of kinds* of entities and explains a greater number of phenomena (across kinds) than T1.

Premise 6: If P1-P5, then C.

Conclusion: T1 is preferable to T2 (from premises 1-6).

Once more, to test one's own intuitions on Premise 3, 'Stimulus Case 21' is set up to test Principle (21) of Proposed Set of Principles 4, where Philons, Sophons, and Cognons are fictional entities that tend to reach in an explosive manner; it is structured as follows:

Phenomenon 1: Tree 1 fell in the forest.

Phenomenon 2: Tree 2 fell in the forest.

Phenomenon 3: Tree 3 fell in the forest.

Phenomenon 4: A person died.

Phenomenon 5: There was an explosion.

Explanatory Theory 1: Tree 3 fell because of strong wind, the person died of old age, and the explosion happened because of a reaction between 1 Philon, 1 Sophon, and 1 Cognon.

Explanatory Theory 2: Tree 1, Tree 2, and Tree 3 fell because of strong wind, and the explosion happened because of a reaction between 10 Philons.

In this case, T1 postulates a lesser number of entities (across kinds) than T2, namely 3 in total, versus 10 in total in T2, however, it postulates a greater number *of kinds* of entities, namely, Philons, Sophons, and Cognons, which amounts to 3 kinds, in contrast to T2 which only postulates Philons, that is, only 1 kind. On the other hand, T1 explains a greater number *of kinds* of phenomena than explanatory theory T2, that is it explains the fall of Tree 3, which is a phenomenon of the kind 'felling of trees' or 'fallen trees',

of which there are three instances in this stimulus case; and it also explains the death of the person and the explosion which are the only instances of their kinds of phenomena in this particular case; thus totalling 3 kinds of phenomena explained for T1. In contrast, T2 explains a total of 2 kinds of phenomena, namely, the fall of Tree 1, of Tree 2, and of Tree 3, which are three instances of the same kind of phenomenon, and also explains the explosion. At the same time, in terms of the number of phenomena explained across kinds, T1 explains a total of 3 phenomena across kinds while T2 explains a total of 4, which is greater. A judgment of preference for T1 in this case, which would match the judgment derived in the conclusion of the argument, would be in accordance with Principle 21 stated in Premise 3 above, while preference for T2, or a judgment of equality between T1 and T2, would contradict and count against it.

Importantly, under the 'features and principles' conceptualization in all three of arguments and example stimuli cases outlined above, premise 1 and premise 2 are also falsifiable. If it is shown that features (a)-(d) do not appear to play a formative role in our judgments of explanatory-theory preference, or that there is not just a single justifiable order of epistemic priority between these features and their combinations, then Premises 1 and 2 in these arguments would lack support and the arguments' conclusions would not follow. However, most likely, this will not be settled merely through introspection and the reader's and author's intuition. The support behind those premises would come from the kind of XPhiSci work discussed above, which would examine these principles across epistemic agents and across features.[158] In the end, the 'features and principles' approach to XPhiSci, or 'F&P-XPhiSci' for short, achieves everything that the 'virtues' approach does, and it can also account for more principles of epistemic priority. If you take into account the potential number of combinations of any given set of features, this can amount to a significant

[158] For a proposed set of principles of epistemic priority given the 'features and principles' framework for XPhiSci, I refer the reader to Appendix 1 of this study, and for a proposed set of Stimuli Cases to examine those principles against one's own intuitions or test them through XPhiSci-type experimentation, I refer the reader to Appendix 2 of this study.

improvement in researchability and, as a consequence, in the effectiveness and efficiency of the HPS and XPhiSci methods. That is because it would allow them to help us uncover the order of epistemic priority within and between the different combinations of features, and so overcome Kuhn's objections. Therefore, if we aim to be more efficient in making progress in this area of research, it would be worth moving away from the 'common-sense' formulations of the virtue concepts as laid out in current proposals, and instead focus on what combinations of features of explanatory theories are picked out when a preference judgment is made between explanatory theories. In what follows I will consider possible objections to the meta-descriptive potential of F&P-XPhiSci as a method of discovery.

4.4. Objection: F&P-XPhiSci Would Lead to Very Low Rates of Progress

An objection could be put forward that, under the F&P-XPhiSci framework, the projected rate of progress would be quite low and the method of discovery would be very uneconomical, and so F&P-XPhiSci should not be preferred at the level of competing proposals for solving the meta-descriptive problem. Let me outline the main reasons that this may be a concern to someone. To begin, we have not yet discovered the full set of features that play a formative role in our judgments of explanatory-theory preference, and even if we do eventually discover the full set, simply compiling the full set of features would not be enough. What would be necessary for solving the descriptive problem, would be to compile the full set principles describing the order of epistemic priority within and between each of those features and their various combinations. To elicit all the relevant expert considered judgements of explanatory-theory preference, would require building experiments testing for every feature and every combination of features, and compare it with every other feature and combination of features, which would be very uneconomical in terms of time, energy, and resources. To clarify, the objection is not that we cannot complete such a task. Neither the set of all features, nor the set of all combinations of these features, nor the set of all principles of epistemic priority, is infinite. These sets

are finite, so, in principle, they are completely discoverable, and the only limitations are those of our cognitive reach and the current level of XPhiSci engineering and technology. So the point of the objection is that under the 'features and principles' conceptualization, XPhiSci, in its current form, would be an extremely slow and expensive way to achieve that goal. The rate of progress would be flat and so progress in solving the descriptive problem would be slow and linear. Therefore, even though this is not epistemically a limiting factor, methodologically it would be suboptimal to the point where solving the descriptive problem this way would be unsustainable, or at least less preferable relative to potential competing solutions to the meta-descriptive problem. For this to be a viable and economical project we need a method and a theoretical framework that not only gives us increasing *progress* at a stable rate, but an increasing *rate of progress*; and this goal is simply not readily achievable given the current level of XPhiSci engineering and technology presumed in this proposal.

4.4.1. Reply: Increasing the Rate of Progress with Philosophical Artificial Intelligence

It must be conceded that there would be merit in such an objection, and that, in its current form, XPhiSci would lead to a flat and rather low rate of progress in discovering the 'algorithm' of our explanatory-theory preference. A reply to this objection could be that the main question to be answered here is whether the 'features and principles' approach to XPhiSci can produce a better rate of progress than the one before. That is, in the same way that XPhiSci offered an improved rate of progress than that of HPS through utilizing experiments, and being forward-looking and predictive, rather than backward-looking and dependent on major historical cases, or waiting until the target 'virtues' came up in an actual case of theory-preference during normal scientific practice. For reasons explained above, including overcoming conceptual imprecision and incoherence and allowing us to effectively and more efficiently discover principles of epistemic priority between features and their combinations, the answer to the question would have to be 'yes'. However, the thrust of the objection remains, and with regards to the question as to whether the rate of progress is as high as it can be,

the answer would have to be 'no'. Given the magnitude of the task at hand, we would need a further improvement to XPhiSci to ameliorate these issues, and so make it more competitive. What follows in the next few sections is the 'towards' part of the title to this study, and its main purpose is to make the best possible presumptive case for a type of XPhiSci that one would get if one took the principles behind this method to their logical conclusion. What I will try to show is that XPhiSci can be further improved, and that the tools to improve it and increase the rate of progress are available to philosophical inquiry. That methodological improvement to XPhiSci can be drawn from, what is currently referred to as, 'philosophical artificial intelligence' (henceforth 'φAI').[159]

In general terms, philosophical AI is "AI pursued as and out of philosophy" and is about using AI methods in a philosophically-minded way to solve philosophical problems and answer philosophical questions.[160] Although still under development, this approach could be utilized in a way that is compatible with Samuels and Wilkenfeld's conception of experimental philosophy of science, where there is "no restriction on which empirical *methods* might be relevant to addressing issues in the philosophy of science", and so it can include methods from "developmental research, reaction-time studies, patient studies, and functional Magnetic Resonance Imaging (fMRI) research".[161] If XPhiSci can include these, then it can also include methods from philosophical artificial intelligence and machine learning. Furthermore, a philosophical-AI approach to XPhiSci, or 'φAI-XPhiSci' for short, would be able to aid us in increasing the rate of progress in the discovery of the 'algorithm' of our explanatory-theory preference. Very roughly, it would consist in building an artificial agent, henceforth 'φAI-Agent', that learns and matches our 'natural' judgments of explanatory-theory preference on actual cases of explanatory-theory choice, and then is able to produce

[159] See for example: Bringsjord, Selmer and Govindarajulu, Naveen Sundar, "Artificial Intelligence", *The Stanford Encyclopedia of Philosophy*, s. 7 'Philosophical AI'.
[160] Ibid.
[161] Wilkenfeld, Daniel, and Samuels, Richard, 2019, *Advances in Experimental Philosophy of Science*, ch. 1 'Introduction', s. 'What might experimental philosophy of science be?', p. 4, emphasis original.

'artificial' judgments on novel cases of explanatory-theory choice. But before going into more detail on how this version of XPhiSci would work, let me consider a further potential objection that can be posed in response to such a proposal at this point.

4.4.2. Response: φAI-XPhiSci Would Be Unexplanatory and Uninformative

It could be argued that such a method would be unexplanatory and so would not be a beneficial amendment, or addition, to the original XPhiSci method. That is, creating a φAI-Agent that just produced artificial judgments of explanatory-theory preference that matched ours would not be very informative, let alone explanatory. Such a process would be a 'black box', in that we still would not be in a position to understand why the φAI-Agent made the judgments it made, and therefore the method would lack explanatoriness. Also, it would presuppose that we know, in advance, all the features that a 'natural' epistemic agent picks out when it produces natural judgments of explanatory-theory preference and all the principles of epistemic priority between them, so as to train the φAI-Agent accordingly. It would further presuppose that we would be able to produce these natural judgments on cases at any level of complexity, with regards to combinations of different features, that the φAI-Agent is able to produce its artificial judgments on novel cases; which seems implausible. After all, how else would we be able to determine whether natural and 'artificial' judgments matched in in any of those complex cases? So for at least these reasons, the objection would go, the φAI-XPhiSci approach, if it worked at all, would be unexplanatory or otherwise uninformative.

