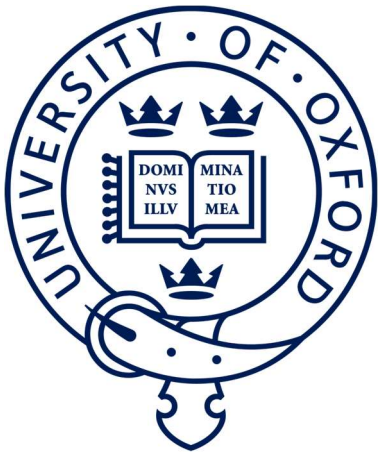
Effortless Elegance

A Metacognitive Account of Aesthetics in Science

Mariona Eiren Miyata-Sturm

New College

University of Oxford



Thesis Submitted for the Degree of

Doctor of Philosophy

July 2025

Supervised by Prof. Timothy Williamson and Prof. Nicholas Shea

Abstract

Scientists regularly appeal to broadly aesthetic qualities such as elegance, beauty, and simplicity when evaluating theories. For example, the geologist Émile Argand said that "the elegance with which drift theory explains these significant facts … is certainly a strong point in its favour" and the physicist Paul Dirac said about general relativity that its "real foundations come from the great beauty of the theory". But how can aesthetics be epistemically relevant?

In this thesis, I present an account according to which aesthetic experiences can function as signals of epistemic value. Very roughly, it says that aesthetic feelings concerning scientific theories are the upshots of metacognitive processes which respond to features of epistemic value and signal the success of one's cognitive engagement with theory and evidence. Through a detailed case study of the role of aesthetics in the development and acceptance of the theory of plate tectonics and by providing an explanation of how aesthetic experiences can be epistemically valuable, I show that there is reason for optimism about the role of aesthetic considerations in science: aesthetic pleasure is a moderately good indicator of epistemic success.

In memory of Andreas Eliassen (1992 – 2018)

And then there's nowhere you can look for us,
bird tracks left behind in the empty sky.

– Bai Juyi
transl. B. Watson

Acknowledgements

First and foremost, I want to thank Timothy Williamson for sharing so generously of the joys of doing philosophy. Tim has been an incredible supervisor over the past five years, providing insightful and detailed comments and criticisms, finding the time to talk until we ran out of things to say, always pushing me to do better yet still making me feel that I have something valuable to contribute. Early in the DPhil, Tim told me to "feel free just to follow your instincts and interests," advice it was easy to take to heart. I hope I have not given him reason to regret it, even when I spent a good amount of time thinking about geology instead of writing this thesis.

Nicholas Shea became my supervisor two years into the project and has been an invaluable support ever since. Nick has a knack for understanding what I'm trying to say better than I do and for asking just the right questions. His comments, questions, and criticism has greatly improved this thesis, and his willingness to supervise this project purely pro bono is greatly appreciated.

In their very different ways, both Tim and Nick provide excellent examples of how to do philosophy well. I have learnt much from both: thank you!

For discussion of parts of this thesis, in formal or informal settings, I thank Madelaine Angelova-Elchinova, Ajantha Abey, Leif Alberg, Marina Bakalova, Kabir Bakshi, Corine Besson, Alexander Bird, Bernard Bugjerde, Barnaby Burleigh, Adam Caulton, Brian Christian, Arian Currie, Will Davies, Erica Colman-Denstad, Krasimira Dimitrova, Dominik Ehrenfels, Artur Harris, Jonas Hertel, Nick Hughes, Milena Ivanova, Andrej Jandrić, Radmila Jovanović, Bernard Katz, Josefine Klingspor, Anette Kolderup, Simon Hasselø Kline, Silke Körber, Boaz Laan, Ofra Magidor, Jacob Mortimer, Stephen Mulhall, Jennifer Nagel, Miriam Natvig, Mehrana Raeisian Nejad, Maike Nowatzki, Madeleine Parkinson, Sybilla Pereira, Sepehr Razavi, James Read, Imogen Rivers, Sam Roberts, Matt Rosen, Maja Rausch, Bernhard Salow, Elisabetta Sassarini, Samuel Schindler, Hans Robin Solberg, Amia Srinivasan, Linea Sturm, Maciej Tarnowski, Alina Wang, and the anonymous referees of the two papers that grew out of this thesis (Miyata-Sturm (2024a, 2024b)). I have undoubtedly forgotten someone who should be listed here: if that is you, please accept my apologies.

For comments and discussion I thank the audiences at the 17th International Congress on Logic, Methodology, and Philosophy of Science and Technology at the University of Buenos Aires, the 11th European Congress of Analytic Philosophy at the University of Vienna, the Knowledge vs Understanding in Science, Fiction, and Art workshop at the University of Sofia, the Institute for

the History and Philosophy of Science and Technology graduate conference at the University of Toronto, the Women in Philosophy graduate conference at the University of Oxford, multiple DPhil seminars at the University of Oxford, the Forum for Science Studies at the University of Oslo, Bernard Katz' retirement conference at the University of Toronto, and Timothy Williamson's not-quite-retirement conference at the University of Oxford.

The amazing gift of literacy is that we do not need spatiotemporal proximity to be in conversation with each other. Every citation should be read as an acknowledgement of the debt I owe these authors for the points they raise and the thoughts they inspire.

I lost more than two years of the DPhil to Covid-19 and its long-term effects. Thanks to the Clarendon Fund, the Jacobsen Foundation, and the Aristotelian Society for providing me with the funding needed to finish the thesis despite this delay. Thanks to the Oxfordshire Post Covid Rehabilitation Service for getting me back on my feet, and to family and friends who provided invaluable support during this time: Corine Besson, Bernhard Bugjerde, Brian Christian, Tom Davis, Jon Holter, Anette Kolderup, Simon Hasselø Kline, Boaz Laan, Adam Mahdi, Mehrana Raeisian Nejad, Maike Nowatzki, Sybilla Pereira, Matt Rosen, Elisabetta Sassarini, Kelli Skinner, Eva-Maria Sturm, Fredrik Sturm, Juliane Sturm, Linea Sturm, Adam Swift, and Siyuan Yan. I could, quite literally, not have done it without you.

This research was generously supported by a Clarendon Scholarship, a Jacobsen Studentship from the Institute for Philosophy (University of London), an Aristotelian Society student bursary, and a research stipend from the Forum for Science Studies (University of Oslo).

Contents

Abstract … *3*

Acknowledgments … *6*

Contents … *8*

Chapter 1: Introduction … *10*
Chapter 2: Theoretical virtues … *14*
2.1. Introduction … *14*
2.2. Why care about theoretical virtues? … *16*
2.2.1. Inference to the best explanation … *16*
2.2.2. Scientific realism and IBE … *20*
2.3. What are theoretical virtues? … *24*
2.3.1. Defining theoretical virtues … *24*
2.3.2. No canonical list of theoretical virtues … *28*
2.3.3. Examples of theoretical virtues … *30*
2.4. (Broadly) aesthetic virtues … *50*
2.4.1. Beauty, sense perception, and formalism … *51*
2.4.2. Are aesthetic theoretical virtues genuinely aesthetic? … *56*
2.5. Conclusion … *64*
Chapter 3: Epistemic virtues … *65*
3.1. Traditional account of epistemic theoretical virtues … *66*
3.1.1. What are epistemic virtues? … *66*
3.1.2. Truth-conduciveness I … *67*
3.1.3. Truth-conduciveness II … *70*
3.1.4. A dilemma for the traditional account … *71*
3.2. Factive alternatives to truth-conduciveness … *72*
3.2.1. Alternative I: greater potential for truth … *72*
3.2.2. Alternative II: knowledge-conduciveness … *76*
3.2.3. An aside: models, idealisations, and abstractions … *80*
3.3. Non-factive alternatives to truth-conduciveness … *81*
3.3.1. Non-factive understanding … *81*
3.3.2. Are aesthetic virtues conducive to non-factive understanding? … *81*
3.3.3. Option 1: genuinely non-factive understanding … *86*
3.3.4. Option 2: grasping conceptual relations … *88*
3.4. Conclusion … *92*
Chapter 4: "The elegance…is certainly a strong point in its favour"… *93*
4.1. The need for a detailed case study … *94*
4.2. Overview of plate tectonics and aesthetic properties … *96*
4.2.1 Brief overview of plate tectonics … *96*
4.2.2. The aesthetic virtues of plate tectonics … *96*
4.3. Aesthetic considerations in the development and acceptance of plate tectonics … *100*

4.3.1. Continental Drift … *100*
4.3.2. Plate Tectonics … *106*
4.4. A 'harmonious major picture', troublesome rocks, and elegant fit… *115*
4.5. The epistemic significance of aesthetic experiences … *119*
Chapter 5: The metacognitive account of aesthetics in science … *124*
5.1. Desiderata for accounts of aesthetics in science … *125*
5.1.1. "Aesthetic criteria … of explanatory success": is explanatory power an aesthetic virtue? … *128*
5.2. The metacognitive account of aesthetics in science … *132*
5.3. Metacognition … *136*
5.4. The epistemicity of aesthetic virtues … *141*
5.4.1. Some theoretical virtues are both aesthetic and epistemic … *143*
5.4.2. Simplicity and overfitting … *144*
5.4.3. The epistemicity of elegance … *146*
5.5. Aesthetic virtues as metacognitive cues … *148*
5.5.1. Simplicity and fluency … *148*
5.5.2. Familiarity, fluency, and background knowledge … *151*
5.5.3. Fluency and aesthetic pleasure … *152*
5.5.4. Fit/coherence … *154*
5.5.5. Conflict detection … *156*
5.5.6. Informativeness/richness of detail … *158*
5.5.7. Cue integration … *161*
5.6. The accuracy of metacognitive judgements … *162*
5.6.1. Does it feel like aesthetic or epistemic pleasure? … *167*
5.7. From metacognitive cues to compelling theories … *168*
5.7.1. Applications … *170*
5.8. Alternative metacognitive accounts … *173*
5.8.1. Arcangeli and Dokic (2020) … *173*
5.8.2. Todd (2017) … *175*
5.9. Conclusion … *177*
Chapter 6: Conclusion … *178*
References … *184*

CHAPTER 1:
INTRODUCTION

This thesis is about the curious fact that scientists regularly appeal to a theory's elegance, simplicity, and other broadly aesthetic qualities when evaluating theories. For example, the central hypothesis of the theory of plate tectonics, the Vine-Matthews-Morley hypothesis, was accepted in part because "[it] offered an elegant explanation" (Mason (2003:41)). And plate tectonics itself – one of the most successful theories in science history – is regularly praised for its elegance, simplicity, and beauty. "[It] is an elegant and comprehensive synthesis of Earth's geodynamics" (Frisch et al. (2011:4)), and "[m]uch of the beauty of plate tectonics lies in the geometric exactness and simplicity of [its] geometry of movement" (Cox and Hart (1991:9)).

What does elegance and beauty have to do with scientific hypotheses and theories? Why does it matter whether a geometric system is simple and beautiful, or an explanation elegant? Is an aesthetically pleasing theory thereby epistemically better? If so, why? These are the kinds of questions considered in this thesis, and to give away the conclusion, I argue that aesthetic considerations are moderately epistemically valuable.

As the quotes above suggest, aesthetic delight in science is closely connected to epistemic evaluation. This is nicely illustrated in an interview with developmental biologist Nicole Le Dourian:[1]

> [Lewis Wolpert] 'You spoke about the aesthetic delight when you saw those sections with your quail cells. Is that part of the pleasure that you get?'
> [Le Dourian] Yes, absolutely, certainly. This is something that is always renewed because when you do an experiment, and when you find a result that you expected, or you find something that you did not expect, you have the impression that you are approaching

[1] The interview was conducted by Wolpert for BBC radio and later published in Wolpert and Richards (1997).

> something real, approaching the realities that you are seeking. It is something which is unique.
> [LW] 'But what about the aesthetic pleasure? Do you actually think embryos are beautiful?'
> [NLD] Embryos are beautiful and if you have the chance to look down the microscope you have a fantastic pleasure … This is especially so now because we have new techniques to look, not at cells as we did twenty years ago, but to look at molecules. We look at receptors, we look at gene activity.
> [LW] 'Do you think it's as beautiful as the old, or the more classical preparations?'
> [NLD] I think the beauty is not only in the colours, the beauty is what is intellectually behind what you see. (1997:180)

Le Dourian clearly finds aesthetic pleasure in the physical objects she studies – in what she sees when she looks through the microscope – and in 'what is intellectually behind' her observations: the facts and relationships her research aims to uncover. This pleasure is closely connected to epistemic success. The experiences of beauty she has when experimenting on her quail embryos is connected to her feeling that she is 'approaching something real', that she is discovering what the world is like. This aesthetic pleasure in making discoveries is heightened by techniques uncovering even more of the nature of quail development, such as looking at gene activity rather than the structure of cells. Understanding, uncovering reality, is aesthetically satisfying.

Le Dourian's views are common among biologists. A recent study by Vaidyanathan, Haraburda, and Jacobi (2023) on 1381 biologists from the US, UK, Italy, and India found that "most biologists in the four countries encounter beauty in the phenomena they study, and associate beauty primarily with the inner logic of systems. Most also find beauty relevant to presenting and analysing results, and see it as a motivation for teaching as well as for pursuing scientific careers" (2023:1). The aesthetic pleasure is inseparable from engagement in these epistemic activities, tied up with attempts at attaining epistemic aims. But why is scientific reasoning accompanied by aesthetic delight at all?

In this thesis, I present a novel account of the cognitive role of aesthetic experiences and judgements in scientific reasoning which explains (i) why aesthetic considerations appear in scientific reasoning in the first place and (ii) how they can be moderately epistemically valuable. This 'metacognitive account

of aesthetics in science' draws heavily on empirical research on metacognition – that is, the monitoring and evaluation of the adequacy of first-order cognitive states and processes – and takes the positive affect accompanying beautiful theories to be generally useful feedback from working with the theory and the evidence that bears on it. In brief, the metacognitive account says that aesthetic feelings in science are moderately reliable psychological signals of epistemic value.

I present the metacognitive account in chapter 5. But getting to the point where we can provide an explanation requires some stage setting. We start by looking at the features which characterise good theories, that is, theoretical virtues. In chapter 2, I discuss what theoretical virtues are and why we should care about them, with a special focus on broadly aesthetic theoretical virtues such as simplicity, elegance, and beauty. The main question discussed in this thesis is: are such broadly aesthetic virtues epistemically valuable? To answer that question, we need to know not only what these aesthetic virtues are but what it would mean for them to be epistemically valuable. A traditional answer to the second question is that epistemic virtues are the features of a theory which makes it more likely to be true. Chapter 3 discusses problems with this account and some of its prominent alternatives, in particular, that aesthetic virtues are conducive to non-factive understanding. I argue against this view and in favour of the view that epistemic virtues are conducive to knowledge. While chapter 2 presents many brief examples of how aesthetic considerations are appealed to in theory evaluation, to both get a good sense of what exactly we are trying to explain and show how widespread this phenomenon is (and therefore how important it is to provide an explanation of it), in chapter 4 I show that aesthetic considerations played a significant role in both developing and justifying the theory of plate tectonics. This case study gives the first detailed look at the role of aesthetics in plate tectonics, and it reveals that aesthetic considerations played a significant role in research from field geology to highly abstract theoretical geophysics. Without a clear picture of the role of aesthetics in theory evaluation we do not fully understand why some theories are accepted and others rejected, hence it is important that we explain this aspect of scientific reasoning. In chapter 5, I present a novel explanation which connects aesthetic experiences to

epistemic theoretical virtues. On this view, the experiences underlying aesthetic assessment of theories are understood as upshots of metacognitive monitoring of the success of one's epistemic engagement with theory and evidence. This provides a rough signal of the epistemic value of theories, which explains why aesthetic considerations are often, though by no means always, epistemically valuable guides in scientific reasoning.

The metacognitive account brings together previous work in philosophy of science, aesthetics, and empirical studies on the role of feelings in reasoning and decision-making to explain the epistemic value of aesthetic considerations in science. While many aspects of the metacognitive account are well supported, it should be noted from the outset that this is a speculative and to some extent programmatic account. Testing it will be left to future work.

[B]eauty, as a general rule, implies simplicity; I do not mean
the simplicity of monotony, but the simplicity of unity.
That which is highly and nobly beautiful always conveys an impression of
balance, harmony, or rhythm;
the parts, however various they may be, are related
in a way which produces an intellectual satisfaction.
– T. H. Huxley, in Blinderman (1962:52)

CHAPTER 2:
THEORETICAL VIRTUES

This chapter gives an overview of theoretical virtues, setting the stage for the following discussion of the role of broadly aesthetic theoretical virtues in scientific theory evaluation. After discussing the role of theoretical virtues in abductive reasoning, I give detailed examples of some of the major theoretical virtues. They are commonly divided into epistemic and non-epistemic, including aesthetic, features of theories. While I ultimately reject the distinction in favour of the view that some virtues are both aesthetic and epistemic, it will be useful to discuss separately in virtue of what virtues are aesthetic and in virtue of what they are epistemic. Chapter 3 discusses epistemicity. The last part of *this* chapter concerns broadly aesthetic theoretical virtues. I argue that we ought to take scientists' apparent aesthetic judgements at face value and that we should count simplicity and unity as at least broadly aesthetic. That said, the following discussion does not stand or fall on this categorisation: the puzzling role played by what looks like aesthetic virtues in the evaluation of theories demands explanation regardless of whether these qualities are genuinely aesthetic.

2.1. Introduction

From the 1910s, three theories of mountain building, rock formation, landmass distribution, and other geological features competed for the status of being the accepted framework for explaining the large-scale geological features of the earth. They were contraction theory, permanence theory, and the theory of continental drift. Drift was the newcomer, presented for the first time by Alfred

Wegener in 1912.[2] Central parts of it were later incorporated into the modern theory of the earth – plate tectonics – and in hindsight we can see that it was on the right track all along. But until the mid-1960s the theory was mostly ignored or made fun of in North American geological circles. Elsewhere it fared much better. This might seem surprising, for it could not be directly tested at the time and each of the phenomena it explained could also be explained by at least one of the other competing theories. What Drift did was recombine and unite features of permanence and contraction theories into one overarching framework, "producing simplification and coordination in the place of previous complexity and contradiction" (Wegener, in Greene (2015:240)) and explaining "significant facts" with "elegance" (Argand, in Wegener (1966:87)). It elegantly unified and simplified explanations of the earth's large-scale geological features. As I show in chapter 4, these broadly aesthetic qualities of the theory made a significant difference to how it was evaluated: those who accepted it did so partly because of these qualities, and those who rejected it tended to deny that the aesthetic merit of a theory has epistemic value.

One of Drift's major selling points was that it could give a simple, unifying explanation for both the data supporting the contraction theory and the data supporting permanence theory. Opponents of the theory, however, criticised it for not being informative or detailed enough and for not providing a plausible explanatory mechanism for drift.

Qualities like these – simplicity, elegance, unity, informativeness, and so on – are called 'theoretical virtues' (or 'explanatory virtues' or 'theoretical values'). Among other things, a theory is virtuous when it fits the data, makes novel and testable predictions, is simple, unifying, and doesn't include *ad hoc* hypotheses. I discuss these and other virtues at some length in section 2.3 below. But why should we care about theoretical virtues in the first place? Because they are symptoms of epistemically good theories. They play a particularly large role in abductive reasoning (inference to the best explanation, 'IBE' for short): when we infer that because a particular theory gives the best explanation of the facts it is probably true, 'the best explanation' is the theory with the most, or the best

[2] See Chapter 4 below for details and references.

balance of, theoretical virtues.[3] IBE is a common form of reasoning in both science and everyday life, and it plays a prominent role in defences of scientific realism, as we'll see in section 2.2 below. This thesis is not about theoretical virtues in general, however, but about broadly aesthetic virtues. Building on the characterisations of some oft-invoked aesthetic virtues in section 2.3, in section 2.4 I discuss why we should count these as aesthetic virtues. While it is not important for our purposes whether these virtues are genuinely aesthetic or not, there are good reasons to count them as aesthetic. As I argue in chapters 3 and 5 (and show with examples in chapter 4), virtues can be simultaneously aesthetic and epistemic. So counting, say, elegance and simplicity as aesthetic does not preclude them from playing an epistemically valuable role in theory evaluation.

But we are getting ahead of ourselves: let us start by situating the debate about theoretical virtues within the larger literature on inference to the best explanation and scientific realism.

2.2. Why care about theoretical virtues?

2.2.1 Inference to the best explanation

The primary reason to care about theoretical virtues is their central role in inferences to the best explanation. IBE, or abduction,[4] is a form of ampliative reasoning, which means that the conclusion goes beyond what is contained in or guaranteed by the premises.[5] It consists of inferring from effect to cause, or accepting the hypothesis which, if true, would best explain the data; hence it's sometimes simply known as 'explanatory reasoning' (Douven (2022:29)).

[3] Details will be filled in in the next section.

[4] IBE and abduction are sometimes thought of as distinct inference patterns, especially in the historical literature on Pierce's account of abduction, where abduction is understood as *generating* potential explanations and IBE is about *evaluating* them (e.g., Minnameier (2004)). Following Lipton (2004), IBE is commonly understood as a two-step process of first generating potential explanations and then choosing between them. Here I treat 'IBE' and 'abduction' as synonyms.

[5] Gilbert Harman (1965) started the contemporary conversation on IBE when he famously argued that enumerative induction (another kind of ampliative reasoning) is a kind of abduction. Here we don't need to take a stand on whether abduction is the most fundamental or simply *a* kind of ampliative reasoning.

Abduction is a familiar form of reasoning, famously relied upon in detective stories and everyday life. Noticing how disappointingly flat your newly baked bread is, you may infer that you forgot to add yeast to the dough. That would certainly explain why it didn't rise. While it's a good explanation, it does not guarantee that the hypothesis is true (perhaps you added dead yeast or the temperature was too high).

Much scientific reasoning is abductive. Take, for example, Wegener's central argument for continental drift: the fact that his hypothesis of moving continents could explain the geological, palaeontological, and palaeoclimatic data better than could the alternatives (the contraction and permanence theories) was in his eyes a strong reason to accept it as true (e.g. (1966:21); see chapter 4 below). His theory did not follow deductively from the evidence – the evidence was quite compatible with alternative explanations – nor was it based on an inductive generalisation from other planets with moving continents: the support for continental drift came from the explanatory superiority of the theory. From the belief that his theory best explained the evidence he inferred that it was true, or nearly so.[6] What, then, is the *best* explanation?

This is where theoretical virtues enter the picture: the best explanation is the one with the most theoretical virtues, or perhaps the best overall balance of theoretical virtues (Lipton (2004:61, 122), Douven (2022:54), Psillos (1999:171), Beebe (2009:611)).[7] This is directly relevant to theoretical goodness

[6] This is by no means an uncommon case. See, for example, Dellsén (2024) and Thagard (1978) for more examples of abductive reasoning in science.

[7] Theoretical virtues often pull in somewhat different directions, so increasing one virtue often means decreasing another. For example, a simple theory is often correspondingly low on strength, and a theory fitted to all the evidence will likely be ad hoc and disunified, prone to overfitting and unstable in the face of new evidence, always needing new adjustments to fit incoming evidence. There is generally a limit to how much we can increase one virtue while simultaneously securing a decent amount of another: push the theory too far in one direction, and it falls behind on other important criteria. This is to say that theoretical virtues are not entirely mutually independent of each other. We cannot give a complete account of what each contributes to the goodness of a theory without considering how they are balanced, or traded off, against each other. A virtuous theory occupies a sweet spot between the many oft-conflicting desiderata, getting as much of each virtue as it can without compromising the others. I do not know whether there is an optimal trade-off between virtues, but see Kuhn (1977) and Okasha (2011) for discussions on whether there could be an

in so far as theories provide explanations and indirectly relevant because most, if not all, virtues are shared between explanations and theories. In terminology made famous by Peter Lipton, the best explanation is the *loveliest* explanation, and theoretical virtues are "features that make one explanation lovelier than another" (2004:122).

Lipton takes the loveliest explanation to be the one that would, if true, provide the deepest understanding (2004:59–61). A more neutral description is that the loveliest explanation is epistemically successful (in one way or another); in chapter 3 we discuss whether we should take an epistemically virtuous theory to thereby be more likely to be true, give us more knowledge, or provide non-factive understanding.[8] This is connected to the epistemic aim of science: if the aim of science is to provide knowledge, say, then features which contribute to attaining this aim are epistemically virtuous. More generally, whether a theory is virtuous depends on what we aim to achieve with it: comprehensiveness, say, might be a virtue when mapping out a new phenomenon but not when presenting a theory in a textbook format. Clearly, scientific research has multiple aims (Potochnik (2017); but see Bird (2022)). For example, we can aim both at developing a vaccine and at learning how a disease spreads or develop a model both to aid us in making predictions and as an aid for communication. Regardless of the multitude of aims science can have, across different fields and circumstances, the primary aim of science is epistemic (however exactly we spell that out). Reflecting this, when we talk unqualifiedly of the best explanation as the one having the highest score of theoretical virtues, we typically only have in mind virtues that are at least partially epistemic. This can conceal substantial disagreement; as we'll see in the next section and in chapter 3, what one takes the epistemic aim of science to be makes a difference to which virtues are taken to be epistemically valuable. In particular, whether one takes (some) non-

---

algorithm for theory choice (which would require a specification of an optimal trade-off).

[8] Note that Lipton is a factivist about understanding: "only actual (i.e. true) explanations provide actual understanding" (2004:134). Thus his account of what lovely explanations provide should not be confused with an account of genuinely non-factive understanding. Chapter 3 discusses whether prominent self-proclaimed 'non-factivist' accounts of understanding nonetheless demand a high enough level of truth to be classified as factivist.

empirical (or, supraempirical) virtues to be epistemic tend to reflect one's stand on scientific realism (Schindler (2018), van Fraassen (1980)).

While which features are theoretical virtues depends on the aim(s) of science, epistemologically speaking the arrow tends to point the other way: which qualities scientists take to be theoretical virtuous is evidence for what is the aim of science. If scientists, say, consistently value scientific theories for giving us knowledge, that's a strong indicator that knowledge is an aim of science. A strong, but not a perfect indicator, since we might value knowledge as a means to something else – say, to enable effective interventions – and since the aim(s) of *scientists* can diverge from the aim(s) of *science* (Bird (2022:17), van Fraassen (1980:8)).

IBE is often criticised for being more of a slogan than a clear description of an inferential pattern, and many more precise variations are proposed (see Douven (2022) and Dellsén (2024) for good overviews). For example, it's common to precisify that IBE only warrants the belief that the explanation is probably approximately true: since it is an ampliative form of reasoning the conclusion is not guaranteed to be true (hence 'probably true'), and in many cases the best we can hope for may well be that it is in the ballpark of the truth (hence 'approximately true'). In addition to specifying the structure of abductive inferences,[9] an account which fills out the slogan of 'inferring to the best explanation' must give a nontrivial description of theoretical virtues and answer thorny questions about how to measure theoretical goodness (which is simpler: Copernican circular planetary orbits with epicycles or Keplerian elliptical orbits, given that circles are mathematically simpler than ellipses?) and how to weigh conflicting virtues against each other (how much fit with data is elegance worth? What is an acceptable ratio for loss of explanatory power to gain in predictive power?). A satisfactory account of IBE hinges on a satisfactory account of

[9] Douven argues that "abduction is best thought of as a slogan such as that explanatory power is of confirmation-theoretic significance or similar circumscriptions … and in each situation the slogan has to be fleshed out in the best possible way for that situation. Finding the best elaboration for a given situation may be a matter of experience or talent" (2022:15). If so, it may not be possible to spell it out in much detail, though this contextualist approach itself goes some way towards spelling out the slogan.

theoretical virtues insofar as they are central to explaining what the *best* explanation is.

However, most work on IBE has been concerned with other foundational issues. What Paul Thagard noted half a century ago (1978:76) is still true today: most work on IBE does not have much to say about theoretical virtues. (See, for instance, Lipton (2004) and Douven (2022) for book-length treatments of IBE which barely touch upon theoretical virtues – though see Dellsén (2024) for a short book on IBE which discusses theoretical virtues at some length, and Beebe (2009) and Cabrera (2023) for papers on IBE which are unusually clear on which virtues characterise best explanations.) The lack of discussion of theoretical virtues in the literature on IBE is presumably due to a division of labour in research on, respectively, the structure of IBE and the nature of theoretical virtues. It might – reasonably, in my eyes – be hoped that whatever theoretical virtues are, they can be seamlessly fitted into work on IBE as a way of filling out important details about what makes an explanation good. It's hoped that the discussion of aesthetic theoretical virtues in this thesis will go some way towards filling the gap in our knowledge of aesthetic theoretical virtues.

2.2.2. Scientific realism and IBE

Theoretical virtues are more of a concern for scientific realists than for anti-realists, because realists must shoulder the heavier burden of explaining why theoretical virtues can legitimately allow us to infer that a theory having them is true, or nearly so. Like most philosophical umbrella terms, many different accounts are labelled 'scientific realism' and 'scientific anti-realism' (see Chakravartty (2017) and Psillos (1999) for overviews). For our purposes the following will do: a realist believes that science aims to describe what the world is like and explain what made it so, even when this goes beyond what we can observe directly (i.e., non-inferentially), and that our best theories tend to be approximately true. ('Best', as there is no pressure on the realist to accept every scientific theory. 'Approximately', since the realist need not think we have succeeded in giving a perfectly true account of the world yet. 'Tend', since even

the best science is not infallible.) Importantly, if the theory is good enough to be accepted as at least approximately true then the theory as a whole is accepted, not just the claims it makes about the observable features of the world.[10]

In contrast, anti-realists of the popular 'constructive empiricist' flavour (following van Fraassen (1980)), think that we are, at best, justified in believing what the theory claims about observable features of the world. The observable/unobservable distinction plays a central role in constructive empiricism, but it is highly contested (Musgrave (1985), Dicken and Lipton (2006), and van Fraassen (1980)). In particular, it faces the problem of spelling out the distinction between unobservable and unobservable entities in a principled way: unobservable to whom – humans at a well-equipped lab today? Someone with average eyesight? This will not make a difference here, so when talking about constructive empiricism I assume for the sake of argument that we can give in unproblematic description of this distinction. Note that the difference between the scientific realist and the constructive empiricists does not concern what epistemic good science can deliver (truth), but a difference in the *scope* of what the realists and constructive empiricists think we can legitimately take to be true based on our best scientific theories (Bird (2022:16)).

Constructive empiricism is not the only alternative to realism, though it is by far the most popular contemporary alternative. Other antirealist positions are arrived at by denying that scientific theories are truth-apt, through scepticism about the existence of a mind-independent, external world, or a scepticism more radical than the constructive empiricists' about our epistemic access to the world (e.g., Chakravartty (2017: sect. 4)).

[10] This is a rough characterisation, and the realist is not necessarily committed to even the approximate truth of our best theories. It's not always clear what's part of a theory and what are background assumptions, nor that each and every statement the theory makes is essential enough to it to be part of the theory proper (for example, given that theories are typically reformulated and modified over long stretches of time it's debatable which version should be taken as the canonical formulation of it), giving realists a lot of wriggle room. The point to note here is that the realist does not withhold acceptance of unobservable features posited by our best theories.

Why claim that theoretical virtues are more of an issue for realists than anti-realists? Because IBE plays a central role in arguments for realism,[11] and, more importantly, because realists are typically both more focused on the role of abductive reasoning in science and take it to justify a stronger conclusion than anti-realists do: namely that the best explanation is at least approximately true, not just that its observational consequences are accurate. Realists typically view it as legitimate to infer that a theory is true (or, approximately true, or that it will give us more knowledge, etc.), based on explanatory considerations. The best explanation doesn't just get it right about observable features: if unobservable phenomena are appealed to as an essential part of the explanation, and the explanation is good, then we have reason to believe that those unobservables exist and have the properties the theory says they have. Antirealists typically reject the view that explanatory success warrants inferring that the explanation – in so far as it goes beyond what is observable – is true (e.g., van Fraassen (1980:20)).

Antirealists can accept that all the virtues invoked by realists play a central role in theory evaluation without taking the additional step of claiming that these virtues make the theory epistemically better.[12] Beyond a small handful of virtues, such as fitting the data, making testable predictions, and being internally coherent (see section 2.3. below), none of the 'virtues' that figure so prominently in abductive inferences are signs of an epistemically good theory. These 'virtues' are only of pragmatic value, concerning "the use and usefulness of the theory; they provide reasons to prefer the theory independently of

[11] In particular in the so-called 'no-miracles argument', which, briefly, says that the success of science would be a miracle if our best scientific theories were not (approximately) true; but since there are no miracles the best explanation for the success of science is that scientific theories are (approximately) true (Putnam (1979), Boyd (1984): see also Psillos (1999: chapter 4)). This is countered by another famous abductive argument, 'the pessimistic meta-induction,' which has it that since most theories in science history has been false, even when they enjoyed significant empirical success, we should believe that our best current theories are also probably false (see Psillos (1999: chapter 5), Laudan (1984), Bird (2022:232–9) for more on this).

[12] See chapter 3 for a discussion of which epistemic good the realist should think theoretical virtues confer on theories.

questions of truth" (van Fraassen (1980:88)). So these pragmatic virtues do not, according to the antirealist, licence believing that the theory is true (or nearly so). Note, however, that the constructive empiricist can claim that the pragmatic virtues are *indirectly* epistemically valuable. van Fraassen says that "it might be arguable that, for purely pragmatic (that is, person-and context-related) reasons, the pursuit of explanatory power is the best means to serve the central aims of science" (1980:89). Now, if these virtues contribute to reaching the aims of science (whether that is truth, empirical adequacy, knowledge, or understanding) then they are epistemically valuable. There is room for different positions here, but to the extent that theoretical virtues are (directly or indirectly) epistemically valuable they are not 'merely pragmatic'. Drawing the line depends on what one takes to be the distinguishing feature of epistemic virtues (discussed in chapter 3), but it's worth noting already that realists do not (should not!) take all epistemic virtues to be truth-conducive, so realists and constructive empiricists may agree on more regarding theoretical virtues than they seem to at first blush. Furthermore, the constructive empiricist can match the realist's abductive reasoning step-by-step, claiming that the best explanation justifies believing that the theory is empirically adequate (i.e., that everything it says about past, present, and future observable phenomena is true) (van Fraassen (1980:12); see Lyons (2025:12–5) for discussion). This will strike a realist as an arbitrary limitation of a legitimate inference pattern: if you are agnostic about what cannot be directly (non-inferentially) observed, you ought to be equally cagey about past events.[13] Realists are wont to take it that the important epistemic difference is between observ*ed* and unobserved entities, not observ*able* and unobservable ones, so if we reason abductively we should infer that the explanation as a whole is (probably, approximately) true.[14]

What is important to note here is that realists face the burden of explaining why theoretical virtues can legitimately allow us to infer that a theory

[13] See Bird (2022: chapter 6) for discussion of why we should not equate observation with perception: this provides a different reason for rejecting (constructive) empiricism.

[14] Again, this is a rough characterisation: not considering variants such as structural realism (Worrall (1989)) or (even) more middle-ground positions such as Chirimuuta's (2023) haptic realism.

having these virtues is thereby more likely to be true (or provide more knowledge, etc.), and that they cannot fall back on the claim that these virtues are merely pragmatic. Accordingly, while the issue of *aesthetic* theoretical virtues in science concerns realists and antirealists alike, they don't face the same epistemic challenges. Defenders of explanatory reasoning face the problem of explaining why and how the aesthetic qualities of a theory make it more likely to be epistemically successful. Its critics, meanwhile, faces the challenge of *explaining away* historical evidence for the usefulness of aesthetic judgement in science. (That might well be an equally daunting task, as suggested by the extended case study in chapter 4.)

Because it is predominantly realists who are committed to abductive reasoning and thus face the challenge of explaining how aesthetic virtues (*qua* features appealed to in IBE) can be epistemically valuable, I will for the purposes of this thesis assume that realism is true. That way we can focus on what is rightly viewed as a weak point in defences of IBE and thus of realism, namely the epistemic relevance of aesthetic theoretical virtues, without longwinded caveats. Thus I will continue to talk about truth without further ado and only discuss antirealist alternatives when that makes an important difference to the point being made. That said, I believe most of the following – including the positive account defended in chapter 5 – can be rephrased to fit at least constructive empiricism without substantial loss.

2.3. What are theoretical virtues?

2.3.1 Defining theoretical virtues

So, what are theoretical virtues? They are the features which characterise good theories (Kuhn (1977:357), Thagard (1978:76), Schindler (2018:1),[15] Keas (2018:2761), Lipton (2004:122), Cabrera (2023:2)); features which, all else being equal, makes a theory having them better than a theory lacking them. Good in which respect? Here background commitments immediately reveal themselves:

[15] But compare Schindler (2018:5), which is in line with the neutral definition from Schindler (2020) discussed below.

no informative definition of theoretical virtues can stay neutral on what is the epistemic aim of science, or what role non-empirical virtues play.

It's common to take theoretical virtues to be those features which makes the theory more likely to be true. Thus a theory is good when it's true, and features of theories which characterise true theories are theoretical virtues. As Finnur Dellsén writes, "[t]he simplest and perhaps most popular … view [is] that the explanatory virtues that are operative in abductive reasoning are *truth-conducive*: if a hypothesis possesses these virtues to a greater extent than alternatives, that hypothesis is thereby more likely to be true, other things being equal" (2024:35–6). And Frank Cabrera writes that "the explanatory virtues play an indispensable role in any given application of IBE, and so if IBE is to be an inductively valid form of inference, then we need to be sure that the explanatory virtues are appropriately connected to the truth" (2023:23). This follows rather straightforwardly from the view that IBE licences inferences to the *truth* of the best explanation. It also reveals the realist commitments common among defenders of IBE, for antirealists can – and do – make a sharp distinction between those virtues they take to be truth-conducive (or rather, conducive to empirical adequacy) and the rest. Bas van Fraassen famously argued that only some theoretical virtues are epistemically valuable and dismissed the rest as merely pragmatic:

> In so far as [theoretical virtues] go beyond consistency, empirical adequacy, and empirical strength, they do not concern the relation between the theory and the world, but rather the use and usefulness of the theory; they provide reasons to prefer the theory independently of questions of truth. (1980:88)[16]

Such reasons include that it is more convenient to use or that it answers the kind of questions we are interested in. So on this view, only some of the virtues makes

[16] Note that empirical adequacy should not be on this list: first, it is an aim of science, according to van Fraassen, and thus cannot also be a useful signal that we have attained this aim, any more a theory's truth can be used to pick out the best explanation. Second, a theory is empirical adequate when it saves *all* the phenomena (whether past, present, or future), and since we will never be in a position to compare the theory to future empirical evidence we cannot tell whether the theory is empirically adequate, hence cannot use it as a theoretical virtue to identify the best theory. Thanks to James Read for stressing this point.

the theory epistemically good, and the other 'virtues' are only of pragmatic value. Thus we cannot define theoretical virtues without either taking a stand on which virtues are epistemic or drastically change which features are normally thought of as virtuous.

Now, realists and antirealists alike may accept that only some explanatory virtues are truth-conducive (or otherwise epistemic: see chapter 3) so it will be useful to hold on to the distinction between epistemic and non-epistemic virtues. With this distinction in hand, we can ask the question which occupies us in this thesis: are aesthetic virtues epistemically valuable? That is, are aesthetic virtues also epistemic virtues? This distinction, however, is slightly artificial: insofar as all these virtues make a theory good, and given that theoretical goodness is primarily an epistemic matter, all theoretical virtues are epistemically valuable. (One can of course argue that not all commonly invoked 'virtues' are truly virtues, but insofar as theoretical virtues are the identifying marks of the best explanation they are epistemically valuable.)[17] Assuming that broadly aesthetic qualities like elegance, beauty, simplicity, and so on, are theoretical virtues, our challenge is to explain *how* these aesthetic virtues can also be epistemically valuable, but *that* at least many of them are epistemic follows from their status as theoretical virtues together with the legitimacy of justifying theory choice through explanatory reasoning (i.e., inferring that the best explanation is true, or truthlike). In other words, if you are committed to IBE, you are also committed to saying that theoretical virtues (whatever exactly they are) are features which characterise (approximately) true theories.

Theoretical virtues are not only "the characteristics of a good scientific theory" (Kuhn (1977:357)) but the criteria by which "one hypothesis [is] judged to provide a better explanation than another hypothesis" (Thagard (1978:76)):

---

[17] This is one area where the literature on the role of theoretical virtues in IBE and values in science diverge, insofar as the question of values in science goes beyond which theories to accept and considers e.g., which research questions are worthy of pursuit: this is not relevant to whether the resulting theory is true (though it can make a huge difference to what is good science in a broader sense, for example by being socially responsible). The literature on values in science overwhelmingly concern social and political values, not aesthetic ones, and so is beyond the scope of this thesis (see Longino (1990) and Douglas (2009) for the classical discussions of values in science).

they are used comparatively to evaluate theories and single out the best explanation. They "guide [scientists] in their choice to adopt one theory or another" (Schindler (2018:5)). These virtues matter insofar as they help identify epistemically successful theories. It is therefore somewhat misleading to say that theoretical virtues are those qualities which *make* a theory good: they do not *cause* theories to be epistemically successful but are symptoms of epistemic success. We rely on epistemic considerations – in particular, the explanatoriness of theories as measured by their theoretical virtues – as proxies to identify true theories. If we could detect the truth or falsity of a theory directly, there would be no need to go via features which indicate, but cannot guarantee, the theory's truth. Theoretical virtues are used to discover what really matters (i.e., whether the world is as the theory says it is), not primarily features of intrinsic interest.

Schindler (2020) defines theoretical virtues as qualities of scientific theories which are considered desirable. This neutrality throws the net too wide: we end up catching all kinds of features that influence theory evaluation and not only features of true theories. Surely scientific communities can be misguided and desire qualities that do not make a theory more likely to be good, whatever exactly this good is. For example, a psychosociological theory which supports dominant gender stereotypes might be reassuring and therefore desired, but that doesn't make being reassuring a theoretical virtue. It is not even a pragmatic virtue; it is, if anything, a vice. Other desirable qualities can be useful: if the theory sounds intriguing or can be given a nice acronym it may be more likely to attract grant funding. Still, these features are not theoretical virtues. An alternative definition which retains the neutrality of Schindler (2020) without falling into this problem of misclassifying any (for some) desirable feature as a virtue is given by Miriam Solomon (2001). Solomon subsumes theoretical virtues under what she calls 'decision vectors', making it clear that her focus is on what actually influences theory evaluation whether or not the influence is virtuous. However, since it is precisely the normative issues we are concerned with here, such neutrality is not desirable.

Lastly, before we consider examples of theoretical virtues, it should be noted that 'theoretical' virtues also apply to models, hypotheses, and other scientific representations. For example, Steven Downes says that "models are

valued for their predictive power, their explanatory power, their productivity in hypothesis generation, their troubleshooting capabilities and, in some cases, their similarity to a real-world system" (2021:69). With some slight modifications (e.g., exchanging 'true' for 'accurate representation', which is merely a linguistic difference, and perhaps with a greater focus on predictions in the case of models and on explanations in the case of theories), what I say about theoretical virtues applies equally to models, but for conciseness I will continue to talk about theories and *theoretical* virtues. That said, models are typically idealised or abstracted in ways which on purpose diverge from the truth in certain respects. Is that incompatible with taking at least some of these virtues to be truth-conducive? Some would argue so, as discussed in chapter 3. However, idealisations and abstractions are used to arrive at literally true beliefs, as I argue in section 3.3.3, so while this complicates the picture somewhat it does not show that truth is irrelevant or unattainable. We come back to this in the next chapter.

2.3.2. No canonical list of theoretical virtues

There is no one canonical list of theoretical virtues in the literature. The closest we get is Thomas Kuhn's (1977) proposal: he lists accuracy, internal and external consistency (or, coherence), broadness of scope, simplicity, and fruitfulness (1977:357). Samuel Schindler, who with one book and several papers on theoretical virtues (e.g., 2014, 2017a, 2017b, 2018, 2022) is one of the leading specialists in the field, takes Kuhn's list to be "what is now considered a standard list of theoretical virtues" (2022:543). It's no stretch to say that the items on Kuhn's list are still the best-known theoretical virtues, but they are far from the only ones regularly invoked. Bas van Fraassen lists empirical adequacy,[18] strength, mathematical elegance, simplicity, scope, and that the theory is 'complete in certain respects', unifying, and explanatory (1980:87). Schindler (2018) lists testability, non-ad hocness, empirical accuracy, internal and external consistency, unification, simplicity, and fertility (2018:6–19). Finnur Dellsén (2024:32) lists scope, parsimony, unification, plausibility, and analogy. (Note that Dellsén's 'plausibility' is our 'external coherence': whether a hypothesis is

[18] Note that empirical adequacy does not belong on a list of theoretical virtues, even by van Fraassen's own lights (see fn. 16 on p. 26 above).

plausible depends on "[h]ow well … the hypothesis fit[s] what one already takes oneself to know" (*ibid.*).) Devon Brickhouse-Bryson (2020) lists fit with data, fit with other truths, evaluating the theory's truth-value via its modal properties (noncontradiction and necessity) or via its various ways of being beautiful (which he takes to include unity, simplicity, elegance, and coherence) (2020:xii–xix). And early proposals, by Robert Boyle and Christiaan Huygens, respectively, include internal and external consistency, absence of ad hoc-ness, and simplicity (Boyle) and "the variety and the novelty of the predictions it could generate" (Huygens) (McMullin (2014:562)). These are all variations on the theme of *explaining as much as possible with as little theoretical machinery as possible while both staying true to the evidence and opening up new avenues of research*. (We take a closer look at these virtues in section 2.3.3 below.)

In feminist philosophy of science it's common to list quite different features as virtues, many of which conflict with the traditional lists – and with each other. For example, Helen Longino (1996) lists empirical accuracy,[19] novelty, ontological heterogeneity, mutuality of interaction, applicability to human needs, and decentralisation of power (1996:44), to which one could add feminism and non-racism (Crasnow (2013:416)). In Miriam Solomon's (2001) long list of 'decision vectors' – features that influence theory evaluation whether or not they are virtuous – we find novel predictions, a preference for theories for which data is already available, conservativeness, radicalism, elegance, and simplicity (2001:57–8).

The absence of a canonical list of theoretical virtues has led to some double counting, with different labels for the same qualities (e.g., 'scope' and 'consilience' for the fact that the theory covers many disparate phenomena), and sometimes the same label for different qualities (e.g., 'coherence' for both internal consistency and fit with other theories). As the virtues listed by feminist philosophers of science suggests, it might be that many important theoretical virtues are not standardly listed. While it would be great to have a comprehensive list of all important theoretical virtues, providing that goes far beyond the scope

[19] Longino treats empirical accuracy and empirical adequacy as equivalent, but the first is a theoretical virtue and the second is (perhaps) an aim of science. See fn. 16 on p. 26 above.

of this thesis. Our focus is solely on broadly speaking aesthetic theoretical virtues and epistemic virtues (which, again, may overlap but which we for the moment treat separately). Even this restriction leaves us with too many virtues than we can comfortably discuss here, hence I will restrict the discussion to some of the most commonly invoked aesthetic and epistemic virtues.