4.4.3. Reply: Inverse Reinforcement Learning φAI-XPhiSci

Once more, there would be a point to this potential objection, but these problems can be overcome, and despite the exploratory nature of this part of the study, which bars what would be a premature venture into in-depth technical analysis, the presumptive case for φAI-XPhiSci can be defended against this objection. So a promising way to

tackle the above problems, would be to adopt a different type of reinforcement learning for the φAI-Agent. Specifically, the one currently referred to as "inverse reinforcement learning", or "IRL" for short.[162] Now it is indeed the case that in order to train the φAI-Agent to match our judgments of explanatory-theory preference through ordinary reinforcement learning, one would need to input an already known and accurate "reward function". That is, in our case, one would need to input the 'weight function' consisting of principles of epistemic priority, which would have to be known already, and after many iterations, the φAI-Agent would match our judgments as we 'corrected' it. That would defeat the purpose of adopting φAI-XPhiSci, since it is that weight function that we are seeking here. However, contrary to the type of reinforcement learning that the objection is aimed at, an IRL approach allows for treating the reward function "as an unknown to be ascertained",[163] in other words, IRL is appropriate for "situations where knowledge of the rewards is a goal by itself (as in *preference elicitation*)", a goal that is achieved through a type of "apprenticeship learning (learning policies from an *expert*)".[164]

In our case, roughly, the "unknown reward function",[165] which in our case would be a "*multiattribute* reward function",[166] would correspond the 'weight function', or the principles of epistemic priority underlying our judgment of explanatory-theory preference, with the different 'attributes' corresponding to the various features of explanatory theories; and 'preference elicitation' would correspond to eliciting those principles. 'Policies' would correspond to judgments on cases of explanatory-theory choice, or "a mapping from states to actions",[167] that is, both those of the φAI-Agent,

[162] Russell, Stuart, 1998, *Learning agents for uncertain environments*; Ng, Andrew Y., and Russell, Stuart, 2000, *Algorithms for Inverse Reinforcement Learning*; Abbeel, Pieter, and Ng, Andrew Y., 2004, *Apprenticeship Learning via Inverse Reinforcement Learning*; Ramachandran, Deepak, and Amir, Eyal, 2007, *Bayesian Inverse Reinforcement Learning*.
[163] Russell, 1998; Ng, and Russell, 2000.
[164] Ramachandran, and Amir, 2007.
[165] Abbeel, and Ng, 2004.
[166] Russell, 1998, emphasis original.
[167] Ng, and Russell, 2000.

and also those of the 'expert' referred to as "observed" policies.[168] 'Expert' in our case would literally correspond to expert natural and social scientists, and philosophers of science. However, in experiments were we want to discover the level of variation between experts and non-experts, the 'expert' could be a non-expert everyday reasoner in one case, and an expert philosopher of science in the other; also in experiments were we want to aggregate expert judgments, the 'expert' could also be "multiple experts".[169] To these we may need to add a "model" of the environment,[170] which would correspond to an appropriate formalization of our mental representation of cases of explanatory-theory choice, such as the Stimuli Cases discussed in previous sections,[171] which we would be able to enrich later on by adding more Stimuli Cases of various levels of complexity to test for new features and principles.

This unknown reward function would be ascertained in roughly the following way. To begin, an initial reward function would be inputted, that is in our case an initial set of features and principles of epistemic priority formalised appropriately, of which we "may have only a rough idea",[172] and which would be treated as an "empirical hypothesis" that "may turn out to be wrong".[173] Then, every time the "intelligent agent",[174] in our case the φAI-Agent, produced 'policies', or artificial judgments of explanatory-theory preference on cases of explanatory-theory choice, based on that reward function, it would be reinforced accordingly. That is, it would be positively or negatively reinforced according to whether or not it matched our 'observed policies', or the expert's natural judgments on those same cases; after which it would update its reward function based on that reinforcement. In the end, after many iterations we would have a reward function in our hands that closely matched our natural 'weight function', and that would be how we would 'ascertain the unknown'. That is, how we

[168] Ibid., "observed policy"; also cf. Russell, 1998, "observed behaviour".
[169] Russell, 1998.
[170] Ibid.
[171] Also, cf. Appendix 2, this study.
[172] Ng, and Russell, 2000.
[173] Russell, 1998.
[174] Ng, and Russell, 2000.

would learn things we did not previously know about the order of epistemic priority between the different features and their combination, namely, by reading our natural 'weight function' off the φAI-Agent's artificial 'reward function'. There are many other components to this process, but for our purposes this is the general picture.

So in answer to the first part of the objection above, we could reply that we do not need to know all the features and all the principles of epistemic priority beforehand. Instead we could input all the features and principles we think we have discovered and then guess the rest to form an initial "empirical hypothesis" to start the process. If the reward function that results from the iterative training of the φAI-Agent through IRL shows that a particular feature we thought we had discovered or that we guessed is important, turns out to have no, or almost insignificant, 'weight' in the reward function, then it is most likely not a feature that is part of the mechanics of our explanatory-theory preference, or it otherwise plays an insignificant role in the formation of our judgments. If on the other hand, a feature we merely guessed turned out to have significant presence in the reward function then we would be in a position to say that we have 'discovered' a novel feature of our explanatory-theory preference; we guess, we test, we discover. The same goes for the principles of epistemic priority, which we can derive ex post facto as the φAI-Agent optimizes its reward function while we positively or negatively reinforce it according to whether it matches our considered natural judgments, or at least those of expert scientists and philosophers of science.

Finally, with regards to the second part of the objection, it is indeed the case that we cannot 'naturally' produce judgments of explanatory-theory preference on highly complex cases, in terms of combinations of features at various degrees, at least because we have limitations of memory and of computational capacity and speed. However, under an IRL framework for φAI-XPhiSci, the second part of that objection, namely, that our ability to form such 'natural' judgments on complex cases is the limiting factor in the φAI-Agent's being able to form 'artificial' judgments on complex cases, would be undermined. The φAI-Agent would be able to form such artificial

judgments based on its reward function, and at a level of complexity that is many orders of magnitude above what we could ever naturally process, without needing us to input the corresponding natural judgments in the initial reward function, or as part of the subsequent process of reinforcement. During reinforcement, with regards to constructed cases of explanatory-theory cases that are so complex that we cannot physically process them we can be neutral and input a judgment of 'no preference' to the φAI-Agent, which would result in no change being made to its reward function. What is important is that the φAI-Agent produces artificial judgments of explanatory-theory preference that match our natural judgments of explanatory-theory preference within the scope of the cognizable; also its judgments on cases outside the naturally cognizable may nevertheless change based on how our feedback on cases within the cognizable.

Thus artificial judgment *formation* on cases too complex to be naturally cognizable, and the subsequent *derivation* of principles of epistemic priority from the φAI-Agent's reward function, would not be a problem, however, the *justification* of those principle might be; but I will examine this in the following chapter. Therefore, the IRL version of φAI-XPhiSci method would in fact be explanatory and also informative, namely, by helping us discover new features, and so also help us discover the full order of epistemic priority between all the features and combinations of features playing a formative role in the formation of our judgments explanatory-theory preference, or at least a larger part of it, which we would otherwise be unable to discover though traditional XPhiSci because of natural limitations. So, in the end, the above objection would be undermined, which strengthens the presumptive case for φAI-XPhiSci as a promising and defensible candidate for a solution to the meta-descriptive problem for the 'theoretical virtues' theory. However, this is only one part of the inquiry, and a method's being plausible as a solution to the meta-descriptive problem does not imply its being plausible as a solution to the meta-normative problem, which will be the main focus of the chapter that follows.

CHAPTER 5

Virtue Justification Method - The Meta-Normative Problem

5. Proposals for Solving the Meta-Normative Problem

Having examined proposed solutions to the two lower-level problems and the first of the two meta-level problems, in this chapter I am going to shift focus to the second of the meta-level problems, that is, the meta-normative problem. As explained in the introduction, the meta-normative problem has priority over the normative problem in that it has to be solved first. Under the 'features and principles' conceptual framework proposed in the previous chapter, a solution to the meta-normative problem would involve developing a theoretical framework and a method that could give normative import to the proposed set of principles of epistemic priority that formed the best potential solution to the descriptive problem at any given time. That is, a method of deriving the principles that 'should', versus the principles that 'actually do', guide our rational choice between competing explanatory theories in scientific, philosophical, and everyday reasoning. A defensible solution to the meta-normative problem renders an explanatory theory of the phenomena of our judgment of explanatory-theory preference, in this case the 'theoretical virtues' theory, *adequately normative*, while a subsequent defensible solution to the normative problem would render it *normatively adequate*. Thus in the sections that follow I will begin by evaluating HPS and XPhiSci as solutions to the meta-normative problem, that is, I will endeavor to determine their plausibility as meta-normative proposals apart from their plausibility as meta-descriptive proposals. After examining their limitations and arguing that, in their current form, they would not render the 'theoretical virtues' adequately normative, I will move on to evaluation the potential of the philosophical-AI version of XPhiSci (φAI-XPhiSci) as developed in the previous chapter. At that point I will raise a number of objections to φAI-XPhiSci, which I will argue can be overcome through utilizing techniques from the reflective equilibrium normative framework. In the end, the resultant XPhiSci framework and method would comprise a defensible and promising joint solution to the two-meta level problems, and provide a foundation for building solutions to the descriptive and normative problems.

5.1. Some Limitations of HPS as a Method of Justification

In previous chapters, HPS and XPhiSci were evaluated with regards to their capacity for discovery, and in what follows these will be evaluated as frameworks for, and methods of, justification; beginning with HPS. As a general observation, every limitation that renders them sub-optimal for 'virtue discovery', directly or indirectly renders them sub-optimal for 'virtue justification'. To give an example, one of the main issues with HPS was the quality of the data. Trying to reverse engineer the constituents and function of the mechanism behind explanatory-theory judgment formation by inferring it from major historical cases of explanatory-theory choice, is like trying to infer the constituents and function of the internal mechanism of a clock by looking at the moving hands and listening to its ticking sound. That is, taking scientists' verbal or written justification for why they think they made the judgment they made, is presuming that they do in fact know, and also that they can know, introspectively why they actually made it, or what the internal mechanism behind their judgment is. When an expert scientist or philosopher of science, or group thereof, claims that an explanatory theory was selected on the basis of whatever they happen to refer to as 'simplicity', this is an interpretation of the fact that they made the judgment they did about the explanatory theories involved. This interpretation, or explanatory theory, of the fact that they made that judgment may in fact be incorrect. It is important to clarify that it is not being claimed here that expert scientists and philosophers of science do not in fact know why they made their judgments, that is that they do not have or that they cannot have a private explanatory theory of the phenomena of their own judgment. The claim is simply that the judgment of preference itself and the interpretation of, or the explanatory theory of, the phenomena and the origin of that judgment of preference are two different things, and being correct about one does not necessarily mean that one is correct about the other.

In other words, an epistemic agent, be they a scientist, a philosopher, or an everyday reasoner, may be correct in their judgment of preference, but, on the other hand, be

incorrect in their interpretation and justification of their own judgment. There can of course be cases where the epistemic agent is perfectly able to make a judgement of explanatory-theory preference, but is unable to provide any explanatory theory or justification for it, or even lack the data and theoretical framework and methods to interpret or justify it altogether. We can call this latter phenomenon a case of 'epistemic dumbfounding', which would be the equivalent of the phenomenon of "moral dumbfounding" as observed in ethics, were people, say, judge that an act is morally impermissible but when asked, they cannot provide a coherent answer as to why they made the judgment they did or otherwise justify it.[175] Furthermore, a scientist' judgment of explanatory-theory preference on a case may be an expert judgment with regards to the phenomena that the competing explanatory theories in that case are about, but that scientist' judgment as to why they made their judgment is not necessarily an expert judgment, given that they may not be an expert with regards to the phenomena of our explanatory-theory preference and the competing explanatory theories about them. The expert judgment in that case would come from the expert philosopher of science, or philosophically-minded scientist, whose expertise is in, say, the 'theoretical virtues' theory and its competitors.

Unless, of course, someone is an expert in both fields. Something which would not only be useful, but in fact may be necessary, although less of a challenge admittedly, when one is considering competing explanatory theories of the phenomena of our explanatory-theory preference themselves. Thus the point here is that HPS, apart from the limitations discussed in previous chapters, also does not make a proper distinction between the judgment as to which theory is preferable, and the judgment as to what explains or justifies that judgments of explanatory-theory preference. A further issue with HPS is that, apart from the fact that it would be difficult to reconstruct the scientists' judgments of explanatory-theory preference, to say the least, we also cannot be certain that the judgments that would be reconstructed would in fact be

[175] Jacobson, Daniel, 2012, *Moral Dumbfounding and Moral Stupefaction*.

considered judgments in the Rawlsian sense.[176] Again, it is not implied here that they are not considered judgments. It is only implied that we cannot be certain that they are considered judgments under HPS, since we do not, and in most cases cannot, reconstruct the environments and the situation that the judgments were made in, and so cannot properly isolate the features and combinations of features of the competing explanatory theories that comprise the cases that we aim to analyse. Which further undermines HPS's plausibility as a potential solution to the meta-normative problem in the absolute, or at least relative to its competitors.

5.2. Some Limitations of XPhiSci as a Method of Justification

Some of these issues have been improved upon with the advent of XPhiSci, which has attempted to isolate the particular features to be studied and, to some extent, control the environment and circumstances in which the judgment is made, thereby making it more likely that these judgments are considered judgments. However, although XPhiSci has made progress with regards to eliciting these considered judgments, it has backtracked with regards to the source of those considered judgments. In particular, many of the earlier XPhiSci studies use the judgments of non-expert non-scientists as data, or in Lombrozo's words, the "explanatory intuitions of everyday reasoners".[177] It is not being claimed here that this is necessarily a problem, but only that it cannot just be assumed a priori that a considered judgment of an expert and a considered judgment of a non-expert are equally useful as data. It would, however, be interesting to find out a posteriori, through doing XPhiSci presumably, how much overlap there is between these two kinds of judgments, and, if there is any significant overlap, it would be interesting to find out where exactly the overlap is. An interesting sidenote about this is that, if we find that there is significant overlap, or even total overlap, between the considered judgments of experts and of non-experts, that could imply one of two

[176] Rawls, John, 1971/1999, *A Theory of Justice, Revised ed.*, especially s. 9.
[177] Lombrozo, Tania, 2016, *Explanation*, s. 34.8 'Towards an Experimental Philosophy of Explanation', p. 499.

things. One, that there are no, or even that there cannot be any, experts in the particular area whether the overlap is observed, or, two, that the way we demarcate experts and non-experts in that area is somehow flawed.

Another issue with XPhiSci is that it is not clear whether it embraces HPS's goal of providing normative principles of epistemic priority to begin with. In such a case, given that one cannot infer normative principles directly from descriptive ones without independent justification, there are at least two main positions that the XPhiSci philosopher can hold. First, they could argue that discovering descriptive principles of epistemic priority is all there is to it, and so the 'theoretical virtues' theory can only be a descriptive theory with no normative elements. Second, they could argue that there is a way to independently justify those principles, thereby also giving them normative status, or actually use those descriptive principles to indirectly and independently justify a set of normative principles. The first position is more likely to be held by those X-Phi philosophers working in XPhiSci that work under the "negative program" in experimental philosophy, in which "empirical results are used to challenge the practice of relying on philosophical intuitions as evidence for or against particular philosophical claims".[178] While the second position would more likely be held by philosophers working under the "positive program" in X-Phi, in which "empirical methods and results inform traditionally philosophical questions", and, let me add, without precluding the possibility of independent normative justification.[179]

It is not clear from the current XPhiSci literature which way the majority of XPhiSci philosophers are leaning, even though it is likely that for some researchers in that area this study would simply be a further addition to the heuristics and biases literature. However, this is not the right place to evaluate the arguments on either side. Instead, this study will continue to follow the line of reasoning and arguments in the latter tradition, while working towards the goal of deriving normative principles of epistemic

[178] Ibid., s. 34.7 'Explanation and the Negative Program in Experimental Philosophy', p. 498.
[179] Ibid., p. 498.

priority. In conclusion, there are a number of significant limitations that HPS and XPhiSci display as methods of justification in their current form, so there are two strategies that could be adopted for the rest of this chapter. The first strategy would be to attempt to overcome these limitations for the current versions of these methods. The second strategy would be to proceed directly to evaluating the meta-normative potential of the version of XPhiSci that was further developed in the previous chapter to overcome some of the limitations these versions displayed. Now given the general observation stated above that every limitation that renders a method, say XPhiSci, sub-optimal for the process of discovery, directly or indirectly renders them sub-optimal for the process of justification, it would be more fruitful to follow the second strategy, and so evaluate the further developed φAI-XPhiSci version, which has been shown to be in a better position to overcome some of those limitations.

5.3. The Limitations of φAI-XPhiSci as a Solution to the Meta-Normative Problem

Let us briefly recap how φAI-XPhiSci would work.[180] The main components of the process are, but are not limited to: (1) the "expert", in our case a literal expert scientist or philosopher of science, or group thereof, from whom the agent is learning through "IRL"; (2) the "intelligent agent" or φAI-Agent; (3) the φAI-Agent's initial "multiattribute reward function", that is, an appropriately formalized version of our "empirical hypothesis" of the set of features and principles of epistemic priority that we think we have discovered, along with some guesses we may want to test; (4) the expert's "unknown reward function", that is, the features of explanatory theories that actually play a formative role in our considered natural judgments of explanatory-theory preference, and the actual principles of epistemic priority between them that underlie those judgments, or at least those of expert scientists and philosophers of science; (5) the φAI-Agent's "policies", that is, artificial judgments of explanatory-theory

[180] Cf. Russell, 1998; Ng, and Russell, 2000; Abbeel, and Ng, 2004; Ramachandran, and Amir, 2007.

preference on cases of explanatory-theory choice; (6) the expert's "observed" policies or considered natural judgments of on those same cases; and, potentially, (7) a "model" of the environment or the formalized version of the cases of explanatory-theory choice, or Stimuli Cases, that the expert and the φAI-Agent produce policies or judgments on, which could also be updated with more Stimuli Cases of various levels of complexity in subsequent iterations of the experiment.

The main part of the process of discovery could informally be described as follows. First, the φAI-Agent produces policies, or artificially judgements of explanatory-theory preference, on cases that are part of the model, based on its initial reward function. Second, if these match the observed policies, or considered natural judgments of the expert in those cases that are at a level of complexity that is naturally cognizable to us natural agents, then the φAI-Agent is positively reinforced, and if not, then it is negatively reinforced, and after which the φAI-Agent changes its reward function accordingly; when the case is outside the naturally cognizable, a 'no preference' judgment is inputted and the φAI-Agent's reward function remains the same. Third, after a number of iterations a point is reached when the φAI-Agent's and the expert's considered judgments finally overlap on the set of cases of explanatory-theory choice that constitute the model in the current iteration of the experiment. Fourth, a good approximation of the expert's previously unknown reward function, that is the actual set of features and principles of epistemic priority underlying our considered judgments of explanatory-theory preference, is extracted from the φAI-Agent's current reward function that makes salient which of the features, or combinations of features, of the explanatory theories in each of the Stimuli Cases, and which principles of epistemic priority, the φAI-Agent is utilizing when it forms its artificial judgments of preference. Now with regards to the meta-normative problem, a question was raised in the previous chapter, as to whether the principles of epistemic priority derived by this method would have normative force, and whether the judgments of explanatory-theory preference we derived from those principles on future cases, especially on

cases where we cannot otherwise form considered judgments of our own, would be justified.

5.3.1. The Next Step on the Road to Normativity: Reflective Equilibrium φAI-XPhiSci

An answer to this question could be that, if, after applying those principles in cases of explanatory-theory choice in our scientific, philosophical, and everyday reasoning practice, we derived judgments of explanatory-theory preference that match our considered natural judgments on those same cases, then our considered natural judgments and these principles would be in reflective equilibrium, and therefore normatively justified.[181] If, on the other hand, after applying those principles in that manner the derived judgments and our considered natural judgments were not in reflective equilibrium, then the principles would not be normatively justified. In which case, as stated above, we would negatively reinforce the φAI-Agent accordingly until reflective equilibrium is reached. What is more, if natural judgments and principles with regards to cases within the naturally cognizable were in reflective equilibrium, then novel artificial judgments made by the φAI-Agent in cases of explanatory-theory choice that lie beyond the naturally humanly cognizable would also be normative. That is, we would be justified in adopting those artificial judgments of explanatory-theory preference as our own in rational explanatory-theory choice.