2.3.3. Examples of theoretical virtues

As the example I opened this chapter with shows, scientific communities sometimes have different opinions on which features are virtues and which are vices (or neutral qualities). This is especially true for aesthetic virtues: there is much disagreement on whether they have a legitimate role to play in theory choice and abductive inferences generally, as we discuss below. There are nonetheless some features which are often considered theoretical virtues.

The following long list gives an overview of many oft-invoked epistemic and aesthetic virtues, but it doesn't try to be comprehensive. I discuss empirical accuracy, strength, internal coherence, external coherence, broadness of scope, ad hocness (a theoretical *vice*), simplicity, elegance, unity, and beauty. These virtues (i) are important in their own right, (ii) are representative of epistemic and aesthetic virtues, and (iii) many of them play a role in the subsequent discussion of the epistemic value of aesthetic virtues. This thesis presents a new model of the epistemic significance of aesthetic virtues; if it's on the right track, it should account for all aesthetic virtues. But before attempting to show that it does, we should investigate whether the model holds for the most central aesthetic virtues: if it does, we can work on expanding the account. It is premature to worry about comprehensiveness until we have worked out the central details, and so when presenting the account in chapter 5 I will focus on only a selection of the virtues listed here.

1. *Empirical accuracy*: what the theory states and entails (together with auxiliary hypotheses) fit[20] observational and experimental data. "A theory is

[20] Steven Downes (2021:70–5) separates *accuracy* from *fit*, perhaps accidentally. When talking about models as representations he talks about accuracy, and when

empirically accurate when what it says about the world is consistent with observations", as Schindler (2018:7) says. A theory fits the data when its predictions correspond to the data or when it accurately describes its target phenomenon (what the theory is about), as measured by empirical data.[21] Note that accuracy is about *empirical* fit, not fit with *all* our evidence, since the evidence need not take the form of observational or experimental data (see 'external coherence' below). It's also different from fit with all past, present, and future empirical data (see below) and requires only fit with the data *currently* had by the scientific community. This means that empirical accuracy is a property theories have at a certain time which can be lost when we acquire new, disconfirming data (Schindler (2018:7)). In practice, however, empirical accuracy is a matter of degree (we typically have data both in favour of and counting against the theory) and thus not overly sensitive to small fluctuations in incoming data. Perhaps a better way of capturing this is with Bogen and Woodward's (1988) well-known distinction between phenomena and data. They argued that there is an important middle layer between observations and theories: these are the phenomena, the relatively stable processes, events, or objects which are evidenced by the particular observations constituting the data. The classic example of a phenomenon is the melting point of lead: it's 327°C, which does not

talking about confirmation of models he talks about fit, but nothing he says indicates that he takes accuracy and fit to be distinct phenomena.

[21] We get a stronger version of empirical accuracy by exchanging evidence for data: a theory is empirically accurate when it fits the empirical evidence, i.e., the *accurate* data, or the data known to be accurate. The way I'm thinking of it here, evidence entails truth (nothing can be evidence unless it's true, though it can of course be mistakenly classified as evidence) but data (the results of measurements, observations, and the like) may include errors, so the weaker and the stronger version of empirical accuracy are importantly different. However, this is contentious territory. The way I have drawn the distinction between data and evidence is more common in epistemology than in philosophy of science. It fits Williamson (2000) and Bird's (2022) account that our evidence is what we know. On the other hand, Leonelli writes that "I view data as mobile pieces of information, which are collected, stored and disseminated so as to be used as evidence for claims about specific processes or entities" (2013:505). On this account, evidence is what is *used to support* a given claim, and the difference between data and evidence is a matter of how the data is used not a matter of whether it's accurate.

correspond to any particular temperature measurement (data point). It is such phenomena that theories should fit: not any particular data point, but the phenomenon we have access to through a large and varied data set (1988:314). However, this distinction will not make a difference to our concerns so for ease of presentation I will continue to talk of data. When using 'phenomena,' it is with the everyday meaning of *thing*, *process*, *event*, or similar in mind, not the technical sense.

Not just the theory itself, but all that follows from it,[22] should accord with the data. As Kuhn put it, "consequences deducible from a theory should be in demonstrated agreement with the results of existing experiments and observations" (1977:357). Contrast empirical accuracy with empirical adequacy: that the theory fits all the past, present, and future data. In van Fraassen's words, "a theory is empirically adequate exactly if what it says about the observable things and events in this world, is true" (1980:12), which includes all observ*able* phenomena, not just observed phenomena. Empirical adequacy entails empirical accuracy for data available at any given time. Note that empirical adequacy is neither the same as empirical accuracy nor a theoretical virtue: it is rather a possible epistemic aim of science, on par with knowledge or truth.

1.1. For example, the standard model of particle physics is remarkably accurate. This is all the more impressive because many of its predictions are so precise that only a narrow band of possible data could fit them. The most spectacular example is perhaps the standard model's (SM) prediction of the 'magnetic moment' of electrons, which is "a measure of the strength of the magnetic field carried by the particle" (Conover 2023). The magnetic moment was recently measured by Gerald Gabrielse and colleagues; as Fan et al. says, "in [this] most precise confrontation of theory and measurement, the SM prediction agrees to 1 part in 10^{12}" (2023:1). Neither predictions nor measurements are

[22] In practice, only some consequences will be evaluated against the data: there are infinitely many consequences of any theory which will remain untested for a number of reasons (perhaps because they haven't been thought of or testing them would require too many resources).

usually this precise. A more typical example comes from using the GPS network to test the predictions plate tectonics make of the movements of plates. GPS measurements largely show that plates are rigid and move relative to one another with the expected velocity and direction. As Peter Molnar writes, "[a]mong tests of plate tectonics, none is more convincing than the GPS measurements of changing positions of tens to hundreds of points on most plates; their relative velocities corroborate the rates inferred from magnetic anomalies, orientations of transform faults, and fault plane solutions of earthquakes along plate boundaries. Moreover, numerous predictions of rates or directions of present-day plate motions and of large displacements of huge terrains have been confirmed many times over" (2015:88). However, some measurements count against plate tectonics, which means that it's not perfectly empirically accurate. For example, "the eastern parts of Siberia, Mongolia, and China … does not behave as a single rigid plate [as predicted by plate tectonics], for South China and North China are separated by a wide zone of deformation, where many destructive earthquakes have occurred" (Molnar (2015:104)).

2. *Strength (or informativeness)*: the theory is informative, usually through being detailed, precise, and general. Strictly, a statement A is stronger than B just in case A entails B but not vice versa; here I'll understand strength also in the looser sense of being committed to more. Strength is a matter of excluding more ways the world could have been; or, as is sometimes said, that the theory has a high level of content (Lipton (2004:116)). For example, 'the apple is red' is stronger than 'the apple is coloured' since it follows from its redness that it's coloured but it doesn't follow from its being coloured that it is red (it could have been yellow). An informative theory says more and can therefore potentially increase our knowledge by more than a weaker theory can.

Strength and wideness of scope can, but need not, go hand-in-hand. For example, Newton's three laws of motion are both highly informative and concern all massive objects, so his theory is therefore both stronger and

wider-scoped than a theory of gravitation applicable only to earthbound objects; our coloured apple shows how two hypotheses can differ in strength without differing in scope. As famously argued by Karl Popper, the stronger a theory is, the less probable it is ([1960] 1987:176) – and therefore it is more impressive when a strong theory's predictions fit the data than when a weak theory's predictions do. Neither strength nor scope are legitimately increased by arbitrarily conjunction; or rather, these virtues should not be increased at the cost of unity, simplicity, elegance, or non-ad hocness.

2.1. For example, compare continental drift and plate tectonics. Plate tectonics is the stronger theory: whereas continental drift give approximate descriptions and hints at possible mechanisms for drift, and doesn't make specific predictions about, for example, the relative drift rate of specific plates, the heat flow values of the seafloor, or the thickness of the mountain chains, plate tectonics is committed to specific values of each of these phenomena.[23] Note that the theories have the same scope, as they are both concerned with explaining the large-scale geological features of this planet: this is a difference in strength without a difference in scope.

3. *Broadness of scope*: the theory is about a wide range of phenomena. The wider the scope, the better. Remember that wideness of scope is one of the ways in which a theory can be informative: a wide-scoped theory entails more than a narrower theory does, all else being equal. Broadness of scope is roughly what Paul Thagard (1978), following William Whewell, calls 'consilience': "Consilience is intended to serve as a measure of *how much* a theory explains, so that we can use it to tell when one theory explains *more* of the evidence than another theory. … [O]ne theory is more consilient than another if it explains more classes of facts than the other does" (1978:79; italics original).

   This focus on type rather than token phenomena to be covered is typical: what matters to a theory's scope is that the theory explains or describes many different kinds of phenomena, not that it explains or describes many different phenomena of the same kind. The classical example is Darwin's

[23] We discuss this in more detail in chapter 4.

arguments in The Origin of Species, where he argued that "his theory explains "several large classes of facts" (Darwin, 1962, 476), suggesting that his argument was deliberately driven by the desire to demonstrate his theory's superior scope (Dellsén (2024:37)). And a botanical theory is wide-scoped if it can explain the morphology of many different plant species rather than the morphology of the individuals within a given species. However, as this example makes clear, the distinction between types and token phenomena to be explained is not as clear-cut as it may seem at first. Can an account of plant morphology be wide-scoped even if it cannot account for the growth patterns of individual plants? Perhaps not; but if so, we do care about particular instances and not just kinds. On the other hand, is it perhaps the class (flowering plants, say) or family (*Rosaceae*) level, rather than the species (Rosa gallica) level, a wide-scoped theory should account for? Perhaps this is to be decided by the level of explanation we are considering. In any case, for the purposes of this thesis I take it that (i) wideness of scope is a measure of the kinds of entities (or processes, or whatever) that a theory can explain, and (ii) exactly which level is relevant is determined by the context.

A wide-scoped theory preferably covers more phenomena than those considered when developing the theory. Thus wideness of scope blends into the virtues of fertility, or that the theory opens up new productive or interesting lines of research, and novel predictions, or that it makes new testable predictions (as opposed to only accommodating already-known data).[24] As Kuhn has it, "a theory's consequences should extend far beyond the particular observations, laws, or subtheories it was initially designed to explain" (1977:357). Wideness of scope is also connected to unity, discussed below.

3.1. For example, the theory of plate tectonics covers all large-scale geological phenomena on this planet. As Frisch and colleagues write, "[p]late tectonics … orchestrates all known tectonic phenomena including earthquake zones, mountain building, structural patterns,

[24] See Schindler (2018: chapters 3 and 4) for a rich discussion of these other virtues, which I don't have the space to discuss here.

nature of sedimentary basins, magnetism, and metamorphism" (2011:4). This makes its, by far, the most widely scoped geological theory in science history.

4. *Internal consistency*: the theory is free of contradictions. That a theory is free of contradictions would seem to be a minimal requirement for it to be taken seriously: after all, any claim whatsoever can be derived from a contradiction,[25] which further means that an inconsistent theory would not gain any support from having its predictions confirmed. (If your theory predicts both that the sun will rise tomorrow and that it won't, the fact that it does rise does not support your theory.)

   However, this is not a trivial virtue. It has been argued, for example, that Niels Bohr's planetary model of the atom is inconsistent: "according to classical electrodynamics, accelerated charged particles emit a constant stream of radiation. Yet Bohr postulated that electrons would emit no energy on their orbits around the nucleus, but only when transitioning from one orbit to another" (Schindler (2018:8)). This is perhaps not a particularly disturbing example, since we can treat the Bohr atom as an idealised model that gets something important right (the rough relationship between an atom's nucleus and electrons) and, when the overall truth of the account matters, we can use quantum mechanical models of atoms.[26] Be that as it may, this example shows that inconsistencies can quite easily become part of a theory, especially if we count everything that follows from the theory as part of it (it might, say, entail both that a solar eclipse, as viewed from a given position, will begin at 12:11:43 and at 12:11:46, which can't both be true at the same time). Since it is not always easy to determine what's part of a theory and what's a background assumption, and if testing the theory's predictions is not done in isolation but always together with background assumptions

[25] In classical logic, which I take for granted here.

[26] The Bohr model's pragmatic and aesthetic virtues provide reasons for using it in circumstances where it doesn't matter that it's not completely true: its attractive simplicity makes it easy to understand the general picture, thus also making it easier to understand the superseding quantum mechanical models when primed with the Bohr model.

(as per the Duhem-Quine thesis),[27] there is much wriggle room in locating where the inconsistency originated.[28] Even without this wriggle room, inconsistencies following from a theory's peripheral claims may well be tolerated: it need not have any impact on the central claims or predictions of the theory. Sometimes inconsistencies may simply be a sign that the theory is pushed beyond its area of application. For example, using effective field theories for sufficiently high energies will result in infinities, but that may be simply because effective field theories break down at high energies.[29] Effective field theories are approximations to high-energy theories such as quantum field theories and so it is not surprising that they break down when pushed beyond their areas of application; the problem seems to be not with the theory itself but what we are using it for.

That said, an inconsistency shows that some part of the theory is false. This is an obvious problem. We may tolerate it for a period while developing the theory (in the hope that it will be sorted out in due course), if the inconsistency follows from a peripheral claim, or if it is part of an idealised model not expected to be true *as long as* the inconsistency does not touch the features of the model which are supposed to be accurate.[30] But if we discover that an important part of a theory is inconsistent, the theory must be abandoned.

5. *External coherence*: the theory fits well with everything else we know, or, alternatively, it fits well with other well-supported theories. "The theory or hypothesis is consistent with presently accepted theories in other fields," as Helen Longino has it (1996:43). Ernan McMullin describes this virtue as "consistency with other theories and with the general background of expectation" (1982:15). These two versions of *what* it is the theory should

[27] See Bird (2022:186–8) for an argument against the Duhem-Quine thesis.
[28] Peter Vickers (2013) is a book-length discussion of this and other cases of purported inconsistency in science. He argues that actual cases of inconsistencies in science are 'usually harmless' when they do occur, but that a clarification of what a theory is shows that actual theoretical inconsistencies are few and far between.
[29] Thanks to James Read for the example.
[30] See chapter 3 for a discussion of whether false theories can give us (non-factive) understanding.

cohere with both say that a theory should not contradict what we have independent reasons to believe is true. They differ in whether we, for the sake of theory evaluation, should only count our best theories as true or include everything we take ourselves to know to be true.

McMullin says that external coherence "draws attention to itself mainly by its absence, by a dissonance between the theory and some part of its intellectual context" (2014:566). This virtue highlights that it's not just the data which the theory can be directly tested against which counts as evidence for or against the theory, but also what we know from other areas (Williamson (2000:193)).

5.1. For example, southern continents contain many late Palaeozoic and Mesozoic fossils that are too similar to plausibly be the remains of independently evolved plants and animals. Their distribution requires a former land connection, and in the early 1900s this was explained by postulating land bridges which had sunk into the seafloor after the Mesozoic era. However, Alfred Wegener argued (rightly) that this contradicted the findings on isostatic equilibrium. Isostacy require seafloor and continental crust to be of different densities, which means that the lighter continents could not have sunk into the denser seafloor (Hallam (1983:119)). That is, the land bridge hypothesis was inconsistent with the (external) principle of isostasy. The contradiction is not internal to either hypothesis; they contradict each other.

6. *Ad hoc-ness.* this is the only theoretical *vice* on this list. An ad hoc (or, gerrymandered) theory typically contains elements added to save the theory from refutation by adjusting it to fit the otherwise disconfirming evidence. These elements (hypotheses, values of variables, etc.) are neither independently motivated nor required by the theory. An ad hoc theory seems contrived, contains hypotheses either glued together by brute force – that is, without any unifying common factor – or included only to save the theory from refutation, or contains parameters which need to be fixed in light of the evidence in order for that same evidence to bear on the theory.[31]

[31] See Schindler (2018: chapter 5) for an overview of the state of the art on ad hocness and a presentation and defence of his account (discussed below).

'Gerrymandering' is a term borrowed from politics, where it refers to a manipulation of electoral districts for the benefit of a particular political party, for example by drawing district boundaries in a particular way. Like gerrymandered district boundaries, ad hoc theories are artificially adjusted to avoid the consequences of an honest test (fair elections in one case, comparison with evidence in the other).

Everyone agrees that ad hocness is bad, but it's not easy to pin down exactly what it is or why we should avoid it. Lipton argues that what makes ad hoc theories bad is that they are prone to fudging: "[t]he scientist knows the answer she must get, and she does whatever it takes to get it" (2004:170).[32] On this view, we distrust an ad hoc theory because the personal and social reasons a scientists may have to save her favourite theory from disconfirming evidence makes it reasonable to suspect that she introduced errors when she adjusted the theory 'just so'. Schindler (2018) takes the objectionable feature of ad hoc theories to be their lack of coherence, rather than any particular motivation the scientists may have for introducing ad hoc hypotheses. On his account, "a hypothesis H is ad hoc, iff theory T and H do not cohere *or* H does not cohere with the accepted background theories B": "i.e., neither T nor B provides *good reason* to believe that H … rather than non-H" (2018:133; italics original). The unpleasant tension in ad hoc theories is due to their lack of, or low levels of, coherence. And the lack of coherence

[32] It's not unproblematic to define ad hocness partially in terms of the scientists' motivations for including the ad hoc element: the resulting theory is either bad or good regardless of the motivation (Forster and Sober (1994)). Schindler's (2018) definition avoids this problem, but as Lipton (2004) notes, the motivation can be relevant because it may reasonably make us suspect that the theory has been fudged. Lipton writes about accommodation of data versus prediction and invokes fudging to explain why we think that a theory which successfully predicts some data is better than one which merely accommodates the same data. Ad hocness can be viewed as a particularly pernicious kind of accommodation, either because of its motive (to save the theory from honest refutation) or the resulting gerrymandered mess. Lipton's account doesn't quite get to the heart of the issue, since the motivations for postulating a certain theory does not make it ad hoc. As Forster and Sober nicely puts this point: "We do not rule out the possibility that historical or psychological circumstances may sometimes be a reliable *indication* of *ad hocnes*. Our only point is that these circumstances do not *make* a theory *ad hoc*, any more than a barometer makes it rain" (1994:17, fn. 25).

constraints means that no particular content of, or value for, H is specified by the theory: it might as well have been non-H as H, so preferring one or the other (without further reasons) is arbitrary.[33]

Schindler's account erases any differences between a disunified and an ad hoc theory: in both cases, there are parts of the theory not required or supported by other parts of it (or by auxiliary hypotheses or background knowledge). It's not clear whether this captures everything there is to ad hocness: despite the unpleasantness of disunity, an ad hoc theory seems even worse. Perhaps an ad hoc theory is just exceptionally disunified, or perhaps the solution is to combine Lipton's and Schindler's accounts: an ad hoc theory is a disunified theory we have reason to suspect has been fudged. There is more to be said here, but this is not the place for it.

Lastly, let me mention that there is a distinctly aesthetic flavour to the distaste we have for ad hoc theories. Ad hocness is often mentioned when scientists discuss what they find ugly about theories and evokes a reaction more akin to what a repulsive smell does than a theory which simply falls short of an epistemic goal. Ad hocness, or something very like it, also appear in clearly aesthetic contexts: it's like a narrative full of coincidences. Reading a novel full of such 'just so' saves is frustrating and trust-eroding in a way strikingly similar to what we experienced when working with ad hoc theories. For these reasons, I take ad hocness to be a broadly aesthetic quality (we come back to this in section 2.3 below).

6.1. For example, at the time when Alfred Wegener developed his theory of continental drift, it was known that many present-day species and fossils are curiously distributed on either side of continents divided by oceans. A popular explanation at time was the 'land bridge' hypothesis, which

[33] This account has a forerunner in McMullin's account: "A second criterion is internal coherence. The theory should hang together properly; there should be no logical inconsistencies, no unexplained coincidences. … There were too many features of the Ptolemaic orbits, particularly the incorporation in each of a one-year cycle and the handling of retrograde motions, that seemed to leave coincidence unexplained and thus, though predictively accurate, to appear as ad hoc" (1982:15). Schindler lists McMullin's later work (1998) as the first connection of non-ad hocness and coherence he is aware of.

said that these continents had once been connected by a narrow landmass which later sank into the ocean and that animals and plants used these landmasses as bridges to travel from one continent to another. Wegener argued that this hypothesis was "impossible … unless we introduce further hypotheses which are 'ad hoc' improbabilities; for example, that the mass of ocean water was exactly the required amount less at the former period than it is today, or that the deep-sea basins remaining at the time were precisely the required amount deeper than today" (1966:13). Otherwise, "[t]he water displacement of the intercontinental bridges would be so enormous that the level of the world's oceans would rise above that of the whole continental area of the earth and all would be flooded, today's continents and the bridges alike. The reconstruction would not therefore achieve the desired end, i.e., dry land bridges between continents" (1966:13). That is, to save the land bridge hypothesis from being refuted by the fact that there simply was not enough space in present day ocean basins for the hypothesised displaced water, one would have to add a 'just so' story of there being less water in the ocean basins, or the basins being deeper, by just the required amount. Wegener didn't have to state explicitly that saving the hypothesis with an ad hoc move would be impermissible: saying that it is ad hoc is tantamount to saying that it is both distasteful and epistemically bad.

7. *Simplicity*. Simplicity, as is often remarked, is complicated. This is in part because there are many different ways in which a theory can be simple and partially because simplicity has some features in common with pragmatic, epistemic, and aesthetic virtues without clearly belonging to either one camp. Paul Thagard (1978) take simplicity to be a matter of the number of auxiliary assumptions needed by a theory to explain some the data (1978:86), "and perhaps the number of entities posited by the theory" (Cabrera (2023:22)). James Beebe (2009) distinguishes between ontological, explanatory, and psychological simplicity: a theory is ontologically simple when it posits either (i) few entities or (ii) few kinds of entities; explanatory simple when (i) it is "elegant or straightforward", (ii) raises few further explanatory questions, or

(iii) posits few primitive explanatory notions; and psychologically simpler when it "presents an easier to understand relation between explanans and explanandum" (2009:609). Ontological simplicity is often called parsimony, though many use 'simplicity' and 'parsimony' interchangeably: here I reserve 'parsimony' for ontological simplicity. Alan Baker (2022), echoing Quine, distinguishes syntactic and ontological simplicity: "Syntactic simplicity, or elegance, measures the number and conciseness of the theory's basic principles. Ontological simplicity, or parsimony, measures the number of kinds of entities postulated by the theory". Note that I reserve 'elegance' for the aesthetic virtue discussed below and will use Baker's 'syntactic simplicity' or Beebe's 'explanatory simplicity' for uncomplicated theoretical machinery; elegance and simplicity are certainly connected, but a theory can have few moving bits without those bits working together in a particularly elegant way. Explanatory and psychological simplicity are closely connected, though they are not reducible to each other: for example, the only constant needed for propositional logic is Sheffer's stroke, but while it requires less conceptual machinery, doing logic with only Sheffer's stroke is not at all psychologically simpler.

Yet another dimension of simplicity is a theory or model's number of adjustable parameters (Sober (2015:88–9)): the fewer moving bits, the better. Described at a high level of generality, a theory can be simple by either containing few components (basic principles, adjustable parameters, fundamental equations, etc.) or by describing a simple world (postulating few (kinds of) entities or straightforward relations between them). The mini-theory "The world is extremely complex" illustrates how explanatory simplicity and parsimony can come apart. Either kind of simplicity can make a theory psychologically simple, i.e. making it easy to understand and play around with.[34]

[34] Simplicity considerations are widely known as 'Ockham's razor', named after William of Ockham. This principle of parsimony is often formulated as "entities should not be multiplied beyond necessity". Sober (2015) argues that there are two principles in play here which should be kept separate: 'the razor of silence', which urges you to avoid postulating what is not necessary to explain something,

Simplicity cuts across the standard categories into which theoretical virtues are divided. It is variously counted as an aesthetic virtue (e.g., McAllister (1996), Keas (2018), Elgin (2020), Ivanova (2020), Currie (2023), and Todd (2017)), an epistemic virtue (e.g., Ivanova (2020), Schindler (2018), Currie (2023), Todd (2017), Forster and Sober (1994), Swinburne (1997), and, tentatively, McAllister (1996:122)), and a pragmatic virtue (van Fraassen (1980)). I count simplicity as an epistemic value, chiefly because it helps avoid overfitting the theory to noise in the data (see chapter 5 below), a pragmatic virtue (a simple theory is certainly nicer and easier to work with), *and* a broadly aesthetic virtue. By 'broadly' I mean that it is not purely aesthetic but is still distinctly aesthetic. This is because (i) it is almost always mentioned when scientists describe in virtue of what they find a theory beautiful, and (ii) because simplicity gives rise to aesthetic pleasure in both straightforwardly aesthetic contexts and in scientific contexts. 'Simplicity' is typically used in an evaluative way (in the context of theory evaluation), and then with clear aesthetic overtones. (I come back to this in section 2.3 below.)

7.1. For example, Cantor's 'diagonal' proof that the set of real numbers is an uncountable infinite is a classic example of a simple proof, illustrating both explanatory/syntactic simplicity and psychological simplicity.[35] Cantor showed that the set of real numbers $\mathbb{R}$ (like those between 0 and 1) does not stand in a one-to-one correspondence with the set of natural numbers $\mathbb{N}$ (positive integers like 1, 2, 3, …,) Therefore, $\mathbb{R}$ is a larger infinity than $\mathbb{N}$. His proof is regularly praised for its simplicity and elegance (e.g., Franks (2010)).[36] Very roughly, it goes like this: list all the numbers between, say, 0 and 1, counting them all:

i) **0**,0000…
ii) 0,**1**111…

and 'the razor of denial' which urges you to reject the least parsimonious of two hypotheses (2015:12).

[35] This discussion is based on Montaño (2013).

[36] See Franks (2010) not just for aesthetic praise of Cantor's proof but for a nice discussion of it and the two previous proofs Cantor published for the uncountability of $\mathbb{R}$, together with historical details on why it took him (only) three attempts to discover the diagonal proof.

iii) 0,0**1**01…

…

This list of numbers contains a number not listed, namely the one you get by taking the first digit of the first sequence, the second of the second, the third of the third, etc. (highlighted in bold). Had you listed the highlighted number, you would thereby also have added a new, unique diagonal number. This generalises, so given any countable set of real numbers, there is a real number not listed in that set. Montaño says that "Cantor's diagonal argument is simple not only in the sense that it is very parsimonious, involving only a few steps, but also in the sense that understanding it requires very little mathematical background" (2012:24). Note that he uses 'parsimonious' for what is better described as explanatory/syntactic simplicity. (Indeed, if it makes sense to say that the proof postulates entities, then it is *infinitely* unparsimonious.)

8. *Elegance*: an elegant theory skilfully combines simplicity and strength, balancing a simple theoretical machinery with an ability to explain much. As Steven French says, "a typical way of cashing this [elegance] out in the scientific context is via a combination of parsimony and power" (2020:204). Examples include Steven Weinberg's claim that "an elegant proof or calculation is one that achieves a powerful result with a minimum of irrelevant complication" (1993:107), and E. O. Wilson's that "[t]he elegance, we can fairly say the beauty, of any particular scientific generalization is measured by its simplicity relative to the number of phenomena it can explain" (1978:11). An elegant theory occupies a sweet spot between explaining much and doing so in a simple way, between strength and simplicity.

   A further feature of an elegant theory is that it makes its explanations seem easy and straightforward, or even inevitable: everything falls into place so precisely that it seems like it couldn't have been otherwise. Or, to quote Weinberg quoting Einstein, "to modify [an elegant theory] without destroying the whole structure seems to be impossible" (1993:108). Such inevitability or naturalness is characteristic of highly skilful work, and praising a theory for its elegance might well contain an element of admiration for the skilfulness of the scientists who formulated the theory (see Ivanova

(2021)). More importantly, such naturalness is an expression of the coherence of the overall framework: everything fits together, nothing sticks out. The tighter the coherence constraints, the more inevitable the theory will appear: if all the parts are anchored to each other, you cannot modify one part without disturbing the whole structure. This holds for both internal and external coherence. (Theories scoring low on coherence, on the other hand, can range from false (if its (central) claims contradict each other) to containing many unexplained coincidences).

Weinberg's definition of beauty comes closer to what is usually meant by elegance in science: it's the combination of "a simplicity of ideas, not simplicity of a mechanical sort that can be measured by counting equations or symbols" and "the sense of inevitability that the theory may give us" (1993:107). (Here I take elegance to be one of the ways in which a theory can be beautiful: more on this below.) Ian Glynn, discussing a proof the details of which we don't need to worry about here, says "[t]hat is a very elegant proof indeed. Once one has understood the argument, it can be seen at a glance and one has no doubts about its validity. It is simple, ingenious, concise, and persuasive; it has an unexpected quality; and it is also very satisfying" (2010:3–4). This gives a good summary of the properties and effects of elegance. The fact that an elegant argument, proof, or theory makes it easy to gain an overview of what it states or entails ('it can be seen at a glance') plays a large role in why it's so satisfying. This ease is well known from research on both metacognition and aesthetics and will play a large role in later chapters. An elegant theory is graceful and economical, without superfluous or unmotivated elements, and is structurally (or narratively) so clear that it's easy to gain an overview of what the theory entails. It is simple yet powerful, and its wide range of explanations follow with ease from its central claims.

8.1. *Example 1*: The Vine-Matthews-Morley (VMM) hypothesis, a cornerstone of the theory of plate tectonics, is often described as elegant. "The hypothesis offered an elegant explanation", wrote Ron Mason, which is part of why he "had absolutely no doubt as to the correctness of [it]" (2004:41). The VMM combines the hypothesis that

new seafloor is created at and spreading away from midocean ridges and the hypothesis that the earth's magnetic field changes polarity from time to time, and says that the upwelling magma which becomes new seafloor takes on the polarity of the earth's magnetic field at the time.[37] The seafloor reflects the geomagnetic field at the time it was created, rendering it a 'magnetic record keeper' for the age of the seafloor. What is elegant about it? Two things in particular stand out: its simplicity, and the way it combined wide-ranging ideas into one harmonious picture. The VMM is an upshot of the hypothesis of seafloor spreading, the hypothesis of geomagnetic reversals, and the hypothesis that remanent magnetisation is widespread in the ocean floor: it does not add anything new, merely spells out how these hypotheses together can explain the magnetic lineations of the seafloor. So it is simple in the sense that it is extremely parsimonious, not adding new content, and also simple in that it's very easy to understand. The VMM also showed clearly how nicely the three hypotheses fit together, how they can be combined into one harmonious picture.

8.2. *Example 2*: We used Cantor's diagonal proof of the uncountability of the set of real numbers to illustrate simplicity, but it can also be used to illustrate elegance – again, this is because simplicity and elegance are intimately connected. Ulianov Montaño says that "since the storytelling of Cantor's diagonal proof is pleasingly simple yet effective, it is natural to describe it in the same way we describe other simple yet effective pleasing things: as elegant, which is indeed the way the diagonal proof is often described" (2013:24-5).

9. *Unity (and harmony)*: a unifying theory combines simplicity and broadness of scope, revealing the underlying connections between its disparate target phenomena. "Accounts of unification often appeal to simplicity. For example, Kitcher's influential (and still dominating) account states that a theory is unifying if it allows one to derive a large number of phenomena from a small number of argument patterns" (Schindler (2022:546)).

[37] See chapter 4 below for more on the VMM.

(Kitcher's (1981) account of unification is an account of explanation, not an account of one of the signature features of good explanations. Still, the idea that unification is a matter of connecting disparate phenomena to a small number of underlying features also describes unification as a theoretical virtue.)[38]

Other accounts take unity and broadness of scope to be the same virtue. For example, Adolfas Mackonis (2013) says that "[t]here are two main requirements for a hypothesis to be unifying, broad, consilient, have a wide scope, or explain a variety of facts. It has to account for at least two different kinds of facts (the more the better) and unification cannot be a simple conjunction of hypotheses or observational statements" (2013:983).[39] I think it's more useful to treat the second criterion as a way of distinguishing good from bad wide-scoped theories: one of the main ways in which an ad hoc and a unified theory differ is exactly by whether the parts of the theories unified or rather glued together by brute force, not in how many phenomena they cover.

I take unity to be closely connected to harmony, or the aesthetically attractive structuring of explanatory relationships characterising many beautiful theories. For example, Henri Poincaré talks about the "intimate beauty which comes from the harmonious order of its [nature's] parts" (in Ivanova (2017:2584)). (We come back to this in chapter 4 below.) As discussed shortly, a beautiful object is often characterised as one which shows unity in disunity. It should come as no surprise, then, that one of the

[38] Kitcher's account is inspired by Hempel's "unofficial account" of explanation (1981:508). Hempel wrote that "[w]hat scientific explanation … aims at is … an objective kind of insight that is achieved by a systematic unification, by exhibiting the phenomena as manifestations of common underlying structures and processes that conform to specific, testable, basic principles" (1966:83). A unifying theory shows the underlying connections between the phenomena it is about.

[39] Here he follows Thagard, who wrote that "[c]onsilience is intended to serve as a measure of *how much* a theory explains" (1978:79; italics original) and "[a] consilient theory unifies and systematizes" (1978:82). Thagard's notion of consilience is closer to what I above called broadness of scope, though as the second quote shows, it goes beyond the sheer number of phenomena explained by a theory.

features of theories most often praised as what makes a theory beautiful is its ability to connect a wide range of phenomena (disunity) through a simple explanation (unity).

9.1. For example, as we saw above, the theory of plate tectonics is good example of a wide-scoped theory. It is an even better example of a unified theory: not only does it explain all large-scale geological phenomena and processes, but it does so through a small number of hypotheses which tie together this variety of features. McKenzie and Parker (1967) and Morgan (1968) presented a simple geometrical framework which together with a small number of assumptions and hypotheses (e.g., that plates are rigid pieces of crust resting on a layer of molten rock, that their movements can be described as rotation on a sphere, and that seafloor is denser than continental crust) showed how mountain building, seafloor spreading, subduction zones, the presence of marine fossils in mountain ranges, and the nature of metamorphic rocks, to take a few examples, are all connected through the same process of plates moving relative to one another across the face of the earth. (We discuss this at some length in chapter 4.)

10. *Beauty.* 'Beauty' seems to be used in one of two ways in science: either as a catchall term for aesthetic value, or as synonymous with 'elegance'. When used as an overarching term for aesthetic value, it covers a range of more specific aesthetic properties, such as elegance, harmony, or simplicity. This is for the most part how I will use it here.

"That beauty evokes an experience of pleasure is probably the only point on which all participants in the continuing debate on beauty agree," as Jennifer McMahon says (2005:1). Still, there are some common themes to claims that a scientific theory is beautiful. For example, Steven Weinberg highlights the simplicity and 'sense of inevitability' typical of beautiful theories (1993:107), where the sense of inevitability comes from a fit so perfect that (here he's quoting Einstein) "to modify it [the theory] without destroying the whole structure seems to be impossible" (1993:108). G. H. Hardy famously wrote that "the mathematician's patterns, like the painter's or the poet's must be beautiful; the ideas … must fit together in a

harmonious way" ([1940] 2005:14). Poincaré, as we saw above, wrote that the kind of beauty relevant to science "comes from the harmonious order of its [nature's] parts" (in Ivanova (2017:2584), and Thomas Huxley wrote that "beauty, as a general rule, implies simplicity; I do not mean the simplicity of monotony, but the simplicity of unity. That which is highly and nobly beautiful always conveys an impression of balance, harmony, or rhythm; the parts, however various they may be, are related in a way which produces an intellectual satisfaction" (quoted in Blinderman (1962:52)).

This focus on simplicity combined with a harmonious, inevitable, or otherwise excellent fit makes it clear that what these scientists are talking about is the same property aestheticians are talking about when they describe beauty as involving 'unity in diversity' or the subsumption of different parts under a simple, unifying pattern (McAllister (1996: 19, 110), Cochrane (2021: 42), Engler (1990: 27), Blinderman (1962), Hutcheson (1726): we come back to this in section 4.2.2 below). Here I take beauty to be a broad aesthetic category, subsuming many more specific aesthetic qualities, though beautiful theories typically combine simplicity with an excellent fit or otherwise exemplify the traditional aesthetic ideal of unity in diversity.

10.1. For example, the theory perhaps most often praised for its beauty is general relativity. I'll let Steven Weinberg describe the beauty of this theory: "there is another quality beside simplicity that can make a physical theory beautiful – it is the sense of inevitability that the theory may give us. In listening to a piece of music or hearing a sonnet one sometimes feels an intense aesthetic pleasure at the sense that nothing in the work could be changed, that there is not one note or one word that you would want to have different. … The same is partially true (it is never more than partially true) of general relativity. Once you know the general physical principles adopted by Einstein, you understand that there is no other significantly different theory of gravitation to which Einstein could have been led. … Einstein could not have incorporated an inverse-cube law in his theory without scrapping its conceptual basis. Thus Einstein's fourteen equations have an inevitability and beauty that Newton's three equations lack" (1993:107–8).

This concludes our overview of some of the central theoretical virtues. As we have seen through multiple examples, scientists appeal to broadly aesthetic virtues when evaluating theories (see chapter 4 below for more examples from the earth sciences). Since this happens regularly, whether scientists are justified in appealing to broadly aesthetic considerations when evaluating theories concerns the legitimacy of a common feature of scientific practice and methodology. The central point of contention concerning the role of aesthetic virtues in theory evaluation is exactly whether aesthetic virtues are 'merely' aesthetic or also epistemically valuable. The first step towards solving this issue is to understand what it is that makes aesthetic virtues aesthetic and epistemic virtues epistemic. We discuss aesthetic virtues in the next section, and epistemicity in chapter 3 below. In chapter 5 I argue that aesthetic virtues are also epistemic.

2.4. (Broadly) aesthetic theoretical virtues

The broadly aesthetic theoretical virtues include beauty, elegance, harmony, unity, simplicity, and non-ad-hocness. Some of these virtues are more straightforwardly classified as aesthetic than the others: it's difficult to doubt that scientists make aesthetic judgements when they talk about the great beauty of a theory, but what about when they talk of the simplicity, unity, or ad hocness? Are these properties genuinely aesthetic?

To start with the conclusion, in this section I argue that there are many good reasons to count the broadly aesthetic virtues as *aesthetic* virtues. However, my argument in this thesis does not turn on whether they are genuinely aesthetic or not: the question of what epistemic role, if any, is played by these virtues arise regardless of what exactly we call them. For our purposes, the central question is whether scientists' preferences for (say) simple and unified theories are in part based on aesthetic experiences, not whether simplicity and unity are genuinely

aesthetic qualities. That said, we still need a rough grasp on what aesthetics is, so let us start there.[40]

2.4.1. Beauty, sense perception, and formalism

'Aesthetics' is sometimes narrowly understood as a synonym for 'beauty' and mainly concerned with art and the experiences interactions with objects of art give rise to. This is a relatively recent and geographically specific understanding of aesthetics, roughly speaking limited to the Western philosophical thinking about aesthetics after Kant's *Critique of Judgement* (1790).[41] But this focuses on only one aspect of our aesthetic lives: clearly, it is not limited to experiences of *beauty* (consider deliciousness or messiness) or the creation and appreciation of *art* (consider facial beauty or the drama of a thunderstorm), so this narrow conception of aesthetics is at best severely limited.[42] Our present concern, the aesthetics of science, clearly falls outside its scope. While scientific theories are often appreciated for their beauty and there are commonalities between science and art,[43] aesthetic evaluation of theories is not limited to beauty and science is not art.

---

[40] The following is not meant to be a comprehensive discussion of aesthetics, only to give us a good enough understanding of aesthetic qualities, judgements, and experiences to make sense of scientists' claims about the aesthetic value of theories. Accounting for aesthetic value is among the most challenging philosophical tasks, up there with accounting for moral value, and there is much more to be said about all the points raised in the following. However, there is not space here to both dive deeply into the philosophy of aesthetics and push forward on accounting for the role of aesthetics in science, so the following brief overview will have to do.

[41] Kant's *Critique* is the most prominent example of a then-recent trend rather than the first work to focus the study of aesthetics on beauty. Hegel's later focus on art was influential in narrowing down the study of aesthetics even more (Cooper (1997:1–2)).

[42] For discussions of everyday aesthetics, see Saito (2007), for aesthetic life more generally, se Lopes, Nanay, and Riggle (2022) and Cochrane (2021), for environmental aesthetics, see Carlson (1979) and Berleant (1978), for aesthetics of games, see Nguyen (2020), and for the aesthetics of food, see Plakias (2021).

[43] Many of the papers collected in Bueno et al. (eds.) (2018) highlight the similarities between art and science. See also Ivanova (2023), Engler (1990), Wylie (2021), and French (2020) for more on aesthetics in art and science.

Alexander Baumgarten, who coined the term 'aesthetics' in the 1750s, defined it as "the science of sense perception" (*scientia cognitionis sensitivae*) (Nanay (2014:101), Wragge-Morley (2020:5)), building on "the Greek *aisthēsis*, which means perception by the senses" (Berleant (2015:2)). This better captures the wide range of topics that aestheticians are concerned with than does a definition of aesthetics as the philosophy of art, especially if we make explicit that aesthetics concerns the *affective responses* to sense perception and not the workings of visual, auditory, etc., systems. It's also more fruitful starting point for considering aesthetics in science than the narrow focus on art and beauty is, as Wragge Morley shows in his excellent (2020) account of the tight links between aesthetic appreciation and science in the early days of the Royal Society of London. It can account for the aesthetic and artistic aspects of much hands-on scientific work, such as fossil preparation (Wylie (2021)), field sketches (Turner (2019)), and experimental design (Ivanova (2023)). As Milena Ivanova says,

> Experimenters, like artists, utilise their creativity, imagination, and aesthetic sensibility in diverse stages of experimental process. (2023:26)

And Caitlin Wylie writes that

> Fossil preparators do not sculpt fossils from matrix like artists sculpt statues from marble, nor do they collect them from the ground as pristine, stable, complete skeletons. Instead, preparators draw on creativity, sensory judgment, and notions of beauty to craft evidence from nature. (2021:204)

It seems that an account of aesthetics focused on sense perception can work well for such cases. That is not surprising, since these cases are about the role of aesthetics in scientific practice – here, in the design and performance of an experiment, and in the preparation (*sculpting*) of fossils. It does less well when the object of aesthetic appreciation is a scientific theory.

In fact, the focus on sense perception is too narrow to account for all (or even most) aesthetic experiences, both within science and without. Among other things, it excludes literature, conceptual art, scientific theories, and mathematical proofs from the aesthetic domain. Although experiencing these objects require sensory perception, the sensory experience itself is not the object of the aesthetic appreciation. We may need to write down each step of an

argument to see how ingenious it is, but what we appreciate is the argument itself and not any one token representation of it. Compare this to aesthetic experiences outside research contexts. For example, a crisp font and the thick, powdery feel of high-quality paper contribute to the aesthetic experience of reading a novel, just as background music contributes to the aesthetic experience of a meal: it can impact the overall atmosphere, but does not change the aesthetic properties of the novel or meal and typically[44] does not significantly change the aesthetic experience of reading the novel or eating the meal. That is, the novel is accessed through sense perception (how else?) and aspects of the sensory situation makes a difference to the overall aesthetic experience of reading it, but the aesthetic qualities of the novel itself are independent of how it is presented to the senses. The same goes for scientific theories. Although we cannot appreciate a theory's aesthetic properties unless we have seen or heard a representation of the theory, the aesthetic appreciation concerns the theory itself, not a representation of it.

James McAllister (1996) argues that it's primarily the theory *qua* abstract object, not any representation of the theory, which is the object of aesthetic appreciation, and further that a 'faithful rendering' of a theory can reveal the theory's aesthetic properties:

> Scientific theories are abstract entities, and abstract entities are not accessible to our senses of sight and hearing. Accordingly, we acquire information about theories not by seeing or hearing them, but by contemplating representations of them. These representations are accessible to our senses, since they are concrete entities such as publications and lectures.… [W]e apprehend certain properties in an abstract entity upon perceiving certain other properties in some concrete

[44] Perhaps it does in some cases: eating what would otherwise have been a nice meal in a foul-smelling house, for example, impacts the experience of the aesthetic properties of meal (such as its deliciousness or even whether it's too salty). This may be because taste is particularly multimodal. As Barry Smith says, "[w]hat we call taste encompasses the combined sensory inputs of taste, touch and smell, as influenced by sight and sound" (2012:6). But we find a similar case with colour juxtaposition, such as the difference the colour of a wall and the frame makes to how we experience a painting hung on it. In this case, the colour of the wall is part of the aesthetic experience because it is part of the same visual input (just like how the sound of traffic can be part of the experience of listening to a song because the auditory inputs are combined) and because colours look differently depending on nearby colours, as is well known from visual illusions. There is much more to be said here, but this is not a place for it.

> rendition of the entity. … [S]ome properties of a faithful representation will be owed to the mathematical structure, logical parsimony, or other properties of the theory, that show through in the rendering. (1996:28–9)

In other words, it is what we apprehend (understand, comprehend, or 'get' in one way or another) through perceiving a concrete representation of a theory which is the proper object of aesthetic[45] evaluation of theories. And "the aesthetic properties of theories' representations [should] be given no weight in evaluating theories" (1996:26). But this is too conservative: as the recent literature on aesthetics in science clearly shows, both concrete outputs such as illustrations and prepared fossils and aspects of scientific practice such as experimental design, field drawing, and fossil preparation are evaluated aesthetically.[46] That said, like McAllister, here we are concerned with the aesthetics of theories, and it is theories, not their representations, which are the bearers of theoretical virtues.

Has McAllister given us a complete account of how we get to know the properties of abstract objects? No, but a discussion of the aesthetics of science is also not the place to settle the age-old question of the epistemology of abstracta.[47] For our purposes, McAllister's account will do. Spelling out the difference in detail between the aesthetic appreciation of an abstract object and the representations through which we experience it is a huge task better left for another day. Also, since the account I present in chapter 5 focuses primarily on aesthetic *experiences* of theories rather than the aesthetic virtues themselves, sorting this out is not particularly pressing for our concerns. The aesthetic qualities of a representation of a theory may play a role in the experiences underlying the aesthetic judgements of the theory, that is, in detecting the theory's aesthetic qualities. Just as it can make a difference to how we judge the aesthetic value of a poem whether it is read silently or recited, or whether it's

---

[45] And other kinds of evaluation, of course: this is also supposed to be how we get to know that the theory is about muons, for example, not just that it is a lovely theory of muons.