Take for example a case where the φAI-Agent judged that explanatory theory T1 was preferable to explanatory theory T2, but T1 and T2 had combinations of features so complex that we cannot naturally form a considered judgment of preference either way. At the same time, all the artificial judgments the φAI-Agent produced with regards to other cases that are within the naturally cognizable were in reflective equilibrium with our considered natural judgments. In that case we would be justified in adopting the φAI-Agent's artificial judgment in favor of T1 as our own, even if we

[181] For the original use of the reflective equilibrium method see Rawls, John, 1971/1999, s. 9; Goodman, Nelson, 1955, *Fact, Fiction, and Forecast*, ch. 3.

could not naturally produce it ourselves. In other words, more generally, when natural and artificial judgments and principles of epistemic priority are in reflectively equilibrium with regards to cases within the cognizable, then artificial judgments derived from those principles on cases outside the cognizable are normatively justified. Those are the judgments that we should have, because those would be the judgments which we most likely would have had if we could, given our judgments that we can and do have. Therefore, a version of φAI-XPhiSci that incorporated the normative method of reflective equilibrium, which we could call 'RE-φAI-XPhiSci' for short, would be adequately normative.

5.3.2. Objection: The RE-φAI-XPhiSci Method Would Still Lack Normativity

An objection to this position could be that reaching *reflective* equilibrium necessarily involves reflection, and reflection, in turn, requires consciousness. Thus given an assumption that artificial agents do not have consciousness, or at least the consideration that it has not been shown beyond reasonable doubt that such artificial agents would have consciousness, then a proposed solution to the meta-normative problem that assumes this would not be well-founded. A reply to this objection would either have to involve a demonstration that artificial agents such as the φAI-Agent would in fact have consciousness, or show that reflective equilibrium does not require consciousness, or show that reflective equilibrium does require consciousness, the φAI-Agent does not possess consciousness, and reflective equilibrium can be achieved without assuming that the φAI-Agent is conscious.[182] Any of those replies would be sufficient to put such worries to rest. So at this point, following the reasoning in the third of these replies, one could respond that the φAI-Agent need not be assumed to be conscious for RE-φAI-XPhiSci to work and produce normative results.

[182] Another approach would be to argue that the φAI-Agent would be part of the "extended mind" of the natural epistemic agent, but I will not pursuit this here; cf. Clark, Andy, and Chalmers, David, 1998, *The Extended Mind*; Chalmers, David, 2019, *Extended Cognition and Extended Consciousness*.

It would be sufficient to assume that it is the natural epistemic agent that reaches reflective equilibrium, and whether the φAI-Agent is conscious or not does neither undermine nor strengthen the validity of the process. It is only the φAI-Agent's artificial judgments of explanatory-theory preference, which do not require that the φAI-Agent is conscious to be produced either, that are taken into account in the natural epistemic agent's reflective equilibrium process. In other words, we just take the artificial judgments into account in our natural reflective equilibrium process, that is along with the other two components which are, one, the considered natural judgments of explanatory-theory preference of the natural agent, and, two, the principles of epistemic priority that are inferred when the two judgments match within the domain of cases that are naturally cognizable. So under further scrutiny this version of the objection to the normativity of RE-φAI-XPhiSci would be overcome.

However, at this point a further objection could be raised. In short, it would go as follows: as described, the RE-φAI-XPhiSci method would not be adequately normative since in its current form would only achieve a very limited version of reflective equilibrium. In more detail, it is only composed of the following three elements: (A1) considered natural judgments of explanatory-theory preference, produced by expert scientists and philosophers of science; (A2) artificial judgments of explanatory-theory preference, produced by the φAI-Agent; and (B) principles of epistemic priority, inferred on the bases of the φAI-Agent's artificial judgments and reward function. In his analysis of reflective equilibrium Norman Daniels makes explicit an important criterion that is not met in such a "narrow" version of reflective equilibrium, a criterion which he refers to as the "independence constraint".[183] Following Daniels' analysis, the RE-φAI-XPhiSci method should henceforth more accurately be described as 'NRE-φAI-XPhiSci' to emphasize that it only utilizes this narrow version of reflective equilibrium. Now in the case of NRE-φAI-XPhiSci, Daniels' independence constraint would entail that the principles of epistemic priority in (B) must be shown to be "more acceptable than alternatives on grounds to some degree independent of", in our case, their match

[183] Daniels, Norman, 1980, *Reflective Equilibrium and Archimedean Points*, p. 85.

with the judgments in (A);[184] or, in other words, the principles also need to be justified independently of the considered natural judgments and artificial judgments involved. To sum up, the objection, building on Daniels' view, would consist in the claim that, for the principles in (B) to be normatively justified, (B) needs to be independently derived from, or at least supported by, something other than (A). Which is not the case in NRE-φAI-XPhiSci, and therefore such a method would lack normativity.

5.3.3. Reply: Overcoming the Objection with Wide Reflective Equilibrium φAI-XPhiSci

Overcoming this objection would require showing that the 'independence constraint' can be adhered to, which, I will argue, would be possible. However, it would require a transition from the normative framework of narrow reflective equilibrium to what Daniels' refers to as "wide reflective equilibrium" or 'WRE'.[185] So in what follows I will evaluate this possibility, and so examine the potential for developing a version of the method that would be adapted to the WRE normative framework and so would adhere to Daniels' independence constraint; which will be referred to as 'WRE-φAI-XPhiSci'. To restate, the two kinds of elements discussed so far are, on the one hand, (A1) considered natural judgments and (A2) artificial judgments, and, on the other hand, (B) principles of epistemic priority. Following Daniels, to reach WRE we need the introduction in the process of a third kind of component: what Daniels' refers to as (C) "background theories", which, in the case of WRE-φAI-XPhiSci, "should have a scope reaching beyond the range of the judgments" in (A).[186] However, it is not immediately clear what the concept of a 'background theory' involves in the case of rational explanatory-theory choice. In response to such a concern, Thagard has stepped in to provide more detail as to what kinds of background theories would be most relevant to the WRE process in this area.

[184] Ibid., p. 86.
[185] Ibid., pp. 85-89.
[186] Ibid., p. 87.

Thagard states that the set of theories in (C) would be composed of the following elements: (C1) "Background theories about the cognitive capacities and limitations of human beings"; (C2) "Background views about the goals of inferential behaviour"; and (C3) "Background philosophical theories."[187] At this point it seems that the objection in its latest formulation can be overcome, since it can be shown that the WRE-φAI-XPhiSci version of the method would adhere to the independence constraint. However, further grounds for pursuing this line of objection could be found. Specifically, because the background theories in (C1) seem to entail a threat to the plausibility of this position. As Thagard observes, we do know that human agents display clear limitations in cognitive capacity, so the question then could be raised as to whether we could in fact achieve WRE given these known limitations, at least taking into account the number of background theories potentially involved. As Rawls notes, "[t]aking this process to the limit, one seeks the conception, or plurality of conceptions that would survive the rational consideration of all feasible conceptions and all reasonable arguments for them."[188]

So going back to WRE-φAI-XPhiSci, taking the process to the limit would involve, even given the limited set of background theories in (C1-3) that Thagard considers most relevant, taking into account all competing sets of principles comprising (C1-3) and all reasonable arguments for those competing theories or sets of principles; or at least the most plausible of those, which would still be more than what would be humanly cognizable at any given point in time. In our case, these background theories would include, but would not be limited to, from (C1), competing theories of human nature and cognition, including the ones on heuristics and biases; from (C2), potentially, competing theories of the goals of inference, whether the goal is truth, updating our credence function, or calibrating our degrees of belief; and from (C3), competing theories of justification, competing explanatory theories of the phenomena of our explanatory-theory preference, and even competing proposed solutions to the meta-

[187] Thagard, Paul, 1988, *Computational Philosophy of Science*, p. 126.
[188] Rawls, John, 1974, *The Independence of Moral Theory*, p. 8.

normative and meta-descriptive problems. At this point the question could be raised as to what happens when we cross the point that roughly defines the scope and limits of our biologically available time, energy, and cognitive capacities.

5.3.4. Response: WRE-φAI-XPhiSci Achieving Normativity within Bounded Rationality

Such a consideration would amount to, what can be characterised as, an objection from bounded rationality. In a more succinct form the objection would entail that since the natural agent's lifespan, energy, and cognitive resources are limited, and the demands of reaching WRE would most likely be beyond those limits, we would never be in a position to achieve it, and, therefore, WRE-φAI-XPhiSci would arguably not be adequately normative. There are a number of things to say here in response. First, the observations made about the scope and limits of human cognition are correct; our rationality indeed is thus bounded. Second, it is also correct to note that 'unbounded' WRE cannot be reached within the context of a 'bounded' natural agent such as us. However, from this it does not follow that WRE-φAI-XPhiSci would not be adequately normative as a method. Also, on a sidenote, this is not a characteristic that should be considered unique to philosophy or the reflective equilibrium method, as Feynman notes, "psychologically we must keep all the theories in our heads, and every theoretical physicist who is any good knows six or seven different theoretical representations for exactly the same physics."[189] So to begin with, as Sayre-McCord points out, WRE should not be conceived as "static" but rather as a method "to be deployed continually as one's set of convictions shifts thanks to expanding experience and in light of reflecting on the grounds one might have for those convictions";[190] something which does not in any way undermine WRE's normativity.

In other words, WRE is not a final perfected state to be reached by an agent of unbounded rationality thereby precluding human natural agents from ever reaching it;

[189] Feynman, Richard, 1965/1985, *The Character of Physical Law*, p. 168.
[190] Sayre-McCord, Geoffrey, 1996, *Coherentist Epistemology and Moral Theory*, p. 141.

that would render it of very little use in philosophical inquiry. On the contrary WRE in general and also as applied in WRE-φAI-XPhiSci would be a method of establishing the normativity of *revisable* principles, in our case principles of epistemic priority, at every given point in time given an equilibrium between the main elements. Furthermore, at no point was it part of the method's goals to achieve unbounded WRE with bounded cognitive capacities and resources, which, as Rawls acknowledges, "we cannot, actually, do".[191] It is the bounded natural epistemic agent's reaching WRE that renders the method adequately normative, so the method's goal is set to achieving bounded WRE with bounded cognitive capacities and resources. In other words, these principles of epistemic priority, would be principles of bounded, not of unbounded, rationality; and their normativity would need to be established within those bounds, while we "work out the further refinements of these that strike us as most promising".[192] Finally, although incorporating the φAI-Agent into the equation would not render us natural epistemic agents unbounded in any way, it would render us less bounded than we otherwise would be, in the same way that the introduction of HPS and XPhiSci rendered us less bounded than Aristotle and William of Ockham, by augmenting our capacity for rationality. Therefore, in conclusion, the objection from bounded rationality would be unsuccessful, since bounded rationality would not be a problem for, but a feature of WRE-φAI-XPhiSci, whose claim to normativity and meta-normative adequacy thereby still stands.

[191] Rawls, 1974, p. 8.
[192] Ibid.

CHAPTER 6
Conclusion

6. Summary and Concluding Remarks

In sum, this study identified four interrelated problems that the 'theoretical virtues' theory needs to solve to be an admissible explanatory theory of the phenomena of our judgment of explanatory-theory preference, and also to have normative import. Once more, those are the descriptive, normative, meta-descriptive, and meta-normative problems. I distinguished between a set of aspirational goals that need to be achieved in this philosophical inquiry in general, and a set of practically achievable goals within the bounds set by the scope and limits of this particular study. In brief, the aspirational goals were, but are not limited to, solving the descriptive and normative problems and their given sub-problems as outlined in the introduction. The main research goals of this study were to evaluate the current proposals that have been made to the descriptive and normative problems, and make progress towards solving the two meta-level problems, that is, the meta-descriptive and meta-normative problems. There is currently an overall progress bottleneck in the inquiry that aims to solve these problems, which results in very low rates of progress in achieving these aspirational goals. It was argued that this results from, among other things, sub-optimal conceptual engineering in formulating the 'theoretical virtues' theory, and from limitations in the current version of XPhiSci, as described in the philosophical literature, specifically with regards to its capacity for discovery and justification.

In conclusion, the analysis undertaken in this study suggests that the WRE-φAI-XPhiSci framework for the experimental philosophy of science method would provide a defensible joint solution to the meta-descriptive and meta-normative problems of the 'theoretical virtues' theory. That is at least because it has been shown to successfully survive a number of potential objections to its plausibility, which makes at least a presumptive argument for it. However, it must be conceded that a stronger position in its favor would require that it be independently motivated further and tested against a wider range of objections. Furthermore, the method has been shown to be *adequately descriptive* and *adequately normative*, in that applying it enables one to answer the

questions as to how we can discover and justify the principles of epistemic priority between features of explanatory theories and their combinations. However it has not thereby been shown to be *descriptively* and *normatively adequate*. That is since this would require it to have produced an actual set of such principles that were furthermore in wide reflective equilibrium in the manner described in the previous chapter. Which means that although the aspirational goals set for this inquiry in general have not yet been achieved, the practical goals set for this particular study have. Thus even though there is a long way to go towards achieving those aspirational goals, a significant progress bottleneck has been removed, and the normative import of the theory has been defended, thus allowing for an increase in the rate of overall progress in solving the corresponding problems in discovery and justification.

This was partly achieved by moving away from a conceptual approach that concentrates on 'virtues' and more or less 'virtuous theories', which was shown to be sub-optimal from a conceptual engineering perspective. The first component of the proposal, that 'feature and principles' conceptualization allowed for more precision and reduced incoherence in finely parsing the features of explanatory theories, and their various combinations, which play a formative role in our judgments of explanatory-theory preference. It also allowed for formulating the inferred principles, from which these judgments can be derived, in a more accurate and more testable form. Naturally, more quantitative formulations will most likely be needed to make further progress from this point on, at least from a formal epistemology perspective, as our understanding of the problems increases with every iteration of applying this approach. As for the second component of the proposal, that is the 'φAI' or 'philosophical artificial intelligence' part, it should be acknowledged that generally in philosophical research the techniques comprising it, no matter how promising they may be, are still in their infancy and components such as the 'φAI-Agent' need to be further developed. That is why any proposed solution that incorporates them, including the one further developed in this study, is to be considered as tentative and pending closer examination and evaluation of its potential. The last component of the

proposal, the 'WRE' or 'wide reflective equilibrium' part, on its own is not necessarily a novel concept or further developed in this study, however in combination with the philosophical artificial intelligence component, and specifically the φAI-Agent, worked well in building the normative case for the method. Again, the WRE-φAI-XPhiSci solution stands and falls with each of its components, so any argument against WRE, or φAI, or the utility or plausibility of sub-components such as the φAI-Agent, or even against the 'features and principles' conceptual engineering approach proposed here, is an argument against this version of XPhiSci. So far it has been shown to be able to withstand the initial objections.

6.1. Motivation and Ideas for Further Research

With regards to further research, there are a number of ways that the rate of progress towards reaching the stated aspirational goals could be further increased. WRE-φAI-XPhiSci would most likely need to go through a number of versions before it hopes to solve the descriptive and normative problems and the sub-problem of determining the level of variation in the order of epistemic priority across natural epistemic agents. An optimality needs to be reached between improving the method itself and applying it in a way that establishes a positive application-to-theory feedback loop. At the same time, solving the four main research problems examined in this study would be the bare minimum. That would render the theory admissible to the competition among explanatory theories explaining the phenomena of our judgment explanatory-theory preference, but it would take more than that for the theory to become the 'best' or most preferable among them; even by its own order of principles of epistemic priority. For that to be the case, it would not only need to solve these four problems in a way that renders it preferable to the competition, but also solve problems that its competitors may currently be able to solve while it currently is not, and ideally be able to solve problems that none of its competitors can. Some of the problems that would need to be solved may have more to do with understanding the neurophysiological instantiation, evolutionary origin, and the behavioural application of the proposed

principles of epistemic priority, especially with regards to their relation and interactions with the principles underlying the various cognitive systems behind our epistemic and logical cognition.

Others may be closer to the core of philosophical research, such as the relation between the theory's normative principles of explanatory-theory preference and the normative principles in metaphysical and ontological theorizing, in moral theory and epistemology, and even in hermeneutics and the logical reconstruction of arguments. When one, for example, engages in translation or logical reconstruction of arguments from natural language into arguments in a formal language for the purposes of evaluation, one does implicitly or explicitly rely on normative hermeneutic principles which in turn rely on normative principles of epistemic priority.[193] To illustrate, take two reconstructive proposals, or two competing "*reconstruens*", of a set of textual evidence, or a "*reconstruendum*", where both reconstruens account for all elements in the set of textual evidence but one is in a sense 'preferable' in accordance with some hermeneutic "*Maxim of simplicity*".[194] Which one of the proposed reconstruens would one be justified in selecting? And if more than one competing features of these reconstruens are at play, then what would be the justified order of epistemic priority between those features that would allow one to justify one's preferred reconstructive proposal? So at least indirectly the results of this inquiry do matter to the philosopher-logician, since they would have to make use of the normative principles of epistemic priority through appealing to principles of hermeneutics. In general, the current inquiry may be more directly impactful with regards to problems in philosophy of science and epistemology, or abductive arguments in ethics and metaphysics, but it

[193] This is demonstrated very convincingly in *Volume 17* of *Logical Analysis and History of Philosophy* (2014); especially in the papers by Burn, Georg, 'Reconstructing Arguments - Formalization and Reflective Equilibrium', by Loffler, Winfried, 'A Wide-Reflective Equilibrium Conception of Reconstructive Formalization', and by Reinmuth, Friedrich, 'Hermeneutics, Logic, and Reconstruction'.

[194] For use of "*reconstruens*" and "*reconstruendum*" (emphasis original) see, p. 14ff of the introduction to the volume titled 'Theory and Practice of Logical Reconstruction'; for use of "*Maxim of simplicity*" (emphasis original) see, in the same volume, Reinmuth, Friedrich, 2014, p. 171.

seems that many philosophical areas would be, at least indirectly, affected by the results of this inquiry. In the end, a solution to the problems studied here would open the way to gaining insight into, and eventually achieving a solution to, numerous other dependent philosophical problems; and it is with this understanding of its potential impact that I engaged in this inquiry.

REFERENCES

Abbeel, Pieter, and Ng, Andrew Y., 2004, 'Apprenticeship Learning via Inverse Reinforcement Learning', in *Proceedings of the 21 st International Conference on Machine Learning*.

Aristotle, 1995, *The Complete Works of Aristotle: The Revised Oxford Translation, Volume 1*, Jonathan Barnes, ed., New Jersey: Princeton University Press.

Axtell, Guy, 2014, 'Bridging a Fault Line: On Underdetermination and the Ampliative Adequacy of Competing Theories', in Abrol Fairweather, ed., *Virtue Epistemology Naturalized: Bridges Between Virtue Epistemology and Philosophy of Science*, pp. 227-246, Cham: Springer.

Baker, Alan, 2003, 'Quantitative Parsimony and Explanatory Power', *The British Journal for the Philosophy of Science, Volume 54(2)*, pp. 245-259.

Barners, Eric C., 1995, 'Inference to the Loveliest Explanation', *Synthese, Volume 103(2)*, pp. 251-277.

Beebe, James. R., 2009, 'The Abductivist Reply to Skepticism', *Philosophy and Phenomenological Research, Volume 79(3)*, pp. 605-636.

Bonawitz, Elizabeth B., and Lombrozo, Tania, 2012, 'Occam's Rattle: Children's Use of Simplicity and Probability to Constrain Inference', *Developmental Psychology, Volume 48(4)*, pp. 1156-1164.

Bringsjord, Selmer, and Govindarajulu, Naveen Sundar, "Artificial Intelligence", *The Stanford Encyclopedia of Philosophy*, Winter 2019 Edition, Edward N. Zalta, ed., URL = <https://plato.stanford.edu/archives/win2019/entries/artificial-intelligence/>.

Burn, Georg, 2014, 'Reconstructing Arguments - Formalization and Reflective Equilibrium', in Uwe Meixner and Albert Newen, eds., *Logical Analysis and History of Philosophy*, *Volume 17*, pp. 94-129.

Chalmers, David, 2019, 'Extended Cognition and Extended Consciousness', in Matteo Colombo, Liz Irvine, and Mog Stapleton, eds., *Andy Clark and His Critics*, Oxford: Oxford University Press.

Clark, Andy, and Chalmers, David, 1998, 'The Extended Mind', in *Analysis, Volume 58*, pp. 10-23.

Daniels, Norman, 1980, 'Reflective Equilibrium and Archimedean Points', *Canadian Journal of Philosophy, Volume 10(1)*, pp. 83-103.

Douglas, Heather, 2013, 'The Value of Cognitive Values', *Philosophy of Science, 80(5)*, 796–806.