[46] See, for example, Wylie (2021), Turner (2019), the papers in Ivanova and Murphy (eds.) (2023), Currie (2023), and Kozlov (2023).

[47] See Benacerraf (1973) for a good formulation of the problem of our epistemological access to mathematical facts and objects, a problem which extends to other abstract objects.

written in a hard-to-read font, a theory may seem much less aesthetically satisfying if presented in such a way that it is difficult to understand what it says. We come back to that later; for now, it is enough to note that aesthetic theoretical virtues are properties of theories, not representations of them.

That, however, will be resisted by formalists. Formalism is the view that only the formal properties of an object are aesthetic (or aesthetically relevant: see Nanay (2016:92)). Formal properties are features like the intrinsic surface properties of a painting (lines, perhaps colours,[48] paint texture), the pitch and timbre of a piece of music, and the texture and weight of a sculpture. It is only qualities internal to the piece itself which makes a difference to its aesthetic value,[49] and the only internal qualities which count are features which can be perceived.

Although formalism "has been criticized severely for several decades" (Nanay (2016:93)) – "[a]t best it receives unsympathetic discussion and swift rejection. At worst it is the object of abuse and derision" (Zangwill (1999:610) – some moderate forms of formalism may manage to avoid implausibly restricted views of aesthetic qualities.[50] Bence Nanay (2016) argues for semi-formalism, according to which "the only aesthetically relevant properties of an artwork are [the] properties that depend constitutively on the formal properties of the artwork" (2016:113). A surprising turn in a poem, for instance, may depend upon its line break and metre in addition to its imagery, and will thus be aesthetic on Nanay's account. What about a surprising turn in a proof? Here it is far less clear that its aesthetic properties (here: surprise)[51] depend constitutively on any formal properties.

---

[48] Although colours are generally taken as a formal property, whether they are intrinsic properties depends on the metaphysics of colour and so it is contentious whether formalists can count colours as part of an artwork's formal qualities. For a good discussion of the metaphysics of colour, see Chirimuuta (2015).

[49] Excluding, among other things, knowledge of its production history.

[50] See also Kivy: "the first stumbling block to the plausible extension of aesthetic criteria to the evaluation and appreciation of science is the result of adhering to aesthetic formalism in its more extreme manifestations" (1991:184).

[51] Surprise is not always an aesthetic property, but it can be, as is clear in the case of poetry. Literature (and other art forms) are not infrequently criticised for being too predictable or clichéd, and in those cases it seems clear that surprise –

I take this to be a problem for the moderate formalists to sort out: either they can account for the aesthetic value of poetry, narrative fiction, proofs, and theories – and if so, we may use them to explain in virtue of what aesthetic qualities are aesthetic – or they cannot, but then they are either incomplete or false. An account of aesthetic qualities which requires us to reject large swathes of commonplace aesthetic evaluation is unacceptable: if we have to choose between the well-established fact that novels (and scientific theories) are objects of aesthetic appreciation and a particular philosophical account of aesthetics, the latter will have to go.

2.4.2. Are aesthetic theoretical virtues genuinely aesthetic?

Elegance. Grace. Beauty. Ugliness. Messiness. These are common aesthetic qualities we recognise from our (undoubtedly aesthetic) evaluations of clothing, architecture, people, cars, paintings, music, and much more. They are also commonly invoked in describing scientific theories. Given how ubiquitous these qualities are in our aesthetic lives – which means that most people, scientists included, are familiar with them[52] – it would be perfectly fine to take scientists at face value when they *appear* to talk about the aesthetic qualities of theories as in fact talking about aesthetic qualities of theories. Or at the very least, that they are expressing aesthetic experiences, such as a feeling of elegance or delight in finding a surprising solution. The default position, the one we should assume unless and until we get a specific reason to think otherwise, should surely be that

---

or rather, the lack of it – is an aesthetic property (thanks to Tim Williamson for the example).

[52] You are not only familiar with aesthetics from visits to art galleries or national parks. Consider a normal morning: you get dressed in a newly ironed shirt because it looks nicer than the crumpled one (although for practical purposes, both are equally fine), you comb your hair or put on lipstick to look prettier, you make a cup of coffee which taste too bitter so you resolve to buy different beans next time, you put on your favourite rock band or lose yourself in a novel on your commute. These are all examples of everyday aesthetic evaluations and experiences and is a reminder of how readily we make judgements about what's neat, sophisticated, dingy, pretty, messy, sweet, sour, bitter, cool, exhilarating, and so on. See Saito (2007) and Lopes, Nanay, and Riggle (2022) for more on aesthetics in everyday life.

scientists use aesthetic vocabulary in standard ways to express their aesthetic judgements.

It's fine to defer to aestheticians on the nature of aesthetics when discussing what role aesthetic judgement plays in one limited area, such as scientific reasoning. Given that the qualities, judgements, and feelings we are studying are the kinds of qualities, judgements, and feelings *they* are studying, our understanding of the aesthetic can ride piggyback on what aestheticians take them to be. Compare: a discussion of how moral values impact scientific judgement does not need to start with a discussion of which ethical theory is correct but can rely on a combination of our everyday familiarity with ethics and a deference to moral philosophers. The familiarity ensures that we are picking out moral evaluations also in the scientific case, and the deferral is an acceptable division of labour. Similarly, our everyday familiarity with aesthetics makes us good judges of which evaluations are aesthetic also in the scientific case, and here too deferring to the experts for the finer points is an acceptable division of labour.[53]

The crux, then, is whether what we have hitherto identified as broadly aesthetic virtues (such as elegance, beauty, harmony, unity, simplicity) are indeed the same kinds of qualities that aestheticians, *qua* aestheticians, study.

Cain Todd (2008) has argued against taking apparently aesthetic evaluations in science to be genuinely aesthetic. According to him, scientists may instead carry out epistemic assessments and then express the result, or their endorsement of the theory, in aesthetic terms. Todd argues that the data on how scientists invoke aesthetic considerations in science does not tell us whether these judgements are genuinely aesthetic evaluations. And so,

> [a]t best, some positive account of the relation between epistemic and aesthetic criteria is needed, but as these are not generally forthcoming one must, I think, be at least initially suspicious that such claims are simply conflating aesthetic and epistemic criteria of assessment. In light of such

[53] In both the ethical and the aesthetic case, the experts are far from agreeing on the nature of their objects of study. That makes the claims we can make more tentative, but the disagreement within the fields does not make someone from outside a better judge (rather the contrary, since the disagreement in all likelihood reflects a complexity of the subject matter not obvious to the uninitiated) and so does not give us a reason to avoid deferring to the experts.

> suspicions, coupled with the problem posed by aesthetic judgements for rational theory choice, I suggest that there is a burden on defenders of the claims [that apparent aesthetic evaluation in science is in fact aesthetic] to provide clear evidence, in the form of distinguishing criteria, that the supposed pleasure expressed is genuinely aesthetic. (2008:65)[54]

In other words, until we have clear evidence that scientists are *not* making epistemic judgements when they praise a theory for (say) its simplicity and elegance, we should think that "the supposedly aesthetic claims are not genuinely aesthetic but are in fact 'masked' epistemic assessments" (2008:61).

I disagree. Todd's argument is built on the assumption that aesthetic virtues cannot *also* be epistemic, but I do not think we have any reason to accept that. On the contrary, as I argue in chapter 5, aesthetic virtues are also epistemic. It's a case of *yes, both*, not *either–or*. Thinking that theoretical virtues cannot be both epistemic and aesthetic is based on a false dichotomy between epistemic and aesthetic virtues. We also have an abductive reason to think that some virtues are both aesthetic and epistemic: if they are, much of what seems puzzling about scientists' aesthetic evaluation in epistemic contexts makes sense. We do not have to explain away either the fact that they make aesthetic assessments or that these aesthetic assessments also function epistemically, but can account for all the data which are so hard to reconcile when we try to fit it into either one of these categories. To accept the claim that aesthetic virtues are also epistemic as more than a working hypothesis we need an explanation of how they are connected (as Todd rightly says, "some positive account of the relation between epistemic and aesthetic criteria is needed" (2008:65)); I provide one in chapter 5. In fact, my explanation is similar to one provided by Todd himself in his later work.

Todd (2008) states that "[c]learly the criteria at issue will not be functioning aesthetically if they are in fact reducible to 'clearly' non-aesthetic

[54] The fact that we are not able to provide criteria distinguishing aesthetic from non-aesthetic properties does not show that we don't recognise the difference when we see it. I take it that Todd is after necessary and sufficient conditions for when a theoretical virtue is aesthetic, but there are vanishingly few properties of which we can provide an informative definition. Despite this, we typically have a good grip on what belongs within the categories we employ, except when it comes to borderline or highly unusual cases.

criteria" (2008:71), but that is not so. *All* aesthetic qualities are based on features which can be described non-aesthetically. (Whether they can be fully reduced to such properties is another issue, and one which I cannot take a stand on here.) For example, the aesthetic qualities of a song are based on the frequency of sound waves, and the well-balanced sweetness of a honey vinaigrette on its chemical composition. Does that mean that sad music and sweet food are not really aesthetic? Or, more cautiously, that they do not give rise to aesthetic experiences? Of course not! As Dominic McIver Lopes puts it, "[e]very aesthetic value is 'realized' by nonaesthetic features of the item" (Lopes (2022:62)). For example,

> When you're disagreeing with a friend about the valedictorian's [graceful] speech, what you do is point out its nonaesthetic features. You note how she sneaked in a sly reference to the school motto, how a single phrase was repeated several times, and how she chose plain talk over fancy words with Latin origins. … The speech is graceful, and so is the swoop of a great blue heron across the water. They share that in common. But what realizes grace in one is totally different from what realizes grace in the other. (2022:62–3)

The non-aesthetic qualities underlying aesthetic qualities are perhaps not 'totally different': since they can give rise to the same aesthetic qualities they must have something in common, but the slyness of a speech or the swoop of a wing are of course *very* different. So are the number of free parameters in one theory and the sparse conceptual framework of another, though they both make the theories simple.

This brings us to a more general worry one may have about the supposedly aesthetic nature of some of the broadly aesthetic virtues. Some aesthetic virtues seem unproblematically aesthetic: beauty, elegance, and harmony are paradigmatic aesthetic qualities. Simplicity and unity, however, provide less clear-cut cases.

One reason why simplicity, and perhaps unity, differ from the clearly aesthetic properties is that they can be given purely descriptive characterisations. An aesthetic judgement seems to be inherently evaluative, giving a verdict on the aesthetic value, good or bad or in between, of the object judged. Judging that a theory is beautiful is thereby to judge that it is aesthetically valuable. But this is

not necessarily the case when we are making judgements of simplicity. Noting that a theory only posits three fundamental forces or that a model has few free variables is in itself no more evaluative than noting that a plant has a deep green colour. But, again just like colour, simplicity often plays a central role in aesthetic experiences and judgements. A painting can be powerful in virtue of its bold colours, tacky due to its poor combination of colours, or it can elegantly combine a challenging palette. Similarly, a theory can be elegant in virtue of the simple way it combines various hypotheses, boringly simplistic, or it can beautifully unite disparate phenomena under one simple common principle. A theory can be weak because it is *too* simple: it can be impoverished and oversimplified, lacking the resources to explain or predict the facts and features it set out to explain.

But note that in such cases the complaint is about *over*simplification. Calling a theory simple or unified is almost always meant as praise and is often part of an explanation of why one finds the theory beautiful.[55] A typical expression of this is found in this quote by T. H. Huxley (the 19th century biologist better known as "Darwin's bulldog"):

> beauty, as a general rule, implies simplicity; I do not mean the simplicity of monotony, but the simplicity of unity. That which is highly and nobly beautiful always conveys an impression of balance, harmony, or rhythm; the parts, however various they may be, are related in a way which produces an intellectual satisfaction. (In Blinderman (1962:52))

Or consider Werner Heisenberg's description of scientists' aesthetic appreciation of simplicity and unity:

> "Beauty is the proper conformity of the parts to one another and to the whole." That this criterion applies in the highest degree to a structure like Newtonian mechanics is something that scarcely needs explaining. The parts are the individual mechanical processes – those which we carefully isolate by means of apparatus no less than those which occur inextricably

[55] Categories which combine descriptive and evaluative features are often called 'thick concepts'. Classical examples of thick concepts are *friend*, *duty*, *dumpy*, and *dainty*: it seems like simplicity also belongs to this list. Thick concepts are well known from ethics (see Williams (1985)) but can equally well be applied to aesthetics, as Zangwill (1995) does. Perhaps the aesthetic/evaluative aspect of theoretical simplicity can be explained by reference to thick concepts, but that investigation will have to be left for later work.

> before our eyes in the colorful play of phenomena. And the whole is the unitary principle of form which all these processes comply with and which was mathematically established by Newton in a simple system of axioms. Unity and simplicity are not, indeed, precisely the same. But the fact that in such a theory the many are confronted with the one, that in it the many are unified, itself has the undoubted consequence that we also feel it at the same time to be simple and beautiful. (1970:6–7)

This is one reason to think of simplicity and unity as aesthetic virtues of theories: 'simplicity' is typically used in an evaluative way, and then with clear aesthetic overtones. While the simplicity of a theory *can* be a purely descriptive matter, in practice judging a theory to be simple is a value-laden matter. The same goes for unity.

In a clearly aesthetic context – say, when discussing art – both colours and the simplicity or complexity of a piece count as clearly aesthetic. But it would be awkward to count colours as aesthetic when discussing the physiology of vision, exactly because they are then not used in an evaluative way. It is similarly awkward to count simplicity as aesthetic in some contexts. The quotes above gives us good reason to think that science is *not* one such context, but clearing this up beyond a doubt would require a way of firmly categorising borderline cases of aesthetic qualities – a daunting task, to say the least, and one which is far beyond the scope of this thesis.

Whether themselves aesthetic or not, simplicity and unity certainly play a major role in scientists' aesthetic experiences, as we have seen multiple examples of and will see even more examples of in the following chapters. For our purposes, we don't need more than this to count simplicity and unity as broadly – that is, at least partly – aesthetic: our question is not whether these properties are *really* aesthetic, but whether scientists' preferences for simple and unified theories are in part based on their aesthetic experiences. These virtues are part of what makes a theory more aesthetically attractive than another, and it is the fact that this attractiveness not only influences theory choice but seemingly can do so to great effect which makes the role of broadly aesthetic considerations in science so puzzling. Whatever we call these virtues, it is clear that they are at least similar to aesthetic qualities and that they give rise to what seems to be straightforward aesthetic experiences – experiences which, in turn, guide theory evaluation.

So, what are these aesthetic experiences?[56] Noël Carroll (2002) groups theories of aesthetic experiences into axiological, content-based, and affect-based accounts (2002:146). These accounts focus on, respectively, the value that an aesthetic experience is meant to have, where the "usual suspect" is that "something is taken to be an aesthetic experience only if the experience is valued for its own sake or is intrinsically valued" (2002:147); "the features towards which the experience is directed" (2002:147), such as unity in disunity or significant form; and the phenomenological feel or experiential qualia of the experience, such as delight in contemplation or disinterested pleasure (2002:146). (Of course, and as Carroll also stresses, many accounts combine some or all approaches.)

The focus we have had so far on aesthetic qualities lends itself quite naturally to a content-based account of aesthetic experiences: an aesthetic experience is an experience *of* an aesthetic quality, such as elegance, symmetry, simplicity, messiness, etc. All experiences, aesthetic or not, fall somewhere on a spectrum of valence – from positive to negative – and intensity – from high to low arousal. Combined with a focus on the aesthetic qualities of theories, we can account for many aesthetic experiences in science, from mild liking *of a theory's simplicity* to intense delight *in its beauty*, etc. This fits well with the examples we have seen of how scientists describe their aesthetic experiences. It is also a simple and straightforward way of accounting for aesthetic experiences. This approach will of course have to grapple with the question of what aesthetic qualities are, though as I argued above, we have good reason to think that the qualities scientist point to in their (apparent) aesthetic judgements are of the same kind as the qualities pointed to in undoubtedly aesthetic contexts, and that for our purposes we can defer to aestheticians on what exactly these qualities are.

While this simple content approach seems to me to get much right, it is neither without its problems (what exactly are aesthetic qualities? How do experiences get their intentionality?) nor without strong competition. The axiological and affect-based accounts also focus on important aspects of aesthetic experiences. We often take the objects of aesthetic appreciation to be

[56] Thanks to Milena Ivanova for pressing me on this.

valuable for their own sake and strive to experience this value. Thus we are willing to spend considerable resources on adding decoration to objects of use, revel in the delicate petals of a flower, or put other concerns aside while the beauty of a richly lit starry sky washes over us. And many aesthetic experiences are quite distinct in their phenomenological feel (though it's not so clear that it's possible to isolate an aspect of these experiences which is *unique* to aesthetic ones, and which therefore can be used to define aesthetic experiences (see Carroll (2002:147–53) and Beardsley (1982) for discussion). The in my eyes most promising version of the affect-based account is some form of attitudinal theory: what is special about aesthetic experiences is the *attitude* taken towards certain objects, or the way in which we engage with them (Stolnitz (1960), Cochrane (2021)). Such accounts are typically inspired by Kant's disinterested pleasure: we enjoy experiencing, say, Chopin's *Fantasie impromptu* for the experience alone, not for any downstream benefits (including practical, social, or cognitive benefits). Bence Nanay's 'distributed attention' is an empirically informed refinement of disinterestedness, which has it that "our attention is focused on one perceptual object, but it is distributed among a large number of this object's properties" (2016:13) in the kind of open, savouring way you may recognise from visits to the museum.[57] Notably, this does not preclude the aesthetic experience to also have cognitive or some other kind of value.

Tom Cochrane develops an account which sits between Carroll's affect-based and axiological approaches: the central feature of aesthetic experiences is that we actively engage with the object. We *value* it, where 'value' is treated as a verb. "When I value something, I engage with it, I take intentionally directed pleasure in it, or I act in a nurturing manner towards it" and "I experience the object of aesthetic value as *just good* in its own right" without reference to how it may be good *for me* (2021:14). (See chapter 4 for more on Cochrane's account.)

I do not think that anything I say in this thesis commits me to any specific account of aesthetic experiences. Whether such experiences are

[57] Note that this is meant to highlight a central feature of many aesthetic experiences, not give necessary and sufficient conditions for an experience to be aesthetic; Nanay does not think there is something all and only aesthetic experiences have in common (2016:12).

characterised by the value of the object, by the attitude we take towards it, by being about aesthetic qualities, or by some combination of the above, what is clear is that scientists report having aesthetic experiences in their scientific work (Vaidyanathan et al. (2023), Ivanova et al. (2024)). Aesthetic judgements are typically based on aesthetic experiences (as opposed to checking off a list of aesthetic qualities), and scientists apparent aesthetic judgements can in most cases be understood straightforwardly as an expression of their aesthetic experiences. Given the continuity of scientists' psychological states and processes *at* work and *off* work, their everyday competency with aesthetic vocabulary, and their presumably normal level of introspective access to their own experiences, the explanation which makes sense of the data in the simplest way is certainly that scientists report genuinely aesthetic experiences when they express, say, delight in a theory's beauty. That said, these are thorny issues and working out the metaphysics of aesthetic experiences requires much more work; luckily, this is not the place for it.

2.5 Conclusion

Theoretical virtues play a central role in abductive inferences as the criteria by which we judge which is the best theory and thus the one we should accept. In this chapter, I have given a fairly comprehensive overview of the nature of theoretical virtues, and looked closer at why we should count some of them as broadly aesthetic virtues. As discussed earlier in this chapter, defenders of IBE face the challenge of explaining why broadly aesthetic virtues make a theory more likely to be true, or otherwise epistemically successful. In other words, what is the epistemic value of these virtues? Answering that question is more difficult than it seems at first blush, as we see in the next chapter.

> [W]e could draw a more complex curve that passes exactly through those data points. The practice of science is to not do this. Even though a hypothesis with more adjustable parameters would fit the data better, scientists seem to be willing to sacrifice goodness-of-fit if there is a compensating gain in simplicity. The problem is to understand the rationale behind this behaviour.
> Aesthetics to one side,
> the fundamental issue is to understand what simplicity has to do with truth.
> – Malcolm Forster and Elliott Sober (1994:4–5)

CHAPTER 3:

EPISTEMIC VIRTUES

This chapter concerns the epistemicity of epistemic virtues: in virtue of what are these virtues *epistemic*? After discussing some problems with prominent truth-conducive accounts, I argue that theoretical virtues are *knowledge*-conducive. Given our larger concern – whether aesthetic virtues can be epistemically valuable – the details matter less than the fact that these virtues are in some way truth-directed. Several recent accounts of aesthetics in science take *aesthetic* virtues to be conducive to non-factive understanding, rather than truth or knowledge. I argue that plausible versions of this proposal nonetheless require a high enough level of truth for something to count as understanding that these versions do not challenge the truth-directedness of epistemic virtues.

In section 3.1, I discuss two accounts of epistemicity (i.e., the epistemic property of theoretical virtues) as truth-conduciveness and argue that they misclassify central epistemic virtues. In section 3.2, I discuss two factive alternatives to truth-conducive accounts and argue that we should take epistemic virtues to be conducive to knowledge. Section 3.3 discusses a radical departure from traditional accounts of epistemicity, according to which some virtues are conducive to *non-factive* understanding. I argue against this view; but more importantly, even if some virtues are quasi-epistemic (conducive to non-factive understanding), this would not provide a substitute for epistemic virtues. What I am interested in exploring here is whether aesthetic virtues can be epistemic –

whether they can help us understand what the world is like – and whether they also provide genuinely non-factive understanding is orthogonal to this question.

3.1. Traditional account of epistemic theoretical virtues

3.1.1. What are epistemic virtues?

Which theoretical virtues are epistemic? The clearest example of an epistemic virtue is empirical accuracy (fitting the data). Other uncontroversial epistemic virtues include being testable (what the theory states and entails is clear enough that one could in principle devise a test which could show it if the theory were false), that it makes novel predictions, that it accommodates already known facts,[58] that it is internally consistent, and that it fits what we know from other fields.

Many other theoretical virtues lack a clear epistemic rationale. These virtues are typically negatively defined as whatever virtues are left when we have accounted for the epistemic ones, typically taken to include pragmatic, political, ethical, aesthetic – and perhaps other – virtues. The division into epistemic virtues and the rest has more to do with philosophers' focus on epistemic virtues than with any kinship among 'the rest': there is no more similarity between political and aesthetic virtues than there is between pragmatic and ethical ones, and far less than there is between epistemic and aesthetic virtues. (In chapter 5 I argue that at least some aesthetic virtues are also epistemic.) Still, the focus on epistemic virtues is far from arbitrary: the primary goal(s) of science is (are) epistemic, so the virtues most central to theory evaluation are those that take scientific theories closer to attaining this goal. Although there is disagreement on which epistemic good is the main goal of science – contenders include knowledge (Bird (2022)), true belief (Khalifa (2017, 2020)), and understanding (Potochnik (2017)) – it is generally taken to involve truth in some way.[59]

---

[58] There is a large literature on whether successful novel predictions are epistemically more important than accommodating old data, and if so, why. We do not need to go into this debate here, but see Horwich (1982), Worrall (1989a), Lipton (2004: chapter 10), Schindler (2018: chapters 3 and 4) for discussion.

[59] Even constructive empiricists, as we saw in chapter 2, take truth *about a limited set of phenomena* to be the epistemic aim of science. Nothing here hinges on

So what are epistemic theoretical virtues? They are usually taken to be the truth-conducive features of a theory – we come back to what *that* means in a second. The net is sometimes cast wider, as Derek Turner and Ahmed AboHamad (2023) do when they say that "[w]e take it that epistemic [virtues] have to do with truth, evidence, justification, and understanding" (2023:54). I take it they would also count knowledge among these central epistemic concerns. While evidence, justification, understanding, and knowledge are commonly taken to be essentially tied to truth, some philosophers take it that understanding does *not* require truth. Thus, on their account, epistemic virtues could increase understanding without being truth-conducive. I discuss this view in section 3.3 below. But first, let's discuss the far more commonly held view that epistemic virtues are conducive to the truth of a theory.

3.1.2. Truth-conduciveness I

As mentioned, epistemic virtues are often glossed as the *truth-conducive* features of theories (e.g., Schindler (2022:545), Longino (1996:41), Kosso (1992:28)).[60] What does that mean? One proposal comes from Larry Laudan (2004), who takes epistemic virtues to be features which are either necessary or sufficient for the truth of a theory. In arguing against the epistemicity of various virtues, for example, he says that "[n]one of these rules [such as "explain different kinds of facts", i.e., wideness of scope] can have an epistemic rationale since it is neither

---

whether we adopt a realist or a constructive empiricist stance: if you do not like talk of the truth of theories *simpliciter*, feel free to restrict the following talk of truth to observable phenomena. This will not make a difference to the central arguments of this section, which concern the way epistemicity is related to truth (regardless of the scope of the truth-apt claims of a theory).

[60] Sometimes, epistemic virtues are instead taken to be truth-*indicative* – that is, symptoms or signals of truth (e.g., Bueno and Shalkowski (2015:2)). Often the labels are used interchangeably. A typical example is from Peter Kosso, who describes some qualities as "relevant to the likelihood of the theory being true. That is, … reliable indicators of truth" and then again as "truth-conducive" (1992:28). For our purposes it doesn't matter much which label we use: if having *F* makes the theory more likely to be true, then *F* will signal that the theory is more likely to be true; and if *F* is a symptom of truth then it tells us that something or other (though not necessarily *F* itself) has made the theory more likely to be true.

necessary nor sufficient for the truth of a statement that it exhibit any of these attributes" (2004:18).[61]

But this is too demanding. We look to the epistemic virtues as symptoms of truth, as reasons to believe that a theory is true in the absence of any direct access to the truth (or falsity) of the theory. If epistemic virtues are this closely tied to truth, their use in abductive reasoning is severely limited, if not epistemically useless: Laudan's account of epistemicity in effect asks us to infer that a theory is true just in case we already know that it's true.

It also misdescribes the virtues. Not even perfect empirical accuracy lives up to Laudan's criterion for epistemicity: it's neither necessary nor sufficient that a theory fits observational and experimental data for it to be true.[62] It's not necessary, because whether something is true clearly doesn't depend on whether the relevant observations have already been made. For example, in 2017, scientists behind the Event Horizon Telescope Collaboration gathered "the first visual evidence of the presence of [a supermassive] black hole" at the centre of the Milky Way galaxy.[63] That this black hole, Sagittarius A*, existed was equally true in 2016 as it was in 2017: we did not have to wait for the observations to be made for a theory postulating its existence to be true. So theories can be true without being empirically accurate.

---

[61] Kuhn (1977) objects to talking about rules or criteria for theory choice and urges us to think of them as values instead. "I am suggesting … that the criteria of choice … function not as rules, which determine choice, but as values, which influence it" (1977:362). While these labels are now often used interchangeably (as they are in this thesis) and while it is commonly acknowledged that theoretical virtues seldom if ever uniquely determine theory choice, it is perhaps revealing that Laudan talks of rules given his unusually strong interpretation of truth-conduciveness.

[62] Laudan's point is precisely that good theories are characterised by more than epistemic virtues, and so scientific reasoning is not wholly an epistemic matter (2004:17). For example, he says that "[f]ew would quarrel either with the notion that a theory is, all else being equal, better if it can explain or predict facts from different domains or if it can show its rivals to be limiting cases. This is, nonetheless, a form of argument *that has, and can have, no epistemic foundation*" (2004:17; italics inserted). This severely restricted account of epistemicity is too implausible to serve as a target for his argument.

[63] https://eventhorizontelescope.org/blog/astronomers-reveal-first-image-black-hole-heart-our-galaxy accessed June 27th, 2024.

Empirical accuracy is also not sufficient for the truth of a theory. There might be errors in the data, so a perfect fit between the theory's predictions and the data is not a guarantee of truth. For example, in 1859 the meticulous hobby astronomer Edmond Lescarbault made careful observations that seemed to confirm the existence of the hypothesised planet Vulcan, the existence of which was postulated by Urban Le Verrier to explain the way Mercury's orbit diverged from the orbit predicted by Newton's mechanics (Levenson (2015:73–8)). (This was after Le Verrier had used the very same method to successfully predict the presence of Neptune.) Lescarbault's observations, however, were faulty, and the postulated planet does not exist. The fact that the theory fitted the observational data was not sufficient for its truth: theories can be empirically accurate without being true. Furthermore, as is well known from the problem of underdetermination of theory by data, not even fitting all the *true* (empirical) data is sufficient for truth. Consider two theories that are equivalent in all the ways which can be tested empirically: no observation or experimental result could distinguish between these two theories. The classic example, due to van Fraassen (1980), is Newtonian mechanics and Newtonian mechanics plus the hypothesis that everything in the universe moves in the same direction at a uniform velocity. To the extent that these alternatives are incompatible, at most one of them can be true. So, since they fit the data equally well, fitting the data is not sufficient for truth. Indeed, this problem is often used to motivate the reliance on theoretical virtues beyond empirical accuracy (e.g., Tulodziecki (2025), Psillos (1999:171)): *how well* the theory explains the data, not just whether it fits it, matters to theory evaluation.[64]

Hence Laudan's (2004) account of what makes epistemic virtues epistemic will not do: it is too demanding to be a plausible candidate for epistemicity.

---

[64] There is a large literature on the problem of underdetermination, and not everyone agrees that it poses a real challenge (for discussion, see Laudan and Leplin (1991), van Fraassen (1980), and Psillos (1999)). See Tulodziecki (2025) for a recent book on theoretical virtues in relation to the problem of underdetermination. We do not need to go deeper into this issue here: even if theories are never underdetermined by the data, Laudan's account of epistemicity would still not get off the ground.

3.1.3. Truth-conduciveness II

According to what is probably the most widely accepted account of epistemic virtues, epistemic virtues are truth-conducive, and truth-conduciveness is taken to be a matter of increasing the probability that the theory is true. That is, an epistemically virtuous theory is not guaranteed to be true, but having those virtues makes it more likely to be true. This is of course compatible with the virtues being neither necessary nor sufficient for truth and so avoids the problems facing Laudan's (2004) account. However, as I show in this section, this too is an untenable characterisation of epistemic virtues.

This traditional view of the epistemicity of epistemic virtues is accepted by Schindler (2022:545), van Fraassen (1980:88), Kosso (1992:28), and Swinburne (1997:1), among others. A typical example of someone taking this account of epistemicity for granted is found in Alan Baker's *Stanford Encyclopaedia of Philosophy* article on simplicity, where he writes that "[j]ustifying an epistemic principle requires answering an epistemic question: why are parsimonious theories more likely to be true?" (2022: section 2). This increased-probability account seems to be what most people have in mind when talking about epistemic virtues, whether or not they make it explicit (e.g., Longino (1996:41)).

Taking truth-conduciveness to be a matter of increased probability makes good sense of many epistemic virtues. The more of a theory fits the data, for example, the more likely it is to be true. One way of arguing this point is by noting that each data point we compare the theory to is a little test of the theory and the more we expose the theory to potentially disconfirming data the more likely we are to discover a potential mismatch. If its predictions fail, then it is likely to be false; if its predictions are borne out, the best explanation of this fact is that the world is just as the theory says it is.[65] Of course, this depends on the quality of the data: fitting *noise* in the data does not ensure that the theory

[65] *Pace* Popper, who famously argued that a theory should 'stick its neck out' and risk refutation but equally famously argued that a theory can never be confirmed and so a theory that has withstood many tests is *not* better supported than one not yet tested. There is much more to be said here, but this is not the place for it: this thesis is built on inductive optimism, and if we are truly never justified in accepting an ampliative inference then everything I say here is moot.

matches up with the world. Also, the conflict may well be between the data and the auxiliary hypotheses rather than the theory. (Fitting the data only fallibly raises the probability of truth, so strictly speaking we should say that epistemic virtues *tend* to increase the probability that a theory is true. Such fallibility is taken for granted throughout – we are, after all, discussing human science and human cognition, where infallibility is simply not in the cards.) The same goes for external coherence, or fitting our best theories from other fields or other background knowledge, rather than observational or experimental data. Since the inconsistency of a theory tells us that at least some part of it is false, internal consistency makes the theory more likely to be true. And since an ad hoc theory may have been fudged by scientists trying to save their pet theory in the face of disconfirming evidence, we have some reason to expect a non-ad hoc theory to be more likely to be true. (There is more to be said about all these points, but the problem I want to focus on lies elsewhere, so let's press on.)

Thus there is something importantly right about the traditional definition of epistemic virtues as features of a theory which increase the probability that it is true, and it can account for the epistemicity of many epistemic virtues. Unfortunately, it cannot account for them all.

3.1.4. A dilemma for the traditional account

Although there is no canonical list of theoretical virtues nor a universally accepted way of drawing the line between epistemic and non-epistemic ones, there is a great deal of agreement on where to place some of the central theoretical virtues. Among the virtues that seems to clearly belong on the epistemic side of the divide, we find virtues that increase the theory's informativeness. Theories that are stronger, wider of scope, more precise, or otherwise have greater content than a weaker theory does are thereby epistemically better, because they answer more of our questions about what the world is like.

But this creates a problem for the traditional account of epistemic virtues: these ostensibly epistemic features *decrease* the probability that the theory having them is true, which is the opposite of what the traditional account requires of epistemic virtues. The more detailed the theory is (i.e., the more ways

in which it might be false), the less likely it is to be true. The same goes for wideness of scope: the more the theory says, the more ways it can be false, and the less likely it is to be true. This has long been recognised by philosophers of science. Peter Lipton, for instance, says that "[s]cientists also have a preference for theories with great content, even though that is in tension with high probability, since the more one says the more likely it is that what one says is false" (2004:116). And Karl Popper said that "if growth of knowledge means that we operate with theories of increasing content, it must also mean that we operate with theories of decreasing probability … Thus if our aim is the advancement or growth of knowledge, then a high probability … cannot possibly be our aim as well: *these two aims are incompatible*" (1987:174).

Taking epistemic virtues to increase the probability that a theory having them is true leads us to a dilemma: either some virtues standardly taken to be epistemic (strength, wideness of scope) are not epistemic virtues after all, or these virtues can be epistemically valuable despite not increasing the probability that the theory is true. The first option is not attractive: these virtues belong to the epistemic side of the divide because it clearly matters to the epistemic value of a theory that it can give us a rich and detailed account of the world. Giving good, informative answers to our questions is, after all, one of the main functions of scientific research. Strong (informative) and wide-scoped theories do this, and it is precisely because of this that they have epistemic value (this is reflected in scientific practice, as the quote by Lipton suggests).

So the second option must be the right one: increased probability is not the only source of epistemic value for theoretical virtues. The traditional account is at best incomplete.

3.2. Factive alternatives to truth-conduciveness

3.2.1. Alternative I: greater potential for truth

Instead of taking epistemicity to be a matter of how likely the theory is to be true, one could take it to be a matter of how much of the truth it states or entails. On this view, an epistemically virtuous theory has greater potential for truth: if true, the theory would give us more of the true descriptions of the world than a

less virtuous theory would. A stronger theory has a greater potential for truth than a less committed theory has, so this way of cashing out truth-conduciveness does not have a problem with accounting for why being more committed is epistemically valuable.

This account of epistemicity has something going for it. As mentioned earlier, many epistemic virtues – like empirical accuracy, internal and external coherence, and non-ad hocness – make the theory more likely to be true than a theory lacking these virtues. If we think of this in terms of a space of possible worlds where the *range* or *region* of a theory – the area in possibility space it covers – is all the possible worlds in which the theory is true, then virtues which increase the probability that the theory is true characterise theories that are more likely to include the actual world in its range. A theory is true if the actual world (or part of the actual world)[66] is in its range. (Assuming that we can arrange possible words by similarity, we can use proximity to the actual world as a measure of truthlikeness: the closer a theory is to the actual world, the closer it is to the truth. A false theory *a* can be better than another false theory *b* by being closer to the truth than *b* is, and *a* may even be better than a true theory *c* if *c* has an unduly large range.)[67] The strength of a theory is inversely related to the size of its range: the smaller the theory, the more committed it is to precisely how the world must be for it to be true. If the theory ranges only over the actual world, it gives *the truth, the whole truth, and nothing but the truth*, as the saying goes. The larger its range, the more likely it is to be true, but it will also be less informative and specific in its predictions. We can increase the chance that a theory is true at the cost of weakening it, i.e. taking it further away from the truth. So the account of epistemicity as greater potential for truth recognises that increasing the likelihood of a theory by making it less informative can incur a significant epistemic loss. On this view, epistemicity is roughly a matter of how much of the total, true description of the world the theory states or entails.[68]

---

[66] Thanks to James Read for pointing this out.

[67] As with all accounts of truthlikeness, this account is controversial. See Oddie and Cevolani (2022) for discussion and further references.

[68] This is not meant as a complete account of a theory's potential for truth. It has some clear problems: for example, a theory which ranges over a small but disunified area (say, by being the union of two separated small regions) is not

It's a strength of this account that it does not conceal the tension between increasing the probability that the theory is true and increasing how much truth we can get from it, or, in more general terms, the tension between getting things right and taking the chances necessary to potentially get even more things right. This reflects a tension common to most epistemic endeavours, not just scientific ones. We regularly trade off the probability that something is true with informativeness, trying to find an optimal balance between what we already know or can easily find out (e.g., there are berries in this field; how long will it take for this stone to reach the ground if thrown from this tower?) and the potential for expanding our knowledge (will the next field provide a more abundant supply of berries? What is the value of the earth's gravitational constant?). For example, in a series of studies, Ilan Yaniv and Dean Foster (1995, 1997) found that people prefer to sacrifice some accuracy for a comparative gain in informativeness when making or evaluating judgements under uncertainty (see also Goldsmith and Koriat (1999), Proust (2013:315)).[69] The more uncertainty we're faced with, the more important it becomes to find a good trade-off; and scientists, especially in the early stages of theory building, face a high level of uncertainty.

The tension between accuracy and informativeness can be seen as a special case of the 'exploration-exploitation trade-off': the need to find a good

---

epistemic virtuous. Or, more cautiously, this kind of gerrymandering counts against its virtuousness. Think of it is way: if you want to classify a plant you have never seen before, it isn't much help to hear that it's either *Sphagnum Capillifolium* or *Deschampsia alpina*, even though this excludes all other options. It narrows down the options quite a bit: you'll know that the plant isn't a pine tree, for example. But it is more helpful to hear that it is one or another *Sphagna* – at least then you will know that it's a moss and that it grows in bogs. That is, here the theory which excludes larger areas of possibility space but does so in a disjunctive manner is less epistemically virtuous than one which holds open a larger space in the same neighbourhood. I'm not sure how deep this problem runs, as it seems like a problem which can, and should, be solved by insisting that the theory should *also* be simple, unified, and elegant. One criterion one could invoke to avoid this problem is to minimise the maximum distance between any two points in the theory's range (thanks to Tim Williamson for the suggestion). Going further into this issue is beyond the scope of this thesis.

[69] This trade-off is paralleled in the pragmatics of speech. Following Grice (1975), it is there known as a trade-off between quality (accuracy) and quantity (informativeness), two out of four of his 'conversational maxims'.

balance between exploiting familiar, good options (going back to the same field in search of berries; settling down with your acceptable partner) and exploring potentially better options (perhaps you'll find more berries in the next field; perhaps your next date would be the partner of your dreams). This trade-off is paralleled in all situations where the need to reduce false negatives (here: excluding truths) is balanced against the need to reduce false positives (being mistaken).[70] In other words, this account mirrors a central aspect of our epistemic lives.

However, this account of epistemicity also faces significant problems. In particular, it falsely classifies lucky guesses as epistemically virtuous. Had it only mattered to us that our theories are true, it would not be a problem if we were to arrive at true theories through luck. But, clearly, lucky guessing does not make for good theorising. Consider the following example:

The opponents of continental drift accused Alfred Wegener of cherry-picking his evidence, misinterpreting data, and relying on second-hand evidence (which the geological community at the time considered epistemically risky).[71] In short, they accused him of doing bad science. Parts of this criticism were personal attacks based on a dislike of the brash style in which Wegener presented his results (fuelled by ignorance of continental drift and the displeasure of seeing their own cherished theories attacked). But Wegener was in fact less careful in handling evidence than he ought to have been, especially in his tendency to draw generalisations far beyond what the evidence supported. Now suppose, contrary to fact, that Wegener's methods were as unreliable as some of his critics argued: cherry-picking evidence, massaging the data to fit with his view, ignoring

[70] It seems that curiosity plays a central role in helping us achieve an optimal trade-off between accuracy and informativeness. It comes as no surprise that accurate representations are useful, but so is the drive to satisfy one's curiosity – i.e., seeking out new knowledge – as Jennifer Nagel (forthcoming) shows. She argues that "creatures who are curious benefit from an interaction between their reactive prediction-error correction processes and the active surprise-seeking force of their curiosity; this internally adversarial interaction accelerates knowledge gain in ways that are very helpful for biological agents in environments like ours" (forthcoming, p. 3). I.e., the tension between (the processes we rely on for achieving) accuracy and informativeness help us gain new knowledge in the kinds of uncertain environments we find ourselves in.

[71] See next chapter for details.

defeaters in the form of counterarguments and counterevidence presented by his (let's suppose) more diligent colleagues, drawing hasty generalisations, and so on. Let's further suppose that he nonetheless arrived at the same theory. Continental drift is, in fact, largely true. But would it have been epistemically virtuous, in this counterfactual scenario? Of course not! Just happening to fit the data is not epistemic virtuous (this is also connected to why ad hocness is a vice, as I discuss in the next section). Truth, then, is not enough.

3.2.2. Alternative II: knowledge-conduciveness

In the foregoing sections, we have considered ways of expanding and adjusting an account of epistemicity as truth-conduciveness, but this has not been entirely successful. Let us therefore consider an alternative basis for epistemicity which is also truth-directed but which avoids the problem facing the other accounts we have considered so far (see section 3.3 below for a discussion of a very different alternative).

The proposal is that epistemic virtues are conducive to knowledge, not merely truth. This gives us a unified account of the ways in which a virtue can be epistemic: we can gain more knowledge about x both by making a theory of x more likely to be true and by broadening or deepening the knowledge it would give us, if true. For example, Arthur Holmes' addition to the theory of continental drift – explaining how convection currents in the asthenosphere could partially drive the drift – deepened the knowledge that drift theory could give us. (In general, explanation of causes deepens knowledge.) Holmes' version was not more likely to be true than Wegener's (because the probability of a conjunction is equal to or less than the probability of its conjuncts (i.e., prob(T&h) ≤ prob(T)): it added an extra hypothesis to the theory and so could not possibly be more likely to be true than the original theory. Hence the increased epistemic value of Holmes' version of continental drift cannot be due to truth-conduciveness. But it can easily be accounted for in terms of giving us more knowledge.

Similarly, discovering that nonhuman primates, rodents, and corvids use tools broadened our knowledge of instrumental rationality – it's the kind of thing possibly had by more than humans (e.g., Camp and Shupe (2017)). That's

certainly epistemically valuable, and this epistemic value is easily explained as knowledge gain: we now know more about instrumental rationality. Our knowledge can be increased both by a theory that is likely to be true (assuming we have arrived at it in a reliable or otherwise justificatory way) and when we gain more evidence for it (as that would make us better justified in believing it, which could, if the theory is true, take us from merely believing a true theory to knowing it).

And, lastly, taking epistemicity to be knowledge-conduciveness avoids the problem of merely lucky true beliefs. You can of course not gain knowledge by happening to gain a true belief by luck: you must have arrived at the belief in the right way. (Whether the knowledge-producing way of forming beliefs is following a reliable method, by having access to reasons in favour of the belief, or something else, is not something we need to take a stand on here.)[72] If it only matters to us that our theories are probably true, or that they have greater potential for truth, then it would not be a problem if we arrive at this through sheer luck. But, clearly, lucky truth is not what we're after. Since we cannot gain knowledge by lucky truth, the knowledge-conducive account gives the right verdict in this case.

As Alexander Bird says about an example similar to the Wegener example I gave in the previous section:

> the truth of [a theory] T is not enough. Clearly what is lacking is adequate justification. Good science requires not just that our theories be true but

[72] For epistemic reliabilism, a prominent 'externalist' account, according to which knowledge is true beliefs produced by a reliable belief-forming process, see Goldman (1976) and Nagel (2016); for a different kind of externalism according to which knowledge is a propositional attitude had only to truths, see Williamson (2000); for epistemic internalism, according to which knowledge requires introspective access to justifying reason(s), see BonJour (1980) and Feldman and Conee (2001). While everyone agrees that knowledge isn't just justified true belief (Gettier (1963), Dharmottara (770; see Stoltz (2007)), what it is (or whether we can give an informative analysis of it: see Williamson (2020)) is highly disputed. Perhaps knowledge requires that the true belief is safe, i.e. that it couldn't easily have been false, or that it is sensitive, i.e. that if the proposition were false, one wouldn't believe it (Nozick (1988)). Deciding on these fundamental issues about the nature of knowledge is beyond the scope of this thesis: we have a good everyday grasp on what knowledge is and can leave it to epistemologists to work out the details.

> that we have adequate grounds for believing them, that our methods are reliable, and that they do not rest on self-deception or other irrational causes. The best explanation for those additional conditions is that they are required for science to [produce] knowledge. (2022:20)

Although Bird is here talking about knowledge as the aim of science, which is (related to but) not the same as epistemic virtues of theories increasing knowledge,[73] the same considerations hold for our counterfactual scenario. Wegener's sloppy methods would not have resulted in an epistemically virtuous theory, and the best explanation of that is because even though it happened to be true it, could not contribute to our knowledge. The truth-conduciveness account will have to draw on additional resources to explain this fact, whereas it follows directly from the knowledge-conduciveness account. Those accepting truth-conduciveness are free to say that the best way of arriving at truth is by using reliable methods, only accepting claims we have evidence for, and so on, so epistemic virtues are the features which increase the likelihood of giving us justified, true beliefs. Yes, but whereas this desirable connection between epistemicity and justification falls out of the knowledge-conduciveness account, in the truth-conduciveness account it must be written in by hand.