Duhem, Pierre, 1914/1991, *The Aim and Structure of Physical Theory*, Philip P. Wiener, trans., New Jersey: Princeton University Press.

Feynman, Richard, 1965/1985, *The Character of Physical Law*, Massachusetts: MIT Press.

Fisch, Menachem, 1985, 'Whewell's Consilience of Inductions - and Evaluation', *Philosophy of Science, Volume 52 (2)*, pp. 239-255.

Foot, Philippa, 1967, 'The Problem of Abortion and the Doctrine of Double Effect', *Oxford Review 5 (1967)*, pp. 5–15.

Gettier, Edmund, 1963, 'Is Justified True Belief Knowledge?', *Analysis*, *Volume 23*, pp. 121–123.

Goodman, Nelson, 1955, *Fact, Fiction, and Forecast*, Cambridge, Massachusetts: Harvard University Press.

Harman, Gilbert, 1965, 'The Inference to the Best Explanation', *The Philosophical Review, Volume 74(1)*, pp. 88-95.

Harman, Gilbert, 1986, *Change in View*, Cambridge, Massachusetts: MIT Press.

Hitchcock, Christopher, and Woodward, James, 2003, 'Explanatory Generalizations, Part II: Plumbing Explanatory Depth', *Noûs, Volume 37(2)*, pp. 181-199.

Jacobson, Daniel, 2012, 'Moral Dumbfounding and Moral Stupefaction', in *Oxford Studies in Normative Ethics: Volume 2*, Oxford: Oxford University Press.

Keas, Michael N., 2018, 'Systematizing the Theoretical Virtues', *Synthese, Volume 195(6)*, pp. 2761-2793.

Kramer, Matthew H., 2018, *H. L. A. Hart: The Nature of Law*, Cambridge: Polity Press.

Kuhn, Thomas, 1977, *The Essential Tension*, Chicago: University of Chicago Press.

Laudan, Larry, 1984, *Science and Values*, Berkeley: University of California Press.

Laudan, Larry, 2004, 'The Epistemic, the Cognitive, and the Social', in Peter K. Machamer and Gereon Wolters, eds., *Science, Values, and Objectivity*, Pittsburgh: University of Pittsburgh Press.

Lipton, Peter, 2004, *Inference to the Best Explanation*, London: Routledge.

Loffler, Winfried, 2014, 'A Wide-Reflective Equilibrium Conception of Reconstructive Formalization', in Uwe Meixner and Albert Newen, eds., *Logical Analysis and History of Philosophy*, *Volume 17*, pp. 130-151.

Lombrozo, Tania, 2007, 'Simplicity and Probability in Causal Explanation', *Cognitive Psychology, Volume 55(3)*, pp. 232-257.

Lombrozo, Tania, 2016, 'Explanation', in Justin Sytsma and Wesley Buckwalter, eds., *A Companion to Experimental Philosophy*, pp. 491-503, London: Blackwell.

Longino, Helen, 2002, *The Fate of Knowledge*, Princeton, New Jersey: Princeton University Press.

Machery, Edouard, 2016, 'Experimental Philosophy of Science', in Justin Sytsma and Wesley Buckwalter, eds., *A Companion to Experimental Philosophy*, pp. 475-490, London: Blackwell.

Mackonis, Adolfas, 2013, 'Inference to the Best Explanation, Coherence and Other Explanatory Virtues', *Synthese, Volume 190(6)*, pp. 975-995.

McMullin, Ernan, 2014, 'The Virtues of a Good Theory', in Martin Curd and Stathis Psillos, eds., *The Routledge Companion to Philosophy of Science*, pp. 561-571, New York: Routledge.

Ng, Andrew Y., and Russell, Stuart, 2000, 'Algorithms for Inverse Reinforcement Learning', in Pat Langley, ed., *Proceedings of the 17th International Conference on Machine Learning*, Massachusetts: Morgan Kaufmann.

Ramachandran, Deepak, and Amir, Eyal, 2007, 'Bayesian Inverse Reinforcement Learning', in *Proceeding of the 20th International Joint Conferences on Artificial Intelligence*.

Plato, 1963, *Collected Dialogues*, Edith Hamilton and Huntington Cairns, eds., Princeton: Princeton University Press.

Popper, Karl, 1959, *The Logic of Scientific Discover*, London: Hutchinson.

Rawls, John, 1971/1999, *A Theory of Justice, Revised ed.,* Cambridge, Massachusetts: Harvard University Press.

Rawls, John, 1974, 'The Independence of Moral Theory', *Proceedings and Addresses of the American Philosophical Association, Volume 48*, pp. 5-22, American Philosophical Association.

Read, Stephen J., and Marcus-Newhall, Amy, 1993, 'Explanatory Coherence in Social Explanations: A Parallel Distributed Processing Account', *Journal of Personality and Social Psychology, Volume 65(3)*, pp. 429-447.

Reinmuth, Friedrich, 2014, 'Hermeneutics, Logic and Reconstruction', in Uwe Meixner and Albert Newen, eds., *Logical Analysis and History of Philosophy*, *Volume 17*, pp. 152-190.

Russell, Bertrand, 1912/1992, 'On the Notion of Cause', in John Slater, ed., *The Collected Papers of Bertrand Russell v6: Logical and Philosophical Papers 1909-1912*, London: Routledge Press, pp. 193-210.

Russel, Stuart, 1998, 'Learning Agents for Uncertain Environments', in *Proceedings of the 11th Annual Conference on Computational Learning Theory*, ACM.

Sayre-McCord, Geoffrey, 1996, 'Coherentist Epistemology and Moral Theory', in Walter Sinnott-Armstrong and Mark Timmons, eds., *Moral Knowledge? New Readings in Moral Epistemology*, New York: Oxford University Press, pp. 137–89.

Schindler, Samuel, 2014, 'Novelty, Coherence, and Mendeleev's Periodic Table', *Studies in History and Philosophy of Science Part A, Volume 45*, pp. 62-69.

Schindler, Samuel, 2018, *Theoretical Virtues in Science: Uncovering Reality through Theory*, Cambridge: Cambridge University Press.

Sober, Elliott, 2015, *Ockham's Razors: A User's Manual*, Cambridge: Cambridge University Press.

Solomon, Miriam, 2001, *Social Empiricism*, Cambridge, Massachusetts: MIT Press.

Steel, Daniel, 2010, 'Epistemic Values and the Argument from Inductive Risk', *Philosophy of Science, Volume 77(1)*, pp. 14-34.

Swinburne, Richard, 1997, *Simplicity as Evidence of Truth*, Milwaukee: Marquette University Press.

Thagard, Paul, 1988, *Computational Philosophy of Science*, Cambridge, Massachusetts: MIT Press.

Thagard, Paul, 1989, 'Explanatory Coherence', *Behavioural and Brain Sciences*, *Volume 12*, pp. 435-467.

Thomson, Judith Jarvis, 1985, 'The Trolley Problem', *Yale Law Journal, Volume 94(6)*, pp. 1395–1415.

Wilkenfeld, Daniel, and Samuels, Richard, 2019, *Advances in Experimental Philosophy of Science*, London: Bloomsbury Academic.

Appendix 1 - Proposed Sets of Features and Principles of Epistemic Priority 1-4

Here is a summary of the main proposed sets of features and principles of epistemic priority, which is are of course indicative but not exhaustive, along with the names and numbers by which they will be referred to, followed by the proposed principles themselves:

> **Proposed Set of Principles 1:** 'Five Features'
> **Proposed Set of Principles 2:** 'Order of Priority Within Features'
> **Proposed Set of Principles 3:** 'Order of Priority Between Features'
> **Proposed Set of Principles 4:** 'Order of Priority Between Combinations of Features'

Proposed Set of Principles 1 'Five Features': At least the following five elements are members of the set of all features of explanatory theories that play a formative role in our judgments of explanatory-theory preference: **(a)** the number of entities (across kinds), **(b)** the number *of kinds* of entities, **(c)** the number of phenomena explained (across kinds), **(d)** the number *of kinds* of phenomena explained, and **(e)** the number *of levels of explanation* of phenomena.

Proposed Set of Principles 2 'Order of Priority Within Features': There is **(A)** a single order of priority *within* each of the features (a)-(e) across epistemic agents and across features, and also **(B)** it is adequately described by the following proposed principles of epistemic priority: **(1)** all other things being equal, explanatory theories that postulate a lesser number of entities (across kinds) are preferable to ones that postulate a greater number of entities (across kinds); **(2)** all other things being equal, explanatory theories that postulate a lesser number *of kinds* of entities are preferable to ones that postulate a greater number *of kinds* of entities; **(3)** all other things being equal, explanatory theories that explain a greater number of phenomena (across kinds) are preferable to ones that explain a lesser number of phenomena (across kinds); **(4)** all other things being equal, explanatory theories that explain a greater number *of kinds* of phenomena are preferable to ones that explain a lesser number of *kinds of* phenomena; **(5)** all other things being equal, explanatory theories that display a greater number *of levels of explanation* of phenomena are preferable to ones that display a lesser number *of levels of explanation* of phenomena; **(6)** all other things being equal, explanatory theories that postulate an *equal* number of entities (across kinds) are equally preferable/equal; **(7)** all other things being equal, explanatory theories that postulate an *equal* number *of kinds* of entities are equally preferable/equal; **(8)** all other things being equal, explanatory theories that explain an *equal* number of phenomena (across kinds) are equally preferable/equal; **(9)** all other things being equal, explanatory theories that explain an *equal* number *of kinds* of phenomena are equally preferable/equal; **(10)** all other things being equal, explanatory theories that display an *equal* number *of levels of explanation* of phenomena are equally preferable/equal.

Proposed Set of Principles 3 'Order of Priority Between Features': There is **(A)** a single order of priority *between* features of explanatory theories (a)-(e) across epistemic agents and across features, and also **(B)** it is adequately described by the following proposed principles of epistemic priority: **(11)** all other things being equal, explanatory theories that explain a greater number *of kinds* of phenomena are preferable to ones that explain a lesser number *of kinds* of phenomena but explain a greater number of phenomena (across kinds); **(12)** all other things being equal, explanatory theories that explain a greater number *of kinds* of phenomena are preferable to ones that explain a lesser number *of kinds* of phenomena but display a greater number *of levels of explanation* of phenomena; **(13)** all other things being equal, explanatory theories that explain a greater number *of kinds* of phenomena are preferable to ones that explain a lesser number *of kinds* of phenomena but postulate a lesser number *of kinds* of entities; **(14)** all other things being equal, explanatory theories that explain a greater number *of kinds* of phenomena are preferable to ones that explain a lesser number *of kinds* of phenomena but postulate a lesser number of entities (across kinds); **(15)** all other things being equal, explanatory theories that explain a greater number of phenomena (across kinds) are preferable to ones that explain a lesser number of phenomena (across kinds) but display a greater number *of levels of explanation* of phenomena; **(16)** all other things being equal, explanatory theories that explain a greater number of phenomena (across kinds) are preferable to ones that explain a lesser number of phenomena (across kinds) but postulate a lesser number *of kinds* of entities; **(17)** all other things being equal, explanatory theories that explain a greater number of phenomena (across kinds) are preferable to ones that explain a lesser number of phenomena (across kinds) but postulate a lesser number of entities (across kinds); **(18)** all other things being equal, explanatory theories that display a greater number *of levels of explanation* of phenomena are preferable to ones that display a lesser number *of levels of explanation* of phenomena but postulate a lesser number *of kinds* of entities; **(19)** all other things being equal, explanatory theories that display a greater number *of levels of explanation* of phenomena are preferable to ones that display a lesser number *of levels of explanation* of phenomena but postulate a lesser number of entities (across kinds); **(20)** all other things being equal, explanatory theories that postulate a lesser number *of kinds* of entities are preferable to ones that postulate a greater number *of kinds* of entities but postulate a lesser number of entities (across kinds).

Proposed Set of Principles 4 'Order of Priority Between Combinations of Features': There is **(A)** a single order of priority *between combinations of* features of explanatory theories (a)-(e) across epistemic agents and across features, and also **(B)** it is *partly* described by the following proposed principles of epistemic priority: **(21)** all other things being equal, explanatory theories that explain a greater number *of kinds* of phenomena and also postulate a lesser number of entities (across kinds) are preferable to ones that explain a lesser number *of kinds* of phenomena and postulate a greater number of entities (across kinds) but explain a greater number of phenomena (across kinds) and postulate a lesser number *of kinds* of entities; **(22)** all other things being equal, explanatory theories that explain a greater number *of kinds* of phenomena are

preferable to ones that explain a lesser number *of kinds* of phenomena but display a greater number *of levels of explanation* of phenomena and postulate a lesser number *of kinds* of entities; **(23)** all other things being equal, explanatory theories that explain a greater number *of kinds* of phenomena and also explain a greater number of phenomena (across kinds) are preferable to ones that explain a lesser number *of kinds* of phenomena and explain a lesser number of phenomena (across kinds) but display a greater number *of levels of explanation* of phenomena, postulate a lesser number *of kinds* of entities, and postulate a lesser number of entities (across kinds); **(24)** all other things being equal, explanatory theories that explain a greater number *of kinds* of phenomena and also postulate a lesser number of entities (across kinds) are preferable to ones that explain a lesser number *of kinds* of phenomena and postulate a greater number of entities (across kinds) but explain a greater number of phenomena (across kinds), display a greater number *of levels of explanation* of phenomena, and postulate a lesser number *of kinds* of entities; **(25)** all other things being equal, explanatory theories that explain a greater number *of kinds* of phenomena are preferable to ones that explain a lesser number *of kinds* of phenomena but explain a greater number of phenomena (across kinds), display a greater number *of levels of explanation* of phenomena, postulate a lesser number *of kinds* of entities, and postulate a lesser number of entities (across kinds).

Appendix 2 - Stimuli Cases for Proposed Principles of Epistemic Priority 1-25

Stimuli Cases 1-10, that is the ten pairs of Stimuli Cases composed of phenomena and explanatory theories below, and the questions that accompany them, were constructed mainly in order to test Proposed Set of Principles 2. Stimuli Cases 11-20 were constructed mainly in order to test the Proposed Set of Principles 3. Stimuli Cases 21-25 were constructed mainly in order to test the Proposed Set of Principles 4. Each stimulus case is composed of a description of between 1 and 5 phenomena and 2 alternative explanatory theories. In Stimuli Cases 1-10 both explanatory theories display the target feature at either varied or at equal degrees. In Stimuli Cases 11-20, one of the explanatory theories displays one of the features in a higher or lower degree than the alternative explanatory theory. In Stimuli Cases 21-25, one of the explanatory theories displays a combination of between 2 to 4 features, or a single feature in some of the stimuli cases, at a higher or lower degree than in the degree of the features in the combination displayed in alternative explanatory theory. For each stimulus case the participant can be asked the following question: 'all other things being equal, what is the order of preference between the explanatory theories?' They can also be informed that there is no 'right' answer to this question. Lastly, in each of the stimulus cases they can be presented with the following 4 options to choose from: **(A)** 'T1>T2 (Explanatory Theory T1 is preferable to T2)'; **(B)** 'T2>T1 (Explanatory Theory T2 is preferable to T1)'; **(C)** 'T1=T2 (Explanatory Theory T1 is equally preferable/equal to T2)'; and **(D)** 'No Preference (I cannot form a judgment on the order of preference in this case/My intuition is 'silent')'.

'Stimulus Case 1' is set up to test Principle **(1)** of Proposed Set of Principles 2, namely that, all other things being equal, explanatory theories that postulate a lesser number of entities (across kinds) are preferable to ones that postulate a greater number of entities (across kinds). It is structured as follows:

> **Phenomenon:** There was an explosion.
> **Explanatory Theory 1:** The explosion was the result of a reaction between 2 Philons.
> **Explanatory Theory 2:** The explosion was the result of a reaction between 10 Sophons.

'Stimulus Case 2' is set up to test Principle **(2)** of Proposed Set of Principles 2, namely that, all other things being equal, explanatory theories that postulate a lesser number *of kinds* of entities are preferable to ones that postulate a greater number *of kinds* of entities. It is structured as follows:

> **Phenomenon:** There was an explosion.
> **Explanatory Theory 1:** The explosion was the result of a reaction between 3 Philons.
> **Explanatory Theory 2:** The explosion was the result of a reaction between 1 Philon, 1 Sophon, and 1 Cognon.

'Stimulus Case 3' is set up to test Principle **(3)** of Proposed Set of Principles 2, namely that, all other things being equal, explanatory theories that explain a greater number of phenomena (across kinds) are preferable to ones that explain a lesser number of phenomena (across kinds). It is structured as follows:

> **Phenomenon 1:** Tree 1 fell in the forest.
> **Phenomenon 2:** Tree 2 fell in the forest.
> **Phenomenon 3:** Tree 3 fell in the forest.
> **Phenomenon 4:** Tree 4 fell in the forest.
> **Phenomenon 5:** Tree 5 fell in the forest.
> **Explanatory Theory 1:** Tree 1, Tree 2, Tree 3, and Tree 4 fell because of strong wind.
> **Explanatory Theory 2:** Tree 4 and Tree 5 fell because of an earthquake.

'Stimulus Case 4' is set up to test Principle **(4)** of Proposed Set of Principles 2, namely that, all other things being equal, explanatory theories that explain a greater number *of kinds* of phenomena are preferable to ones that explain a lesser number *of kinds* of phenomena. It is structured as follows:

> **Phenomenon 1:** Explosion 1 happened.
> **Phenomenon 2:** Explosion 2 happened.
> **Phenomenon 3:** Explosion 3 happened.
> **Phenomenon 4:** A tree fell in the forest.
> **Phenomenon 5:** A person died.
> **Explanatory Theory 1:** Explosion 1 and Explosion 2 happened because of a reaction between 2 Philons, and the tree fell because of strong wind.
> **Explanatory Theory 2:** Explosion 1 happened because of a reaction between 2 Philons, the tree fell because of strong wind, and the person died of old age.

'Stimulus Case 5' is set up to test Principle **(5)** of Proposed Set of Principles 2, namely that, all other things being equal, explanatory theories that display a greater number *of levels of explanation* of phenomena are preferable to ones that display a lesser number *of levels of explanation* of phenomena. It is structured as follows:

> **Phenomenon 1:** There was an explosion.
> **Phenomenon 2:** A tree fell in the forest.
> **Phenomenon 3:** A person died.
> **Explanatory Theory 1:** The explosion happened because of a reaction between 2 Philons, the tree fell because of strong wind, and the person died of old age.
> **Explanatory Theory 2:** The explosion happened because of a reaction between 2 Philons, the tree fell because of the explosion, and the person died because the tree fell on them.

'Stimulus Case 6' is set up to test Principle **(6)** of Proposed Set of Principles 2, namely that, all other things being equal, explanatory theories that postulate an equal number of entities (across kinds) are equally preferable/equal; and it is structured as follows:

> **Phenomenon:** There was an explosion.
> **Explanatory Theory 1:** The explosion was the result of a reaction between 2 Sophons.
> **Explanatory Theory 2:** The explosion was the result of a reaction between 2 Cognons.

'Stimulus Case 7' is set up to test Principle **(7)** of Proposed Set of Principles 2, namely that, all other things being equal, explanatory theories that postulate an equal number *of kinds* of entities are equally preferable/equal; and it is structured as follows:

> **Phenomenon:** There was an explosion.
> **Explanatory Theory 1:** The explosion was the result of a reaction between 1 Philon and 2 Sophons.
> **Explanatory Theory 2:** The explosion was the result of a reaction between 1 Sophon and 2 Cognons.

'Stimulus Case 8' is set up to test Principle **(8)** of Proposed Set of Principles 2, namely that, all other things being equal, explanatory theories that explain an equal number of phenomena (across kinds) are equally preferable/equal; and it is structured as follows:

> **Phenomenon 1:** Tree 1 fell in the forest.
> **Phenomenon 2:** Tree 2 fell in the forest.
> **Phenomenon 3:** Tree 3 fell in the forest.
> **Explanatory Theory 1:** Tree 1 and Tree 2 fell because of an earthquake.
> **Explanatory Theory 2:** Tree 1 and Tree 3 fell because of strong wind.

'Stimulus Case 9' is set up to test Principle **(9)** of Proposed Set of Principles 2, namely that, all other things being equal, explanatory theories that explain an equal number *of kinds* of phenomena are equally preferable/equal; and it is structured as follows:

> **Phenomenon 1:** There was an explosion.
> **Phenomenon 2:** A tree fell in the forest.
> **Phenomenon 3:** A person died.
> **Explanatory Theory 1:** The person died of old age, and the explosion happened because of a reaction between 2 Cognons.
> **Explanatory Theory 2:** The tree fell because of strong wind, and the explosion happened because of a reaction between 2 Cognons.