One objection one may have to the foregoing is that this draws conclusions about the epistemicity of theories from the epistemically responsible way of arriving at theories: this blends together the context of justification (whether the theory is epistemically good) with the context of discovery (how we arrived at it). However, we routinely exclude merely accidentally successful theories from being epistemically valuable (though often implicitly so). Indeed, one of the reasons why we value precision is that the more precise the theory's predictions are, the less likely they are to accidentally fit the evidence. A noncommittal theory may get it right more often than a precise theory will, but there are also many more ways in which its predictions can happen to be true for reasons unrelated to the theoretical framework. If your horoscope tells you that this week will "bring unexpected yet exciting developments in love, finances,

[73] One of the virtues of taking epistemicity to be knowledge-conduciveness is that it dovetails nicely with taking the aim of science to be knowledge, not (mere) truth. See Bird (2022) for a good defence of this view, and the discussion of non-factive understanding below for references to a currently popular alternative.

or self-expression"[74] and there is an unexpected yet exciting development in your life, the horoscope gets this right even if the exciting development has nothing to do with "supportive trines to your planetary ruler, Venus, and Uranus in your sign" (whatever that means). What astrology gets right it gets right by luck, and astrologers maximise their luck by being noncommittal. On the other hand, it is much less likely that a precise prediction will be successful just by luck. Getting (say) the spreading rate of tectonic plates right is impressive because one is unlikely to just happen to pick the right numbers unless what one bases the prediction on is mostly true. (This is of course the guiding thought of the 'no miracles argument' for scientific realism: it would be incredible if science were as successful as it is without mostly getting it right. See chapter 2 (p. 22, fn. 11).)

There is something of the same rejection of accidental accuracy in our distaste for ad hoc theories: although such theories fit the evidence well (indeed, often better than the non-ad hoc alternatives since they have been designed to fit those pieces of evidence) they are not arrived at in a knowledge-producing manner. In this case, it's not luck but intentional adjustment which is the culprit, but neither arriving at a 'just so' story by luck nor forcing our luck helps knowledge production.

So, to sum up, here I take epistemic virtues to be the features of theories which are conducive to knowledge. This allows us to account for the epistemic value of epistemic virtues in a unified manner, and it avoids the problem of luck, so knowledge-conduciveness avoids the problems with the other accounts we have considered thus far. Let me be clear that this is not meant to be the last word on what makes epistemic virtues epistemic. We only need a good enough grip on epistemicity to make sense of the subsequent discussion of whether aesthetic virtues can have epistemic value: nailing down a definition of epistemic virtues which goes beyond that also goes beyond the scope of this thesis and will be left for future work. That said, taking epistemic virtues to be knowledge-conducive captures the central features of epistemicity and correctly categorises strength and scope as epistemic virtues, which should make us confident that it is at least on the right track.

---

[74] Weekly Overview for Taurus - Astrology.com accessed 02.07.2025.

My response to the problems facing accounts of epistemicity as truth-conduciveness is to add something: epistemic virtues are conducive to knowledge, not mere truth. But perhaps we should instead subtract something: perhaps truth is *too strong*, rather than too weak, to account for all epistemically valuable theoretical virtues? In recent years, many philosophers of aesthetics in science have taken aesthetic virtues to have epistemic value insofar as they are conducive to non-factive understanding. In the rest of this chapter, I discuss and ultimately argue against this view. Some of the non-factivist accounts are motivated by the widespread use of abstractions and idealisations in science, and particularly in models. So before we discuss understanding-based accounts, let us take a brief detour to consider whether the present account of epistemicity as knowledge-conduciveness also applies to scientific models.

3.2.3. An aside: models, idealisations, and abstractions

Does this way of understanding epistemic virtues also apply to models? Models typically make gross simplifications in the form of idealisations and abstractions, thus openly detracting from truth. "An abstract description of a system leaves a lot out. But it is not intended to say things that are literally false. An idealized description of a system is a description that fictionalizes in the service of simplification," as Godfrey-Smith (2009:48) puts it. In the first case, we leave out details needed to paint a complete picture (e.g., models of the gravitational attraction of the moon on the earth which do not take other celestial bodies into account); in the second case we introduce more-or-less well-chosen falsehoods (e.g., frictionless planes or perfectly rational economic agents). Either way, we move away from what's literally true. As discussed below, some philosophers of science think that the widespread use of idealisations and abstractions in science shows that epistemic virtues cannot be truth-conducive; the same argument would show that they cannot be knowledge-conducive.

We come back to that below. The current question is whether models pose a particular problem for taking epistemic virtues to be truth-directed (whether by being knowledge or truth-conducive). I think not: first, useful falsifications are not particular to modelling – though they might be more pronounced in models than in theories – and second, because these falsifications

are used to get at (approximately) true representations (see section 3.3.3. for more on this). Either truth is too far removed from even our best theories and models for science to successfully aim at truth or anything entailing it, like knowledge, and thus for the epistemic virtues to be truth-directed, or it is not. In either case, an account of the epistemic virtues of models stands and falls together with an account of the epistemic virtues of theories.

3.3. Non-factive alternatives to truth-conduciveness

3.3.1. Non-factive understanding

In recent years, a number of philosophers of science have argued that (some) epistemic virtues are conducive to understanding but not truth – and, of particular interest to us, that *aesthetic* virtues are epistemically valuable by being understanding-conducive (e.g., Elgin (2020), Ivanova (2020), Arcangeli and Dokic (2020), Meynell (2020), Kosso (2007)). Non-factive understanding is so different from how epistemicity is normally understood that it will be useful to mark this difference explicitly. Let's call virtues that are conducive to non-factive understanding and not truth, knowledge, or factive understanding 'quasi-epistemic' virtues. As mentioned, this provides another alternative to taking all epistemic virtues to increase a theory's likely truth. If, that is, understanding is taken to be non-factive: if understanding entails truth, then understanding-conduciveness will not provide the desired alternative to truth-conduciveness.

In this section, we discuss whether aesthetic virtues could be conducive to non-factive understanding rather than truth – that is, be quasi-epistemic rather than epistemic. For ease of presentation, I assume here that aesthetic virtues can be of (quasi-)epistemic value: this is discussed at length in chapter 5 below.

3.3.2. Are aesthetic virtues conducive to non-factive understanding?

Why think that aesthetic virtues are quasi-epistemic? Prominent understanding-based accounts focused on aesthetics in science are motivated by the uneasy combination of i) historical cases where aesthetic pleasure and truth diverged with ii) scientists' rather frequent claims that beauty can be a guide to truth (and historical cases seemingly vindicating these claims) (Ivanova (2020), Elgin

(2020)). As Milena Ivanova says, "we can find plenty of examples to base an inductive argument on, some in favour of the link between beauty and truth and some against it" (2020:95). Faced with this inconclusive track record, instead of trying to show how aesthetic virtues can be conducive to discovering the truth, Ivanova wants to sidestep the issue by arguing that aesthetic virtues can be conducive to non-factive understanding. Since, on her view, non-factive understanding is one of the epistemic aims of science (2020:96), this allows aesthetic virtues to play a significant epistemic role without requiring us to postulate an implausibly direct link between beauty and truth.

According to Ivanova, understanding doesn't "require true theories, but rather an ability to manipulate and use a theory in a certain theoretical domain" (2020:98). It is not like propositional knowledge but rather "a skill or ability to grasp how the facts fit together" (2020:97–8). You understand a theory when you can skilfully work with it, apply it to different situations, and 'see' how various phenomena are interrelated.

It is difficult to state what non-factive understanding is without relying on factive constructions, i.e. verbs which presupposes that the compliment clause is true.[75] An example of a factive construction is 'see': you cannot see what is not there. (Compare the infelicitousness of "I saw that Monika walked by, but I was mistaken": here one would need to say "*I thought* I saw that Monika walked by, but I was mistaken".) 'Know' is another factive verb: you cannot know what isn't true. Even assuming for the sake of argument that 'understanding' is *not* factive, 'how the facts fit together' certainly is. (Compare the infelicitousness of "The facts fit together in the way described by Milena, but she is wrong about how the facts fit together".) By using factive constructions when stating their accounts, non-factivists conceal how radical such accounts are. Or perhaps this just reveals that these accounts are not genuinely non-factive after all: as we'll see in section 3.3.3, self-proclaimed non-factivists such as Catherine Elgin and Angela Potochnik, whom Ivanova refers to, do not in the end deny that truth

[75] As this presupposition is not cancellable, it's perhaps better to say that the inference from (say) 'she saw that *p*' to '*p*' is deductively valid (Williamson (2000:34–5)).

plays a significant role in understanding. But let's set this aside for the moment and get back to Ivanova (2020).

According to this view, instead of tracking the truth of a theory aesthetic virtues "offer us understanding of relations between the phenomena" (2020:99), thereby allegedly giving us the skill which enables us to manipulate the theory without requiring what the theory says to be true.

But this goes too fast. You can of course understand a false theory: after talking to creationists for a while you could get how the 'facts', according to their view, fit together. But the primary object of scientific understanding is the phenomenon the theory is about, not the theory itself. Understanding creationism doesn't help you understand the world that creationism attempts to explain (i.e., the facts): understanding a false theory and understanding the world through a false theory are significantly different. Skilfully manipulating a theory completely disconnected from the actual world does not have the epistemic benefits that quasi-epistemic virtues are supposed to have, namely to offer us 'an understanding of the relation between the phenomena' or give us 'a grasp of how the facts fit together'; it is more akin to engaging with a science fictional narrative. Ivanova moves freely between *understanding a theory* and *understanding the world through the theory*, but they are not the same. Let's consider this in light of her example of Newtonian mechanics:

> One's ability to use Newtonian mechanics to calculate the gravitational pull of the moon requires an understanding of how to employ Newton's laws together with a set of initial conditions in order to derive an answer. If truth were a presupposition for having understanding of gravitational effects, one would lack understanding in this case, given that the theory in play is strictly speaking false. (2020:98)

One's ability to use Newtonian mechanics to *try to* calculate gravitational effects does not require it to contain any truths (I can also try to categorise people according to their Hogwarts houses; it does not mean I will succeed). But to "calculate the gravitational pull of the moon [on the earth]" presupposes that we get the right number. Saying that we are using Newtonian mechanics to do so creates a context where the expected and appropriate value is approximate, rather than one which gets it right to the last decimals (not the least because we're ignoring gravitational effects of other objects in this example). The

presupposition is still that the value is accurate, as is clear if we compare it to a theory which is not only strictly speaking false (i.e., where much of what the theory states is nonetheless true) but simply false. "James calculated the gravitational pull of the earth on the moon using numerology" is a cleaner example of whether understanding is non-factive, since there is no truth content to numerology which can obscure the issue. And here, I think, it is clear that being able to "use and manipulate" numerology will not give us an "understanding of [the gravitational] relations between the [moon and the earth]".

This case is related to another common way of motivating non-factive accounts of understanding. The idea is that many, if not most, scientific theories of the past were false and yet scientists of the past did have some understanding. So it must be possible to gain understanding in lieu of truth. Henk de Regt, who has written extensively on understanding, puts it this way:

> many scientific theories from the past were quite successful at the empirical level: they could describe and predict a host of observable phenomena in their domain. But most of these theories were later rejected and were replaced with new theories that describe and predict the phenomena even better. These new theories are often in direct contradiction with the old ones, which means that they cannot both be true. So, despite their empirical success, the old theories turned out to be false after all. Does that mean that the scientists who developed and defended these theories lacked *any* understanding of the phenomena that they could describe and predict so well with their theories? I think that this conclusion does not do them justice. (2024:26; italics inserted)

Note that de Regt takes the alternative, factive accounts of understanding to imply that the reliance of past people on false theories meant that they had no understanding *at all*. If that is the contrasting view, one can understand why some are drawn to non-factive accounts of understanding. But, again, this goes too fast. There is no direct line from false theories to *no understanding at all*. People adhering to Aristotelian biology may have known and understood a great deal about, say, the life cycle of cats and the growing patterns of tomato plants. But their (limited) understanding, and the theory's (limited) empirical success, is based on what they got right, not what they got wrong. A theory is no monolith: it contains many hypotheses, some of which can be true even if the theory overall

is false. And careful observations may give you much knowledge even if the theory you rely on to explain the observations is itself false: you may know *how* tomato plants grow, even if you don't know why (whether it's *telos* or gene expression that drives the growth). Pre-modern theories that were able to successfully predict solar eclipses did not rely on their mistaken views to arrive at their predictions, like Anaximander's hypothesis that the sun is but a hole in the heavenly shell through which primordial light shines through (Kutrovátz (2024:2)). Instead, they relied on correctly observed regularities. To the extent that they made successful predictions, they relied on at least approximately true inferences from at least approximately accurate observations. And even when the theory is completely false, in many cases there are significant pieces of background knowledge one can rely on: one's understanding of how to make tomatoes grow can be based on one's general knowledge about plants, even if one also holds a false belief about the *telos* of the tomato plant. So saying that because people of the past relied on false theories they had no understanding at all is wrong on at least two accounts: first, their understanding may have been based on the theory's true hypotheses, and second, it may have been based on their background knowledge and accurate observations.[76]

In other words, when spelt out in slightly more detail than de Regt provides, it becomes clear that factive accounts can account for the ways in which understanding and adherence to false theories can co-occur. This undercuts a major motivation for non-factivity: we do not need to posit non-factive understanding to account for these phenomena, but can give a perfectly reasonable explanation in terms of (factive) understanding or knowledge.

Given that the default position regarding understanding is that understanding is factive (e.g., Sliwa (2015), Hills (2016), Lipton (2004), and Baumberger et al. (2017): see Hannon (2021) for discussion), by undercutting

[76] This understanding will be limited or shallow, precisely because their adherence to a false theory, but I take it that de Regt and others would concede that someone accepting (say) Aristotelian biology has a poorer understanding than someone who accepts Darwinian biology. Though if that's the case, wouldn't it undermine the non-factive accounts? After all, the salient difference between the two theories which presumably grounds the difference in levels of understanding is that one is mostly false and the other is mostly true. See the next section for discussion.

the motivations for postulating non-factive understanding we shift the burden of proof to the non-factivist.

However, I appreciate that the foregoing will not suffice to convince a non-factivist that she is wrong. So in the next two sections, I argue that non-factive accounts of understanding do not provide the desired alternative to taking the epistemic value of aesthetic virtues to be truth-directed.

To provide a real alternative to truth-directed accounts by showing that aesthetic virtues are (at best) quasi-epistemic, one of the following options must be true (cf. Kvanvig (2018:697–8)):

1. Understanding does not entail truth, i.e., it is non-factive. As we have seen, this is the option Ivanova favours ("understanding can be achieved in the absence of truth" (2020:86)).
2. Understanding consists of 'a skill or ability' which neither reduces to know-how nor is about access to further facts: instead, understanding is to 'grasp' conceptual relations over-and-above factual relations.

I discuss these options in turn in the following sections.

3.3.3. Option 1: genuinely non-factive understanding

Angela Potochnik (2017) argues that the fact that idealisations are central to the success of science shows that truth cannot be the epistemic aim of science, since idealisations are by definition partially false. She argues that since falsehoods can contribute to the epistemic success of science although falsehoods can't contribute to the truth of a theory or model, truth is not necessary for epistemic success (2017:93). Instead, she claims, understanding is one of the epistemic aims of science, and "idealizations positively contribute to generating understanding by revealing causal patterns and thereby enabling insights about these patterns that would otherwise be inaccessible to us" (2017:95).[77] Importantly, they do so because they are partially false, not in spite of it, because the divergence from truth can make it easier to see patterns and causal

[77] 'Reveal' is a factive verb: you cannot reveal what isn't there. As noted above, it is not clear whether one can even state a genuinely non-factive account coherently. I argue below that these accounts are *not* genuinely non-factive.

dependencies that a more accurate representation would drown in detail. This is what Catherine Elgin (e.g., 2019) calls 'felicitous falsehoods': representations that are false but nonetheless help us understand the phenomenon represented.

However, as Michael Hannon rightly notes, that felicitous falsehoods are useful *routes* to understanding does not show that what is understood is false (2021:278). Potochnik and Elgin argue convincingly that partial falsehoods in the form of idealisations and abstractions are 1) widespread in science and 2) partially responsible for its epistemic success, but this may simply be because they help us arrive at literally or approximately true beliefs. For example, the Lotka-Volterra model of predator-prey relationships falsely represents predators as insatiable and the two populations as insulated from any influences other than that of each other, but because it abstracts away from irrelevant features it helps us form the true belief that the population sizes of predators and prey directly influence each other and gives us an approximately true picture of how they do so.[78]

Potochnik says that "[a]chieving genuine understanding requires *some* type and degree of accuracy in virtue of the requirement of epistemic acceptability, but this requirement is weaker and variable in a way that truth and even approximate truth are not" (2017:117). But on the contrary, the role of truth in her account of epistemic acceptability should be acceptable to those thinking that truth is the aim of science (or that it follows from whatever is the aim of science, such as knowledge or factive understanding) and that epistemic virtues are truth-directed. Here's Potochnik: "A posit is epistemically acceptable when its divergence from truth is insignificant, taking into account (a) the posit's role in the representation and (b) the epistemic purpose to which that representation is put" (2017:100; italics inserted). Insignificant divergence from

[78] We find similar clear examples of how felicitous falsehoods enable understanding and communication in art. Artists regularly distort representations to help the observer focus on the central features of the phenomenon in question, as in Edvard Munch's painting *Jealousy* (1907) where the green face and desperate, round eyes of the man facing the viewer conveys the sickly feeling of jealousy or Ada Palmer's *Terra Ignota* series which, precisely because it is set in a science fictional world whose worldbuilding allows us to sidestep many of our ingrained assumptions about societal structure, helps us focus on the injustice of building an utopia by sacrificing a few people.

truth is clearly acceptable also if you think science is after the literal truth about things. It also seems right that whether the model's divergence from truth is insignificant or not depends partially on what we try to do with it. If we used the Lotka-Volterra model to make predictions about the influence of climate change on predator-prey relationships, its divergence from how the world is (i.e. not taking into account any influences on either predator or prey except for their population sizes) would be significant and make it unsuitable for its purpose. (What exactly the epistemic purpose demands is surely a question best left to the scientist themselves to decide.)

Similarly, Elgin says "[t]o accept that *p* … is to take it that *p*'s diversions from truth, if any, is negligible. In that case, *p* is true enough.… A sentence can be true enough in some contexts but not in others" (2019:390). If the divergence from truth is negligible and insignificant, I see no problem with so-called non-factive accounts of understanding – beyond the misleading label. It also seems right that the level of falsity tolerated depends on the context (or its 'epistemic purpose', as Potochnik puts it: I take it that this is the relevant feature of the context also in Elgin's account).

So it is not by denying that truth plays a significant role in understanding that an account of epistemic virtues as conducive to understanding is an alternative to truth-directed accounts, for, as we have seen, they do not claim this. One could develop an account of understanding that is genuinely non-factive, according to which, say, you can understand that the butler did it even though it was the professor who committed the murder. The sheer implausibility of such a view together with the fact that even self-proclaimed non-factive accounts of understanding don't go this far should be enough for us to set this option aside.

3.3.4. Option 2: grasping conceptual relations

Let's consider option (2), i.e., that understanding partially consists in "a skill or ability to grasp how the facts fit together" (Ivanova (2020:97–8)) which does not reduce to knowledge or anything else entailing truth, and that it is this feature of understanding aesthetic virtues track. Note that this skill or ability cannot be about accessing further facts if the point is to avoid the connection to truth,

since facts are true propositions. In particular, it cannot involve facts about the relations between various phenomena, facts about how the facts fit together. So how should we understand this ability to 'grasp' relational facts?

In a survey paper of contemporary accounts of understanding, Michael Hannon notes that "[t]heorists have tended to use the term ['grasping'] in an intuitive and largely metaphorical way; for instance, when one 'sees" how various bits of information 'hang together'" (2021:280). Perhaps the most promising way of unpacking the metaphor is Alison Hills' (2016) account of cognitive control: that you can actively use the information in drawing inferences, providing explanations, or otherwise manipulate it. For example, if you understand why tree rings form[79] you will be able to explain it in your own words (it reflects the trees' cyclic growth pattern), relate it to other phenomena, and draw counterfactual inferences (if our nails grew in layers, they would have a similar pattern). On this view, understanding is an ability to manipulate facts, which is supposed to go beyond the abilities which follow from knowledge (typically, if you know something you are also able to explain it, draw inferences from it, and otherwise manipulate it, so exactly how understanding-as-manipulation is supposed to differ from knowledge must be spelt out).

Peter Kosso, who argues that understanding is the aim of science and that aesthetic virtues contribute to achieving it (see his (2007) and (2002)), says that understanding is 'apprehending' relations among a coherent network of *x*'s (2007:179). It is unclear what this *x* is: he mentions theories (2007:174), facts (2007:187), and conceptual relations (2007:179). This apprehension is supposedly not knowledge of facts about the coherence relations but a 'sense' of the conceptual relations: "knowledge of many facts does not amount to understanding unless one also has a sense of how the facts fit together" (2007:173).

On the contrary, I think the correct thing to say is that this 'sense' is further knowledge: you don't understand a body of information unless you *know*

[79] This is of course also a factive construction: understanding why tree rings form presupposes that they form in the way understood. Again, the way non-factive accounts are typically presented may not be compatible with genuine non-factivism.

how the pieces fit together. (See the following paragraph.) Kosso's arguments that knowledge and understanding are distinct states are unconvincing. For instance, his argument (2007:175) that multiple choice exams test knowledge but not understanding is far from conclusive. Multiple choice tests can test lucky guesses, true beliefs, knowledge, and understanding, and a successful test does not necessarily discriminate between these states;[80] and these tests are also not designed to test knowledge-why or understanding-why. His mosaic metaphor (2007:177), that one cannot 'see the bigger picture' regardless of how much knowledge one has about individual mosaic tiles or all of the tiles considered individually, is not an argument for this 'seeing' being distinct from *knowing* the bigger picture.

The claim that grasping is an ability to use or manipulate what it is that you understand is typically combined with the claim that understanding is primarily a matter of grasping how various facts are related to each other, rather than a matter of grasping single facts. For example, Jonathan Kvanvig (2018) says that "[u]nderstanding seems to involve conceptual and explanatory connections between various items of information that are seen or grasped by the person in question" (2018:697). Understanding is here taken to be some sort of first-hand appreciation of how a body of information is interrelated: it's what you get when you yourself fully appreciate 'conceptual and explanatory connections'. And Kosso says that "[t]he achievement of understanding is in apprehending the connections between theories and the global coherence among concepts" (2007:179).

Now, one could simply say that this is *knowledge* of explanations and conceptual relations. As Hannon says, "when one 'grasps' a correct explanation for why an event occurs, perhaps one simply knows the correct explanation" (2021:281). The distinction Kvanvig and Kosso draw seems to be between different objects of, or perhaps degrees of, knowledge rather than between knowledge and something else: it's a distinction between

[80] Though a well-designed test can be good at discriminating between lucky guesses and substantial epistemic achievements; say, if it presents new scenarios, forcing the test subject to extrapolate from what they already know. However, this does not show that understanding is distinct from knowledge.

understanding/knowing a particular fact and understanding/knowing a set of interrelated facts and how they are interrelated.[81] The Kvanvig/Kosso position assumes that the primary unit of understanding is a network of propositions, not a single proposition (in the jargon, that understanding is objectual and not propositional).[82] Berit Brogaard (2005) argues convincingly that the reason why understanding seems more valuable than knowledge in Kvanvig (2003) is that he takes understanding to be objectual and knowledge to be propositional, and it is more valuable to have a systematic body of knowledge than to know unconnected facts. That is, the important difference is between the two units of understanding/knowledge, not between understanding and knowledge (Brogaard argues further that understanding reduces to knowledge). See also Elgin: "If the primary unit of understanding is the proposition, then the difference between knowledge and understanding seems slight" (quoted in Hannon (2021:283)).

Be that as it may, I'll grant that *if* understanding is what Kosso and Kvanvig takes it to be, then aesthetic evaluations in science could be (quasi-)epistemically useful for very different reasons from those I argue for in this thesis. If understanding is an ability to manipulate conceptual relations distinct from knowledge, and aesthetic virtues track this ability, then aesthetic virtues might be (quasi-)epistemically valuable without being truth-directed. This would make quasi-epistemic (non-factive understanding-directed) virtues radically different from epistemic (truth-directed) virtues, to the extent that the former is no substitute for the latter. Insofar as those who argue for understanding-based accounts of aesthetic virtues are not in the business of changing how we think about uncontroversial epistemic virtues such as empirical accuracy, the claim they should make is that there is another, additional epistemic good – non-factive understanding – which aesthetic virtues can help us attain. So we can still ask the question: are aesthetic virtues epistemic, regardless of whether they are *also* quasi-epistemic?

---

[81] For arguments that understanding reduces to knowledge, see e.g. Sliwa (2015), Brogaard (2005), and Khalifa (2017).

[82] See Elgin (2017) and Kvanvig (2003) for defences of objectual understanding.

What I am interested in exploring in this thesis is whether aesthetic qualities have epistemic value, regardless of whether there is an additional quasi-epistemic value in the vicinity. Thus for the rest of this thesis, I will set non-factive understanding and quasi-epistemic virtues aside and focus on whether aesthetic virtues can be epistemic.

3.4. Conclusion

In order to understand whether aesthetic virtues can be epistemically valuable, i.e. whether they are also epistemic virtues, we need to know what it means for a virtue to be epistemic. This chapter has discussed some prominent accounts of epistemicity. We have found that the most popular accounts of epistemicity, according to which theoretical virtues are truth-conducive, misclassify central epistemic virtues. If strength and scope are epistemic virtues, and lucky guesses doesn't make a theory epistemically good, then epistemic virtues cannot be truth-conducive. I considered some ways of avoiding this problem while retaining the claim that epistemic virtues are truth-directed, and argued that epistemic virtues are conducive to knowledge. This entails that epistemic virtues are signals of truth, since you cannot know what isn't true, but it opens up for theories to be epistemically valuable also by broadening or deepening our knowledge. Lastly, we discussed whether we should instead take some virtues to be conducive to non-factive understanding. I argued against this view. But even if some virtues were conducive to non-factive understanding, this would be so different from their being truth-directed that we can still sensibly ask whether the quasi-epistemic virtues are also epistemic.

So, the question is whether broadly aesthetic virtues can be conducive to knowledge. I will argue that they can: aesthetic virtues are epistemically valuable. In chapter 5, I present an account which explains why aesthetic virtues are epistemically valuable. But before attempting to justify the appeal to aesthetic virtues in theory evaluation, we need to know more about exactly what role they play in the development and defence of scientific theories. We turn to that next.

Joly and Holmes together have a beautiful theory,
and I believe it will be epoch-making.
– Chester Longwell, letter to Charles Schuchert (1926),
in Oreskes (1999:120)

Below all silken soil-slip, all crinkled earth-crust,
Far deeper than ocean, past rock that against rock grieves,
There at the globe's deepest dark and visceral lust,
Can you hear the *groan-swish* of magma as it churns and heaves?
– Robert Penn Warren (1980)

CHAPTER 4:
"THE ELEGANCE … IS CERTAINLY A STRONG POINT IN ITS FAVOUR"[83]

This chapter discusses how aesthetic considerations influenced the development and acceptance of the theory of plate tectonics, the modern theory of the earth's major geological features and the unifying framework for the earth sciences. I present a detailed case study of how aesthetic considerations influenced plate tectonics from its forerunner, continental drift, to the 'plate tectonic revolution' in the 1960s, giving an overview of how aesthetic factors influenced plate tectonics before looking in detail at a handful of examples from this history where elegance and simplicity tipped the balance in favour of a given hypothesis.

This case study gives us multiple examples of the various ways in which aesthetic considerations are explicitly and implicitly appealed to when evaluating theories, from Argand, Vine, and others' claims about being convinced of the truth of a theory because of its elegance and beauty to the implicit but strong preference in some research communities for simple and unifying theories. As I show here for the first time, aesthetic considerations played a substantial and positive role in 20th-century earth science. This gives us a detailed picture of what it is an account of aesthetics in science should explain.

[83] Émile Argand (quoted in Wegener (1966:87)). See section 4.3.1. below.

4.1 The need for a detailed case study

This chapter shows how aesthetic factors influenced the development and acceptance of the theory of plate tectonics. Although the history of plate tectonics has been written about by philosophers and historians before, the role of aesthetics in this history is underexplored – when mentioned at all – in previous work (Laudan (1978), Glen (1982), Le Grand (1988), Giere (1988), Hallam (1989), Oreskes (1999), Frankel (2012), Schindler (2018), Šešelja and Weber (2012), Pellegrini (2019), Weber and Šešelja (2020)).[84] This is surprising, for as we will see in this chapter, 20th-century earth science is awash with examples of aesthetic considerations influencing science. There has long been a widespread impression that aesthetic factors only play a role in highly abstract fields like physics and mathematics. Although this is changing (see especially Turner (2019), Ivanova (2021), Wylie (2021), Currie (2023), and Kozlov (2023)), more work is needed on aesthetics in 'messier' sciences like biology, psychology, and geology. Intriguingly, much of the recent work on aesthetics in less abstract sciences focuses on the earth sciences (see Wylie (2021), Turner (2019), Currie (2023)). However, this work is more concerned with the practices of science rather than with scientific theories. So even though good work has been done on both the aesthetics of earth science and on the history of plate tectonics, there has until now not been any work on the role of aesthetics in the history of plate tectonics. Thus we need to also look at the role of aesthetics in theory evaluation in these less abstract fields to get a full picture of how aesthetics influences scientific reasoning. The present case study is, to the best of my knowledge, the first in-depth discussion of how aesthetic factors influenced the development and acceptance of plate tectonics.[85]

---

[84] Anderson (2002) is the only paper-length discussion of aesthetics and plate tectonics I've come across, but it does not touch upon the history of plate tectonics. Interestingly, it criticises the 'standard model' of plate tectonics for *not* being simple and elegant enough (2002:56, 63).

[85] Much of the material of this chapter is published as Miyata-Sturm (2024b). I owe special thanks to Elizabeth Crowley, head librarian at Oxford University Department of Earth Sciences, for allowing me access to their collection.

Looking at a detailed case study is instructive, as it shows various ways in which aesthetic considerations are appealed to in practice. As the previous chapters illustrate, discussions of theoretical virtues easily become abstract and removed from actual scientific practice. It requires some work to connect the philosophical discussion of theoretical virtues with historical cases, but it's important that the philosophising is anchored in the historical facts that we are attempting to explain. To see how and when aesthetic factors influence science we need to consider historical cases in detail. Although much of the literature on aesthetics in science contains case studies, the cases are typically briefly described, giving more a taste of how aesthetics influence scientific reasoning rather than an in-depth picture of it (Engler (1990), Kivy (1991), Breitenbach (2013), Todd (2017), Ivanova (2021), the papers in Bueno et al. (2018) and in Ivanova and French (2020)). That said, there are some notable exceptions and the trend is towards more case studies (Chandrasekhar (1987), McAllister (1996), Glynn (2010), Kragh (2011), Ivanova (2017), Turner (2019), Wragge-Morley (2020), (2023), Wylie (2021), Stuart (2023), Ritson (2023)). The history of plate tectonics provides many examples of how aesthetic factors influence scientific reasoning, giving us compelling evidence that aesthetic virtues are treated as having epistemic significance by many scientists, and therefore making it an excellent case to study to understand this puzzling phenomenon.

I begin with a brief overview of plate tectonics and its aesthetic virtues (section 4.2) before discussing how the theory was guided by aesthetic considerations from its beginning as continental drift in the 1910s to the theory as we know it today (section 4.3.). In section 4.4., I discuss three cases where elegance and simplicity tipped the balance in favour of a given hypothesis, showing that aesthetic judgements played a substantial and positive role in the evaluation of continental drift and plate tectonics. In section 4.5, I consider ways of explaining this intriguing fact which are compatible with the historical evidence that aesthetic factors are sometimes misleading (Ivanova (2020), Hossenfelder (2018)) and briefly discuss recent accounts which posits an indirect link between beauty and knowledge; this discussion continues in chapter 5.

4.2 Overview of plate tectonics and aesthetic properties

4.2.1 Brief overview of plate tectonics

The theory of plate tectonics is one of science's big success stories, and aesthetic considerations played a significant role in its development and acceptance. Plate tectonics is *the* grand unified theory of the earth sciences, developed over the past century in two 'growth spurts', or periods of rapid development. The first, in the 1910 and '20s, gave rise to the theory of continental drift – the immediate precursor to plate tectonics – and the second, in the 1960s, gave rise to the modern theory. As a contemporary textbook has it,

> Plate tectonics remains the first and only global geodynamics theory which orchestrates all known tectonic phenomena including earthquake zones, mountain building, structural patterns, nature of sedimentary basins, magnetism, and metamorphism – *plate tectonics is an elegant and comprehensive synthesis* of Earth's geodynamics. (Frisch, Meschede, and Blakey (2011:4), italics inserted)

Briefly put, plate tectonics says that the lithosphere (the uppermost, brittle layer of the earth) is broken up into rigid plates separated by narrow zones where the plates move apart from or under or slide past each other. Underneath the plates, there is a layer of hot, plastic rock (the asthenosphere), which rises to the surface in the cracks left by diverging plates, mostly along midocean ridges, and becomes part of the lithosphere as it cools. Plates are moved by gravitational 'slab pull' from subducted plate segments, 'ridge push' as newly minted seafloor is pushed aside to make room for more upwelling magma, and convection currents in the asthenosphere driven by radioactive heat from the earth's core. Based on two simple hypotheses – that plates are rigid and that their relative movements can be described as rotation around a pole – the theory predicts plate movements with astonishing accuracy.

4.2.2. The aesthetic virtues of plate tectonics

Frisch et al. claim that the theory is elegant, and such praise is common. Two features of the theory are particularly often praised for its aesthetic appeal: its geometrical simplicity and its unification of a wide variety of phenomena. Plate tectonics explains much by making few and uncomplicated claims about the planet: "[m]uch of the beauty of plate tectonics lies in the geometric exactness

and simplicity of this geometry of movement" (Cox and Hart (1991:9)), "[t]he beauty of plate tectonics lies, in large part, in the simplicity with which the kinematics can be described – as rigid plates" (Molnar (2003:305)). The overall picture of horizontally mobile pieces of lithosphere is simple, in many of the ways a theory can be simple: in its mathematical description (using Euler's theorem about rotation around a pole), the number of basic assumptions it's built upon (plates as rigid pieces of crust moving on top of a plastic layer), in the picture it paints of the processes responsible for how the earth's geological features change over time, and in the ease with which this can be visualised and understood.

In other words, both *what* plate tectonics states and entails and *how* it states it is simple. Plate tectonics is also simple at lower levels of description, though the details of particular geological or geophysical features necessarily make the theory more complex the closer we get to the ground (so to speak). For example, the Mendocino fracture zone looks extremely complex, but is easily described as the intersection of three plates. "Three plates! Of course. So elegant, so simple, so powerful," as Atwater exclaimed (2003:250; see below).

Plate tectonics also provides an elegantly unifying framework for understanding the earth's large-scale geological features: it's "an elegant and comprehensive synthesis" (Frisch et al. (2011:4)); the "plate tectonic model explains most of the geological and geophysical features of trenches and island arcs with elegant simplicity" (Cox and Hart (1991:30)). It unifies all the earth's major geological phenomena under one explanation (see the full quote by Frisch et al. (2011) in section 4.2.1 for a list of many of the phenomena it unifies). Its explanations and predictions regarding a wide variety of phenomena follow with ease from its central claims, giving a streamlined picture of the workings of the world where most of what the theory states or entails seem almost inevitable given the whole picture. Hypotheses about mobile continents, seafloor spreading, different kinds of plate boundaries, and more, were joined under the simple geometrical framework of McKenzie and Parker (1967) and Morgan (1968) (see below), and this fitting together of wide-ranging ideas in an uncomplicated way evoked great pleasure. Judging by how the scientists express themselves, and how similar their obvious delight is to typical aesthetic

experiences, this pleasure was distinctly aesthetic. As discussed in chapter 2, unless there is a good reason to think otherwise in a given case, we should take aesthetic claims in science to be expressions of the same kind of psychological experiences as people paradigmatically (but far from exclusively) have when they engage with works of art – in short, as expressions of aesthetic experiences.

As we saw in chapter 2, scientists describing what they find beautiful in a theory often talk about the harmony of how parts of it fit together. For example, Henri Poincaré talks about the "intimate beauty which comes from the harmonious order of its parts" (in Ivanova (2017:2584)), and Subrahmanyan Chandrasekhar, following Heisenberg, takes beauty to be the "proper conformity of the parts to one another and to the whole," which causes "wonderment and surprise" (1987:70). This fits well with how plate tectonics is viewed. As we have already seen examples of, and as we will see more examples of shortly, earth scientists regularly praise plate tectonics for its beauty and elegance, connecting these aesthetic properties with its simplicity, unity, and the excellent (even seemingly inevitable) way in which the various hypotheses and assumptions fit together into one compelling whole.

Philosophers also often note that beauty is partially based on harmony or goodness of fit (Scruton (2000:96), Todd (2017:228)). As Cochrane says, "we find objects or events beautiful when many parts seem to harmoniously fit together" (2021:42). This is in line with how beauty is standardly understood by scientists sensitive to the aesthetic delights of science, as we have just seen. Furthermore, beauty is traditionally explicated in terms of unity in diversity (or 'simplicity in complexity'): there is a pattern connecting the different parts at some deeper level (Cochrane (2021:36), McAllister (1996:21, 110), Engler (1990:27), Hutcheson (1726)). Despite differences (in colours, figures, rhythms, explicanda, etc.), we can discern an underlying theme which connects the dots and produces a unifying picture. This traditional principle of unity in diversity finds a modern expression in the claim that aesthetic preference lies at the 'aesthetic middle' between simplicity and complexity. "Most theorists accept Daniel Berlyne's (1971) observation that preference for complexity follows an inverted U shape, where a moderate level of complexity is optimal", as Tom Cochrane (2021:36) says. (Note that this is about *aesthetic* preference; the focus

of the empirical studies is typically on paintings or music.) Many studies support the inverted U-curve model of aesthetic preference, though results are somewhat mixed (see Chmiel and Schubert (2017), Van Geert and Wagemans (2020:135), Delplanque et al. (2019:146–7), Cutting (2020:68–70), Messinger (1998:558), Adkins and Norman (2016:1305) for discussion and further references).

If we tend to prefer a moderate level of complexity, why do scientists praise theories for their simplicity and not for their balance between simplicity and complexity? Presumably because scientific theorising is almost inevitably complex; with some level of complexity guaranteed, what is difficult – and praiseworthy – is pushing the theory towards the simpler end of the scale. Despite scientists' explicit praise of simple and unified theories, a closer look at the theories they prefer shows that they conform to the aesthetic ideal of unity in diversity. It is theories which reveal an underlying connection or similarity between disparate phenomena (like plate tectonics) that are praised for simplicity and unity. This requires both 'diversity' and 'unity': too much of the first and the theory is too messy, but unity without diversity, or simplicity without complexity, leaves the theory with little explanatory strength, interest, or aesthetic appeal (Cochrane (2021:43), Delplanque et al. (2019:146)). As a case in point, consider how the biotechnologist Leroy Hood describes his experiences of beauty in science:

> what I have been impressed with in science over my twenty-one years is the absolute conflict between, on the one hand, as we come to learn more and more about particular biological systems there is a simple elegant beauty to the underlying principles, yet when you look into the details it's complex, it's bewildering, it's kind of overwhelming, and I think of beauty in the sense of being able to extract the fundamental elegant principles from the bewildering array of confusing details. (Wolpert and Richards (1997:44))

A beautiful theory translates the bewildering complexity of the world into the simple underlying 'principles' (features, processes, causes) responsible for this complexity, connecting the diversity of what we observe to the simple principles that unifies them.

4.3 Aesthetic considerations in the development and acceptance of plate tectonics

4.3.1. Continental Drift

Continental drift provided a simple framework for explaining the large-scale features of the earth – such as mountain building and similarities among continents currently divided by oceans – by stating that the continents move horizontally through the seafloor,[86] which is the thin, solid crust of an otherwise molten layer between and under the continents (Oreskes (1999:77–9)). This molten layer is denser and behaves like a highly viscous fluid (Wegener (1966:45)).[87] Large-scale geological features are caused by the way that continents move around. The developer of drift, Alfred Wegener, argued that his theory could explain more phenomena and explain them better than competing theories could (e.g., (1966:76)). Wegener is credited with developing the theory, and deserves that credit, but many of his ideas were not new (Oreskes (1999:54–5), Le Grand (1988:46)).[88] What was new was the way he put those ideas together into a coherent and detailed theory that could explain both geological and geophysical evidence. In other words, continental drift was of wider scope than its competitors, and it unified much of what was right about the other theories by bringing the evidence together into one simple framework.

The various pieces of evidence had seemed incompatible with each other and led to the development of theories which each could explain only half the evidence and was disconfirmed by the other half (Wegener (1966:5)). The theory that best fitted the geological and palaeontological evidence was contraction

---

[86] We now know that the continents are passively dragged along with the seafloor instead of ploughing *through* the seafloor: this is one of the few significant cases of a hypothesis from drift theory *not* being incorporated into place tectonics.

[87] "To put it in rather picturesque terms, the two layers behave like open water and larger ice floes", as Wegener put it (1966:37).

88 Wegener presented the first account of continental drift in two papers in 1912 (Frankel (2012:38)), expanded it into a book in 1915 (*Die Entstehung der Kontinente und Ozeane*, transl. *The Origin of Continents and Oceans*), which was published in expanded and revised editions in 1920, 1922, and 1929. The version relied on here (Wegener (1966)) is a reprint of the English translation of the fourth edition. Independently of Wegener, Frank Taylor proposed a version of Drift in 1910. His account was not widely read, and it is unlikely that Wegener knew of it when developing his theory (Oreskes (1999:82), Frankel (2012, vol I:70)).

theory, which was generally accepted in Europe, and the one that best fitted geophysical evidence was permanence theory, which was generally accepted in North America (Oreskes (1999)).

Contraction theory said that the earth is cooling and that as it does so it gradually shrinks, much like an apple becomes wrinkly when it dries (Suess (1904–24), Wegener (1966:9)).[89] The continuous change and strong lateral forces exerted on the earth's surface could explain the complex folding of the Alps and the distribution of fossils, including the presence of marine fossils in mountains.

Permanence theory said that oceanic and continental crust resulted from minerals solidifying at different temperatures as the earth cooled, causing the continents and ocean basins to be different in kind, which fitted the geophysical data (particularly the fact that the crust is in isostatic equilibrium which, given the mass difference between continental mountains and ocean basins, must mean that they differ in density). The oceans and continents deformed over time but were always in the same relative position to each other, hence 'permanence' (Oreskes (2003:5)).[90] Both permanence and contraction theories struggled to explain the evidence that the other theory was designed to explain, as Wegner was quick to point out (1966:5–17). Drift theory, however, fitted both the geological and geophysical evidence (1966:17–21). It thus unified phenomena explained by both contraction and permanence theories, and it did so through a simple unifying hypothesis: light and thick continents move horizontally through dense seafloor, which acts on a geological timescale like a viscous liquid, and major geological features are formed when continents move apart or collide with each other.[91]

---

[89] "The dislocations visible in the rocky crust of the earth are the result of movements which are produced by a decrease in the volume of our planet. The tensions resulting from this process show a tendency to resolve themselves into … horizontal (i.e. thrusting and folding) and vertical (i.e. sinking) movements," as Suess (1904:107) described it.

[90] Permanence theory was originally a version of contraction theory, but the claim that the earth had contracted was later abandoned (Le Grand (1988:23)).

[91] See also Oreskes (1999:54–5) and Frisch et al. (2011:2–3).

Wegener struggled with both fleshing out the details and finding a plausible mechanism for drift.[92] However, what mattered to Wegener was the plausibility and elegance of the overall idea, not the details. Wegener's argument was abductive: he presented pieces of evidence which were not necessarily convincing when taken separately but which pointed in the same direction, and argued that they were best explained by his account as it united the widespread evidence into one compelling picture.[93] Or as Wegner put it,

> if one starts out from the standpoint of drift theory, and if one maps the fossil evidence for climates in charts developed for the relevant period by means of the theory, these contradictions completely vanish, and all the climatic evidence arranges itself to form the pattern of climatic zones which is familiar to us today. (1966:130))

While his explanation may have provided the best explanation of the evidence, it certainly did not make all contradictions completely vanish, or all evidence arrange itself in just the right way. Here Wegener makes clear that what matters to him is the overall framework, not that it gets all the details right. This can be seen both in the major revisions he made to *The Origin of Continents and Oceans* ("[e]ach edition was a complete revision", as his brother Kurt Wegener wrote in the preface to the editions appearing after Alfred's death (1966:iv)), and in explicit statements, such as where he discusses geological evidence in favour of drift:

---

[92] It's sometimes claimed that this is why drift was rejected in North America, but as Oreskes' (1999) excellent historical account shows, this was not the case.

[93] Oreskes (1999) doesn't comment on this, although she gives an in-depth discussion of Wegener's argumentative style in relation to how methodological and stylistic differences in American and European earth science contributed to the rejection of continental drift in America (1999:126, 153–5, 157). According to Oreskes, the major methodological divide of the earth sciences was between inductivism and deductivism, where induction does not include abduction (1999:145, 303–4) and deductivism is basically the hypothetico-deductive method (1999:141, 157, 284). She says that "Wegener's treatise was explicitly deductive" (1999:154). But Wegener's arguments were often abductive. The issue might be terminological, for Oreskes says that "[h]is was a causal-theoretical account, in which a basic principle was used to illuminate a wide range of geological evidence, not the other way around" (*ibid.*), and 'illuminate' sounds more like 'explain' than that the evidence deductively follows from a basic principle.

> Even though the theory in certain individual cases may still be uncertain, the totality of these points of correspondence constitutes an almost incontrovertible proof of the correctness of our belief that the Atlantic is to be regarded as an expanded rift. (1966:76)[94]

He makes two points related to theoretical virtues here: first, that there were many 'points of correspondence' between the evidence and his theory (i.e., it is highly empirically accurate), and second, that it's the *overall* fit between his explanation and the evidence which matters when evaluating it. This is a hallmark of abductive reasoning: each piece of evidence might give only weak support to the explanation, but the explanation will be strongly supported by evidence taken together if they point in the same direction.

As Wegner makes clear in other places, drift theory is not only supported by the fact that many pieces of evidence fit his theory, but by the particularly nice way they fit.[95] During the first weeks of his work on drift, Wegener wrote in a letter to his collaborator and soon-to-be father-in-law Wladimir Köppen "that his hypothesis of continental displacements was not an imaginative creation or a fantasy but … a radical reorganization of existing observational data into a new picture, *producing simplification and coordination in the place of previous complexity and contradiction*" (Greene (2015:240), italics inserted). Clearly, it's not just the fit between the evidence and theory which matter but the simple and well-coordinated picture that his theory paints.

Some of the virtues Wegener refers to here have clear aesthetic overtones. We do not need to invoke aesthetic preferences to explain why he wanted to avoid contradiction, but a distaste for complexity and desire for simplification and coordination are best interpreted as aesthetic attitudes

---

[94] See also van der Gracht, who organised a 1926 symposium on continental drift and said that the "manner in which *major* facts fit into this theory is very suggestive of the fundamental truth that such a thing as considerable inter- and intra-continental drift occurs. The details of the picture, and particularly the mechanical and physical explanation, will need considerable further research" (1928:75). van der Gracht was a vocal proponent of continental drift and central to the positive reception of the theory in the Netherlands (see, e.g., Barzilay (2010:215)).