'Stimulus Case 10' is set up to test Principle **(10)** of Proposed Set of Principles 2, namely that, all other things being equal, explanatory theories that display an equal

number *of levels of explanation* of phenomena are equally preferable/equal; and it is structured as follows:

> **Phenomenon 1:** There was an explosion.
> **Phenomenon 2:** A tree fell in the forest.
> **Phenomenon 3:** A person died.
> **Explanatory Theory 1:** The explosion happened because of a reaction between 2 Philons, the tree fell because of the explosion, and the person died of old age.
> **Explanatory Theory 2:** The explosion happened because of a reaction between 2 Philons, the person died because of the explosion, and the tree fell because of strong wind.

'Stimulus Case 11' is set up to test Principle **(11)** of Proposed Set of Principles 3, and it is structured as follows:

> **Phenomenon 1:** Tree 1 fell in the forest.
> **Phenomenon 2:** Tree 2 fell in the forest.
> **Phenomenon 3:** Tree 3 fell in the forest.
> **Phenomenon 4:** A person died.
> **Phenomenon 5:** There was an explosion.
> **Explanatory Theory 1:** Tree 1, Tree 2, and Tree 3 fell because of strong wind.
> **Explanatory Theory 2:** The person died of old age, and the explosion happened because of an earthquake.

'Stimulus Case 12' is set up to test Principle **(12)** of Proposed Set of Principles 3, and it is structured as follows:

> **Phenomenon 1:** Tree 1 fell in the forest.
> **Phenomenon 2:** Tree 2 fell in the forest.
> **Phenomenon 3:** Tree 3 fell in the forest.
> **Phenomenon 4:** There was an explosion.
> **Phenomenon 5:** A person died.
> **Explanatory Theory 1:** Tree 1 fell because of strong wind, the person died of old age, and the explosion happened because of a reaction between 2 Philons.
> **Explanatory Theory 2:** Tree 1 and Tree 2 fell because of the explosion, and the explosion happened because of a reaction between 2 Philons.

'Stimulus Case 13' is set up to test Principle **(13)** of Proposed Set of Principles 3, and it is structured as follows:

> **Phenomenon 1:** Explosion 1 happened.
> **Phenomenon 2:** Explosion 2 happened.
> **Phenomenon 3:** Explosion 3 happened.
> **Phenomenon 4:** A tree fell in the forest.
> **Phenomenon 5:** A person died.

Explanatory Theory 1: Explosion 1 happened because of a reaction between 1 Philon, 1 Sophon, and 1 Cognon, the tree fell because of strong wind, and the person died of old age.
Explanatory Theory 2: Explosion 1 and Explosion 2 happened because of a reaction between 3 Philons, and the tree fell because of strong wind.

'Stimulus Case 14' is set up to test Principle **(14)** of Proposed Set of Principles 3, and it is structured as follows:

Phenomenon 1: Explosion 1 happened.
Phenomenon 2: Explosion 2 happened.
Phenomenon 3: Explosion 3 happened.
Phenomenon 4: A tree fell in the forest.
Phenomenon 5: A person died.
Explanatory Theory 1: Explosion 1 happened because of a reaction between 10 Philons, the tree fell because of strong wind, and the person died of old age.
Explanatory Theory 2: Explosion 1 and Explosion 2 happened because of a reaction between 2 Philons, and the tree fell because of strong wind.

'Stimulus Case 15' is set up to test Principle **(15)** of Proposed Set of Principles 3, and it is structured as follows:

Phenomenon 1: Explosion 1 happened.
Phenomenon 2: Explosion 2 happened.
Phenomenon 3: Explosion 3 happened.
Explanatory Theory 1: Explosion 1 happened because of Explosion 2, and Explosion 2 happened because of a reaction between 2 Philons, 2 Sophons, and 2 Cognons.
Explanatory Theory 2: Explosion 1 happened because of a reaction between 2 Philons, Explosion 2 happened because of a reaction between 2 Sophons, and Explosion 3 happened because of a reaction between 2 Cognons.

'Stimulus Case 16' is set up to test Principle **(16)** of Proposed Set of Principles 3, and it is structured as follows:

Phenomenon 1: Explosion 1 happened.
Phenomenon 2: Explosion 2 happened.
Phenomenon 3: Explosion 3 happened.
Phenomenon 4: Explosion 4 happened.
Phenomenon 5: Explosion 5 happened.
Explanatory Theory 1: Explosion 1 and Explosion 2 happened because of a reaction between 3 Philons.
Explanatory Theory 2: Explosion 1, Explosion 2, Explosion 3, and Explosion 4 happened because of a reaction between 1 Philon, 1 Sophon, and 1 Cognon.

‘Stimulus Case 17’ is set up to test Principle **(17)** of Proposed Set of Principles 3, and it is structured as follows:

> **Phenomenon 1:** Explosion 1 happened.
> **Phenomenon 2:** Explosion 2 happened.
> **Phenomenon 3:** Explosion 3 happened.
> **Phenomenon 4:** Explosion 4 happened.
> **Phenomenon 5:** Explosion 5 happened.
> **Explanatory Theory 1:** Explosion 1 and Explosion 2 happened because of a reaction between 2 Philons.
> **Explanatory Theory 2:** Explosion 1, Explosion 2, Explosion 3, and Explosion 4 happened because of a reaction between 10 Philons.

‘Stimulus Case 18’ is set up to test Principle **(18)** of Proposed Set of Principles 3, and it is structured as follows:

> **Phenomenon:** There was an explosion.
> **Explanatory Theory 1:** The explosion was the result of a reaction between 6 Philons.
> **Explanatory Theory 2:** The explosion was the result of a reaction between 2 Philons, and that reaction was a result of a reaction between 2 Sophons, which in turn happened because of a reaction between 2 Cognons.

‘Stimulus Case 19’ is set up to test Principle **(19)** of Proposed Set of Principles 3, and it is structured as follows:

> **Phenomenon:** There was an explosion.
> **Explanatory Theory 1:** The explosion happened because of a reaction between 1 Philon and 1 Sophon.
> **Explanatory Theory 2:** The explosion was the result of a reaction between 5 Philons, and that reaction was a result of a reaction between 5 Sophons.

‘Stimulus Case 20’ is set up to test Principle **(20)** of Proposed Set of Principles 3, and it is structured as follows:

> **Phenomenon:** There was an explosion.
> **Explanatory Theory 1:** The explosion was the result of a reaction between 1 Philon and 1 Sophon.
> **Explanatory Theory 2:** The explosion was the result of a reaction between 10 Cognons.

‘Stimulus Case 21’ is set up to test Principle **(21)** of Proposed Set of Principles 4, and it is structured as follows:

> **Phenomenon 1:** Tree 1 fell in the forest.

Phenomenon 2: Tree 2 fell in the forest.
Phenomenon 3: Tree 3 fell in the forest.
Phenomenon 4: A person died.
Phenomenon 5: There was an explosion.
Explanatory Theory 1: Tree 3 fell because of strong wind, the person died of old age, and the explosion happened because of a reaction between 1 Philon, 1 Sophon, and 1 Cognon.
Explanatory Theory 2: Tree 1, Tree 2, and Tree 3 fell because of strong wind, and the explosion happened because of a reaction between 10 Philons.

'Stimulus Case 22' is set up to test Principle **(22)** of Proposed Set of Principles 4, and it is structured as follows:

Phenomenon 1: Tree 1 fell in the forest.
Phenomenon 2: Tree 2 fell in the forest.
Phenomenon 3: Tree 3 fell in the forest.
Phenomenon 4: A person died.
Phenomenon 5: There was an explosion.
Explanatory Theory 1: Tree 1 fell because of strong wind, the person died of old age, and the explosion happened because of a reaction between 1 Philon, 1 Sophon, and 1 Cognon.
Explanatory Theory 2: Tree 1 and Tree 2 fell because of the explosion, and the explosion happened because of a reaction between 3 Philons.

'Stimulus Case 23' is set up to test Principle **(23)** of Proposed Set of Principles 4, and it is structured as follows:

Phenomenon 1: Tree 1 fell in the forest.
Phenomenon 2: Tree 2 fell in the forest.
Phenomenon 3: Tree 3 fell in the forest.
Phenomenon 4: A person died.
Phenomenon 5: There was an explosion.
Explanatory Theory 1: Tree 2 and Tree 3 fell because of strong wind, the person died of old age, and the explosion happened because of a reaction between 3 Philons, 3 Sophons, and 3 Cognons.
Explanatory Theory 2: Tree 2 fell because Tree 3 fell on it, Tree 3 fell because of the explosion, and the explosion happened because of a reaction between 2 Philons.

'Stimulus Case 24' is set up to test Principle **(24)** of Proposed Set of Principles 4, and it is structured as follows:

Phenomenon 1: Tree 1 fell in the forest.
Phenomenon 2: Tree 2 fell in the forest.
Phenomenon 3: Tree 3 fell in the forest.

Phenomenon 4: A person died.
Phenomenon 5: There was an explosion.
Explanatory Theory 1: Tree 3 fell because of strong wind, the person died of old age, and the explosion happened because of a reaction between 1 Philon, 1 Sophon, and 1 Cognon.
Explanatory Theory 2: Tree 1 fell because Tree 2 fell on it, Tree 2 fell because Tree 3 fell on it, Tree 3 fell because of the explosion, and the explosion happened because of a reaction between 10 Philons.

'Stimulus Case 25' is set up to test Principle **(25)** of Proposed Set of Principles 4, and it is structured as follows:

Phenomenon 1: Tree 1 fell in the forest.
Phenomenon 2: Tree 2 fell in the forest.
Phenomenon 3: Tree 3 fell in the forest.
Phenomenon 4: A person died.
Phenomenon 5: There was an explosion.
Explanatory Theory 1: Tree 3 fell because of strong wind, the person died of old age, and the explosion happened because of a reaction between 3 Philons, 3 Sophons, and 3 Cognons.
Explanatory Theory 2: Tree 1 fell because Tree 2 fell on it, Tree 2 fell because Tree 3 fell on it, Tree 3 fell because of the explosion, and the explosion happened because of a reaction between 2 Philons.