[95] Isn't fit simply fit? No: an elegant and an ad hoc theory can fit exactly the same evidence equally well, but there is an important difference in how nicely these theories fit the evidence.

regarding the goodness of the way his 'new picture' fits together. If we keep in mind that theoretical virtues can be simultaneously aesthetic and epistemic (see chapter 5), Wegener's clear preference for 'simplicity and coordination' looks like the typical aesthetic preference for simplicity discussed in above. Or rather, a preference for a balance between simplicity and complexity of a 'coordinated' sort, where the complexity fits into an overarching, simpler, easy-to-discern pattern (i.e., simplicity in complexity, or unity in diversity): the object is neither boringly simple nor overwhelmingly complex. As mentioned above, the complexity of the earth's geodynamics ensures that any sufficiently detailed account will be complex enough to not be boring. Scientific theories seldom run the risk of being too simple to be aesthetically pleasing: the challenge lies in simplifying them. Wegener's letter expresses the common aesthetic pleasure and satisfaction found in a simple, unified, and 'coordinated' theory. I take his claim about coordination to belong to the same category as elegance and harmony (as in 'a harmonious fit'), which is about the aesthetically pleasing way in which the parts fit together, whether that's the theory in itself (internal consistency) or the theory in relation to the evidence (empirical fit or external consistency). Often this pleasure is expressed in overtly aesthetic terms: drift theory, and later plate tectonics, is often praised for its elegance and the harmonious way in which it fits together a broad range of data.

For example, when discussing the uneven distribution of mountain chains across the globe, Wegener quotes Émile Argand as saying that

> The *elegance* with which drift theory explains these significant facts, which were not known when the theory was originated, *is certainly a strong point in its favour*. Strictly speaking, none of these facts really proves drift theory … but *they all fit in excellently* … to an extent that makes them highly probable. (1966:87, italics inserted)

This praise shows that Wegener was not alone in aesthetically appreciating the theory, or indeed in taking the aesthetic properties of drift as support for the theory. Furthermore, the fact that he published Argand's praise as support for his theory also shows that he took its aesthetic appeal to carry epistemic weight.

One might object that what Argand appreciated was that drift could explain 'significant facts' and that it made novel predictions which turned out to

be correct: the appreciation is strictly epistemic, not aesthetic. But Argand explicitly expresses the aesthetic praise as something additional: the explanations and novel predictions are one thing, but the *elegance* with which the theory explains the facts 'is certainly a strong point in its favour'. Similarly, the 'excellent fit' of facts that 'strictly speaking' don't prove drift – I assume he means there are other possible explanations for the data – is taken to raise the probability of the theory. This isn't just about the likely truth of drift, and it doesn't treat aesthetic virtues as epiphenomenal or as an afterthought with no epistemic import. It's about the aesthetically satisfying way drift explains orogeny (mountain building).[96] The most plausible interpretation of Argand's claim is that he took elegance and goodness of fit to have epistemic value; in other words, that the epistemic value of Drift is partially due to its aesthetic features. Why think that? Because, first, Argand uses standard aesthetic vocabulary – and as discussed in chapter 2 above, unless we have a specific reason to think otherwise, we should assume that he intended to use these terms with their standard meanings. Second, Argand's sounds like an entirely standard aesthetic reaction, namely the appreciation experienced when the parts (whether parts of a painting, music, fictional narrative, or scientific theory) fit together in a harmonious way.

Wegener also quotes the famous South African field geologist Alexander du Toit as saying that it's the geological evidence that will decide the fate of the theory

> because arguments based on such matters as the distribution of fauna are not competent here; they can generally be explained equally well, *even if less neatly*, by the orthodox view that assumes the existence of extended land bridges. (1966:vii, italics inserted)

It mattered not only that drift theory could explain the distribution of fossils and present-day species – so too could other theories – but that it could do so *neatly*. I take 'neatly' here to mean that it provided a tidy, straightforward explanation without unmotivated elements or loose ends. Furthermore, it could explain *both* the geological and palaeobiological data, thereby unifying different fields. Wegener's argument for drift was abductive and he clearly put weight on broadly

[96] In other words, the evaluation is *both* aesthetic and epistemic.

aesthetic virtues in presenting drift theory as the best explanation of geological and geophysical evidence. He allowed considerations of elegance, simplicity, unity, and non-ad hocness[97] to play an epistemically significant role in the case he built for his theory; these virtues were both aesthetically valued and major selling points for the theory of continental drift.

4.3.2. Plate Tectonics

Aesthetic considerations played a similar role in the second growth spurt, when plate tectonics as we know it today was developed. Although the research happened at a much larger scale, with well-funded teams of scientists collecting an amount of data that scientists before the 1940s could only have dreamt of,[98] many of the important theoretical developments were made by scientists who didn't have access to conclusive evidence but relied on their aesthetic sensibilities in extrapolating from what they did have access to.

Continental drift never became the consensus view, but in Europe, South Africa, South America, and India (Frankel 2012, vol. I:31) it became a staple account alongside contraction theory and permanence theory, further developed by luminaries like John Joly, Reginald Daly, Alexander du Toit, and especially Arthur Holmes (Oreskes (1999), Frankel (2012, vol. I)). In North America permanence theory still held sway (Oreskes (1999:124–7), Le Grand (1988:117–8)).

Intriguingly, part of what North American earth scientists disliked about continental drift was what Wegener and others thought of as its virtues: "[i]t was too large, *too unifying*, too ambitious. Features that were later viewed as virtues of plate tectonics were attacked as flaws of continental drift" (Oreskes (2003:11, italics inserted)). The enticement of this unifying theory was resisted in part because it (and Wegener) clashed with the prevailing American methodology of strict empiricism and 'multiple working hypotheses'. Perhaps not incidentally,

---

[97] For example, he criticised the land bridge hypothesis for requiring "further hypotheses which are 'ad hoc' improbabilities" (1966:13) (this case is discussed in the list of theoretical virtues in section 2.3.3, under 'ad hocness').

[98] See Menard (1986) for an excellent account of the research teams and their developing ideas leading up to the formulation of plate tectonics by one of its central players.

the North American earth scientists spent decades looking in the wrong direction. Chamberlin's method of 'multiple working hypotheses' had a huge influence on North American earth science (Oreskes (1999), Chamberlin ([1897] 1970)), and was explicitly "an attempt to obviate the 'unconscious selection . . . of phenomena that fall into harmony' with theoretical commitment and 'unconscious neglect of phenomena that fail of coincidence'" (Oreskes (1999:144)). This could have been a necessary corrective to overenthusiastic generalisations, but, given the scholarly climate in which it was made (which was hostile to wide-scoped generalisations and theory building), it may rather have functioned as a warning against trusting the satisfaction of a harmonious fit between theory and evidence. Instead of taking aesthetic virtues to support a theory, the Chamberlinian line of reasoning is that if the theory is harmonious and elegant, you have probably ignored some important evidence.[99] Chamberlin thought that aesthetic virtues can play an important role in motivating research, but not in evaluating or justifying a theory:

> Important as the intellectual affections are as stimuli and as rewards, they are nevertheless dangerous factors in research. All too often they put under strain the integrity of the intellectual processes. (1970:605–6)

'Intellectual affections' – for example, the pleasure and satisfaction one finds in an elegant explanation – can be valuable by making the scientists eager to engage in the hard and often dreary work of science, but according to Chamberlin they can play no legitimate role in the context of justification because they tend to negatively influence cognitive processes.

It is suggestive that the research communities and individuals who showed the most willingness to treat broadly aesthetic virtues as legitimate considerations when evaluating theories were both central to developing continental drift/plate tectonics and first recognised the potential of continental drift and other central hypotheses of plate tectonics, like seafloor spreading (discussed below). In contrast, the commitment among North American

[99] For more on the methods and norms in early 20th century North American earth science (and to a lesser degree in European and other research communities), see Oreskes (1999), Frankel (2012) Hallam (1983), and Le Grand (1988).

scientists to empiricism, multiple working hypotheses, and descriptive geology together with a distrust of what proponents of IBE view as strong abductive support (wide-sweeping, elegant explanations) was a hindrance not help in discovering the truth about our planet's geological features.[100] Such historical cases contain too many variables to be treated as a proof of any philosophical position and it is not clear to what extent this generalises – for that we would need a much more careful historical analysis across many fields – but it certainly suggests that aesthetic considerations can be of epistemic value.

In the '30s and '40s work on drift nonetheless slowed to a trickle, not least due to the disruption of research by the Second World War. This was a blessing in disguise for plate tectonics, for many earth scientists joined their countries' navies and while at sea collected data on the structure and composition of the seafloor. Some of the most important new evidence for

[100] Frankel argues that the main reason why North American earth scientists were so negative to continental drift was their regionalism: since permanence theory could explain local landforms, they discounted its problems accounting for the geology of the southern hemisphere (and the evidence in favour of drift found there) (2012, vol. I:349). He rather testily (2012, vol. I: 36, 350–1) rejects Oreskes' (1999) argument that among the many reasons why they tended to reject drift was 'epistemological chauvinism'. Oreskes said that "[they] responded to the difficulties in their discipline by relying on and emphasizing the information that was geographically, nationally, and/or disciplinary closest to themselves. They did not necessarily ignore the information, but they did weigh the available evidence differentially, based on their affinities. And while these affinities expressed themselves epistemologically – in terms of differential weightings of evidence – their sources were largely social" (1999:52–3). Here's Frankel: "Oreskes is right to claim that mobilism was rejected by US geologists, in part, because of their adherence to key parts of Dana's permanentism. But why did permanentism last so long as the reigning view? Was it social pressure to conform? Was it pressure to adhere to certain methodology? Was it the passing down of Dana's wisdom to later generations? Or, was it because Dana's permanentism was at the time better equipped than mobilism to solve key problems of North American geology, especially through its later renovation by Kay and others in response to new data?" (2012, vol. I:350–1). I think Frankel misses Oreskes' point: the fact that they weighted evidence about local phenomena more heavily *is* what made them epistemologically chauvinistic. Explaining North American landforms is not more important than explaining South American landforms for understanding geology, but North American geologists' tendency to accept a theory which could explain local phenomena regardless of its overall epistemic standing is treating it as such.

drift/plate tectonics was based on these data.[101] The importance of this research for submarine warfare and detection of nuclear tests opened the eyes of government funding bodies to the usefulness of geophysical research, dramatically increasing resources in the post-war years (Menard (1986:37–42), Oreskes (2003:17)).

Some of the research was a continuation of work done by Felix Vening Meinesz, Harry Hess, and David Griggs in the 1930s, who extended Holmes' argument for convection currents as the mechanism behind drift and orogeny (Le Grand (1988:116–7), Oreskes (2003:13–6)). The Scripps Institution of Oceanography (directed by Roger Revell) and Lamont Geological Observatory (directed by Maurice Ewing) sent scientists on near-continuous expeditions in the two decades immediately following the war. When winter conditions made research in northern seas impossible, they sailed to the southern hemisphere, and vice versa (Menard (1986:41)). Continuously, and aided by recent technological advances, they collected core samples, studied earthquakes, and collected all the data they could gather on bathymetry, heat flow, gravitational variations, and the composition of the seafloor with echograms, magnetometers, and seismic refraction studies (explosives).[102] Gravitational measurements revealed huge structural features of the seafloor, such as the immense mid-ocean ridges, deep-sea trenches, and what were later recognised as transform faults. Seismology showed that the seafloor, far from being the smooth surface

[101] Unfortunately, much of it was classified. One of the best examples of scientists finding a way to share this important but classified data is Bruce Heezen and Marie Tharp's (1977) map of the ocean floor, which was based on – but did not reveal – the data that they were prohibited from sharing, providing the scientific community with an accurate depiction of major structures of the seafloor (Oreskes (2003:23)).

[102] By noting the velocity with which pressure waves travel from an explosion one can discover the thickness and some of the structure of different layers of the seafloor that the waves travel through. The seismologists struggled for many years to get the explosives and the hydrophone (the 'receiver' of the seismic waves) down to the seafloor until an "astonishingly simple technical solution" was found: "[i]f the ocean is considered to be just one more layer of the earth, the problem vanishes. The hydrophones and explosions could be at the surface of the sea" (Menard (1986:35)). Thus started the "standard procedure to throw a 0.2 kg explosive charge overboard every two minutes" which accidentally killed a researcher in 1961 (Menard (1986:42)).

previously imagined, mostly consisted of steep hills relatively unburdened by sediments (Menard (1986:51–2)) – suggesting it was young.

In the early '60s, Hess and Robert Dietz independently proposed what Arthur Holmes had suggested decades earlier (Morley (2003:79), Cox (1973:15–6), Menard (1986:152–61)), namely that new seafloor is created through upwelling magma in cracks left by diverging plates and powered by enormous convection currents in the asthenosphere. The magma becomes part of the 'spreading' seafloor as it cools to solid rock and moves away from the rift to sink beneath continents (where trenches form). When the subducted seafloor sinks into the hot asthenosphere it melts to become part of the mantle material, which in turn becomes part of the lithosphere when it reaches the surface in midocean ridges or volcanic eruptions. If one accepted mantle convection, "a rather reasonable story could be constructed to describe the evolution of ocean basins … Whole realms of previously unrelated facts fall into a regular pattern, which suggests that close approach to satisfactory theory is being attained" (Hess [1962] (1973:27)). Note how much weight Hess here puts on unity and coherence.

Dietz (1961) called this process 'seafloor spreading', and the name stuck. Hess and Dietz's picture, supported by the old considerations in favour of drift together with the recent gravitational and magnetic data, gave new support to the hypothesis of mobile continents. Some of the most intriguing of these data were from Ron Mason and Arthur Raff's magnetic survey of the north-eastern Pacific from the late 1950s, published as maps (Mason and Raff (1961), Morley (2003)).[103] These maps astonished the scientific community, for they showed a clear and remarkably regular pattern of stripes of positive and negative magnetic anomalies parallel to segments of midocean ridges and roughly in line with the

[103] Both the zebra pattern map and the magnetic anomaly graphs which later confirmed VMM are regularly described as beautiful. Tanya Atwater recalls "sorting through [Lamont's] voluminous data sets to plot out magnetic anomaly profiles from all the world's spreading centres. … [W]hile I was at it I conducted a magnetic anomaly "beauty contest". After all these decades of ships collecting new data, that old *Eltanin*-19 crossing [from 1965] still won first prize" (2023:386, note 5). It is "wonderfully clear and symmetric and exceptionally easy to read" (*ibid.*), suggesting that its beauty is connected to its quality as data, and thus that epistemic and aesthetic evaluations are connected.

coast, but the pattern didn't correspond to other features of the seafloor. This came to be known as the 'zebra pattern' and it puzzled the scientific community for some years until explained by the 'Vine-Matthews-Morley hypothesis' (VMM) (Vine and Matthews (1963)).[104] VMM is an upshot of the combination of seafloor spreading and geomagnetic field reversals (discussed below), stating that the new magma created at spreading centres takes on the polarity of the geomagnetic field at the time of cooling. As this field periodically changes polarity, the direction of the magnetic field will be recorded in a pattern of stripes of normally and reversely magnetised rock parallel to the ridge axis and symmetrical on either side of it, showing a linear relationship between the age of the seafloor and its distance from the ridge.

The zebra pattern revealed by the magnetic surveys is exactly what one would expect from constant seafloor spreading and periodic geomagnetic reversals, so Mason and Raff's maps confirmed the central prediction of VMM.[105]

Lawrence Morley, who independently from Vine and Matthews proposed the hypothesis, recounts the immediate confidence he felt about it:

> I had [the zebra map] on my mind for nearly two years before I spotted the Dietz paper on ocean floor spreading. Eureka! (2003:80)
> I never had any doubts about the concept. It locked three disparate and unproven theories together in a mutually supportive way: the theories of continental drift, sea floor spreading, and the periodic reversing of the geomagnetic field. It was like finding the key piece to an enormous jigsaw puzzle that made everything fit together. (2003:83)

The mutual support was abductive: the support for each of the three hypotheses was stronger when they were taken together, because their combination resulted in one simple and unified picture where everything fits together excellently, almost like it couldn't have been another way (cf. Weinberg: "[t]here is another

[104] Morley came up with the idea before Vine and Matthews, but his paper was rejected by two journals and it took some years until the community learnt that he had had the same idea (Vine (2003:57)).
[105] Assuming that the negative anomalies (low intensity magnetism), as measured at surface level, correspond to reversely magnetised rock (Vine (2003:65)) – which it does – as more direct measurements were unavailable in the '60s.

quality besides simplicity that can make a physical theory beautiful – it is the sense of inevitability that the theory may give us" (1993:107)). As discussed in section 4.2.2 above, many scientists explain what they find beautiful in terms of this kind of combination of simplicity and excellent fit.

Morley was certainly not alone in being immediately convinced on seeing this 'key piece of the jigsaw puzzle'. Here is how Mason recalls receiving the hypothesis:

> Vine and Matthews' hypothesis offered an *elegant explanation* of how the magnetic lineations of the northeast Pacific could have come about, although in this case there was no obvious connection with an ocean ridge. … I had absolutely no doubt as to the correctness of their hypothesis. (2003:41; italics inserted)

Note how much weight is put on aesthetic and explanatory factors here. Significant pieces of evidence were missing and so was a central piece of the big picture (the connection with a mid-ocean ridge – Vine and Matthews' data was from the Carlsberg Ridge and lacked the striking symmetry of Mason and Raff's maps (from the Juan de Fuca Ridge) and later 'zebra patterns' from the Reykjanes ridge). And yet Mason was utterly convinced that the hypothesis was true! It elegantly made sense of the available data and *seemed so right* that he was willing to accept the strong universal generalisation expressed by the Vine-Matthews-Morley hypothesis.

The pleasure these scientists felt when engaging with this hypothesis, based on the way it fitted together different ideas in a mutually supportive framework and the elegance of the ensuing explanation of the magnetic lineations, also gave them a great confidence in it. Compare Paul Dirac, who wrote that "one has a great confidence in [a] theory arising from its great beauty, quite independently of its detailed successes" (1980:40, quoted in Ivanova (2020:88)). Luckily, pleasurable feelings don't always cause immediate acceptance of the object of those feelings.[106] But this case shows what can be harder to spot in less extreme cases, namely that the pleasure of seeing how

[106] But see Dechêne et al. (2010) on the 'illusion of truth' effect, discussed in chapter 5: we are more likely to accept what we find aesthetically pleasing.

everything fits perfectly together – a typical aesthetic delight – is closely connected to acceptance.

By 1963, after the publication of the seafloor spreading and the VMM hypotheses, almost all the pieces of plate tectonics were assembled. But no one had yet given a good mathematical description of the plates' movements, and some important geological features – like the fracture zones that offset midocean ridges – were still puzzling. The pieces came together with two additional simple hypotheses: that plates are rigid bodies and that their movements are described by Euler's theorem of motion on a sphere.

In continental drift it was assumed that the continents behaved more or less like rigid plates, and, from Holmes onward, that the seafloor did the same.[107] What was new was the decision to model the lithosphere as broken up into perfectly rigid plates with all the action happening along their boundaries and in the asthenosphere underneath them. Doing that allowed one to use Euler's theorem about rigid-body rotations on a sphere (see below) to describe how the plates moved, and suddenly everything fell into place. Like so many other important discoveries in the history of plate tectonics, this idea was hit upon independently by two groups: Dan McKenzie and Robert Parker (1967) and Jason Morgan (1968).

Euler's theorem says that any movement of a rigid body on the surface of a sphere corresponds to a rotation of that body around some axis that passes through the centre of the sphere. The movement of one such rigid body relative to another traces small circles whose centre is their common rotational axis. Edward 'Teddy' Bullard et al. (1965) used this theorem to fit continents together but failed to see its implications for lithospheric movement in general. For example, it perfectly described the curious fracture zones along the mid-Atlantic ridge that can "almost be fitted with a set of great circles – almost but not quite. Morgan found that the fracture zones could be fitted much more accurately by

[107] "For me the central idea [of plate tectonics] is the rigidity of plate interiors. It is this property that allows the surface motions of the earth to be described *by so few parameters*. Other versions of the theory – continental drift, palaeomagnetic reconstructions, and sea floor spreading – implicitly or explicitly used this property, but did not recognise its importance" (McKenzie (2003:183; italics inserted)).

a set of small circles drawn about a point that he soon realised had the significance of being a pole of relative motion between two plates" (Cox (1973:46)).

McKenzie used a map of earthquake zones to come up with the plate boundaries. In this he was greatly aided by Tuzo Wilson's famous 1965 paper, where Wilson argued that all active plate boundaries are connected as a network spanning the globe and introduced a new, third kind of plate boundary (transform faults). The two previously recognised boundaries are what is now called 'constructive plate boundaries' (like midocean ridges) and 'destructive plate boundaries' (subduction zones). These boundaries are often offset by huge faults (fractures) where the plates slide past each other ('conservative plate boundaries', or 'transform faults'). These faults end abruptly where they intersect constructive or destructive boundaries, 'transforming' into these other kinds. Looking at a map of the epicentres of earthquakes and volcanoes, you see a map of plate boundaries, because most of the seismic activity happens along active boundaries.[108] Such maps were essential for Morgan and McKenzie/Parker when they worked out the geometry of plate movement, as they were the best source of information on the size, shape, and location of tectonic plates. Wilson's "original theory was worked out for flat plates on a plane" (Cox (1973:45)); with the help of Euler's theorem, his global network could be fitted to the globe. This turned out to describe the plate movements perfectly, even where it looks extremely complex. Tanya Atwater recalls questioning McKenzie and Parker about the adequacy of their account for the complex Mendocino fracture zone, and they replied that it lay on the intersection of three plates. "Three plates! Of course. So elegant, so simple, so powerful" (2003:250). Atwater's exclamation could have been made about plate tectonics in general, and it nicely expresses the general attitude towards the theory: it is so elegant, so simple, and so powerful.

So far, we have seen that aesthetic factors influenced the development of continental drift and its transformation into plate tectonics from a bird's eye perspective. Let's have a closer look at some of these cases.

---

[108] This is a central reason for thinking that the plates are rigid.

4.4 A 'harmonious major picture', troublesome rocks, and elegant fit

More than once did individual scientists and the scientific community stick with strong, simple, elegant theories in the face of apparently contradictory evidence, and more than once were they reluctant to incorporate data when this would have made their otherwise elegant theory complicated. That is, sometimes aesthetic considerations were weighted heavier even than fit with data. In this section, we'll see three examples of aesthetic considerations shaping research in this way.

The Squantum 'tillite'. As mentioned, for Wegener what mattered was to paint a compelling big picture, not to be right in all the details, and he was happy to let a "harmonious major picture" (1928:98) overrule fit with all the data. But when the apparent counterevidence diverged too much from his elegant theory, he rejected the evidence rather than revising the theory. For example, central to the case he made for continental drift was palaeoclimatological data showing that there had been extensive glaciation in areas that are now tropical or sub-tropical, which he (rightly) interpreted as evidence that these landmasses had once been located much closer to the poles. However, some pieces of apparent counterevidence threatened to undermine this picture, like the so-called 'Squantum 'tillite".[109] A tillite is a rock formed of sediment deposited by a glacier and consists of pebble-sized and larger rocks contained in a fine-grained matrix. It's central to the rock's classification that it's formed by a glacier, so it must have formed somewhere cold enough for glaciation. The Squantum 'tillite' is found just outside of Boston and dated to the Permo-Carboniferous, at a time when – according to Wegener (and today's view) – North America had a latitude between 10 and 30°: not a location known for being cool.[110] Wegener (and other Drifters; e.g., Holmes (1928:433)) questioned the interpretation of the data,

[109] See Wegener (1928:97–8) for a discussion of other apparent traces of glaciation in North America in the Permo-Carboniferous.

[110] Evidence from other landmasses showed that the whole planet couldn't have been cold enough for glaciation during this period.

arguing that it did not constitute evidence against drift in general or his historical reconstruction in particular, and he used both empirical evidence and fit with drift theory to argue against the classification of the Squantum rocks as tillites:

> We have a multitude of indications for the past geological climate in the United States, and although we admit that several details are uncertain, they yet match and *give us a harmonious major picture. In the main, it is clear and simple* that there is cumulative and concordant evidence that in the upper Carboniferous the United States was within the edge of the pluvial tropical belt, and during the Permian in the hot tropical desert zone. (1928:98–9, italics inserted)

The 'harmonious major picture' he extracted from *much* of the data was enough to overrule the contradicting data (like the Squantum 'tillite'). The evidence for the hot climate of North America during the Permo-Carboniferous was 'clear and simple' but *only after discarding the conflicting data* by arguing that it wasn't correctly interpreted and therefore didn't contradict his reconstruction of continental locations. It's pretty bold to prefer harmony and simplicity to fit with the data, yet they sometimes contain errors – as was indeed the case here: the rocks are not glacial deposits (tillites), but probably deposits from an underwater lava flow (Carto and Eyles (2012)). Wegner's aesthetic preferences worked like a counterbalance to misleading data, pushing him to explain away the data that didn't fit his harmonious major picture instead of (over)fitting his theory to it.[111]

Geomagnetic data. Like Wegener, many of the scientists involved in the development of plate tectonics eschewed a perfect fit with data for an elegant big picture. As mentioned, the discovery of geomagnetic field reversals was essential for understanding the zebra pattern and for developing plate tectonics. But observations were made early on that cast doubt on field reversals. For instance, Takeshi Nagata and Seiya Uyeda discovered that some rocks heated beyond their so-called Curie temperature took on the *reverse* polarity of the magnetic field they were in (Nagata (1952)).[112] If this phenomenon was widespread, then stripes of reversely magnetised seafloor wouldn't support the hypothesis of geomagnetic field reversals. The discovery of these 'self-reversing'

[111] See chapter 5 for more on how aesthetic virtues can counterbalance the risk of overfitting.
[112] See Heller and Petersen (1982:368–9) for more examples.

rocks, as they became known, "contributed greatly to the ongoing scepticism" of continental drift (Pitman (2003:89)). As Opdyke explains, "[s]ince both this newfound support for continental drift [palaeomagnetic evidence of apparent polar wandering][113] and the argument for field reversals rested on palaeomagnetic observations and were often argued by the same people, they were lumped together as radical and unreliable" (2003:98). That's unsurprising. What *is* surprising is that the scientists who were positively inclined towards drift didn't take the data on self-reversing rocks as undermining field reversals and thereby withdraw the support it gave drift.

This is arguably because the drifters valued aesthetic properties like simplicity and elegance higher than perfect fit with data. Faced with the choice between a simple, elegant, and unifying explanation and greater empirical adequacy, they chose the elegant explanation. As Irving and Runcorn said in a paper from 1957, the case for self-reversals "is not strong when reversals of the main field give a simple and general explanation for the varied phenomena observed" (1957:99). That is, compared to the simple and general (that is, strong) explanation, the evidence for self-reversals looked weak. The support provided by the explanation's simplicity and strength was weighted heavier than the evidence for self-reversals. Note that in this case, the interpretation of the data wasn't questioned (Heller and Petersen (1982)). The scientists were genuinely giving up empirical fit in favour of explanatory simplicity: the two cases of magnetic reversals (self-reversing rocks, on the one hand, and magnetic anomalies along spreading ridges and laboratory experiments on the magnetic properties of rocks, on the other) were treated as expressions of the same underlying phenomenon (remanent magnetism), and yet the data on self-reversing rocks were dismissed as 'rare' 'exceptions' (Cox, Doell, Dalrymple (1964:1539–40)) which therefore didn't threaten the elegant explanation.

The Vine-Matthews-Morley hypothesis. This willingness to dismiss what didn't fit simple and elegant hypotheses and extrapolate from what did fit them

[113] Magnetic data on 'apparent polar wandering' showed where the poles would have had to be located if the landmasses from which the data was collected had remained stationary: however, if one accepted that continents had drifted one could instead use it to recreate earlier positions of the landmasses relative to a stationary magnetic pole.

is also seen in the development of the Vine-Matthews-Morley hypothesis. As Menard puts it, "Fred Vine and Drummond Matthews had to believe three hypotheses simultaneously at a time when few scientists believed any one of them. They also had to disbelieve the only observation directly related to the hypothesis" (1986:212). One of these hypotheses was seafloor spreading,[114] to which there was apparent counterevidence in the form of old rocks found near midocean ridges. But as Harry Hess wrote in a letter to Menard in 1966: "if these are disregarded everything fits just as it should" (in Frankel (1982:36)).

Not everyone was convinced. Walter Pitman recalls his reaction to the hypothesis:

> in the end [it] proved to be a brilliant insight. But the data they presented, taken from over the Carlsberg Ridge in the northwest Indian Ocean, were not very impressive with regard to correlatability, linearity, and symmetry. I was sceptical and dismissive, as were most others. (2003:89)

Still, Vine and Matthews (and, independently, Morley) extracted one of the most aesthetically pleasing *and* successful hypotheses in all of earth science from these noisy and underwhelming data. How? A plausible explanation is that their ability to focus on the essentials only and thereby extract the correct big picture from ambiguous data was partially due to – or, an expression of – the central role that aesthetic considerations played in their theorising. Their aesthetic preferences acted as useful restrictions on their interpretations of the data and attempts to explain them (cf. Elgin (2020)).

Vine is unusually open about being guided by aesthetic considerations. He was converted to drift from an early age based on not much more than the perfect fit between the coastlines of Africa and South America, saying that "I decided it had to be true; it was too simple and elegant" not to be true (in Frankel (1982:12)). The data Vine and Matthews relied on lacked the clarity and symmetry that's so striking in Raff and Mason's zebra pattern maps and, even more so, in the *Eltanin*-19 magnetic profile (see section 4.3.2). Still, the

[114] The others were continental drift and widespread remanent magnetism.

inconclusive data from the Carlsberg Ridge combined with a nose for elegant explanations allowed him to develop and accept this celebrated hypothesis.[115]

This concludes the historical case study. There are many more episodes from the history where aesthetic considerations influenced theory evaluation, but a thorough discussion must be left for another time. What I hope is clear is that the history of plate tectonics provide many interesting examples of aesthetic considerations in science and that aesthetics influence research well beyond highly abstract fields such as theoretical physics and mathematics. This helps us get a realistic sense of the variety of aesthetic considerations in science. This case study also provides an additional case against which to test explanatory accounts: an account of aesthetics in science must be able to explain what is going on in the cases discussed above, just as it must explain what is going on when mathematicians aesthetically evaluate their theorems. The history of plate tectonics is, to use an earth science metaphor, a fruitful dig for examples of aesthetic considerations guiding research and it should be considered when trying to explain this puzzling aspect of scientific reasoning.

So what lessons should we take from this case study? We discuss that in the next section and chapter 5 below.

4.5. The epistemic significance of aesthetic experiences

There's recently been a debate about whether plate tectonics was accepted because of the evidence in favour of it or partially because of sociological factors (Šešelja and Weber (2012), Pellegrini (2019, 2022), Weber and Šešelja (2020)).[116]

---

[115] Compare Chandrasekhar on Einstein: "he arrived at his field equations by qualitative arguments of a physical nature combined with an unerring sense for mathematical elegance and simplicity" (1987:71).

[116] Šešelja and Weber and Pellegrini might be talking past each other: Šešelja and Weber wants to show that continental drift was worthy of pursuit (though not acceptance) in the first half of the 20th-century (basically, before the palaeomagnetic evidence was analysed and plate tectonics proper was formulated: after that, *plate tectonics* (not Drift) was worthy of acceptance). They explicitly focus on whether there was enough evidence in favour of it to justify pursuing Drift, not to what extent scientists were sensitive to this (2012:148). Pellegrini, on the other hand, wants to show that sociological factors such as

It might seem that this case study adds another reason to think that its development and acceptance was not sufficiently responsive to epistemic considerations. But that's not what we should take out of this case study. On the contrary, the historical evidence presented here supports the conclusion that aesthetic sensibilities can be helpful in developing predictively successful, strong theories that don't easily succumb to misleading data.

How can the pleasure we find in elegance and other aesthetic properties play such a positive epistemic role?

As we saw in the previous chapter, some philosophers propose that what these positive aesthetic feelings signal is non-factive understanding (Ivanova (2020), Elgin (2020), Kosso (2002)). I argued against non-factive accounts of understanding above and will not repeat those arguments here, but I want to stress that this case study suggests that aesthetic pleasure signals something stronger than non-factive understanding: it helped the scientists to *get it right*, providing valuable guidance in their efforts to explain our planet's large-scale geological structures. This goes beyond grasping or framing information without regard for its truth. Wegener, Vine, and others evaluated explanations in part by appeal to aesthetic theoretical virtues. This turned out to be a highly successful restriction on their theorising: allowing them to focus on the big picture and the central trends, and, relatedly, giving them a fruitful guideline for which data points to take seriously and what to ignore, which helped avoid overfitting the theory to misleading data.[117]

However, you'll remember from the last chapter that Milena Ivanova (2020) argues convincingly that it's hard to reconcile the idea that beauty can lead to truth with the historical track record of both beautiful but unsuccessful theories (like Copernicus', with its beautifully circular planetary orbits) and successful, ugly theories (like the standard model of particle physics).[118]

---

belonging to a research culture which is in favour of mobilism (the view that continents' locations are not permanent; this is not precisely the same as continental drift) also play a role in research communities' acceptance of a theory.

[117] Think of the Squantum 'tillites' and the young rocks found near midocean ridges. Overfitting is discussed in more detail in chapter 2 and chapter 5.

[118] See also Currie (2023) and Hossenfelder (2018).

A nice example of how scientists complain about the ugliness of the standard model while simultaneously praising its epistemic success comes from Michio Kaku and Jennifer Thompson (1997):

> At present, there has been no experimental deviation from the Standard Model. Thus, it is perhaps the most successful theory ever proposed in the history of science. However, most physicists find the Standard Model unappealing because it is exceptionally ugly and asymmetrical. … Because experimentally it is very successful, most physicists believe that the Standard Model is just an intermediate step toward a true theory of everything. The reason why the Standard Model is so ugly is that it is obtained by gluing, by brute force, the current theories of the electromagnetic force, the weak force, and the strong force into one theory. (Think of trying to jam three jigsaw puzzle pieces together that obviously do not fit. Tape can be used to connect the three pieces forcibly. That is the Standard Model.) (1997:75)

The point to note here is that there is sometimes a conflict between a well-supported theory (even "perhaps the most successful theory ever proposed") and our aesthetic values. What should we say to such cases? One option is to say, with Kaku and Thompson, that despite its empirical success, an ugly theory is very unlikely to be true: its empirical success gives us only reason to believe that it's a step in the right direction, not the whole truth. In many case I think this is the right response. But regardless of the details of the specific example, it is certainly possible for a theory to be both ugly and true.[119] A clear example of that is found in the aesthetics of mathematical proofs. You can of course not prove something that isn't true, and yet mathematicians have a great deal to say about how beautiful their proofs are (see Montaño (2012) and Glynn (2010) for examples and discussion.) So another option, which is the one I favour, is that the connection between aesthetic value and epistemic success is contingent, psychological, and no more than moderately reliable. Hence we should expect quite a few exceptions; cases where beauty and truth diverge. This is compatible with aesthetic value *often* being indicative of epistemic value.

The fact that aesthetic value and truth can diverge does not make truth irrelevant: as Adrian Currie notes, one can deny "that truth plays any direct role

[119] Indeed, you can make indefinitely many ugly, true (mini) theories by joining together arbitrary truths into one conjunction.

in constraining aesthetic judgement or appreciation (although no doubt, through epistemic engagement, the two are indirectly linked)" (2023:318, fn. 4).

If they are *indirectly* linked we can explain the fact that beauty turned out to be a useful guide in 20th century earth science while allowing that it's a fallible guide that sometimes leads us astray. In Currie's (2023) account, the link is the sustained, active, 'knowledge-directed' work (which he calls 'epistemic engagement') that 'attunes' aesthetic sensibilities for epistemic purposes – work, we may add, which tends to be a good way of gaining knowledge. An example of attunement is when budding scientists, by making field sketches, learn to "see like geologists" by drawing on and improving aesthetic skills of observation and representation (2023:325; see also Turner (2019), Wylie (2021)).[120]

While this is an important insight,[121] it's an account of the *refinement* of pre-existing aesthetic, non-epistemic sensibilities for epistemic purposes (2023:328). But pre-tuned aesthetic judgements can also be epistemically significant. For example, features that tend to be aesthetically pleasing – figure-ground contrast, simplicity, symmetry, and prototypicality – serve as cues for metacognitive feelings, which play a large role in evaluating and regulating cognitive effort. (We come back to this shortly, and discuss it at length in the following chapter.) And recall Vine's claim that he believed that South America had once fit with Africa because "it was too simple and elegant". This happened "when I was about fourteen or younger" based on reading one page of his school textbook (Frankel (1982:12); the textbook did not take a stand on the truth of continental drift). That gave Vine minimal opportunities for attunement, yet his confidence was already connected to aesthetic appreciation. This suggests that something more basic than attunement connects aesthetic and epistemic experiences.

Intriguingly, recently accounts from philosophical aesthetics (Armstrong and Detweiler-Bedell (2008), Cochrane (2021)), psychological aesthetics (Reber et al. (2004), Schwarz (2018)), 'neuroaesthetics' (Ramachandran and Seckel

[120] Compare Bird (2020), who argues that aesthetic appreciation of theories is based on recognising similarities with exemplars used in science education. We come back to this in chapter 5.
[121] Owing much to Turner (2019), as Currie acknowledges.

(2012)), and accounts focusing on aesthetics in mathematics (Todd (2017), Reber (2018)) have converged on the view that there is a deeper – though still psychological and contingent – link between aesthetic and epistemic judgements.[122] For example, Tom Cochrane sees aesthetic qualities as "exploit[ing] the psychological mechanisms that humans possess for the distal detection of something of potential practical value" (2021:19), in the way that taste helps us detect nutritional value before digestion. Beautiful objects exploit our mechanisms for detecting knowledge by exemplifying patterns that we recognise "will allow lots of details to be reconciled or predicted" (2021:42), giving us "the pleasure of things fitting together" (2021:43).[123] This doesn't mean that beauty always leads to knowledge; far from it. Thus there's no unequivocal connection between beautiful and true theories. However, since we might be rewarded with (aesthetic) pleasure we seek out those patterns that help us 'reconcile or predict' otherwise puzzling phenomena by embedding them in larger, simpler patterns (2021:51), potentially leading us to knowledge (by detecting those patterns) along the way. Applied to our case, their sense of beauty guided the scientists towards patterns (of continental movements, seafloor spreading, field reversals, etc.) that were central to explain 'lots of details', which is why this guidance was so successful.

This is compatible with the account of aesthetics in science I present in the following chapter. So without further ado, let us press on to consider what can explain the substantial and positive contribution aesthetic considerations made to plate tectonics, and, in general, the epistemic value of aesthetic considerations in science.

---

[122] See also Ivanova and Murphy (eds.) (2023) and Kozlov (2023) on how aesthetic and epistemic considerations are connected in scientific experiments.
[123] Compare Ivanova: "[Feynman] argues that for a physicist beauty is felt when one can grasp the 'pattern' of nature" (2017:2587).

When I appreciate the elegance of a theory, I do not have to choose between appreciating its elegance aesthetically or appreciating it epistemically; for in doing the one I am, *ipso facto*, doing the other, and vice versa.
– Peter Kivy (1991:186)

The pleasure of understanding in mathematics and in other aesthetic contexts seem often to have something to do with the notions of harmony or fit rather than fluency *per se*.
– Cain Todd (2017:228)

CHAPTER 5:
THE METACOGNITIVE ACCOUNT OF AESTHETICS IN SCIENCE

As we saw in the previous chapter, scientists regularly appeal to broadly aesthetic qualities such as simplicity, elegance, or beauty when evaluating theories. We also saw that this apparently contributed positively to the epistemic success of plate tectonics: in many cases, the beautiful theory turned out to be the best theory and, as the North American geologists unfortunately discovered, ignoring aesthetic pleasure can rob you of an important source of information about a theory's epistemic value.

In this chapter, I present an account which explains (i) why aesthetic considerations appear in scientific reasoning in the first place and (ii) how they can be epistemically valuable. The guiding thought is that it *feels good* to gain insight or solve a problem, all the more so if the solution seems particularly apt or comes surprisingly easily. The 'metacognitive account of aesthetics in science' relates scientists' aesthetic experiences to epistemic theoretical virtues. In brief, it says that the positive feelings accompanying beautiful theories signal the presence of epistemic virtues. This is because these feelings are generated by metacognitive processes which respond to certain epistemically relevant features of theories and of the scientist's cognitive engagement with the theory and evidence. (The details are spelt out in the rest of this chapter.)

In many cases, the confidence scientists have in beautiful theories is well-placed: for example, an elegant theory is less likely to be overfitted to noise in the data or inconsistent with background knowledge than a messy theory is.

However, it is important that we do not overstate this point. Metacognitive feelings are no more than moderately reliable indicators of one's epistemic situation, and metacognitively generated aesthetic delight can also be seriously misleading, for example when one's relevant background beliefs are mostly false, which renders the pleasure generated by how well they fit the theory worse than useless.

I present the metacognitive account in more detail in section 5.2 below. Before we get there, we should consider the constraints on any acceptable theory of the role of aesthetics in science; I discuss this in section 5.1. Here I also discuss Kivy (1991) and Bird (2020), who provide different ways of meeting the desiderata from what I do here. In section 5.3, I explain what metacognition is, before turning to the connection between aesthetic theoretical virtues and metacognitive cues in the following sections. I describe the epistemic value of the broadly aesthetic virtues of simplicity and elegance in section 5.4, and, in section 5.5, how features of simple and elegant theories can function as metacognitive cues. Working out the details for all broadly aesthetic virtues is beyond the scope of this thesis; the details on these two central virtues will stand as proof of concept. On this account, the epistemic value of aesthetic experiences in science depends on the epistemic value of metacognitive feelings; I discuss what we know about the accuracy of judgements based on metacognitive feelings in section 5.6, before tying it all together and explaining how scientists can be justified in relying on aesthetics in theory evaluation (section 5.7). I end this long chapter by contrasting my account with alternatives also connecting metacognitive and aesthetic feelings (section 5.8).

5.1. Desiderata for accounts of aesthetics in science

The role of aesthetics in science cries out for explanation: why do aesthetic considerations play a role in scientific research at all? It certainly flies in the face of the commonly held belief that scientific reasoning is coldly rational and that there is a sharp distinction between the supposedly subjective nature of

aesthetics and the objective nature of science.[124] As Derek Turner puts it, "[m]any people think that aesthetic value is largely subjective; beauty is in the eye of the beholder, and all that. … The very suggestion that scientific practice has an aesthetic dimension might seem like a threat to the rationality and objectivity of science" (2019:15). Even if one is on board with the view, common in contemporary psychology, that there is no sharp distinction between cognitive and affective processes (e.g., Okon-Singer et al. (2018), Pessoa (2022)), and the further view that emotions are central to good scientific practice (Thagard (2006)), it is still puzzling that *aesthetic* affect should play a role in science.[125]

But what is even more puzzling than the fact that aesthetic considerations are integral to science is that they are apparently integral to *good* science: how can it be that our aesthetic preferences in many cases help guide us towards better theories? In many cases, but far from all: there are many beautiful but false theories and true ugly ones.[126] Aesthetic considerations are only loosely connected to epistemic value, often helpful but sometimes detrimental to scientific research. An account of aesthetics in science should explain why

[124] Helen Longino's (1990) is the classic discussion on scientific objectivity and values, arguing against the view that science must be value free to be objective. Her focus is on political and social values, not aesthetic ones. Whether aesthetic values are objective or subjective is a contentious question. The classic discussion is David Hume (1757); see also Louise Hanson (2018), who argues that aesthetic realism – the view that there are "robustly mind-independent" aesthetic facts – is as defensible as moral realism.

[125] In Miyata-Sturm (2024a), I argued that starting from the psychological research on affective cognition puts us in a good position to explain the connection between *aesthetic* affect and epistemic value; see that paper for discussion and references to the psychological literature on the interplay between affect and cognition. There I argued that we should treat the role of aesthetics in science as a special case of affective cognition and presented an earlier version of the metacognitive account. The present account also denies that there is a sharp distinction between affective and cognitive states and processes, but motivates the hypothesis that aesthetic feelings in science are metacognitively generated with commonalities between metacognitive and aesthetic features (described below, see esp. sections 5.4. and 5.5.) rather than via a general relationship between affect and cognition.

[126] See chapter 4 above for examples of both positive and negative influence of aesthetics on scientific reasoning.

aesthetic evaluation is involved in scientific reasoning at all and why it is moderately epistemically valuable.[127]

There are various ways of meeting these desiderata. In chapter 3, we discussed accounts which say that aesthetic virtues are conducive to non-factive understanding. This is one way of trying to secure the required loose link between aesthetic virtues and epistemic value: since, on these accounts, understanding is non-factive, aesthetic virtues are not directly linked to truth. But since they help us 'grasp' relational 'facts' – where neither the grasping nor the 'facts' concerns how the world is, as that would result in a *factive* account of understanding – they are supposed to help us attain an epistemic goal of science.

Other accounts focus more on why aesthetic evaluation is involved in scientific reasoning at all. In particular, on some accounts aesthetic and explanatory virtues are identical, though perhaps viewed under different modes of presentation: they are two ways of looking at the same phenomenon. This would explain why aesthetic considerations appear in science in the first place: they are just explanatory considerations. (Why they appear under the guise of aesthetics still needs to be fleshed out: we come back to that when discussing Bird's (2020) account in the next section.) In chapter 2, I mentioned that 'theoretical virtue' and 'explanatory virtue' are mostly used interchangeably; explanatory virtues are whatever makes an explanation good, and the theory which scores highest on these virtues is the one IBE urges us to accept as (probably approximately) true. In other words, an abductivist will classify these virtues as at least relevant to the epistemic value of the theory.

In the next section, I discuss two accounts which provide some general reasons for thinking that the boundary between aesthetic and explanatory virtues is porous. While I agree that aesthetic and explanatory/epistemic virtues overlap in important ways, ultimately, I do not think either of these accounts can explain why aesthetic considerations are so prevalent or so useful in science.

---

[127] The second desideratum subsumes Breitenbach's (2013) two desiderata for an account of aesthetics in science: that it "can account for the apparently unstable link of beauty and truth, while also going some way towards showing that this link is not altogether arbitrary" (2013:89).

5.1.1. "Aesthetic criteria … of explanatory success": is explanatory power an aesthetic virtue?

Peter Kivy (1991) wrote that

> when one is asking the grand questions in theoretical physics, or cosmology, or evolutionary biology, it is difficult *not* to see the quest as aesthetically motivated and the answers as aesthetically satisfying, since what I have been calling the Classical aesthetic criteria in science [symmetry, simplicity, order, unity, coherence, elegance, harmony] are some of the very criteria … of explanatory success, where the explanations are at the high theoretical level. (1991:189)

That is, the aesthetic theoretical virtues are the very same features which makes for explanatory success, i.e. epistemic success.[128] Whether we call them *aesthetic* or *explanatory* virtues is only a matter of how we label them.

What are the reasons Kivy gives for thinking that aesthetic and explanatory virtues coincide? The fundamental reason is that if we reject aesthetic formalism, which he thinks we should, then the motivation for postulating a sharp distinction between aesthetic and epistemic concerns disappear. Certain virtues appear aesthetic, by normal criteria for judging whether something is aesthetic (1991:185–6), but they also appear epistemic. "I have often been told," Kivy says, "by philosophers of science with whom I have informally discussed my ideas that these features cannot be functioning aesthetically in science since, clearly, they are functioning *epistemically*" (1991:186). Why choose? Taking these virtues as being both aesthetic and epistemic accounts for all the data, and insofar as the alternative is only motivated by an implausible account of aesthetics (formalism),[129] we should accept that aesthetic and epistemic virtues are two sides of the same coin.

Using the metaphor that science is like (realist) representational art (such as realism in painting), he says, it becomes clear how aesthetic virtues are epistemically valuable:

---

[128] At least for explanations at a "high theoretical level", as opposed to "the level of such questions as 'How does a thermostat work?'" which "seem to have little to do with aesthetic appreciation" (*ibid.*)

[129] Though note that Kivy's view may be compatible with moderate formalism, discussed in section 2.3.1.

> aesthetic appreciation of a representational form cannot be prised apart from appreciation of its representational success. When I appreciate the elegance of a theory, I do not have to choose between appreciating its elegance aesthetically or appreciating it epistemically; for in doing the one I am, *ipso facto,* doing the other, and vice versa. (1991:186)

In other words, since the extent to which a representational work accurately represents its target is part of its aesthetic value, aesthetic and epistemic evaluation are not separable. Just as the extent to which a portrait (when evaluated according to the standards of realism, as opposed to, say, Cubism) is beautiful is partially a matter of how well it represents the sitter, so the beauty of a theory is partially a matter of how well it represents nature.

Note that this isn't a claim about representing beautiful nature truthfully: a portrait of an ugly person can be beautiful partially in virtue of its truthful rendering of their ugliness, and a theory can be beautiful partially in virtue of its truthful rendering of a messy world. But the theory itself shouldn't be messy. Why? Here Kivy seems to require that aesthetic theoretical virtues correspond to aesthetic properties of the target phenomenon after all. He compares aesthetic virtues with linear perspective in paintings, saying that "it may be plausibly thought of as a feature always contributing to realism where it is present … The same might be said, *pari passu*, of the … aesthetic features in science. They are neither necessary nor sufficient, separate or together, for truth. But they always count towards truth where present, and are always of positive aesthetic value" (1991:187). However, perspectivism gets its (representational) value from accurately depicting facts about distance and other spatial properties, so if aesthetic virtues are like this they seem to boil down to accurate depictions of nature. It seems to follow that on Kivy's view, simplicity (say) makes a theory beautiful because the world is simple, or at least because some structural features of the world are simple, just as correct perspective makes a painting beautiful because it shows the world as it is. But a simple theory of a complex phenomenon can be aesthetically attractive, so this doesn't quite capture the aesthetic value of theories (though it may contribute to it). (Compare: the beauty of a portrait does not boil down to the beauty of the sitter, though the latter may contribute to the former.)

Another general reason for thinking that aesthetic and epistemic virtues are two sides of the same coin is found in Alexander Bird's (2020) 'exemplar account'. This is a development of Peter Lipton's (2004) account of 'explanatory loveliness' (i.e., theoretical virtuousness). Lipton cashes out loveliness as the property of being the explanation (or theory) which would, if true, give us the most understanding. As will be remembered from chapter 2, the loveliest theory is the one with the most, or the best balance of, theoretical virtues. Bird takes loveliness to be identical with or at least similar to aesthetic qualities (2020:126–7). According to Bird, theory evaluation is based on perceived similarity between exemplars of successful theories and the evaluated theory. This perceived similarity is what makes a theory lovely: exemplars are models of what a good explanation looks like, so explanations 'feel good' to the extent that they look like the exemplars.

Why think of explanatory loveliness as an aesthetic quality? Some of the reasons he lists are, first, that we use aesthetic terminology – such as 'loveliness', 'beauty' and 'elegance' – to describe what's good about theories, and, second, "[w]e take a certain kind of *pleasure* on grasping a particularly revealing, unifying, apt, or economical explanation. This pleasure is the same as or is very similar to the kind of pleasure one has on experiencing (with understanding) an artistically valuable poem, sonata, or still life" (2020:127). In other words, the intellectual satisfaction which characteristically accompanies lovely explanations is (very similar to) aesthetic pleasure.[130] We describe epistemically good theories – *lovely* theories – in aesthetic terms in part because this is the kind of pleasure and satisfaction we recognise from paradigmatically aesthetic contexts. This would explain why aesthetic considerations appear in scientific contexts: there isn't an additional layer of aesthetic evaluation modifying the epistemic evaluation, just two ways of describing epistemically good theories.[131]

---

[130] Note the difference between taking broadly aesthetic virtues to be part of what makes a theory epistemically virtuous, as I do, and taking explanatory loveliness to be broadly aesthetic, as Bird does.

[131] Doesn't this make the connection between epistemic and aesthetic value too close? No: "An apt analogy with an exemplar that is the correct solution to a problem can be conducive to finding the correct solution to a new problem. But this … does not exclude the possibility of systematic error. If the analogy is inapt or the exemplar is itself an error, then thinking with exemplars will lead to false

It's not clear how the exemplar account can account for specific aesthetic virtues such as elegance, simplicity, and beauty. If the aesthetic appreciation of a theory is a matter of recognising its similarity to other successful theories, what role does (say) elegance play? One could say that an elegant theory is one which is similar to an *elegant* exemplar theory, but then isn't what makes the theory aesthetically virtuous that it (too) is elegant, rather than that it is pleasingly similar to an exemplar? And if so, isn't it the elegance, not the similarity to another elegant theory, which is the basis for its virtuousness? That fits better with many historical cases: for example, the theory of continental drift was recognised as an elegant theory although it didn't look like any other geological theory at the time.

Furthermore, while I think exemplars of successful theories – or, more generally, experience with epistemically successful theorising – can play an important role in refining or training scientists' aesthetic sensibilities, this does not go to the heart of the issue. As briefly discussed in chapter 4, Adrian Currie (2023), Derek Turner (2019), and Caitlin Wylie (2021) all discuss various ways in which aesthetic sensibilities and artistic abilities are honed in science education in epistemically useful ways. The training of fossil preparators, for example, allows them to develop and fine-tune a sense of when a fossil is well-prepared, which is simultaneously an aesthetic and epistemic judgement. As Wylie says, "[fossil] preparators draw on creativity, sensory judgment, and notions of beauty to craft evidence from nature" (2021:204). Exemplars of successful theories probably play a similarly large role in shaping scientists' judgements of what a beautiful and epistemically successful theory looks like, but it doesn't explain the epistemicity of untrained aesthetic experiences (like how the simplicity and elegance of continental drift convinced the young Fred Vine of its truth; see section 4.4.1).

We need an account of aesthetics in science which goes deeper and bases the explanation of why aesthetic experiences appear in science on an aspect of scientific reasoning which is not wholly dependent on scientists' training. Something more fundamental connects scientists' aesthetic experiences and

---

beliefs" (2020:131). In other words, a theory's loveliness can be misleading, but it can also help us to find new, good explanations by building on what has worked before.

their epistemic endeavours, and so we should be looking at a more basic aspect of scientific reasoning: some feature of human cognition which is refined in science education but not wholly based on it. Hence the metacognitive account, to which we now turn.

5.2. The metacognitive account of aesthetics in science

The metacognitive account takes aesthetic feelings to be feedback from one's cognitive engagement with a theory and the evidence that bears on it. When cognitive processes unfold smoothly, we experience pleasure, satisfaction, and confidence. Reflecting the broadly aesthetic nature of the theoretical virtues which makes for effortless cognition (e.g., elegance or simplicity), this is often experienced as aesthetic pleasure.

In a little more detail (I describe the nuts and bolts shortly): the pleasure, satisfaction and confidence experienced when working with a beautiful theory are metacognitive feelings, generated by processes that monitor the quality and success of the cognitive process involved in working with the theory. These feelings reflect properties such as the fluency, or effortlessness, of the cognitive processes; fluency, in turn, tends to reflect epistemically valuable features of the theory being worked with, such as its simplicity or coherence. Thus the feelings we experience as aesthetic are metacognitively generated, and they serve as rough psychological signals that the theory is epistemically virtuous (by being simple or coherent, for example: I come back to what makes these virtues epistemic shortly).

The following flowchart illustrates the relationship between the properties and processes involved in bringing about aesthetic experiences in theory evaluation:

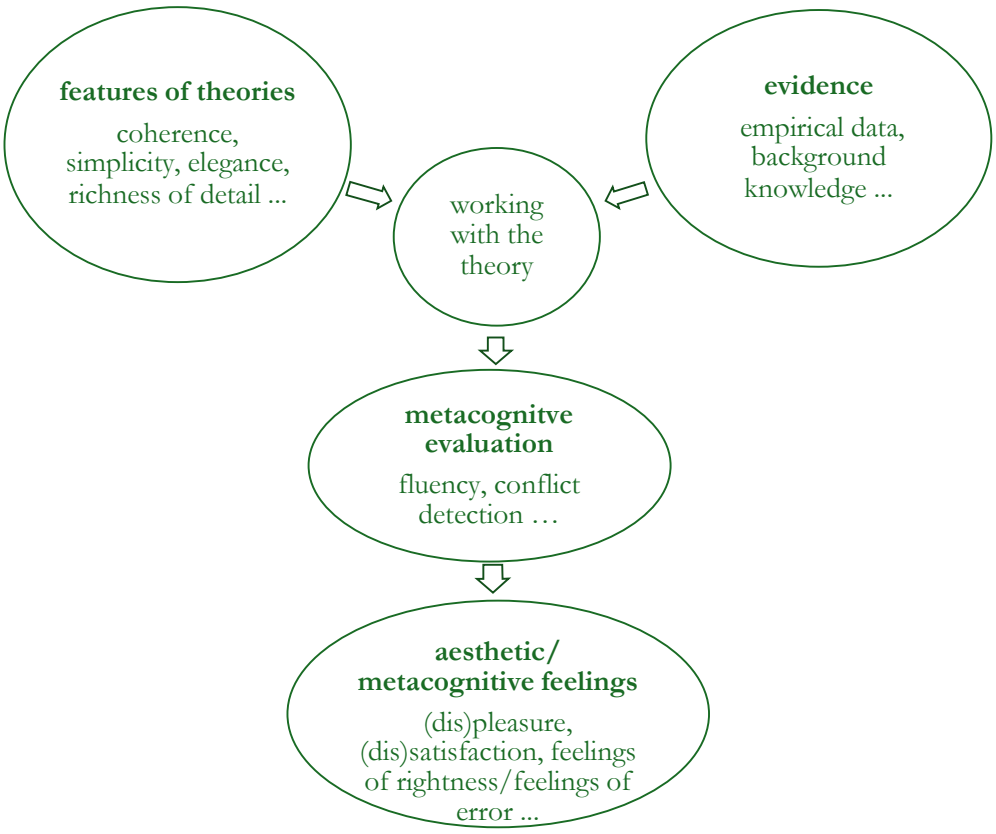


Figure 1: The Metacognitive Account. The pleasure, satisfaction and confidence felt about beautiful theories are metacognitive feelings, generated by cognitive processes that monitor the first-order cognitive engagement with theory and evidence. These feelings reflect properties such as the fluency of the cognitive processes involved. Fluency, in turn, tends to reflect epistemically valuable features of the theory being worked with, such as its simplicity or coherence, allowing aesthetic feelings to serve as a rough psychological signal that the theory is epistemically virtuous.

Typically, when we consider a theory, we do so in the light of evidence: we evaluate it as an explanation of observations, say, and against the background of what we already know, as opposed to evaluating its intrinsic features alone. This is the case also before we have started systematic evidence-gathering: when scientists start theorising about a phenomenon, they already have some knowledge about it – if only the facts which made them interested in looking closer in the first place ("what is this fuzzy, mould-like growth?") – and some background knowledge of what the world is like. Thus, when working with a theory – whether explicitly evaluating it, using it to explain something, deriving predictions from it, play around with it by tweaking it this way or that, etc. – scientists have not only the theory but some pieces of evidence in mind. As I

come back to in a moment, this means that the psychological processes monitoring one's engagement with a theory also monitor engagement with the evidence (and how they combine). Such *metacognitive* processes monitor and evaluate the quality and success of ongoing first-order cognitive processes, including those engaged when working with a theory, such as evaluating, reasoning, remembering, and making decisions. Metacognition monitors how likely the mental action is to succeed (allowing you to change strategy if it's deemed unlikely to succeed), how well the processes unfold (allowing you to keep track and adjust your efforts if necessary), and whether they were in fact successful (e.g., Ackerman and Thompson (2017), Proust (2013)).

Metacognitive monitoring is based on automatic, subpersonal inferences from cues and heuristics from the processing (here: from working with the theory).[132] These cues include the fluency with which the cognitive processes unfold, whether there are any conflicts or tensions between the items considered, and whether the majority of items support one decision. Fluency can be based on many factors, including simplicity, the availability of supporting reasons, and one's familiarity with the topic. These cues reflect the dynamics of the processing – how it unfolds, whether it goes smoothly or encounters problems – without much regard for the content of the represented items. A conflict between items is a cue that something has gone wrong regardless of whether the conflict is between the theory and evidence or between different parts of the theory.

Strikingly, we recognise many of the aesthetic qualities of theories from psychological studies, where the same properties figure as cues for metacognitive monitoring. Simplicity and symmetry, for instance, make for fluent processing, and the same is true for unity and coherence (or harmony, as such goodness-of-fit would be called in aesthetics). Empirical accuracy (that the theory fits observations and experimental results), internal coherence (that the theory is free of contradictions), and external coherence (that the theory fits our background knowledge) all allow the cognitive processes to flow smoothly without snagging on a conflict between the items represented, triggering the feelings of

[132] More precisely, procedural or type 1 metacognition is like this: see section 5.3. below.

effortlessness typical of elegant theories. It is difficult to gain an overview of *dis*unified theories, making them cognitively demanding to work with and triggering feelings of displeasure.

This leads us to metacognitive and aesthetic feelings. Metacognitive monitoring goes on in the background, automatically instigated and largely outside of conscious control. We are made aware of such monitoring through feelings: for example, feelings of confidence or feelings of rightness when everything goes well, the feeling that something is wrong when a problem is detected, feelings of knowing (the answer) when asked a question, or feeling that you have a word on the tip of your tongue. My hypothesis is that the pleasure and satisfaction which are central to the aesthetic evaluation of theories are also metacognitively generated.[133] Such feelings reflect the quality and success of our cognitive engagement with the theory and evidence, signalling (for example) that a smooth solution has been found. The pleasure (or displeasure, as the case may be) is a signal of the quality and success of one's cognitive engagement with the theory, which thereby also tells us the extent to which we should trust our evaluation of the theory (if we are reasoning well, we are more likely to arrive at an accurate evaluation of it). The metacognitive feelings are experienced as aesthetic because of the aesthetic value of some of their cues (such as the theory's elegance or its ability to unify various phenomena in a harmonious way) and because the pleasure we find in intellectual pursuits and in the structure of ideas is not much different from the pleasure we find in artistic pursuits and sensory experiences.

Moreover, since these feelings are triggered by epistemically valuable features of theories, they signal the presence of such features. In other words, they can signal epistemic value. For example, beautiful theories are often

---

[133] A slightly more complicated hypothesis is that the aesthetic feelings are downstream consequences of metacognitive feelings: they depend on, but are not identical to, metacognitively generated feelings (such as feelings of fluency or rightness). (Thanks to Nick Shea for the suggestion.) The account I present here is programmatic and it is likely that some adjustments will have to be made – including, perhaps, abandoning the claim that aesthetic feelings are metacognitively generated in favour of the claim that they follow from metacognitively generated feelings. For now, however, let us see how much we can explain with the simpler hypothesis.

compelling, and feelings of confidence are triggered when most of the information considered pulls in the same direction. This is correlated with accuracy in many real-world situations.[134] In the case of theory evaluation, we can explain the correlation between confidence and accuracy with reference to the epistemic value of the triggering features. For example, simplicity signals that the theory is not overfitted to noise in the data; the rich and detailed picture of the world entailed by a strong theory suggests it would give us more knowledge, if true; and if it fits well with what we know from other areas it's more likely to be true. Thus, the metacognitive account brings together previous work in philosophy of science and psychological research on the role of feelings in reasoning and decision-making to explain how the positive affect accompanying beautiful theories can function as an epistemically useful guide in theory evaluation. But we need to keep in mind that metacognitive feelings are only moderately reliable: they are accurate in many cases, but by no means in all. So although these feelings can signal epistemic value, the signal is weak and easily corruptible. Hence aesthetic feelings are mildly epistemically valuable: often useful in scientific reasoning, but also quite often misleading.

There is a lot to unpack here. Let us start by looking in more detail at metacognition and metacognitive feelings.

5.3. Metacognition

Metacognition is cognition of cognition: metacognitive processes monitor, evaluate, and control object-level cognitive processes (Proust (2013:4), Ackerman and Thompson (2017:607), Shea et al. (2014:186)). For example, when you try to recall the route to your hotel in a new city, the remembering is object-level cognition and your confidence that the remembered route is correct is a meta-level evaluation of it.

Some well-studied metacognitive phenomena are judgements of learning (JOL), feelings of knowing (FOK), tip-of-the-tongue states (TOT), feelings of

[134] See discussion and references in section 5.6 below.

confidence/feelings of rightness (FOR), and feelings of error.[135] JOLs are about predicted recall: it's an evaluation of how likely you are to remember (recall or recognise) some piece of information. FOKs and TOTs are about information retrieval in the moment, not a prediction of how likely you are to remember in the future. A high FOK signals that you can retrieve the requested information. Despite being called a feeling of *knowing*, FOKs are at best understood to be about the accuracy of the retrieved information (Schwartz (1994:359)).[136] TOTs also signal that one has the information stored in memory, but unlike FOKs they arise when attempts at retrieval fail (Schwartz and Metcalfe (2011)).[137] Sometimes these experiences are coupled with partial recall ("the name starts with a 'B'…"). TOTs signal that it's worth investing some effort into trying to retrieve the information: you probably have it stored, although you can't access it at the moment. It's somewhat unclear whether feelings of rightness and feelings of error are the same as high and low confidence, respectively. The terms are often used interchangeably, and for our purposes it will be harmless to assume that there's nothing deeper here than different labels for the same phenomenon. High confidence in a cognitive process signals that it's going right, whereas low confidence signals that it's going awry, that there is an error somewhere. "It is also not clear whether a Feeling of Error is qualitatively different from a low Feeling of Rightness or whether they represent two ends of a single dimension

---

[135] Experimental work on metacognition trace back to Joseph Hart's work on the accuracy of feelings of knowing in the 1960s and John Flavell's work on meta-memory in the '70s. Flavell coined the term 'meta-memory', which then gave rise to the more general 'metacognition' (Dunlosky and Tauber (2015:8–9)).

[136] However, in the psychological literature it is not always required that the retrieved information be true, taking us even further from knowledge. For example, according to Koriat's 'accessibility model', "both correct and incorrect clues contribute to the enhancement of the FOK. What matters is the overall accessibility of information, that is, the number of clues accessed and the ease with which they come to mind" (2000:159). However, "FOK judgements are relatively accurate in predicting actual memory performance [i.e., the accuracy of the retrieved information] … [because] [i]nformation that comes to mind during retrieval is more likely to be correct than wrong" (2000:160). This will not make a difference to our concerns. See also Schwarz and Metcalfe (2011:6).

[137] TOT and FOK are sometimes used interchangeably (e.g., Irak et al. (2019:239), Metcalfe et al. (2017:2), Arango-Muñoz and Michaelian (2014:99–100)).

of certainty," as Ackerman and Thompson (2017:612) say, but nothing I say here will turn on which of these views is correct.

Adult humans can deliberately attend to and direct their cognitive processes. For example, I may explicitly consider how confident I am that the number sequence which comes to mind is the PIN code I am trying to remember, and remembering that I almost never get my PIN wrong I infer that it's likely to be correct again this time. This is known is 'deliberate', 'declarative', 'analytic', or 'system 2' metacognition, paralleling the 'system 1' and 'system 2' distinction in dual-process theories of cognition according to which one mode of processing is deliberate, slow, and computationally costly (system 2) and the other is automatic, fast, and with a large capacity (system 1) (Goupil and Proust (2023)).[138] We also have what is variously known as 'procedural', 'automatic', 'core', or 'system 1' metacognition. For example, in evaluating whether q follows from $(p \rightarrow q)$ I may simply get the strong feeling that it does: I feel confident about the inference (Shea (2024)). Procedural metacognition is based on online feedback from cognitive processing, automatically instigated by subpersonal processes, and typically outside or at the periphery of conscious awareness (Shea et al. (2014), Goupil and Proust (2023), Proust (2013), Arango-Muñoz and Michaelian (2014), Beran et al. (2012), Koriat (2012b)). As Nicholas Shea puts it, "people have epistemic [i.e., metacognitive] feelings about many cognitive processes: perceptual decisions, memory recall, action execution, and so on. These forms of 'procedural' metacognition concern a cognitive process, and affect its results, without requiring the thinker to deliberate on or conceptualize the process" (2024:2).

How can feedback from the process of evaluating, deciding, remembering, etc., monitor the quality of a decision or the accuracy of a retrieved memory and present this in a way useful for cognitive control? There are two main parts to that story: cues and feelings. A cue is a stimulus or signal that can guide behaviour: something which prompts or triggers a response. Metacognition is based on (triggered by) cues and the upshot of the evaluation

[138] For classical works on dual process theories, see Tversky and Kahneman (1974) and Evans (2007).

is conveyed through feelings. This is what Asher Koriat calls the 'experience-based approach' of cue-utilisation accounts of metacognition:

> Unlike information-based approaches, which emphasize the content of domain-specific beliefs and knowledge retrieved from memory, experience-based approaches focus on the contribution of mnemonic cues that derive on-line from task performance. These cues are assumed to give rise automatically and unconsciously to a sheer metacognitive feeling. (Koriat (2012b:214))

(Information-based metacognition corresponds to what I above called deliberate metacognition.) The basis for experience-based metacognition are more-or-less valid cues from the processing experience, and their outputs are metacognitive feelings.

The consensus view in the psychological literature on metacognition seems to be that metacognition is cue-based (Koriat (1997, 2000), Dunlosky and Metcalfe (2009:62–6), Undorf et al. (2018)), involving inferences from cues from the processing experience.[139] As Undorf et al. (2018) puts it, "people do not have privileged access to their cognitions. Instead, people infer the state of their own cognitive systems using cues and heuristics" (2018:507).

These cues can be both features of the object(s) considered and features of our cognitive engagement with those objects. For example, if asked how likely it is that you'll remember a list of words – say, 'rain, shower, droplets', or 'magpies, men, curtains' – the relatedness of the first set ('semantic coherence') is used as diagnostic for future recall (Topolinski and Strack (2009)).[140] Rain, shower, and droplets are all connected to water, whereas the second set has no such unifying theme: this makes the first easier to remember than the second.

---

[139] These inferences are subpersonal in the case of procedural metacognition but can be deliberately instigated and attended to in the case of analytic metacognition (e.g., by remembering that you almost always remember your PIN code). As Carruthers and Richie say, "[s]ince no one in the human metacognition literature thinks that monitoring is direct [as opposed to cue-based], we propose to dismiss this possibility in respect of animals also" (2012:82). The alternative to cue-based inferences, that metacognitive judgements are based directly on the content of the object-level cognition, contradicts some robust experimental results, for example that a feeling of knowing can appear before the answer to a question is retrieved (and so cannot be based on direct access to the relevant memory); see Reder (1987)).

[140] Probably through its effect on fluency: see below.

Other metacognitive cues depend on how one relates to the object(s) that the first-order cognitive processes are directed towards. For example, the speed with which something comes to mind, which is a measure of cognitive fluency (or effortlessness), is a cue for the accuracy of the recalled information. Fluency can be brought about in many ways, for instance by repeated exposure, a clear font, or a simple task, which means that it both responds to features intrinsic to the object(s) of the first-order cognitive processes and to features of those processes. The fluency with which one processes information plays a central role in metacognition, as we get back to shortly.

The outcome of the metacognitive monitoring is conveyed through metacognitive feelings like feelings of error, confidence, and tip-of-the-tongue feelings. Feelings are positively or negatively valenced experiences that can vary in intensity from almost imperceptible to overwhelmingly intense. What sets *metacognitive* feelings apart from other feelings is that they concern one's cognitive states (Goupil and Proust (2023), Arango-Muñoz and Michaelian (2014)). For example, a tip-of-the-tongue-state is the output of an evaluation of whether you have the relevant information stored in memory, and the feeling signals that you do have the information stored even though you can't access it at the moment.

We are generally much more sensitive to the metacognitive feelings than to the *source* of these feelings (Schwarz (2018)) and this sometimes leads us astray. We may, for example, fall prey to the 'illusion of truth' effect which can make us believe that something is more likely to be true if it is written in a font that is easy to read (Dechêne et al. (2010)) (see below). However, "it must be emphasised that metacognitive judgements are often quite accurate … [and that] metacognitive illusions are the exception rather than the rule" (Undorf et al. (2022:309)). We discuss the accuracy of metacognitive judgements in section 5.6.

Before we get back to how metacognition is relevant to explaining the role of aesthetics in scientific reasoning, which involves some of the most complex reasoning humans engage in, I should mention that the empirical results discussed here are typically from lab studies of relatively simple cognitive processes. Furthermore, much of the psychological research on metacognition is focused on meta-memory and judgements relating to learning, memorisation, and knowledge retrieval in educational settings and on confidence in perceptual

tasks (e.g., regarding eyewitness testimony), but what is most relevant here is metacognitive monitoring of complex cognitive tasks such as judgements, complex inferences, and decision-making. The available evidence suggests that results from research on the simpler processes extend to these complex processes, for instance in that metacognitive processes are cue-based subpersonal inferences which are moderately reliable indicators of one's epistemic situation (Ackerman and Thompson (2017); see also Shea (2024), Fields and Glazebrook (2020), Ackerman (2014), Thompson et al. (2011)). Moreover, as these complex processes involve many of the simpler processes (e.g., following a long chain of reasoning requires remembering previous steps), research on the latter are directly relevant to the former.

5.4. The epistemicity of aesthetic virtues

Why does metacognition matter to the aesthetics of science? Recall that the metacognitive account says that beautiful theories are compelling because they make for effortless cognitive processing. Reflecting the broadly aesthetic nature of the features of theories which support cognition, such as simplicity and unity, the feelings reporting the outcomes of this metacognitive monitoring are experienced as aesthetic.[141] A hypothesis of this account is that for the most part, aesthetic pleasure and feelings of rightness concerning a theory coincide and will be experienced as an overall feeling of delight in the theory which is not separable from its perceived epistemic success.

---

[141] A different reason for thinking that aesthetic feelings are metacognitive is found in Dokic (2016). He argues that aesthetic experiences are (roughly) a way of organising non-aesthetic attitudes which is self-sustaining, i.e. it motivates one to stay in the experience (2016:74). "[A]esthetic experience is not intentionally about aesthetic properties or values, and is not a separate aesthetic attitude, but involves a unique combination of epistemic or metacognitive feelings having to with familiarity and novelty" (2016:69). While I think there is much to commend Dokic's view, it is more demanding than the view I present here as it requires a specific view of aesthetic experiences (an adverbialist account). More importantly, I do not think it can account for the variety of aesthetic experiences familiar from both science and other aesthetic contexts as it appeals to only fluency and disfluency. See section 5.8.1. for a comparison between a variant of his view developed for aesthetics in science (Arcangeli and Dokic (2020)) and the metacognitive account I present here.

In the cases where they do come apart (e.g., when scientists lament that a theory is 'so beautiful it's a shame it's not true'), I take it that the theory seems beautiful when isolated from the disconfirming evidence and that this sense of beauty lingers even when you take all the evidence into account and thereby notice the unpleasant cognitive conflict. The theory provides what *could have been* an elegant explanation if only the world had cooperated, but alas, it does not. Whether such beautiful but false theories lose their aesthetic attraction over time (McAllister (1996)) or whether we continue to experience them as beautiful is a contentious issue. It may well vary, perhaps depending on how salient the refuting evidence is. Copernicus' circular planetary orbits probably seem less aesthetically satisfying now than they did before the scientific community accepted Kepler's elliptical orbits, though I wager that most people still find circles more aesthetically pleasing than ellipses. I expect this to be because the conflict with background knowledge is so difficult to bracket off (i.e., too salient): we now know too much for the simplest model to feel appealing, since we know it cannot account for the evidence.

In cases where the theory seems beautiful when viewed in isolation but we do have access to ample disconfirming evidence of it, the metacognitive account would predict that the theory will typically appear much less beautiful when viewed in light of the evidence. Also, working with such a theory should not trigger feelings of confidence,[142] so we see that although the theory is still somewhat attractive this virtuousness is overruled by the low feelings of rightness. An example of this is found in Wegener's discussion of contraction theory. He argued that it was conclusively shown to be false but said apologetically that it had been "applied to a large number of individual research results with such consistency that even today it possesses a degree of attractiveness, with its bold simplicity of concept and wide diversity of application" (1977:30). Wegener strongly approved of the 'boldness' of its simplicity and its wideness of scope, not to mention its success in accounting for 'a large number of individual research results', which lent it an attractiveness that not even its falsity could undo. 'Even today'– that is, even having access to

[142] Because the majority of items that come to mind will disfavour the theory, as per Koriat's model of confidence discussed in section 5.5.4.

disconfirming evidence – contraction theory was somewhat (aesthetically) pleasing, suggesting that when the theory was viewed in isolation from that disconfirming evidence it was a very attractive theory. However, when the theory is viewed not in isolation but together with the relevant evidence, the feelings of rightness do not reach the threshold for theory acceptance. The theory does not seem compelling anymore.

There are certainly cases where the aesthetic and epistemic value of a theory conflict. But in many cases, they pull in the same direction. As I describe in this section, many of the features in virtue of which a theory is aesthetically attractive are also epistemically valuable. And many of these features function as cues for procedural metacognitive monitoring, as I argue in section 5.5. Insofar as the metacognitive monitoring is cued by these features, it allows the metacognitive feelings to provide a heuristic for, or cognitive route to, identifying epistemic value. In other words, the feelings offer a quick and easy alternative to explicit identification of epistemic value.[143]

5.4.1. Some theoretical virtues are both aesthetic and epistemic

As discussed in chapter 2, there is much disagreement on exactly how some of the broadly aesthetic theoretical virtues should be classified. Simplicity, for example, is variously counted as an aesthetic virtue (e.g., McAllister (1996), Elgin (2020), Ivanova (2020), Currie (2023), and Todd (2017)), an epistemic virtue (e.g., Ivanova (2020), Schindler (2018), Currie (2023), Todd (2017), Elliott and Sober (1994), and, tentatively, McAllister (1996:122)), and a pragmatic virtue (e.g., van Fraassen (1980)). I take it to be all the above: a simple theory is easier to work with, thereby pragmatically useful. It is also epistemically useful in that it helps us avoid overfitting the theory to noise in the data (we come back to this shortly). Simplicity is also aesthetically valuable: it is connected to pleasure and liking in both straightforwardly aesthetic contexts and in scientific contexts, almost always mentioned when scientists describe in virtue of what they find a

[143] Compare Schwarz (2018), who argues that processing fluency provides a quick-and-dirty route to identifying truth which parallels the cognitively effortful route of evaluating a claim against the evidence. I discuss how Schwarz' and my account diverges in section 5.5.3 below.

theory beautiful, and typically used as praise in the context of theory evaluation.[144]

In other words, simplicity is aesthetically and epistemically valuable. And theoretical elegance is at least partially a matter of balancing simplicity and strength (informativeness), each of which is epistemically valuable, which means that elegance is also both aesthetic and epistemic. As I argue below, these virtues make for effortless processing and are therefore involved in generating feelings of pleasure, satisfaction, and confidence. This means that in many cases, these feelings correlate with epistemic virtues, making them signals of epistemic value. This is central to explaining the epistemic value of aesthetic considerations in science, and hence to explaining why scientists can be justified in relying on aesthetic considerations in theory evaluation.

Before we get to this, we need to look closer at the epistemicity of some aesthetic virtues. I will discuss two such cases in the following sections. Many other virtues would benefit from a detailed discussion, especially since the points do not immediately generalise to other virtues, but the following will serve as a proof of concept.

5.4.2. Simplicity and overfitting

Recall that a theory is explanatorily or syntactically simple if it contains few components (basic principles, adjustable parameters, fundamental equations, etc.), parsimonious or ontologically simple if it postulates few (kinds of) entities or straightforward relations between them, and it is psychologically simple if it's easy to understand or work with (see chapter 2).

The main reason simplicity is epistemically valuable is that it can help us avoid overfitting the theory to noise in the data, as Malcolm Forster and Elliott Sober (1994) convincingly argued. 'Overfitting' refers to the somewhat surprising phenomenon that if you try to describe the trend in a dataset by fitting a curve to the data points, a complicated curve that fits all the data is typically worse at predicting new data than is a simpler curve with a more approximate fit. A simple, smooth curve better captures the underlying trend because such a curve is less likely to be fitted to errors in the data (such as measurement errors)

[144] See chapter 2 for more on the aesthetic value of simplicity.

and because it avoids outliers which can distort the underlying trend and be misleading even if they are not false (such as random fluctuations). An overfitted model in effect predicts that similar errors or random fluctuations will show up in the future, which, given their randomness, is unlikely to happen.

This can be corrected for by selecting simple models. And indeed, that is an ingrained scientific practice. As Forster and Sober put the point, "[e]ven though a hypothesis with more adjustable parameters would fit the data better, scientists seem to be willing to sacrifice goodness-of-fit if there is a compensating gain in simplicity" (Forster and Sober (1994:5)). So, what justifies scientists' preference for simpler curves, according to Forster and Sober, is the fact that selecting a simple curve is in general a good way of avoiding overfitting; that is, a good way of avoiding fitting the curve to the noise as well as signal in the data.

There are quantitative methods for avoiding overfitting, like the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC).[145] They provide a direct route to identifying the epistemic value of simplicity. But we can also use simplicity – and the feelings of pleasure to which simplicity gives rise, as discussed below – to help us avoid overfitting the theory to noise in the data beyond the narrow concerns of fitting a curve to quantitative data. The threat of overfitting extends well beyond quantitative model selection: it arises in all circumstances where the data we are basing our theory on may contain errors. As Timothy Williamson (2024) says,

> The vaguer but deeper lesson [of the curve-fitting problem] is that if we make fitting the data too easy, by helping ourselves to such a wide range of options that any supposed data will find a match, we also make ourselves easy victims of the data, because the process will not alert us to any defects or outliers in them. (2024:62)

We can use simplicity considerations to avoid overfitting both in the many circumstances where we cannot use AIC or similar methods, and also as a shortcut for those more precise methods of model selection in cases where they do apply.

[145] See Forster and Sober (1994) and Sober (2015) for discussion.

Since the simpler theory is less likely to be adjusted to noise in the data, we can use feelings responding to simplicity as a rough proxy for such methods. Through its effect on fluency, simplicity connects to confidence and pleasure (see below), and by guarding against overfitting, it can be used as a shortcut for predictively more successful theories.

5.4.3. The epistemicity of elegance

An elegant theory explains a range of significant facts in a simple and effective way, without superfluous or unmotivated elements, and its explanations follow with ease from its central claims. Steven Weinberg sums this up nicely when he writes that "an elegant proof or calculation is one that achieves a powerful result with a minimum of irrelevant complication" (1993:107). Theoretical elegance also depends on coherence, where everything fits together. The tighter the fit, i.e. the more each postulate depends on other parts of the theory, the more inevitable or natural the theory will appear, making it impossible to change one part without disrupting the whole. (See chapter 2 for a more detailed description of elegance.)

There may be more to theoretical elegance than this, but its central feature is that it strikes a good balance between simplicity, strength (or, informativeness), and coherence. A theory is strong when it is wide-scoped, rich in content, or otherwise committed to much. This is an epistemic virtue: it pertains to how much knowledge we can gain from the theory.[146] Without being balanced out by simplicity, this richness can easily become baroque. An elegant explanation, on the other hand, achieves this richness in an economical way. An elegant theory is also a highly coherent theory. Coherence, or fit, as I understand it here, ranges from consistency to mutual abductive support: at the very least, a theory must be consistent (i.e., internally coherent) to be true, and the more each part of the theory supports the other parts, the better reason we have for accepting it as true.[147] A theory low on coherence can range from false, if it's

[146] Or, if you like, how much understanding it can give us or how much of the truth it licenses believing: see chapter 3.

[147] Though a coherent system can of course be false, as is well known from criticisms of coherentist theories of justification in epistemology.

self-contradictory, to being full of unexplained coincidences. (The reason for lumping together different kinds of coherence instead of splitting it along its fracture zones is that psychological processes responding to coherence are primarily sensitive to the coherence among mentally represented items and not to their content. This coarse-grained categorisation of coherence is useful when discussing the metacognitive account because we will primarily be interested in the psychological role of coherence here. We come back to this in section 5.5 below.) And simplicity, as we saw in the previous section, is epistemically virtuous because it helps us avoid overfitting the theory to noise in the data. In other words, the central features of elegance are epistemic.

There is also reason for categorising elegance as an epistemic virtue over and above the epistemic virtues it consists of. Since simplicity and strength often pull in opposite directions, trading off informativeness and simplicity helps us balance the risks each of the virtues are associated with. Simplicity brings with it a risk of omission errors (or, false negative/type II error), and strength brings with it a risk of commission errors (false positive/type I error). The elegance of a theory indicates that its trade-off of simplicity and strength falls within a 'sweet spot' range where the theory scores high on both simplicity and strength without compromising either. This is itself epistemically valuable: getting the trade-off right helps us to avoid the errors associated with simplicity and strength.

Someone might object that while the foregoing considerations show that elegance is an epistemic virtue, it simultaneously shows that it is *not* an aesthetic virtue. If theoretical elegance essentially involves simplicity and informativeness, then what is aesthetic about it?

The assumption driving this question is that no virtue is *both* aesthetic *and* epistemic. We should not accept that assumption: as we have just seen in the cases of simplicity and elegance, a theoretical virtue may have both aesthetic and epistemic aspects. Secondly, as I insisted in chapter 2, the default position should be to take scientists' claims at face value: if they use aesthetic vocabulary, that is probably on purpose. Unless we have a specific reason to think otherwise, we should expect that they choose the terms most apt to express their experiences. Thirdly, all aesthetic qualities are based on features which can be described non-aesthetically: just as the aesthetic properties of a song are based on the frequency

of sound waves, so "[e]very aesthetic value is 'realized' by nonaesthetic features of the item" (Lopes (2022:62)) (see section 2.4.2. for discussion.) Be that as it may, the puzzling role of apparently aesthetic considerations in science arises regardless of what we call them. They certainly appear aesthetic, but it makes no difference to my account whether the apparently aesthetic experiences of apparently aesthetic virtues are genuinely aesthetic or just similar to typical aesthetic experiences and aesthetic properties.

Elegant theories are striking: explanatorily powerful but nonetheless constrained by the cohesiveness and simplicity of the overarching pattern and the paucity of elements needed to arrive at the explanations. The resulting effortlessness with which one can connect all the dots, so to speak, is registered as pleasure, satisfaction, and confidence, as I discuss in the next section.

5.5. Aesthetic virtues as metacognitive cues

Some of the features central to the aesthetic and epistemic value of a theory, including simplicity, informativeness, and coherence, appear in the psychological literature as metacognitive cues. Some of these cues are also known to be connected to aesthetic appreciation, providing further support for the claim that aesthetic and epistemic evaluations are closely related. I outline some of these results in the following sections.

5.5.1. Simplicity and fluency

As discussed above, processing fluency – cognitive ease or effortlessness – is central to many metacognitive evaluations. It is also triggered by a variety of features, such as repeated exposure, a crisp font, simple task, semantic coherence, and rhyming. Let us have a closer look at the fluency cues which coincide with features of aesthetically attractive theories.

In general, the simpler something is, the more fluently it will be processed. This holds for visual figures of uncomplicated shape and clear figure-background contrast, straightforward semantic relations, easily parsed linguistic representations, and the psychological simplicity of handling the familiar (Alter and Oppenheimer (2010), Schwarz (2018), Dechêne et al. (2010:238)). Reducing

complexity increases fluency across the board. For example, simplicity makes for both conceptual fluency (how easy it is to classify or find meaning in the object) (Labroo et al. (2008), Kelley and Lindsay (1993), Ball et al. (2018)) and perceptual fluency (for example, how easy it is to visually distinguish the object from the background and pinpoint its spatial orientation) (Reber and Schwarz (1999), Oppenheimer and Frank (2007)). As Adam Alter and Daniel Oppenheimer (2010) say in a review of studies on fluency,

> researchers have manipulated processing fluency using numerous approaches, including semantic priming (e.g., Begg, Anas, & Farinacci, 1992), visual clarity (e.g., Reber & Schwarz, 1999), and phonological priming (e.g., McGlone & Tofighbakhsh, 2000). Each manipulation influences a different proximal cognitive system: Semantic priming induces conceptual fluency, visual clarity induces perceptual fluency, and phonological simplicity induces linguistic fluency. (2010:219–20)

(Priming involves exposing the subjects in the study to some stimuli before the test stimulus is presented, often too quickly to be consciously noticed, for example showing a shape on the screen before a test of visual discrimination.)

The research on fluency relates to the different kinds of simplicity distinguished above: parsimony, explanatory and syntactic simplicity, and psychological simplicity. Note the similarity between fluency and simplicity: a psychologically simple explanation – one that is easy to understand and to work with – is perhaps equally well described as an effortlessly processed explanation. The ease with which we can understand an explanation or a phenomenon (psychological simplicity) is partially due to how simple it is: whether the explanation is straightforward and involves few components (explanatory simplicity) or postulates few objects or events (parsimony). Given that explanatory simplicity and parsimony are epistemically valuable because they help us avoid overfitting (as discussed above), this ease is then a signal that the theory is unlikely to be fitted to noise in the data. In other words, it is a signal of epistemic value.

The empirical evidence from research on meta-reasoning suggests that we can proceed on the assumption that the findings on the role of fluency in relatively simple cognitive tasks can be extended to more complex cognitive

processes, such as reasoning and decision-making (Ackerman and Thompson (2017), Thompson et al. 2011). As Ackerman and Thompson say,

> a robust finding in Meta-Reasoning, as in many other domains, is that fluency … is a pervasive cue to certainty. For example, answers that come to mind quickly engender a strong Feeling of Rightness and Final Confidence, regardless of the accuracy of the answer. (2017:613)[148]

(I return to the question of accuracy below.) Further, indirect support for this is found in studies of explanatory preferences, on the assumption that simpler explanations are more fluently processed. Tania Lombrozo, who has studied explanatory practices among adults and children for two decades, found that

> [w]hen two candidate explanations were equally likely, a majority of participants selected the simpler explanation [one cause]. It was not until the complex explanation [two causes] was *10-fold more likely* than the simpler alternative that a majority of participants selected it as the most satisfying explanation for the symptoms. (2016:751; italics inserted)

In other words, people have a vast preference for simple explanations. They don't necessarily have access to the reasons for their preference: the simpler explanation just *feels* like the better explanation. Given how confidence and simplicity are connected through fluency effects, we have reason to believe that the preference for simpler explanations is also partially based on simpler theories being more fluently processed and on the subsequent effects of fluency on confidence.[149]

There are two things to note in passing before moving on. First, since the preference for simple explanations is prevalent among all participants (who are overwhelmingly undergraduate students, as is typical in psychological experiments) the preference for simplicity is not a phenomenon restricted to trends among scientific communities but something much more basic to human

[148] Ackerman and Thompson distinguish between confidence at different stages of the processing: judgements of initial solvability, which signal whether it's worth even attempting the task, intermediate confidence or feelings of rightness, and final confidence (or final feeling of error, as the case may be) (2017:611). This level of detail will not make a difference to our discussion; talk of confidence in the following applies equally to confidence at any stage of reasoning.

[149] Lombrozo did not study fluency directly.

cognition. This is in line with the view that theoretical simplicity functions as a metacognitive cue, and it tells against McAllister (1996) and Bird's (2020) view that the preference for simplicity in science is solely based on noticing that epistemically successful theories have been simple. Ulianov Montaño (2013) argued convincingly against McAllister that some aesthetic virtues are what he (Montaño) called 'historical constants': they "consistently exhibit high degrees of preference throughout history" (2013:138). He listed simplicity, harmony, and symmetry as examples. Lombrozo's research shows clearly that it is correct that simplicity, at least, is preferred across different contemporary communities without any connection to previously successful theories (though it may be connected to previous success in everyday reasoning), which suggests that Montaño is right in treating it as a historical constant.

Second, more work remains to establish whether all the kinds of simplicity relevant to theory (and model) choice are generally preferred, or whether the preference for some kinds of simplicity are based on something else. As Lombrozo puts it,

> [a]lthough it is likely that multiple metrics for simplicity operate in parallel, the evidence to date supports two proposals: that explanations for individual events are better to the extent they invoke fewer independent assumptions, and that when learning multi-causal systems, learners favor sparse and strong causal structures. (2016:751)

In other words, we know that some kinds of parsimony and some kinds of explanatory simplicity are generally preferred. It would be good to test people's preferences for a greater variety of theoretical simplicity. My working hypothesis is that any kind of theoretical simplicity will increase the fluency of one's cognitive engagement with the theory, but again, this needs to be tested empirically.

5.5.2. Familiarity, fluency, and background knowledge

Familiarity with a topic increases the fluency of processing involving it. For example, Lynn Reder and Frank Ritter (1992) found that familiarity with a term that appeared in the question (but not in the answer) cued feelings of knowing, presumably via its effect on fluency (see also Koriat and Ma'ayan (2005)). That

is, familiarity with the topic triggered subjects' sense that they would be able to answer the question.

It is not surprising that familiarity with a topic makes thinking about it easier, and we should expect that this extends to theory evaluation (i.e., that familiarity with the topic makes processing a theory on it easier). Psychologists induce a false sense of familiarity through priming to isolate the role familiarity plays, but in real-life situations – including in science – it's likely that familiarity will primarily be a matter of prior knowledge of the topic.[150] This is good news. It suggests that a major source of fluency – familiarity – is one's topic-relevant knowledge. And if you know a lot about a topic you will be in a better position to give a correct verdict on a theory about it than if you lack this knowledge. So feelings of fluency can indicate that you are in the positive epistemic situation of knowing much about what the theory is about and thus that you are in a good position to evaluate it.

However, we should be careful not to overstate this point. As the priming methodologies show, it is easy to instil a false sense of familiarity. And fluency does not guarantee that judgements based on it are accurate, regardless of what the sense of fluency is based on (we come back to this below).

5.5.3. Fluency and aesthetic pleasure

Before discussing coherence, conflict detection, and informativeness, we should mention how fluency also triggers aesthetic appreciation. For example, repetition – which increases fluency – increases liking (up to a point). This is the basis for the 'mere exposure effect': we tend to find what we are familiar with more aesthetically pleasing than unfamiliar objects (Zajonc (1968)). And vertical symmetry, which is easier for humans to detect than symmetry across a horizontal or diagonal line, also "received the highest ratings of figural goodness" (Reber et al. (2004:8)). People generally perceive something as more aesthetically pleasing if it is fluently processed, and they also believe that they have seen it before, that it is true, and that they are more likely to remember it if

[150] Gigerenzer (2007) defends the general usefulness of heuristics in ecologically valid situations; that is, in the kind of real-world situations they have developed in.

it was fluently processed (Schwarz (2018), Reber (2018), Alter and Oppenheimer (2009)). Also, if something is presented in a form conducive to fluency it is easier to understand its content (Reber et al. (2004)).

The close connection between fluency on the one hand and positive aesthetic judgements on the other makes some psychologists, most notably Rolf Reber and Norbert Schwarz, argue that experiences of beauty reduce to fluency (Reber, Winkielman and Schwarz (1998), Reber, Schwarz, and Winkielman (2004), Schwarz (2018), and Reber (2018)). (This is a highly contentious issue in empirically informed aesthetics: see Shimamura and Palmer (2012), Cochrane (2021), and Forster (2020) for discussion.) Moreover, the fact that both aesthetic pleasure and many kinds of positive metacognitive evaluations are triggered by fluency is the basis for both Schwarz' (2018) and Reber's (2018) variants of the metacognitive account. Their accounts share the central hypothesis that both judgements of beauty and judgements of truth are in part based on fluency. As Schwarz says,

> judgments of beauty and judgments of truth share a common characteristic: people make them, in part, by attending to the dynamics of their own information processing. When an object is easy to perceive, people evaluate it as more beautiful than when it is difficult to perceive; similarly, when a statement is easy to process, people are more likely to accept it as true than when it is difficult to process. (2018:25)

And Reber:

> It is processing fluency that determines aesthetic pleasure. According to this account, an object is seen as beautiful if it can be processed with surprising ease. (2018:253)

But that doesn't quite cut it: things that are too easy are *boring*, not beautiful. Fluency generally only gives rise to mild feelings of liking, so this account doesn't capture more intense aesthetic experiences – like experiences of beauty (Silvia (2012), Armstrong and Detweiler-Bedell (2008), Cochrane (2021)). Furthermore, beauty is standardly taken to require just the right level of complexity, as captured by the traditional way of cashing out beauty in terms of unity in disunity and psychological studies showing that aesthetic appreciation falls on an inverted U-curve of complexity (see chapter 4). Fluency cannot be the whole story, even

though it plays a large part in it. So it's better to take fluency to be *one* of the cues involved, balanced out by enough richness of detail to sustain interest.

Although Reber focuses solely on fluency, this richness of detail ('disunity') which is so satisfyingly derivable or unified by beautiful theories may be what he alludes to when he talks about the pleasures of *surprising* ease. After all, against the background of a strong, wide-scoped, or unifying theory of a complex phenomenon, effortless cognitive engagement is likely to be surprising. Whether or not that is what Reber has in mind, I agree with him that "the sense of beauty in both the context of discovery and of justification may predict the truth of a proposition at above-chance level if the mental processes underlying the experience of beauty are well aligned to the truth of the proposition" (2018:254). But on my account, 'the mental processes underlying the experience of beauty' depend on more than the fluency of the processes.

This brings us to the other (features of) aesthetic virtues functioning as metacognitive cues we will discuss here: the coherence between either theory or evidence or parts of the theory, and strength. I expect there to be many more cues utilised in the metacognitive processes underlying aesthetic experiences in science, but a comprehensive investigation is beyond the scope of this thesis. The limited discussion provided here will serve as proof of concept; if we can account for central features of the two most common ways for a theory to be beautiful – simplicity and elegance – then it will be worth extending this to other aesthetic virtues in future work. And if we cannot account for these virtues, we will not have wasted our time on a comprehensive account doomed to fail.

5.5.4. Fit/coherence

Another feature which appears both as a theoretical virtue and as a metacognitive cue is the mutual fit or coherence among items. As will be remembered from the discussion of elegance above, theoretical elegance is partially constituted by simplicity, informativeness, and coherence. (Here I use 'fit' and 'coherence' interchangeably.) A good fit between items is sometimes paradigmatically epistemic (e.g., empirical accuracy, or the fit between theory and empirical evidence, and external coherence, or fit with background knowledge), sometimes paradigmatically aesthetic (harmony or elegance, or a particularly

striking or pleasing fit between items), and sometimes it falls somewhere in between (like unity, or that there is a common factor tying together disparate elements). As mentioned above and discussed below, when we're talking about the psychological role of coherence, the central feature is the fit and not what is fitted together.

Coherence makes for fluent processing. For example, in the remote associates test, subjects are presented with three words which are either unrelated or have a fourth associated word in common. For example, the triad 'desk, lamp, computer' is connected through its association with 'office': they form a coherent set. 'Lamp, three, foam' do not. Studies replicated since the 1960s have shown that people are "able to discriminate between coherent and incoherent triads without an explicit retrieval of the common associate for the coherence triads" (Topolinski and Strack (2009:1466)). This is presumably due to fluency effects: coherent triads are easier to process, and this ease is used as a cue for coherence judgements (Topolinski and Strack (2009), Reber and Unkelbach (2010)). Another way of describing this is as semantic or conceptual coherence, which, as we saw in section 5.5.1, is a source of fluency.

Note the similarity between coherent triads and theoretical unity: coherent word triads are bound together by a common object, property, or concept, like the underlying, unifying factor of some scientific theories. Coherence, then, functions as a cue for metacognitive evaluations through its effect on fluency: the fluency with which one processes conceptually unified items, as opposed to disunified items, provides feedback from the object-level processing which is used by meta-level processes to evaluate its quality and accuracy.

Coherence also plays a central role in a prominent psychological model of confidence, Asher Koriat's 'self-consistency model' (SCM) of confidence (e.g., Koriat (2012, 2024)). The SCM says that when we make decisions, we consider a sample of items relevant to the decision and choose the option that is supported by the majority of considerations.[151] The larger the majority is, the

[151] The model was developed for 'two alternative forced choice' situations, i.e. situations where you have to choose one of two options. This simplification is worth bearing in mind, but there is no indication that the model breaks down in

more confident we will be. This confidence is based on cues from the decision-making, like how much deliberation it took to reach the decision and how quickly you reached it, and Koriat takes the main cue for confidence to be 'self-consistency', or the extent to which what comes to mind favours a given alternative – i.e., how great a majority of considerations cohere with the choice option. (This need not be evidence expressible as declarative statements: "many of the clues that come to mind in attempting to make a choice consist of associations, hunches, and images that are not readily expressed in the form of declarative statements but may influence the choice" (2012:82).) When the majority of items considered supports one option, that's the one you choose, and the bigger the majority, the more confident you will be in your choice.

Or, to put it differently, the more items that fit together, the more confident you'll be. Importantly, according to this model it is the fit itself, not the content of what is fitted together, which triggers confidence. Applied to theory choice, whatever you bring to mind when evaluating a theory, whether it is the hypotheses constituting it or the available evidence, the more they pull in the same direction and the less conflict you detect between items, the more compelling it will be. So here we have another feature which is both epistemically valuable, part of aesthetic virtues, and connected to confidence.

5.5.5. Conflict detection

The opposite of fit or coherence is lack of fit, incoherence, or contradiction. There is good evidence that such cognitive tensions are detected, though not always made consciously available: for example, we slow down when making errors even when tricked to believe that no error has been made (Logan and Crump (2010)). In the metacognition literature, this is known as 'conflict detection' (e.g., Thompson and Markovits (2025), Thompson et al. (2011), Thompson and Johnson (2014), De Neys et al. (2008), De Neys et al. (2011)). Participants in studies on conflict detection are typically presented with reasoning tasks where some are "conflict problems" "designed to cue two incompatible responses, e.g., one based on belief and one based on validity,

---

situations which are more complex and thus more difficult to test in a psychology lab.

whereas for non-conflict problems, both cues pointed to the same answer" (Thompson and Johnson (2014:218).[152] A robust finding in these studies is that people report lower confidence in conflict problems: for example, when we go for the intuitive but incorrect answer we are slower to decide[153] and less confident in the decision than when we choose the intuitive and correct answer. As Valerie Thompson and colleagues report from one of the studies,

> Consistent with the hypothesis that metacognitive monitoring is sensitive to conflict …, FOR [feeling of rightness] judgments were lower for the incongruent than the congruent problems; congruent items were also responded to more fluently. (2011:129)

Tension between items considered trigger feelings of error, making us aware that something is off, that there is a conflict somewhere.[154] In other words, metacognitive feelings can signal both coherence and tension: coherence is signalled through feelings of rightness/confidence, and tension is signalled through feelings of error.

It is not clear how a lack of fit will be registered: it will not make for fluent processing, as a good fit will, but it may not make for *disfluent* processing, as will an outright contradiction or incoherence. The lack of fit is presumably

---

[152] An example of a conflict problem is the famous 'Linda the bank teller' conjunction fallacy. Participants are first presented with a vignette: "Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations" (Tversky and Kahneman (1983: 297)). They are then asked to judge what is more probable: that Linda is a bank teller or that she is a bank teller *and* is active in the feminist movement. The intuitive answer, which most participants give, is that she is a bank teller feminist (perhaps because that makes sense of why the personal information is presented in the first place, cf. the Gricean maxim of quantity). This is of course wrong: it's not possible for a conjunction to be more probable than one of its conjuncts. A non-conflict problem gives information that is either congruent with or irrelevant to the correct answer; say, that Linda has always been detail-oriented and interested in finance, or that she is fond of camping.

[153] Remember that response speed is a measure of fluency.

[154] This is further supported by studies showing a neurological and physiological activation related to conflict detection (activation in the anterior cingulate cortex (De Neys et al. (2008)) and increased skin conductivity (De Neys et al. (2010)), respectively) during reasoning.

increasingly important the more complex the object of evaluation is: the more loose ends there are, the slower and more effortful will the cognition be.

These effects of coherence – and the lack of it – are thought to hold generally in reasoning and decision-making (Ackerman and Thompson (2017)). Theory evaluation is a special case of reasoning and decision-making, so we should expect these effects to carry over to the scientific context. In particular, we should expect well-fitting theories (whether the fit is internal or external coherence, empirical accuracy, or elegance) to feel compelling, both through fluency effects and through the consistency among items in favour of a given theory. The coherence among items mentally represented by scientists while they are actively working with a theory is of course an indirect and highly fallible basis for evaluating its epistemic value: perhaps the scientist doesn't consider all of the relevant evidence and so won't notice that something sticks out, or they may not attend to the fact that some parts of the theory contradict each other, or perhaps they consider an unrepresentative sample of the evidence which makes the few pieces of evidence against the theory seem more significant than they really are.

This is a central part of why I say that aesthetic and metacognitive feelings are no more than a moderately reliable route to identifying epistemic value. This slack is also related to why theories which seem elegant when evaluated in isolation can lose their compellingness when relevant evidence is taken into account: what is kept in mind matters. That being said, that we don't *notice* any tension is an indicator that there *is* no tension: it's a fallible and limited, but not a useless, signal.

5.5.6. Informativeness/richness of detail

We have discussed how two features which are both part of aesthetic virtues and independently epistemically valuable – simplicity and coherence – are also known to be metacognitive cues. The last such virtue I will discuss in this thesis is strength, or informativeness. Here the connection to metacognitive cues is more speculative. A strong theory tells us much: it explains a variety of phenomena and detailed predictions can be derived from it. This richness of

detail may function as a metacognitive cue in scientific reasoning just as it does in other kinds of reasoning and decision-making.

Studies on the reliability of eyewitness testimony (a subfield of the metacognition literature) have shown that richness of detail functions as a cue for believability. Or, more generally, that a narrative rich in detail makes us more confident in it. As Galit Nahari and Merav Pazuelo write,

> The richer an account is perceived to be in spatial and temporal information, names of people and places, emotions, descriptions of visions, senses, sounds, tastes and smells, the more likely it is to be believed. (2015:363)

In other words, the more informative a claim or narrative is, the more it is believed. "The belief that genuine statements are typically more detailed than fabricated statements is known as the 'richness-in-detail heuristic'" (Harvey (2021:15)). This heuristic plays a large role in communication and transmission of knowledge by testimony, and it is also a moderately valid cue for truthfulness. For example, when people are interviewed about an event which is meaningful or at least salient to them immediately after it happened, "truth-tellers typically report more detailed statements than liars" and so "richness-of-detail … is a diagnostic cue to deception" (Harvey et al. (2020:1)). This effect is reduced over time in line with the general deterioration of memory, which is not matched by liars' ability to invent details. But inventing details is challenging, and under conditions of cognitive load liars provide fewer details than truth-tellers do (Vrij et al. (2008)). As Adam Harvey sums it up, "richness-in-detail is both a subjective (believed) and objective (valid) veracity cue" (2021:15; see also Nahari and Pazuelo (2015:363)).

Could it be that the role informativeness plays in triggering listeners' confidence in descriptions of events is paralleled in scientists' evaluations of informative theories? It would not be surprising if a theory which could produce richly detailed explanations and predictions thereby triggered confidence in the same way as a detailed story does – and for the same reasons: a detailed narrative is anchored to the world at many points, it contains details one wouldn't expect to be provided unless they were true, and the more detailed it is the more easily it can contradict itself, rendering a detailed *and* coherent story all the more

impressive. Future empirical work is needed to show whether the informativeness of theories has the same effect on confidence as the informativeness of descriptions of events has in eyewitness testimony and other communications.

One could object that a theory's strength is very different from providing lots of details: sure, a strong theory is committed to much, but a simple, powerful equation is not at all like a narrative rich in detail. Knowing that you *can* derive lots of detail is not the same as considering those details, which one would presumably need to do for those details to act as cues for confidence. It is true that the connection between strength and the richness of detail-as-metacognitive-cue is less straightforward than the other cues we have considered, which is why I earlier described their connection as speculative. However, a strong theory can easily be complex, convoluted, baroque: full of details, far from the simple and elegant theories that are so impressive for also being informative. An *elegant* theory – or a simple and strong equation – is not like the rich narrative we recognise from the eyewitness studies (although there is significant overlap in the detailed picture of the world both provide). As will be remembered, I take elegance to involve at least two 'streamliners' (simplicity and coherence) in addition to strength. Thus it is not a problem for my account that strength alone does not make a theory attractive – after all, an ad hoc theory can easily be strong.[155]

[155] That being said, this may be one of the areas where the difference between scientific fields can make a difference to aesthetic standards. Perhaps psychologists and field geologists expect a strong theory to look quite like the detailed narratives explored in the eyewitness testimony literature, because in these areas a simple generalisation would not be informative enough: more details are needed to get from the overarching theory to the particular features of the phenomenon under study. In areas where we can expect to get simple *and* strong theories, one without the other leaves the theory unimpressive and unconvincing. The question of how much variation there is in aesthetic sensibility among scientists from different fields (or across different scientific communities in general) is a big question which must be left to future research. Here I will only mention that the metacognitive account has the resources to provide a flexible framework for different aesthetic standards since fluency is partially a matter of familiarity. Since a scientist's familiarity with a topic and expectations about what a good explanation looks like are influenced by her education, relatively large variations in aesthetic standards can be accounted for on the metacognitive account.

5.5.7. Cue integration

We have now looked at a number of metacognitive cues, and a pressing question is what happens when more than one cue is present. Will one overrule the others, or are they combined? If the latter, do cues with similar effects enhance each other and cues with opposite effects cancel each other out? The research on cue integration is less well-developed than research on individual cues. Although there are many existing studies varying multiple cues without testing cue integration specifically, these cannot necessarily be mined for information on cue integration because "cue effects … tested at the aggregate level … may occur even if each individual participant based his or her [metacognitive judgement] on only one cue but different individuals focused on different cues" (Undorf et al. (2018:509)).

The evidence we do have suggests that effects of individual cues add up so that, for example, thinking through something which triggers fluency in multiple ways makes for even more fluent processing than had there only been one cue for fluency. For example, in a series of studies, Monica Undorf and her colleagues (2018) varied two, three, and four individual cues which are known to affect judgements of learning (JOLs) and showed, through multiple forms of individual-level analysis, that a majority of participants integrated multiple cues.

Their study also suggests that some cues can cancel each other out: for example, whereas JOLs are typically high for words that are both emotional and concrete, in one experiment they found that "the impact of emotionality on JOLs decreased with increasing concreteness" (Undorf et al. (2018:515)). This is congruent with experiments by Rakefet Ackerman (2023) which "suggest that cues may affect confidence in varying, and even opposing, directions on the same items, potentially leading to a complex inference of item difficulty" (2023:32). There is also some evidence that cues are strategically integrated (Undorf and Bröder (2020)) rather than being bundled together "into a single unified feeling of ease or difficulty" (Ackerman (2023:32)). This is further supported by other studies showing that beliefs about the source of a feeling of

(dis)fluency can cancel out the fluency effect (e.g., Schwarz et al. (1991)). Still, it seems that for the most part, the cue effects aggregate (Undorf et al. (2018)).

This is all to say that although the cues we have considered until now all individually impact metacognitive judgements, we do not know enough about their effects when combined. This is a potential weakness for the metacognitive account which can only be addressed through more experimental evidence on the combined effects of simplicity, fit, and richness of detail. I want to flag that my analysis of theoretical elegance as combining simplicity, informativeness, and coherence (which in turn trigger the feelings of rightness which makes elegant theories so compelling) is built on the assumption that these cues do not cancel each other out. Testing this prediction will be left to future work.

5.6. The accuracy of metacognitive judgements

Are metacognitive judgements accurate? Or, as a psychologist might put it, are the cues which trigger the feelings on which metacognitive judgements are based valid? The empirical evidence suggests that metacognition is moderately accurate: we often get it right, but we are also prone to metacognitive illusions.

Studies on the reliability of metacognition have produced mixed results, but they generally point towards a moderate to high level of accuracy across different metacognitive feelings (e.g., Koriat (2000, 2008, 2012), Reder (1987, 1996), Schwartz (1994), Yeung and Summerfield (2012)). As Yeung and Summerfield say, "[p]eople are capable of robust evaluations of their decisions: they are often aware of their mistakes even without explicit feedback, and report levels of confidence in their decisions that correlate with objective performance" (2012:1310). And Koriat: "[n]ot only did the effects of ease of learning on cue utilization parallel those observed for cue validity, but the changes in the extent of cue utilization across repeated presentations were found to closely mirror those that occur in the magnitude of cue validity" (2008:426). 'Cue utilisation' refers to how the cue is used: whether it is relied upon as diagnostic of success in the first-order task. 'Cue validity' refers to whether it is in fact diagnostic of success. In other words, people are quite sensitive to the validity of

metacognitive cues, adjusting how much they rely on one cue or another to match how well the cue correlates with successful mental actions.

For example, Annika Boldt, Vincent de Gardelle, and Nick Yeung (2017) found that confidence in perceptual tasks was positively correlated with correctness. When the participants expressed high confidence they gave very few incorrect answers, and when they expressed low confidence they gave many incorrect answers. This did not just hold for the extreme cases (the 'certainly wrong' and 'certainly correct' judgements): there was a monotonic decrease in error rates with level of confidence (2017:1525). Ana Fernandez Cruz, Santiago Arango-Muñoz, and Kirsten Volz (2016) looked at feelings of error during mental calculations and found that they were positively correlated with arithmetic errors. The participants were successful in almost 90% of the trials by either reporting a feeling of error when they committed a calculation error or by *not* reporting a feeling of error when everything went well (2016:115). In other words, the feelings of error were good indicators of actual error, and low or absent feelings of error were good indicators that everything went well. Fernandez Cruz et al.'s study showed higher accuracy than is commonly found: it's more common that metacognitive judgements are accurate in 60–80% of tasks, with the accuracy falling significantly below that in certain circumstances.

Many studies have shown that we are sometimes embarrassingly wrong in our metacognitive judgements (Undorf et al. (2022), Schwarz (2018)). Some of the ways in which we often go wrong are by being overconfident (Griffin and Tversky (1992)), by being overconfident about difficult things and underconfident about easy things (this is called the 'hard-easy effect': see Dunlosky and Metcalfe (2009:120–4), and by mistaking something being fluently processed for it being true (the 'illusion of truth' effect: see Dechêne et al. (2010) for a meta-study). One example of the hard-easy effect is found in one of Koriat and colleagues' (2009) studies on judgements of learning (JOL), that is, the metacognitive feeling predicting that you will be able to recall some piece of information at a later point in time. The participants were given word pairs to study and asked to predict which of the target words they would later recall if presented with the cue word. For children in sixth grade, the mean judgements of learning for easy word pairs were 70% whereas recall was 86%, and JOLs for

the difficult pairs were 52% with a recall of 32%. That is, they were underconfident about their ability to recall easy word pairs and overconfident about their ability to recall difficult word pairs. Incidentally, note how metacognitive evaluations are sensitive to task difficulty: although the JOLs here over- or underestimate recall, they vary with the object-level task and are relatively stable signals of actual recall.

Even worse for my case is 'the illusion of truth' effect. People tend to believe that what is fluently and effortlessly processed is more likely to be true – even when it's not. An amusing example of this effect is found in a study by Matthew McGlone and Jessica Tofighbakhsh (2000). In this study, participants were asked to judge the accuracy of rhyming aphorisms and their non-rhyming synonymous counterparts. Some examples are 'birds of a feather flock together' and its non-rhyming counterpart 'birds of a feather flock conjointly', 'life is mostly strife' and 'life is mostly struggle'. The participants thought it slightly more likely that 'birds of a feather flock together' is accurate than that 'birds of a feather flock conjointly', plausibly because the first is more effortlessly understood. A meta-study by Alice Dechêne and colleagues (2010) showed that the illusion of truth effect "is of medium size" and "is robust against variations in presentation duration and modality, the proportion of repeated statements, and participants' age" (2010:251).

Another metacognitive illusion which might make one sceptical about the reliability of metacognition is 'the illusion of explanatory depth' (IOED; also known as the illusion of understanding). It will be instructive to discuss this case in some detail. This illusion occurs when people believe that they understand something better than they do and overestimate how well they can explain it (Rozenblit and Keil (2002), Fernbach et al. (2013), Fisher, Goddu, and Keil (2015), Gaviria and Corredor (2021), Meyers et al. (2023)). The paradigm for studying the IOED asks participants to rate their understanding of a certain phenomenon (e.g., how a sewing machine works, why leaves turn yellow in the autumn) before explaining the phenomenon and then rate their understanding again. Typically, the experimenters explicitly ask the participants to explain the phenomena by giving as much detailed information on the causal mechanisms as possible. To the participants' surprise, their explanations tend to be shallow,

and in response to this self-generated evidence of their lack of understanding, the second rating of understanding tends to be lower than the first. Until prompted to explain – which shatters the illusion – our feelings of understanding are inflated and misleading.

Therefore, one might argue, feelings of understanding should not be trusted and, together with other metacognitive illusions (like the illusion of truth effect, the hard-easy effect, etc.), IOED shows that metacognitive feelings ought not guide scientific theorising.

However, studies of IOED show clearly that explaining helps recalibrate those feelings. To take one example, after explaining how a cylinder lock works, participants' self-reported level of understanding plummeted, and after answering specific questions testing their understanding ("how would you pick a cylinder lock?") it dropped even further (Rozenblit and Keil (2002:527–9)). We are able to recognise a shallow explanation, even when we're the ones doing the explaining, and we recognise that if we are not able to verbally explain a phenomenon then we don't understand it very well.[156] There is also some evidence that the good effect of explaining generalises: "participants reported knowing less about how snow forms after attempting to explain how a zipper works" (Meyers et al. (2023:1)). This suggests that the illusion of explanatory depth will not be of much concern in the context we are considering, that is, when working with scientific theories, as this typically involves considering (an) explanation(s) of the target phenomenon. Moreover, there is evidence that people more engaged in analytic or type 2 processing are less susceptible to the illusion (Sloman and Fernbach (2017:80–4), Gaviria and Corredor (2021)).[157]

---

[156] It should be noted that this holds for knowledge-that, not knowledge-how (unless know-how reduces to knowledge-that), such as how to ride a bicycle or pick a lock: something we can have perfectly good tacit knowledge of but be unable to explain verbally. Interestingly, participants didn't report reduced feelings of understanding after explaining procedures like how to tie a bow tie or bake chocolate chip cookies although this tests procedural knowledge for which people are unlikely to provide good verbal explanations (Rozenblit and Keil (2002:542–3)).

[157] The extent to which people tend to be engaged in type 1 or type 2 processing – whether they tend towards more intuitive or more deliberative thinking – is standardly measured by the cognitive reflection test (CRT). The CRT consists of three problems which have an intuitive but false solution, including the well-

Since deliberate thinking characterises scientists' evaluation of theories,[158] one of the contexts in which we are the *least* likely to fall prey to the illusion of explanatory depth is precisely one where we are engaged in scientific theorising.

This is good news. Yes, we are prone to metacognitive illusions. But we're not equally likely to fall prey to the illusions in all circumstances. The research on the illusion of explanatory depth has identified two conditions under which we are unlikely to fall prey to the illusion: when we are engaged in explaining, and when we are engaged in deliberate thinking. Both are characteristic of scientific reasoning. Therefore, the IOED does not show that metacognitive feelings are too unreliable to be useful in scientific theorising.

All in all, the reliability of metacognitive feelings is moderate. This fits well with the riskiness of relying on aesthetic judgements in science: although aesthetic considerations appear epistemically useful more often than not, the history of science is full of examples of beautiful but false theories or true but ugly theories. That the metacognitive account does not overstate the epistemic value of aesthetic considerations is a point in its favour.

---

known bat-and-ball problem. To answer correctly people must suppress the intuitive solution that comes to mind and engage in deliberative thinking. Gaviria and Corredor reports that "participants who had difficulty performing Type-2 processing (i.e., those who obtained a CRT equal to 0) overestimated their knowledge … more than participants who obtained high CRT scores" (2021:821–4), which echoes Sloman and Fernbach's finding that "[p]articipants who did well on the CRT showed no illusion [of understanding]. In stark contrast, participants who scored zero or answered only one question correctly on the CRT showed a big illusion" (2017:83).

[158] Does this undermine the role of metacognitive feelings in scientific theorising? No: as Ackerman and Thompson say in a review of research on metacognition of reasoning and problem-solving, "[w]e have argued that the processes that monitor our cognitions are low level [i.e., subpersonal], implicit, and cue based, *even in activities, such as reasoning and problem solving, that rely extensively on explicit, deliberate thought*. This observation leads to an apparent paradox in that *deliberate, analytic thinking may be initiated and terminated in response to unconscious, implicit cues*. As such, we argue that understanding the processes involved in Meta-Reasoning is crucial to understanding reasoning proper" (2017:615-6; italics inserted). Metacognition is centrally involved also in complex cognitive processes, such as reasoning and decision-making (see also Shea (2024), Fields and Glazebrook (2020), Thompson et al. (2011)). Even when these processes appear deliberate and under conscious control, they are to a large extent directed by subpersonal metacognitive inferences, making it essential to understand these metacognitive processes in order to understand human reasoning.

5.6.1. Does it feel like aesthetic or epistemic pleasure?

We know that people struggle to separate different sources of affect; this is the basis for metacognitive illusions such as the illusion of truth-effect (Dechêne et al. (2010), Undorf et al. (2022)). Background knowledge, attribution of cues (such as fluency) to factors that are irrelevant to the task at hand (Schwarz et al. (1991)), or one's theoretical assumptions about the cognitive processes in play (Koriat et al. (2009)) influence how these feelings are interpreted.

For example, the informational value of feelings of ease changes if people believe that these feelings are based on some irrelevant factor. In one study (Schwarz et al. (1991)), participants were asked to give either six or 12 examples of situations in which they had been assertive and then rate their assertiveness. Surprisingly, the participants who gave 12 examples of assertive behaviour rated themselves as less assertive than did the participants who gave only six examples, even though they had just produced twice as much evidence in support of the claim that they were assertive – plausibly because the disfluency due to the difficulty of the task reduced their confidence. This, however, didn't hold for participants who were told that music played during the experiment would influence ease of recall. The feelings of ease or difficulty were no longer taken as carrying any information relevant to the task at hand. The difference between these two groups was that the second group believed a claim – that the music would influence their recall – and this changed how they interpreted their feelings of (dis)fluency.

The lesson to take from this is that it is difficult to pinpoint the source of feelings, and thus interpret them accurately. In particular, it can be difficult to distinguish standard metacognitive feelings (feelings of knowledge, feelings of understanding, tip-of-the-tongue feelings, etc.) from aesthetic feelings when they appear together. This may go some way to explain why some people take simplicity, elegance, unity, and the other virtues I have called broadly aesthetic virtues to be epistemic, and others take them to be aesthetic: if you attribute your felt satisfaction to the recognition of epistemic value, then the feeling is unlikely to strike you as particularly aesthetic, and vice versa. (Just as the participants in

Schwarz et al.'s study didn't take their felt ease as an epistemic signal when attributed to the music.) Does the theory feel right in an aesthetically pleasing way, or compelling in an elegant way, or both epistemically and aesthetically attractive? If I'm right in taking aesthetic feelings in science to be metacognitively generated, then it will of course not be possible to separate aesthetic from metacognitive feelings: they just *are* metacognitive. Whether they are taken to provide information about the aesthetic or epistemic value of the theory may still vary, perhaps according to, say, the scientist's individual sensitivity to aesthetic value or whether there is so little evidence available that it would be strange to describe the theory as epistemically valuable.

According to both Todd (2017) and an earlier variant of my metacognitive account, (at least some) aesthetic and metacognitive feelings are inseparable: they are the same kind of experience, though they sometimes appear under different guises. In Miyata-Sturm (2024a), whether they are felt as aesthetic or metacognitive depends upon contextual factors and background beliefs (2024a:13–4); in Todd (2017), whether they are felt as aesthetic or not depends upon whether "the object of the experience is categorized under a broadly intellectual/cognitive domain [or] the domain is seen as aesthetic" (2017:229). In both cases, it is prior expectations, not the phenomenological feel or the underlying cognitive processes, which makes us use aesthetic or epistemic terms to describe them.

5.7. From metacognitive cues to compelling theories

According to the metacognitive account, the pleasurable feelings triggered by beautiful theories make such theories compelling, explaining why scientists in many cases stick to them even in the face of apparent counterevidence (as discussed in chapter 4). We have pre-theoretic reasons to think that such a metacognitive experience should play a role in scientific reasoning. Pre-*metacognitive account*-theoretic, that is: it is what we should expect based on studies of meta-reasoning, of which *scientific* reasoning and decision-making is a special case. Metacognitive feelings play a large regulative role in our reasoning. For example, they can signal whether a decision is correct (Koriat (2012)), whether

one has the resources to tackle the problem (Ackerman and Thompson (2017)), whether an inference is valid (Shea (2024)), or whether one has made an error (Fernandez Cruz et al. (2016)). These aspects of reasoning and decision-making are all relevant to theory evaluation. Confidence is central to all these evaluations: it signals that all is good and that we can continue as before. Feelings of error signal that something has gone wrong and that we need to change strategy.

Metacognitive feelings are feedback from the processing experience, triggered by cues from the thinking process and the object(s) towards which the thinking is directed. In the case of theory evaluation, the cues are, respectively, features of the ongoing cognitive engagement with the theory or features of the theory and evidence considered, such as the fluency with which one can work with the theory or the fit between elements of the theory. As we have seen, theoretical elegance is partly a matter of simplicity, strength, and coherence. These features are cues for feelings of rightness, either directly (as in how a good fit between items that come to mind when making decisions trigger confidence) or indirectly (as in how simplicity triggers confidence through its effect on fluency). This extends to other broadly aesthetic virtues beyond elegance, to the extent that they involve the features discussed here.

Since aesthetic virtues play a role in theory evaluation primarily through scientists' affective responses to them, it is important that we understand the cognitive underpinnings and epistemic role of these affect-based judgements, over and above understanding the epistemic value of aesthetic virtues. Aesthetic appreciation of theories is based on affect, on positively valenced experiences, *not* on explicit consideration of whether the theory confirms to some aesthetic ideal. (This holds in general for aesthetic appreciation: compare how you appreciate a book or a song because of how it makes you feel, without explicitly evaluating its aesthetic merits.)

So, the pleasure, satisfaction and confidence felt when working with a theory are metacognitively generated feelings. They reflect properties such as the fluency of the cognitive process of working with the theory or the fit between items considered; properties which in turn tend to reflect epistemically valuable features of the theory (such as simplicity, empirical accuracy, or coherence). Hence what we think of as aesthetic feelings serve as a moderately reliable

psychological signal that the theory is epistemically virtuous (i.e., that it is simple, or empirically accurate, or coherent, etc.).

In other words, scientists can be justified in accepting a theory partly on its aesthetic merits. Since the connection between aesthetic pleasure and epistemic value is rather loose, this raises a crucial question: under what conditions can we rely on aesthetics in science? Answering that question must be left to future work.

5.7.1. Applications

To see how the metacognitive account works in practice and to get a feeling for its explanatory potential, let us work through some of the examples of aesthetic considerations influencing theory evaluation discussed in chapter 4.

Take for instance the sense of elegance which convinced many scientists that the Vine-Matthews-Morley hypothesis was correct. Its elegance is, roughly, due to how it fitted together a range of disparate facts into one simple framework. Working with this theory – thinking through its consequences, using it to explain some data points, evaluating its merit, etc., – should on my account tend to produce pleasurable feelings. The simplicity, coherence, and informativeness in virtue of which it is elegant will trigger feelings of pleasure, satisfaction, and rightness (or, confidence). These feelings may well blend together to one overall feeling that the theory is great, though the feelings of pleasure and satisfaction central to the aesthetic experience may be separable from the compellingness. For example, the scientists who were convinced by the VMM before conclusive evidence was found tied its compellingness directly to its elegance. Recall Ron Mason: "Vine and Matthews' hypothesis offered an elegant explanation … I had absolutely no doubt as to the correctness of their hypothesis" (2003:41).

Let's consider another example, this one in more detail. Remember that continental drift was praised for its elegance (though less frequently than plate tectonics was praised for *its* elegance, which is consistent with my hypothesis of the connection between epistemic and aesthetic evaluation). As Bill Menard wrote, "[t]he basic arguments for drift were so attractive that Wegener might

well have achieved more modest goals, but he wanted it all. He could have emphasised the *simple beauty and integrating power* of the theory without insisting that it offered the *only* explanation for any *particular* observation. It did not, and experts quickly proved it" (1986:28; first italics inserted).[159] If my account is on the right track, and Drift's 'simple beauty and integrating power' means it was an elegant theory, then Drift would tend to trigger feelings of pleasure, satisfaction, and confidence. It is clear from the historical record that many found it to be a pleasing and attractive theory (see, e.g., the quotes by Argand in chapter 4). It would also trigger confidence (as it certainly did for Argand, Wegener, du Toit, van der Gracht, Holmes, and others) – at least when isolated from disconfirming evidence, and when considered in light of the evidence it did explain (many earth scientists knew only one of the areas unified by Drift well enough to be in a good position to judge the theory fairly, and thus were ignorant of much of the evidence in favour of it). But what happened when the scientists considered those 'particular observations' or alternative explanations which Menard mentions? According to the metacognitive account, it is likely that this introduced dissonance which reduced the scientists' confidence to below threshold level of theory acceptance. Note that this lack of coherence could be introduced either by evidence which apparently contradicted the theory or by a conflict with other strongly held beliefs. (Since many of Drift's critics were committed to either contraction theory or permanence theory, this introduced a bias in their reasoning about drift theory. Put in terms of Koriat's SCM, they would require access to much more supporting evidence in favour of Drift than would a neutral arbiter for the reasons they had against Drift not to form the majority of considerations.)

We do not typically evaluate a theory independently of the evidence, as we would a science fictional theory, but only evaluate it in conjunction with whatever observations and background knowledge we have of what it is a theory

[159] It's worth noting that Menard's historical account of why Wegener's theory was rejected is neither well-developed nor well supported, but in this particular claim he was probably right. Menard was a geologist not a historian, and although he handled the details of the history he did have access to with care he lacked the overview of earth science research before the Second World War that we now have access to.

of. This also holds for broadly aesthetic evaluation of theories. Simplicity may be an intrinsic feature of the theory, but we don't appreciate a theory's simplicity independently of the evidence. (I'll bet you a copy of *The Origin of Continents and Oceans* that you do not find a simple theory which cannot account for the data particularly pleasing.) Except in (highly!) unusual circumstances where one sets aside the relevant evidence to focus on the theory in isolation, aesthetic appreciation of a theory is a reaction to the experience of working with the theory and the evidence that bears on it. In particular, feelings of elegance and of confidence are sensitive to the coherence among all the considerations that are brought to bear on the question of the theory's truth, which importantly include the accessed evidence.

Applied to the current example, this means that the scientists' pleasure, satisfaction, and confidence in the theory of continental drift would depend on whether or not the scientists had in mind Menard's 'particular observations', i.e. the observations which fitted poorly with Drift. Considering only the large-scale properties and general facts which continental drift explained so well, no conflict presented itself and the features underlying its elegance triggered confidence. Considering the details the theory had trouble accounting for, a conflict between items considered would be detected, lowering confidence. It would be experienced as feelings of error, making the theory less compelling, and it would additionally be considered less elegant.

Now suppose, contrary to fact, that continental drift was *not* largely true. In that case, the positive feelings triggered by, say, its simplicity and informativeness would be misleading. And so would the simplicity and informativeness themselves (keep in mind that epistemic virtues don't guarantee truth). It is *not* part of my account that pleasure, satisfaction, and confidence infallibly signal epistemic value. On the contrary, as discussed earlier, the empirical evidence shows that metacognitive judgements in general are only moderately accurate. Furthermore, I expect feelings triggered by a theory's simplicity and informativeness to be less reliable indicators of epistemic value than those features themselves, because they add another layer where things can go wrong. For example, working out the theory's predictions may be accompanied by feelings of ease which are due to the familiarity of the topic but

misattributed to the theory's elegance. In other words, on my account we should expect the pleasure and confidence accompanying beautiful theories to not infrequently lead scientists astray. This is in line with the historical data on the usefulness of relying on aesthetic considerations in science: they tend to be helpful but there are also many historical cases where the aesthetically pleasing theory was hopelessly far from the truth (see section 4.5 above).

5.8. Alternative metacognitive accounts

Let me wrap up this chapter with a discussion of some similar accounts and how the metacognitive account presented here differs from them. Given the curious epistemicity of aesthetic experiences in science – or, alternatively, the curious aesthetic value of epistemic concerns – it is not surprising that other philosophers and psychologists also explain the aesthetics of science with reference to metacognitive feelings. I discussed Rolf Reber (2018) and Norbert Schwarz's (2018) fluency-based accounts in section 5.5.3. In a perhaps unfortunate admission that simplicity considerations are not of primary concern, I argued that fluency alone cannot account for all (or even most) aesthetic experiences.[160] Reber and Schwarz are psychologists; the philosophers developing metacognitive accounts of aesthetics in science seem to either find the variability of aesthetic experiences more salient or pressing, or find simplicity concerns less compelling.[161] In any case, the accounts developed by Cain Todd (2017) and Margherita Arcangeli and Jérôme Dokic (2020) go beyond mere fluency. In this section, I discuss how my account differs from theirs.

5.8.1. Arcangeli and Dokic (2020)

I mentioned Arcangeli and Dokic (2020) in section 3.3.1 above as one of the accounts which takes aesthetic considerations in science to be quasi-epistemic,

[160] See Shimamura and Palmer (2012), Cochrane (2021), Bullot and Reber (2013), Dokic (2016), and Forster (2020) for further discussion of the adequacy of a fluency-based account of aesthetic experiences.

[161] This fits Schindler's (2022) finding that philosophers rank simplicity lower than natural and social scientists do.

that is, conducive to non-factive understanding. This is a central difference between our accounts: I take it that aesthetic pleasure can signal the *truth-directed* epistemic success (i.e., that the theory is epistemically valuable by, for example, being simple and thus being more likely to avoid overfitting), whereas they think that "aesthetic considerations may aid scientific *understanding* rather than truth" (2020:116) which may not be so "radical" as entailing truth (2020:117).[162] In chapter 3 I argued against taking understanding to be non-factive; I will not discuss that further here.

Arcangeli and Dokic seem to accept fluency-based accounts when it comes to beauty (2020:114–7) but they argue that they cannot account for experiences of sublimity. This is the kind of experience we can have when looking up at the night sky, a huge mountain, or a powerful waterfall, which most people (apparently) experience as pleasure tinged with fear: "[t]here is an overwhelming vastness, or power, which disturbs and unsettles the mind" (2020:106). This experience "crucially hinges on negative emotions, either fear or terror, or a more general affective experience such as a 'feeling of self-negation'" (2020:117). This does not sound like the pleasure of fluent processing. Instead, they take it that "any aesthetic experience, whether about something beautiful or sublime, involves a proper balance of fluency and disfluency" (2020:109) – enough disfluency to sustain interest, which avoids the boringness of too effortlessly processed objects (see section 5.5.3). The difference between these two kinds of experiences is that sublime experiences are more disfluent (2020:118) and "involve the feeling that the theory or set of hypotheses under consideration has been pushed towards the limit of what we may cognitively

[162] Arcangeli and Dokic state that "[t]he ordinary notion of understanding is truth-independent: for instance, one may understand a sentence (such as "There is life on Alpha Centauri") even if we do not know whether it is true or false" (2020:116). But this is irrelevant to whether scientific understanding entails truth: when we talk about scientific theories giving us understanding, we are not talking (only) about understanding what the theory says but understanding the phenomenon the theory is about. This is one of the main differences between science fiction and science: we can revel in the Bussard engine of Paul Anderson's (1970) *Tau zero*, understanding every sentence of the book and yet not gain any understanding of what actual near-*c* space travel would be like. See section 3.3. above for more on this.

encompass" (2020:120). In other words, sublime theories spell the limits of understanding.

It is difficult to evaluate their proposal without more details on how (dis)fluency gives rise to aesthetic appreciation and, especially, which features of theories or our interaction with theories trigger these metacognitive signals, but these details are unfortunately not given in Arcangeli and Dokic (2020). I take it to be a strength of my account that it specifies specific features of theories which trigger the feelings at the heart of aesthetic experiences in science. In other words, my account is more informative: it gives us more detailed knowledge (if true) than Arcangeli and Dokic's account can provide.

On the other hand, their account is wider in scope since it extends to experiences of sublimity, which I have not discussed. I take it from the examples they provide that there is a need for an account of sublimity in science. Aesthetic experiences are rich and varied: they go beyond experiences of beauty, and they go beyond experiences of sublimity. (Including dramatic, cool, or graceful experiences, and including all kinds of *negative* aesthetic experiences: of ugliness, boredom, messiness, disgust.) A complete account of aesthetics in science must explain a wide variety of aesthetic experiences. More work is needed, but it's great that Arcangeli and Dokic have made a start down this road.

5.8.2. Todd (2017)

Cain Todd's (2017) metacognitive account is a close forerunner to the account I present here. Todd argues that at least some aesthetic experiences in mathematics are metacognitively generated and that these 'aesthetic-epistemic feelings' "serve a genuine cognitive and epistemic function" (2017:211). (He discusses mathematics only, but I assume that his account can be extended to aesthetic experiences in the natural and social sciences without problems.) These feelings are both aesthetic and metacognitive: "we can think of (some) epistemic feelings as possessing aesthetic attributes, or of (some) aesthetic feelings as exhibiting epistemic attributes, or of a range of feelings that are jointly aesthetic-epistemic in nature" (2017:227). The connection between aesthetic and metacognitive feelings is (partially) suggested by how both respond to fluency, including how "symmetry, simplicity, fluency, order, clarity" "occur as the

objects and causes of both aesthetic and epistemic experiences" (2017:227), and how the "pleasure of understanding" is often connected to "harmony or fit" (2017:228).

Like Arcangeli and Dokic (2020), Todd (2017) does not go into detail on which features of theories or of the scientists' cognitive engagement with theory and evidence figure in the metacognitive processes underlying aesthetic experiences. What does the work is the general similarity between metacognitive and aesthetic feelings – including, as we have seen, how fluency gives rise to both positive metacognitive and positive aesthetic feelings. The details provided here may be a way of fleshing out Todd's account: it's difficult to say whether he would approve of my account without knowing how he would have fleshed out his.

However, Todd also presents an intriguing "highly speculative" (2017:230) "strong hypothesis" (2017:229):

> All epistemic feelings and aesthetic experiences are determinate variations of the 'feeling of fittingness', which is a determinable experiential manifestation of certain sub-personal processes of representing and assessing the relation between certain cognitive processes and certain properties of the stimuli at which they are directed. When this relation exhibits, among other things: a) coherence b) breadth of scope, c) simplicity, and d) fluency, we have what I will call 'cognitive consonance'. It is this consonance that is manifested as a feeling of fittingness, and this feeling is jointly aesthetic-epistemic in nature. (2017:229)

The determinate-determinable distinction is best explained with an example: *green* is a determinate of the determinable *colour*. There are many different colour properties and something is coloured only if it has one of these determinate colour properties. Going back to Todd's hypothesis, cognitive consonance is a determinate property (like colour) and can take many forms (like green, blue, red): for example, feelings of understanding, fit/coherence, harmony, elegance, beauty, or wit (2017:228–9).

Note that the features Todd lists (coherence, breath of scope, simplicity, and fluency) are almost the same as the features I take to underlie feelings of elegance (simplicity, fluency, informativeness, and coherence). (Note, further, that neither of us claims to have presented a complete account of these aesthetic virtues and experiences.) Unfortunately, Todd does not expand on his

hypothesis. To the extent that the features of theories and of our engagement with them I have discussed are the same features which figure in Todd's 'feeling of fittingness', this chapter presents one way of fleshing out his hypothesis. That said, an important difference between our accounts is that I aim to explain how aesthetic experiences connect to commonly invoked theoretical virtues whereas Todd's 'strong hypothesis' seems to go more in the direction of explaining aesthetic and epistemic virtues as two sides of the same coin, à la Kivy (1991) and Bird (2020) discussed in section 5.1.1 above.

5.9. Conclusion

In this chapter we have examined the framework I propose to explain the curious role of aesthetics in science. Why do aesthetic feelings appear in scientific contexts in the first place? Because they are outputs of a metacognitive monitoring of the quality and success of the scientist's cognitive engagement with theory and evidence. Why do they appear aesthetic? Presumably, because they are triggered by the kinds of features of theories which makes theories aesthetically valuable. How can they be even moderately epistemically valuable? Because they tend to reflect epistemically valuable features of theories, such as simplicity and informativeness. Is this all we need to know to explain the role of aesthetic considerations in science? No! The metacognitive account as presented here is but a framework with some of the details filled in. Much more work needs to be done to answer outstanding questions I have not had the space to consider, such as under what conditions aesthetic considerations are useful, and to test this account both against a wider selection of historical cases and through psychological experiments. Work on the history and philosophy of aesthetics in science is still in its early stages, and there is so much we do not know yet. If nothing else, I hope the account presented here has shown how empirical studies on affect and cognition can be fruitfully combined with the history and philosophy of science to illuminate the puzzling role that aesthetic experiences play in scientific reasoning.

CHAPTER 6:
CONCLUSION

We have covered quite a lot of ground across different philosophical subdivisions (philosophy of science, epistemology, aesthetics) and academic fields (cognitive science, history of science), and yet we have barely scratched the surface of everything there is to say about the role aesthetic considerations play in scientific reasoning. Let me end by briefly summarising the main threads of this thesis and some lessons we should draw from it, before outlining some avenues for further research.

In this thesis we have explored the role broadly aesthetic theoretical virtues like elegance, simplicity, unity and beauty play in theory choice, leading up to a new explanation of the puzzling fact that aesthetic considerations can provide useful guidance in scientific theorising. We began by locating the discussion on aesthetic and other theoretical virtues – the features of a good theory – in the literature on inference to the best explanation. Since IBE licenses accepting as true or nearly true the theory that best explains our evidence, and aesthetic virtues are often appealed to as signs of explanatory goodness, we should (*ceteris paribus*) accept the most elegant, the simplest, or otherwise most beautiful theory. This in effect treats the aesthetic virtues as being epistemically valuable: whether that is justifiable is the main issue explored in the thesis. To focus on this question without undue hedging or getting sidetracked, here I have taken for granted that the best explanation will in many cases be the true one. However, I think that most of what I say here can be reformulated without other casualties than style to fit more modest approaches to explanatory reasoning. Setting this aside, the question is whether a theory's aesthetic qualities can contribute to making it the best explanation.

Answering *that* requires both understanding what aesthetic virtues are and what it would mean for them to be epistemically valuable. We explored the nature of aesthetic virtues in chapter 2, and I argued that we should take scientists at face value as making aesthetic judgements when they apparently do so. However, even if you doubt that scientists report genuine aesthetic experiences or that scientific theories can have genuine aesthetic qualities, the

question of what role these *apparently* aesthetic considerations play in science arises: what are these apparently affect-based judgements, why do they impact scientific reasoning, and how can they do so to generally good effect?

Chapter 3 discusses what it would mean to say that aesthetic virtues are epistemically valuable. The traditional account has it that virtues are epistemic if they are conducive to truth, that is, if they make a theory having these virtues more likely to be true than a theory lacking them. But that misclassifies virtues which seems to be clearly epistemic. In particular, a strong or informative and wide-scoped theory is epistemically good, but a theory which is committed to more cannot be more likely than a theory which is committed to less. I discuss some alternatives to the traditional account, including my preferred account, according to which epistemic virtues are knowledge-conducive. This is still truth-directed but can account for the many ways virtues can be epistemic, not just by increasing the theory's probability but by broadening or deepening the knowledge we may gain from it. In the recent literature on aesthetics in science it has become common to argue that aesthetic virtues can be conducive to non-factive understanding. This is typically motivated by historical cases showing both that aesthetic considerations can be useful epistemic guides and that they can lead scientists astray: therefore, so the argument goes, we should account for the epistemic value of aesthetic considerations without linking it to truth. I argue that understanding, even by the lights of self-proclaimed non-factivists like Angela Potochnik and Catherine Elgin, require too much truth for this to be a genuine alternative to factive accounts. And even if some virtues can give us genuinely non-factive understanding, that would be sufficiently different from epistemic value as it is normally thought of that we can still ask whether aesthetic virtues can make a theory *epistemically* better even if they are also conducive to genuinely non-factive understanding.

In the rest of the thesis, through a detailed case study and by developing an explanatory account, I argue that aesthetic virtues can be moderately epistemically useful.

The case study connects the quite abstract discussion of theoretical virtues to actual scientific practice, illustrating in detail what it is that we're trying to explain, and show how pervasive aesthetic evaluations of theories are. While

the argumentation, conceptual clarification, and other aspects of classical philosophical methodology used throughout the thesis is valuable (for example, in showing what epistemic good aesthetic virtues may provide), if we want to understand what role aesthetic considerations play in scientific reasoning we must look to actual scientific practice. Here I have given a case study from the history of science, but one could also instead – or, preferably, in addition – carry out surveys of contemporary scientists' aesthetic experiences (as in e.g. Vaidyanathan et al. (2023), Ivanova et al. (2024)) or case studies of the role of aesthetics in current research (e.g., Hossenfelder (2018)). These different methods illuminate different aspects of the role of aesthetics in science, and to fully understand this strange aspect of scientific research we should consider it from all possible angles. Hence the current focus on the earth sciences, to show that aesthetic considerations play a role in theorising also in less abstract sciences.

I present multiple cases of how aesthetic considerations made a difference to both the development and the acceptance of the theory of plate tectonics, from the first formulations of its immediate predecessor until it was generally accepted by the scientific community. For example, Wegener put considerable weight on the simple and harmonious picture of the world his theory presented, Argand and Vine took its elegance to be a reason for believing the theory, and apparently disconfirming evidence was in many cases dismissed when the alternative would have been to modify the elegant theory.

This turned out to be the right move: plate tectonics is one of the most successful theories in science history. But science history also contains examples of aesthetic considerations being misleading, so it seems that aesthetics is only a moderately useful guides in theorising.

To explain this, I develop what I call the metacognitive account of aesthetics in science. This account connects the aesthetic experiences underlying aesthetic judgements of theories to metacognition and metacognitive feelings. Metacognition is the monitoring and evaluation of one's first-order cognitive states and processes. Much of this monitoring is conveyed through feelings such as feelings of certainty, feeling that you know the answer to a question, or feeling that you just made an error. Intriguingly, many of the same features in virtue of

which theories are beautiful are the same features which psychologists' study as cues, or triggers, for metacognitive feelings. This connection is the basis for my explanation for why elegance, simplicity, and other aesthetic virtues play both an epistemic and aesthetic role in science.

Briefly, the metacognitive account says that the pleasure and satisfaction felt when working with a beautiful theory can signal the presence of epistemic virtues, because these feelings are generated by metacognitive processes responding to epistemically relevant features of theories and to one's cognitive engagement with theory and evidence. The feelings reflect the quality of the cognitive process involved in working with the theory, such as the fluency of working with it or the fit between items considered which, in turn, tend to reflect epistemically valuable features of the theory, such as simplicity and coherence. These feelings can then be used as a rough signal of the presence of such epistemically valuable features.

This account, then, connects scientists' aesthetic experiences to epistemic theoretical virtues. This may not account for all aspects of aesthetic considerations in scientific reasoning, but it does provide a promising framework for explaining why scientists appeal to aesthetic virtues in the first place and how it can be that these virtues are moderately useful guides in scientific theorising.

This immediately raises the crucial question: under what conditions are these aesthetic feelings reliable indicators of epistemic value? In the discussion on the illusion of understanding, we discussed some conditions under which metacognitive feelings are reliable indicators of one's epistemic situation. Furthermore, as suggested by the discussion on both familiarity and fit as metacognitive cues, it is likely that one's background knowledge and other available information influences both one's aesthetic and epistemic evaluation of a theory. This is likely to have ramifications for the conditions under which metacognitive feelings can be indicators of epistemic success and the conditions under which they are detrimental to scientific reasoning. The next step is to test the metacognitive account empirically, including testing the conditions under which metacognitive feelings are more likely to have make a positive contribution to scientific reasoning. It would also be good to test through psychological experiments to what extent how a theory is presented changes

how it is evaluated, both aesthetically and epistemically. Testing the account empirically would be a significant piece of research requiring close collaboration with psychologists – showing, again, that we need a multitude of both theoretical and empirical methods to understand the puzzling role of aesthetic considerations in science.

The metacognitive account provides a framework for explaining the role of aesthetics in science. As I have stressed throughout, this thesis presents a programmatic account with only some of the details filled in: filling in the picture will be left for further work. Let me end by briefly mentioning three other areas of further research. First, it would be good to explore whether *all* broadly aesthetic virtues can be accounted for through this framework. Can it account for experiences of sublimity in science? Can it illuminate the role of unity, and the distaste we have towards ad hocness? (I think the answer is *Yes,* but showing that requires further work.) Second, it would be good to explore the extent to which there is variability between individuals and scientific communities in their aesthetic judgements, and to what extent this can be explained as a function of different metacognitive cues (such as what one is familiar with or what evidence is easily available in a community). Third, it would be good to explore the extent to which the metacognitive account can be extended to account for aesthetic evaluation of evidence evidential reasoning. I have stressed that theories are typically evaluated against the available evidence and thus that the metacognitive and hence aesthetic feelings (typically) respond to theory and evidence taken together. To what extent is this compatible with Caitlin Wylie (2021) and Adrian Currie's (2023a) work on the aesthetics of fossil preparation, Derek Turner's (2019) work on field sketches, and Milena Ivanova's research programme on the aesthetics of scientific experiments (e.g. Ivanova and Murphy (2023))? Recent work by Nick Shea (2024) on metacognitive feelings as signals of reliable inferential transitions suggests that we can extend the metacognitive account to evidential reasoning, but these are open questions which require further research.

Judging by the theoretical, historical, and empirical support we have found for the metacognitive account in this thesis, we can reasonably expect that

further development of this framework will help illuminate the value of aesthetic considerations in scientific research.

References

Ackerman, R. (2014) "The Diminishing Criterion Model for Metacognitive Regulation of Time Investment," *Journal of Experimental Psychology: General*, 143, 3, pp. 1349–68.

Ackerman, R. and Zalmanov, H. (2012) "The Persistence of the Fluency-Confidence Association in Problem-Solving," *Psychonomic Bulletin & Review*, 19, pp. 1187–92.

Ackerman, R. and Thompson, V. (2017) "Meta-Reasoning: Monitoring and Control of Thinking and Reasoning," *Trends in Cognitive Sciences*, 21, 8, pp. 607–17.

Adkins, O. C. and Norman, J. F. (2016) "The Visual Aesthetics of Snowflakes," *Perception*, 45, 11, pp. 1304–19.

Alter, A. L. and Oppenheimer, D. M. (2009) "Uniting the Tribes of Fluency to Form a Metacognitive Nation," *Personality and Social Psychology Review*, 13, 3, pp. 219–35.

Anderson, D. L. (2002) "Occam's Razor: Simplicity, Complexity, and Global Geodynamics," *Proceedings of The American Philosophical Society*, 146, 1, pp. 56–76.

Arango-Muñoz, S. and Michaelian, K. (2014) "Epistemic Feelings, Epistemic Emotions: Review and Introduction to the Focus Section," *Philosophical Inquiries*, 2, pp. 97–122.

Arcangeli, M. and Dokic, J. (2020) "A Plea for the Sublime in Science," pp. 104–24, in M. Ivanova and S. French (eds.) *The Aesthetics of Science: Beauty, Imagination and Understanding*. New York and Abingdon: Routledge.

Armstrong, T. and Detweiler-Bedell, B. (2008) "Beauty as an Emotion: The Exhilarating Prospect of Mastering a Challenging World," *Review of General Psychology*, 12, 4, pp. 305–329.

Atwater, T. (2003) "When the Plate Tectonic Revolution Met Western North America," pp. 243–63, in N. Oreskes (ed.) *Plate Tectonics: An Insider's History of the Modern Theory of the Earth*. Colorado: Westview Press.

Baker, A. (2022) "Simplicity," in E. N. Zalta (ed.), *The Stanford Encyclopedia of*

Philosophy.
https://plato.stanford.edu/archives/sum2022/entries/simplicity/

Ball, L., Threadgold, E., Marsh, J. and Christensen, B. (2018) "The Effects of Stimulus Complexity and Conceptual Fluency on Aesthetic Judgments of Abstract Art: Evidence for a Default-Interventionist Account," *Metaphor and Symbol*, 33, pp. 235–52.

Baumberger, C., Beisbart, C., and Brun, G. (2017) "What is Understanding? An Overview of Recent Debates in Epistemology and Philosophy of Science," pp. 1–34, in S. Grimm, C. Baumberger, and S. Ammon (eds.) *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*.

Beardsley, M. (1982) "Aesthetic Experience," pp. 285–97, in D. M. Callen, M. J. Wreen, and M. C. Beardsley (eds.) *The Aesthetic Point of View: Selected Essays*. Cornell University Press.

Beebe, J. R. (2009) "The Abductivist Reply to Skepticism," *Philosophy and Phenomenological Research*, 79, pp. 605–36.

Benacerraf, P. (1973) "Mathematical Truth," *Journal of Philosophy*, 70, 19, pp. 661–79.

Beran, M. J., Brandl, J., Perner, J. and Proust, J. (2012) "On the Nature, Evolution, Development, and Epistemology of Metacognition: Introductory Thoughts," pp. 1–18, in M. J. Beran, J. Brandl, J. Perner, and J. Proust (eds.) *Foundations of Metacognition*. Oxford: Oxford University Press.

Beran, M. J., Brandl, J., Perner, J. and Proust, J. (2012) *Foundations of Metacognition*. Oxford University Press.

Berleant, A. (1978) "Aesthetic Paradigms for an Urban Ecology," *Diogenes*, 26, 103, pp. 1–28. doi:10.1177/039219217802610301.

Berleant, A. (2020) "Aesthetic Sensibility," *Ambiances*, online. Accessible at https://doi.org/10.4000/ambiances.526

Bird, A. (2020) "How Can Loveliness Be a Guide to Truth? Inference to the Best Explanation and Exemplars," pp. 125–45, in Ivanova, M. and French, S. (eds.) *The Aesthetics of Science: Beauty, Imagination and Understanding*. New York and Abingdon: Routledge.

Bird, A. (2022) *Knowing Science*. Oxford: Oxford University Press.

Blinderman, C. S. (1962) "T. H. Huxley's Theory of Aesthetics: Unity in Diversity," *The Journal of Aesthetics and Art Criticism*, 21, 1, pp. 49–55. https://www.jstor.org/stable/427638

BonJour, L. (1980) "Externalist Theories of Empirical Knowledge," pp. 53–73, in P. French, T. Uehling Jr. and H. Wettstein (eds.) *Midwest Studies in Philosophy*, vol. 5. Minneapolis: University of Minnesota Press.

Bogen, J. and Woodward, J. (1998) "Saving the Phenomena," *The Philosophical Review*, 97, 3pp. 303–52.

Boldt, A., de Gardelle, V. and Yeung, N. (2017) "The Impact of Evidence Reliability on Sensitivity and Bias in Decision Confidence," *Journal of Experimental Psychology: Human Perception and Performance*, 43, 8, pp. 1520–31.

Boyd, R. (1983) "On the Current Status of the Issue of Scientific Realism," *Erkenntnis*, 19, 1–3, pp. 45–90.

Breitenbach, A. (2013) "Aesthetics in Science: A Kantian Proposal," *Proceedings of the Aristotelian Society*, 113, 1, pp. 83–100.

Brickhouse-Bryson, D. (2020) *Judgments of Beauty in Theory Evaluation*. Lanham, Maryland: Lexington Books.

Bueno, O. and Shalkowski, S. A. (2015) "Modalism and Theoretical Virtues: Toward an Epistemology of Modality," *Philosophical Studies*, 172, 3, pp. 671–89.

Bueno, O., Derby, G., French, S. and Rickles, D. (eds.) (2018) *Thinking about Science, Reflecting on Art: Bringing Aesthetics and Philosophy of Science Together*. London and New York: Routledge.

Bullard, E. C., Everett, J. E., and Smith, A. G. (1965) "The Fit of the Continents Around the Atlantic," *Philosophical Transactions of The Royal Society*, 258, pp. 41–51.

Bullot, N. J., and Reber, R. (2013) "The Artful Mind Meets Art History: Toward a Psycho-Historical Framework for the Science of Art Appreciation," *Behavioral and Brain Sciences*, 36 2, pp. 123–37.

Brogaard. B. (2005) "I Know, Therefore I Understand," unpublished manuscript.

Cabrera, F. (2023) "Inference to the Best Explanation: An Overview," in L. Magnani (Ed.) *Handbook of Abductive Cognition*. Springer, Cham. https://doi.org/10.1007/978-3-031-10135-9_77

Camp, E. and Shupe, E. (2017) "Instrumental Reasoning in Nonhuman Animals," pp. 100–18, in K. Andrews and J. Beck (eds.) *The Routledge Handbook of Philosophy of Animal Minds*. Routledge.

Carroll, N. (2002) "Aesthetic Experience Revisited," *The British Journal of Aesthetics*, 42, 2, pp. 145–68.

Carruthers, P. and Ritchie, J. B. (2012) "The Emergence of Metacognition: Affect and Uncertainty in Animals," pp. 76–93, in M. J. Beran, J. Brandl, J. Perner, and J. Proust (eds.) *Foundations of Metacognition*. Oxford: Oxford University Press.

Carlson, A. (1979) "Appreciation and the Natural Environment," *The Journal of Aesthetics and Art Criticism*, 37, 3, pp. 267–75.

Carto, S. L. and Eyles, N. (2012) "Sedimentology of the Neoproterozoic (c. 580 Ma) Squantum 'Tillite', Boston Basin, USA: Mass flow deposition in a deep-water arc basin lacking direct glacial influence," *Sedimentary Geology*, 269, pp. 1–14.

Chamberlin, T. C. ([1897] 1970) "The Method of Multiple Working Hypotheses," pp. 604–9, in K. F. Mather and S. L. Mason (eds.) *A Sourcebook in Geology, 1400–1900*. Cambridge: Harvard University Press.

Chandrasekhar, S. (1987) *Truth and Beauty: Aesthetics and Motivations in Science.* University of Chicago Press.

Chakravartty, A. (2017) "Scientific Realism," in E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2017 Edition). https://plato.stanford.edu/archives/sum2017/entries/scientific-realism/

Chmiel, A. and Schubert, E. (2017) "Back to the Inverted-U for Music Preference: A Review of The Literature," *Psychology of Music*, 45, 6, pp. 886–909.

Chirimuuta, M. (2015) *Outside Color: Perceptual Science and the Puzzle of Color in Philosophy*. Cambridge, MA and London: the MIT Press.

Chirimuuta, M. (2023) "Haptic Realism for Neuroscience," *Synthese*, 63, pp.

1–16. https://doi.org/10.1007/s11229-023-04295-z

Cochrane, T. (2021) *The Aesthetic Value of the World*. Oxford: Oxford University Press.

Cooper, D. E. (ed.) (1997) *Aesthetics: The Classic Readings*. Oxford: Blackwell.

Cox, A. (ed.) (1973) *Plate Tectonics and Geomagnetic Reversals*. San Francisco: W. H. Freeman and Company.

Cox, A., and Hart, R. B. (1991) *Plate Tectonics: How it Works*. Blackwell Scientific Publications: John Wiley & Sons, Inc.

Cox, A., Doell, R. R., and Dalrymple, G. B. (1964) "Reversals of the Earth's Magnetic Field," *Nature*, 144, 3626, pp. 1537–43.

Currie, A. (2023) "Epistemic Engagement, Aesthetic Value, and Scientific Practice," *The British Journal for the Philosophy of Science*, 74, 2, pp. 313–34.

Cutting, J. E. (2020) "Goldilocks Aesthetics," *Projections*, 14, 2, pp. 66–74.

Dechêne, A., Stahl, C., Hansen, J. and Wänke, M. (2010) "The Truth About the Truth: A Meta-Analytic Review of the Truth Effect," *Personality and Social Psychology Review*, 14, 2, pp. 238–57.

De Neys, W., Vartanian, O., and Goel, V. (2008) "Smarter than We Think: When Our Brains Detect That We Are Biased," *Psychological Science*, 19, 5, pp. 483–489.

De Neys, W., Moyens, E., and Vansteenwegen, D. (2010) "Feeling We're Biased: Autonomic Arousal and Reasoning Conflict," *Cognitive, Affective, & Behavioral Neuroscience*, 10, 2, pp. 208–216.

De Neys, W., Cromheeke, S., and Osman, M. (2011) "Biased But in Doubt: Conflict and Decision Confidence," *PloS one*, 6, 1, e15954.

Delplanque, J., De Loof, E., Janssens, C. and Verguts, T. (2019) "The Sound of Beauty: How Complexity Determines Aesthetic Preference," *Acta Psychologica*, 192, pp. 146–52.

Dellsén, F. (2024) *Abductive Reasoning in Science*. Cambridge: Cambridge University Press.

De Regt, H. (2024) "Understanding in Science and Beyond," pp. 23–41, in A. I. Mărăsoiu and M. Dumitru (eds.) *Understanding and Conscious Experience: Philosophical and Scientific Perspectives*. New York: Routledge.

Dicken, P. and Lipton, P. (2006) "What can Bas believe? Musgrave and van

Fraassen on Observability," *Analysis*, 66, pp. 226–33.

Dietz, R. S. (1961) "Continents and Ocean Basin Evolution by Spreading of the Seafloor," *Nature*, 190, pp. 854–7.

Dirac, P. A. M. (1980) "Why We Believe in the Einstein Theory," pp. 1–11, in. B. Gruber and R. S. Millman (eds.) *Symmetries in Science*. New York and London: Plenum Press.

Dokic, J. (2016) "Aesthetic Experience as a Metacognitive Feeling? A Dual-Aspect View," *Proceedings of the Aristotelian Society*, 1, pp. 69–88.

Douglas, H. E. (2009) *Science, Policy, and the Value-Free Ideal*. University of Pittsburgh Press.

Downes, S. (2021) *Models and Modelling in the Sciences: A Philosophical Introduction*. Abingdon: Routledge.

Dunlosky J. and Metcalfe, J. (2009) *Metacognition*. California: SAGE publications, Inc.

Dunlosky J. and Tauber, S. (2015) "A Brief History of Metamemory Research and Handbook Overview," pp. 7–22, in J. Dunlosky and S. Tauber (eds.) *The Oxford Handbook of Metamemory*. Oxford: Oxford University Press.

Elgin, C. Z. (2017) *True Enough*. Cambridge, MA: MIT Press.

Elgin, C. Z. (2019) "True Enough," pp. 384–98, in D. Edwards (ed.) *Truth: A Contemporary Reader*. London and New York: Bloomsbury Academic.

Elgin, C. Z. (2020) "Epistemic Gatekeepers: The Role of Aesthetic Factors in Science," pp. 21–35, in Ivanova, M. and French, S. (eds.) *The Aesthetics of Science: Beauty, Imagination and Understanding*. New York and Abingdon: Routledge.

Engler, G. (1990) "Aesthetics in Science and in Art," *British Journal of Aesthetics*, 30, 1, pp. 24–34.

Evans, J. St. B. T. (2007) *Hypothetical Thinking: Dual Processes in Reasoning and Judgement*. Hove and New York: Psychology Press.

Fan, X., Myers, T. G., Sukra, B. A. D., and Gabrielse, G. (2023) "Measurement of the Electron Magnetic Moment," *Physical Review Letters*, 130, pp. 1–7.

Feldman, R. and Conee, E. (2001) "Internalism Defended," *American Philosophical Quarterly*, 38, 1, pp. 1–18.

Fernandez Cruz, A. L., Arango-Muñoz, S., and Volz, K. (2016) "Oops, Scratch

That! Monitoring One's Own Errors During Mental Calculation," *Cognition*, 146, pp. 110–120.

Fernbach, P. M., Rogers, T., Fox, C. R., and Sloman, S. A. (2013) "Political Extremism is Supported by an Illusion of Understanding," *Psychological Science*, 24, 6, pp. 939–46. https://doi.org/10.1177/0956797612464058

Fields, C., Glazebrook, J.F. (2020) "Do Process-1 Simulations Generate the Epistemic Feelings that Drive Process-2 Decision Making?" *Cognitive Processing*, 21, pp. 533–53. https://doi.org/10.1007/s10339-020-00981-9

Finn, B. (2008) "Framing Effects on Metacognitive Monitoring and Control," *Memory & Cognition*, 36, 4, pp. 813–21.

Fisher, M., Goddu, M. K., and Keil, F. C. (2015) "Searching for Explanations: How the Internet Inflates Estimates of Internal Knowledge," *Journal of Experimental Psychology: General*, 144, 3, pp. 674–87.

Forster, M. (2020) "Processing Fluency," pp. 430–46, in M. Nadal and O. Vartanian (eds.) *The Oxford Handbook of Empirical Aesthetics*. https://doi.org/10.1093/oxfordhb/9780198824350.013.21

Forster, M. and Sober, E. (1994) "How to Tell When Simpler, More Unified, or Less *Ad Hoc* Theories Will Provide More Accurate Predictions," *British Journal for the Philosophy of Science*, 45, pp. 1–35.

Frankel, H. R. (2012) *The Continental Drift Controversy*. Vol. I–IV. Cambridge: Cambridge University Press.

Frisch, W., Meschede, M., and Blakey, R. (2011) *Plate Tectonics: Continental Drift and Mountain Building*. Heidelberg: Springer.

Franks, J. (2010) "Cantor's Other Proofs that R is Uncountable," *Mathematics Magazine*, 83, 4, pp. 283–89. http://www.jstor.org/stable/10.4169/002557010X521822

French, S. (2020) "Performance and Practice: Situating the Aesthetic Qualities of Theories," pp. 168–210, in Ivanova and French (eds.) *The Aesthetics of Science: Beauty, Imagination and Understanding*. New York and Abingdon: Routledge.

Gaviria, C., Corredor, J. (2021) "Illusion of Explanatory Depth and Social Desirability of Historical Knowledge," *Metacognition Learning*, 16, pp. 801–32. https://doi.org/10.1007/s11409-021-09267-7

Gettier, Edmund (1963) "Is Justified True Belief Knowledge?" *Analysis*, 23, pp. 121–3.

Giere, R. N. (1988) *Explaining Science: A Cognitive Approach*. Chicago: University of Chicago Press.

Gigerenzer, G. (1991) "How to Make Cognitive Illusions Disappear: Beyond 'Heuristics and Biases'," *European Review of Social Psychology*, 2, 1, pp. 83–115.

Gigerenzer, G. (2007) *Gut Feelings: The Intelligence of the Unconscious*. New York: Viking Penguin.

Glen, W. (1982) *The Road to Jaramillo: Critical Years in the Revolution in Earth Science*. Stanford: Stanford University Press.

Glynn, I. (2010) *Elegance in Science: The Beauty of Simplicity*. Oxford: Oxford University Press.

Goldman (1976) "What is Justified Belief?" pp. 1–23, in G. S. Pappas (ed.) *Justification and Knowledge*. Dordrecht: D. Reidel.

Godfrey-Smith, P. (2009) "Abstractions, Idealizations, and Evolutionary Biology," pp. 47–55, in A. Barberousse, M. Morange, and T. Pradeu (eds.) *Mapping the Future of Biology*. Boston Studies in the Philosophy of Science. Springer.

Goldsmith, M., and Koriat, A. (1999) "The Strategic Regulation of Memory Reporting: Mechanisms and Performance Consequences," pp. 373–400, in D. Gopher and A. Koriat (eds.) *Attention and Performance XVII. Cognitive regulation of performance: Interaction of theory and application*. Cambridge, MA: the MIT Press.

Goupil, L. and Proust, J. (2023) "Curiosity as a Metacognitive Feeling," *Cognition*, 231, 105325. https://doi.org/10.1016/j.cognition.2022.105325

Greene, M. T. (2015) *Alfred Wegener: Science, Exploration, and the Theory of Continental Drift*. Baltimore, Maryland: Johns Hopkins University Press.

Griffin, D. and Tversky, A. (1992) "The Weighing of Evidence and the Determinants of Confidence," *Cognitive Psychology*, 24, pp. 411–35.

Grice, H. P. (1975) "Logic and Conversation," pp. 41–58, in P. Cole, and J. Morgan (eds.) *Syntax and Semantics*. New York: Academic Press

Hallam, A. (1989) *Great Geological Controversies*. 2nd ed. Oxford: Oxford

University Press.

Hannon, M. (2021) "Recent Work in the Epistemology of Understanding," *American Philosophical Quarterly*, 58, 3, pp. 269–90.

Hanson, L. (2018) "Moral Realism, Aesthetic Realism, and the Asymmetry Claim," *Ethics*, 129, pp. 39–69.

Hardy, G. H. ([1940] 2005) *A Mathematician's Apology*. University of Alberta Mathematical Sciences Society. Available at http://www.math.ualberta.ca/mss/

Harman, G. (1965) "The Inference to the Best Explanation," *The Philosophical Review*, 74, 1, pp. 88–95

Harvey, A. C. (2021) *'Richness in Detail' as a Veracity Cue: Examining the Impact of Suboptimal Recall Contexts*. PhD thesis, Department of Psychology, University of Portsmouth.

Harvey, A. C., Vrij, A., Leal, S., Deeb, H., Hope, L., and Mann, S. (2020) "Amplifying Recall After Delays via Initial Interviewing: Inoculating Truthtellers' Memory as a Function of Encoding Quality," *Acta Psychologica*, 209, 103130, pp. 1–10. https://doi.org/10.1016/j.actpsy.2020.103130

Heisenberg, W. (1971) "The Meaning of Beauty in the Exact Sciences," address delivered to the Bavarian Academy of Fine Arts, Munich (1970). Transl. P. Heath. Accessible at https://docslib.org/doc/6615237/the-meaning-of-beauty-in-the-exact-sciences.

Heller, F. and Petersen, N. (1982) "Self-Reversal Explanation for The Laschamp/Olby Geomagnetic Field Excursion", *Physics of the Earth and Planetary Interiors*, 30, 4, pp. 358–72.

Hempel, C. G. (1966) *Philosophy of Natural Science*. Princeton: Princeton University Press.

Hess, H. ([1962] 1973) "History of Ocean Basins," pp. 23–38, in A. Cox (ed.) *Plate Tectonics and Geomagnetic Reversals*. San Francisco: W. H. Freeman & Co.

Hills, A. (2016) "Understanding Why," *Noûs*, 49, 2, pp. 661–88.

Holmes, A. (1928) "Continental Drift," *Nature*, 3073, 122, pp. 431–3.

Horwich, P. (1982) *Probability and Evidence*. Cambridge: Cambridge University

Press.

Hossenfelder, S. (2018) *Lost in Math: How Beauty Leads Physics Astray*. New York: Basic Books.

Hume, D. ([1757] 1965) "Of the Standard of Taste," pp. 231–55, in D. Hume, *Essays: Moral, Political and Literary*. Oxford: Oxford University Press.

Hutcheson, F. ([1726] 2004) *An Inquiry into the Original of Our Ideas of Beauty and Virtue in Two Treatises*. Indianapolis: Liberty Fund.

Irving, E. and Runcorn, S. K. (1957) "Analysis of the Palaeomagnetism of the Torridonian Sandstone Series of North-West Scotland," *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 250, 974, pp. 83–99.

Ivanova, M. (2017) "Poincaré's Aesthetics of Science," Synthese, 194, pp. 2581–94.

Ivanova, M. (2020) "Beauty, Truth and Understanding," pp. 86–103, in M. Ivanova and S. French (eds.) *The Aesthetics of Science: Beauty, Imagination and Understanding*. New York and Abingdon: Routledge.

Ivanova, M. (2021) "The Aesthetics of Scientific Experiments," *Philosophy Compass*, 16, 3, pp. 1–9.

Ivanova, M. (2023) "The Aesthetic Value of Scientific Experiments," pp. 15–36, in M. Ivanova and A. Murphy (eds.) *The Aesthetics of Scientific Experiments*. New York: Routledge.

Ivanova, M. and French, S. (eds.) (2020) *The Aesthetics of Science: Beauty, Imagination and Understanding*. New York and Abingdon: Routledge.

Ivanova, M. and Murphy, A. (eds.) (2023) *The Aesthetics of Scientific Experiments*. New York: Routledge.

Ivanova, M., Ritz, B., Duque, M., and Vaidyanathan, B. (2024) "Beauty in Experiment: A Qualitative Analysis of Aesthetic Experiences in Scientific Practice," *Studies in History and Philosophy of Science*, 104 pp. 3–11. https://doi.org/10.1016/j.shpsa.2024.02.003

Irak, M., Soylu, C., Turan, G., and Çapan, D. (2019) "Neurobiological Basis of Feeling of Knowing in Episodic Memory," *Cognitive Neurodynamics*, 13, pp. 239–56. https://doi.org/10.1007/s11571-019-09520-5

Kaku, M. and Thompson, J. (1997) *Beyond Einstein: Superstrings and the Quest for the*

Final Theory. Oxford: Oxford University Press.

Kant, I. ([1790] 2000) *Critique of the Power of Judgement*. P. Guyer (ed.), transl. E. Matthews. Cambridge: Cambridge University Press.

Keas, M. K. (2018) "Systematizing the Theoretical Virtues," *Synthese*, 195, pp. 2761–93. https://doi.org/10.1007/s11229-017-1355-6

Kelley, C. M., & Lindsay, D. S. (1993) "Remembering Mistaken for Knowing: Ease of Retrieval as a Basis for Confidence in Answers to General Knowledge Questions," *Journal of Memory and Language*, 32, pp. 1–24.

Khalifa, K. (2017) "Understanding, Truth, and Epistemic Goals," *Philosophy of Science*, 87, pp. 944–56.

Khalifa, K. (2020) *Understanding, Explanation, and Scientific Knowledge*. Cambridge: Cambridge University Press.

Kitcher, P. (1981) "Explanatory Unification," *Philosophy of Science*, 48, 4, pp. 507–31.

Kivy, K. (1991) "Science and Aesthetic Appreciation," *Midwest Studies in Philosophy*, XVI, pp. 180–95.

Koriat, A. (1997) "Monitoring One's Own Knowledge During Study: A Cue-Utilization Approach to Judgements of Learning," *Journal of Experimental Psychology: General*, 126, 4, 349–370.

Koriat, A. (2000) "The Feeling of Knowing: Some Metatheoretical Implications for Consciousness and Control," *Consciousness and Cognition* 9, 149–71.

Koriat, A. (2008) "Easy Comes, Easy Goes? The Link Between Learning and Remembering and its Exploitation in Metacognition," *Memory & Cognition*, 36, pp. 416–28.

Koriat, A. (2012a) "Subjective Confidence in Perceptual Judgements: A Test of the Self-Consistency Model," *Journal of Experimental Psychology: General*, 140, 1, pp. 117–39.

Koriat, A. (2012b) "The Subjective Confidence in One's Knowledge and Judgements: Some Meta-Theoretical Considerations," pp. 213–33, in M. J. Beran, J. Brandl, J. Perner, and J. Proust, (eds.) *Foundations of Metacognition*. Oxford: Oxford University Press.

Koriat, A. (2024) "Subjective Confidence as a Monitor of the Replicability of the Response," *Perspectives on Psychological Science*, pp. 1–18.

https://doi.org/10.1177/17456916231224387

Koriat, A. and Ma'ayan, H. (2005) "The Effects of Encoding Fluency and Retrieval Fluency on Judgments of Learning," *Journal of Memory and Language*, 52, 4, pp. 478–92.

Koriat, A., Ackerman, R., Lockl, K., and Schneider, W. (2009) "The Memorizing Effort Heuristic in Judgments of Learning: A Developmental Perspective," *Journal of Experimental Child Psychology*, 102, pp. 265–79.

Kosso, P. (1992) *Reading the Book of Nature: An Introduction to the Philosophy of Science*. Cambridge: Cambridge University Press.

Kosso, P. (2002) "The Omniscienter: Beauty and Scientific Understanding," *International Studies in the Philosophy of Science*, 16, 1, pp. 39–48.

Kosso, P. (2007) "Scientific Understanding," *Foundations of Science*, 12, pp. 173–88.

Kozlov, A. (2023) "Scientific Experiments Beyond Surprise and Beauty," *European Journal for Philosophy of Science*, 13, 38, pp. 1–22.

Kragh, H. (2011) *Higher Speculations: Grand Theories and Failed Revolutions in Physics and Cosmology*. Oxford: Oxford University Press.

Kuhn, T. S. (1977) "Objectivity, Value Judgment, and Theory Choice," pp. 320–39, in T. S. Kuhn, *The Essential Tension: Selected Studies in Scientific Tradition and Change*. Chicago: The University of Chicago Press.

Kutrovátz, G. (2024) "The Hearth of the World: The Sun before Astrophysics," *Universe*, 10, 256. https://doi.org/10.3390/universe10060256

Kvanvig, J. L. (2018) "Knowledge, Understanding, and Reasons for Belief," pp. 685–704, in D. Star (ed.) *The Oxford Handbook of Reasons and Normativity*. Oxford: Oxford University Press.

Labroo, A. A., Dhar, R., & Schwarz, N. (2008) "Of Frowning Watches and Frog Wines: Semantic Priming, Perceptual Fluency, and Brand Evaluation," *Journal of Consumer Research*, 34, pp. 819–31.

Laudan, L. (1984) "Explaining the Success of Science," pp. 83–105, in J. Cushing et al. (eds.) *Science and Reality*. Notre Dame: Notre Dame University Press.

Laudan, L. (2004) "The Epistemic, the Cognitive, and the Social," pp. 14–23, in P. K. Machamer and G. Wolters (eds.) *Science, Values, and Objectivity*. University of Pittsburgh Press.

Laudan, L. And Leplin, J. (1991) "Empirical Equivalence and Underdetermination," *Journal of Philosophy*, 88, pp. 449–72.

Laudan, R. (1978) "The Recent Revolution in Geology and Kuhn's Theory of Scientific Change," *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, Vol. 2: Symposia and Invited Papers*, pp. 227–39.

Le Grand, H. (1988) *Drifting Continents and Shifting Theories: The Modern Revolution in Geology and Scientific Change*. Cambridge: Cambridge University Press.

Leonelli, L. (2013) "Integrating Data to Acquire New Knowledge: Three Modes of Integration in Plant Science," *Studies in History and Philosophy of Biological and Biomedical Sciences*, 44, pp. 503–14.

Levenson, T. (2015) *The Hunt for Vulcan: And How Albert Einstein Destroyed a Planet, Discovered Relativity, and Deciphered the Universe*. New York: Random House.

Logan, G. D. and Crump, M. J. C. (2010) "Cognitive Illusions of Authorship Reveal Hierarchical Error Detection in Skilled Typists," *Science*, 330, 29, pp. 683–6.

Lombrozo, T. (2016) "Explanatory Preferences Shape Learning and Inference," *Trends in Cognitive Sciences*, 20, 10, pp. 748–59. http://dx.doi.org/10.1016/j.tics.2016.08.001

Longino, H. E. (1990) *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press.

Longino, H. E. (1996) "Cognitive and Non-Cognitive Values in Science: Rethinking the Dichotomy," pp. 39–58, in L. H. Nelson and J. Nelson (eds), *Feminism, Science, and the Philosophy of Science*. Kluwer Academic Publishers.

Lopes, D. M. (2022) "Getting Into It: Ventures in Aesthetic Life", pp. 61–86, in D. M. Lopes, B. Nanay, and N. Riggle, *Aesthetic Life and Why it Matters*. Oxford: Oxford University Press.

Lopes, D. M., Nanay, B., and Riggle, N. (2022) *Aesthetic Life and Why it Matters*. Oxford: Oxford University Press.

Lyons, T. D. (2025) *Scientific Realism*. Cambridge: Cambridge University Press.

Mackonis, A. (2013) "Inference to The Best Explanation, Coherence and Other Explanatory Virtues," *Synthese*, 190, pp. 975–9.

Mamassian, P. (2022) "Uncertain Perceptual Confidence," *Nature Human Behaviour*, 6, pp. 179–180.

Mason, R. (2003) "Stripes on the Sea Floor," pp. 31–45, in N. Oreskes (ed.) *Plate Tectonics: An Insider's History of the Modern Theory of the Earth*. Colorado: Westview Press.

Mason, R. and Raff, A. D. (1961) "Magnetic Survey Off the West Coast of North America, 32° Latitude To 42° Latitude," *Bulletin of the Geological Society of America*, 72, pp. 1259–66.

McAllister, J. (1996) *Beauty and Revolution in Science*. New York: Cornell University Press.

McGlone, M. and Tofighbakhsh, J. (2000) "Birds of a Feather Flock Conjointly (?): Rhyme as Reason in Aphorisms," *Psychological Science*, 11, 5, pp. 424–28.

McKenzie, D. P., and Parker, R. L. (1967) "The North Pacific: An Example of Tectonics on a Sphere," *Nature*, 216, 5122, pp. 1276–80.

McMahon, J. (2005) "Beauty," in B. Gaut, D. M. Lopes (eds.) *Routledge Companion to Aesthetics*. 2nd ed. London: Routledge.

McMullin, E. (1982) "Values in Science," *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, Vol. 2: Symposia and Invited Papers*, pp. 3–28

McMullin, E. (2014) "The Virtues of a Good Theory," pp. 561–71, in M. Curd and S. Psillos (eds.), *The Routledge Companion to Philosophy of Science*. New York: Routledge.

Menard, H. W. (1986) *The Ocean of Truth: A Personal History of Global Tectonics*. Princeton: Princeton University Press.

Messinger, S. M. (1998) "Pleasure and Complexity: Berlyne Revisited," *The Journal of Psychology*, 132, 5, pp. 558–60.

Metcalfe, J., Schwartz, B., and Bloom, P. (2017) "The tip-of-the-Tongue State and Curiosity," *Cognitive Research: Principles and Implications*, 2, 31. https://doi.org/10.1186/s41235-017-0065-4

Meyers, E. A., Gretton, J. D., Budge, J. R. C., Fugelsang, J. A., and Koehler, D. J. (2023) "Broad Effects of Shallow Understanding: Explaining an Unrelated Phenomenon Exposes the Illusion of Explanatory Depth,"

Judgment and Decision Making, 18:e24, pp. 1–18. doi:10.1017/jdm.2023.24

Meynell, L. (2020) "Getting the Picture: Towards a New Account of Scientific Understanding," pp. 36–62, in M. Ivanova and A. Murphy (eds.) *The Aesthetics of Scientific Experiments*. New York: Routledge.

Minnameier, G. (2004)) "Peirce-Suit of Truth: Why Inference to the Best Explanation and Abduction Ought Not to Be Confused," *Erkenntnis*, 60, 1, pp. 75–105.

Miyata-Sturm, M. E. (2024a) "Aesthetic Feelings in Scientific Reasoning," *Teorema: International Journal of Philosophy*, XLIII/1, pp. 5–27.

Miyata-Sturm, M. E. (2024b) "Aesthetic Considerations in the Development of Plate Tectonics," *Studies in History and Philosophy of Science*, 108, pp. 1–9.

Molnar, P. (2015) *Plate Tectonics: A Very Short Introduction*. Oxford: Oxford University Press.

Montaño, U. (2012) "Ugly Mathematics: Why Do Mathematicians Dislike Computer-Assisted Proofs?" *The Mathematical Intelligencer*, 34, 4, pp. 21–8.

Montaño, U. (2013) "Beauty in Science: A New Model of the Role of Aesthetic Evaluations in Science," *European Journal for Philosophy of Science*, 3, pp. 133–5. DOI 10.1007/s13194-013-0064-3.

Morgan, W. J. (1968) "Rises, Trenches, Great Faults and Crustal Blocks," *Journal of Geophysical Research*, 73, pp. 1959–82.

Morley, L. W. (2003) "The Zebra Pattern," pp. 67–85, in N. Oreskes (ed.) *Plate Tectonics: An Insider's History of the Modern Theory of the Earth*. Colorado: Westview Press.

Murphy, A. (2023) "Profound Experiments," pp. 135–54, in M. Ivanova and A. Murphy (eds.) *The Aesthetics of Scientific Experiments*. New York: Routledge.

Musgrave, A. (1985) "Realism vs Constructive Empiricism," pp. 196–208, in P. M. Churchland and C. A. Hooker (eds.) *Images of Science*. Chicago: University of Chicago Press.

Nahari, G. and Pazuelo, M. (2015) "Telling a Convincing Story: Richness in Detail as a Function of Gender and Information," *Journal of Applied Research in Memory and Cognition*, 4, pp. 363–67.

Nagata, T. (1952) "Reverse Thermo-Remanent Magnetism," *Nature*, 169,

pp. 704–5.

Nagel, J. (2016) "Knowledge and Reliability," pp. 237–56, in H. Kornblith and B. McLaughlin (eds.) *Goldman and his Critics*. Malden, MA: Blackwell.

Nagel, J. (Forthcoming) "Natural Curiosity," in A. Logins and J. Vollet (eds.) *Putting Knowledge to Work: New Directions for Knowledge-First Epistemology*. Oxford: Oxford University Press.

Nanay, B. (2014) "Philosophy of Perception as a Guide to Aesthetics," pp. 101–20, in G. Currie, M. Kieran, A. Meskin, and J. Robson (eds.) *Aesthetics and the Sciences of Mind*. Oxford: Oxford University Press.

Nanay, B. (2016) *Aesthetics as Philosophy of Perception*. Oxford: Oxford University Press.

Nguyen, C. T. (2020) *Games: Agency as Art*. Oxford: Oxford University Press.

Nozick, R. (1988) Knowledge and Scepticism," pp 21–42, in J. Dancy (ed.) *Perceptual Knowledge*. Oxford: Oxford University Press.

Oddie, G. and Cevolani, G. (2022) "Truthlikeness," in E. N. Zalta and U. Nodelman (eds.), *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/win2022/entries/truthlikeness/

Okasha, S. (2011) "Theory Choice and Social Choice: Kuhn versus Arrow," *Mind*, 120, 477, pp. 83–115.

Okon-Singer, H., Stout, D. M., Stockbridge, M. D., Gamer, M., Fox, A. S., and Shackman, A. J. (2018) "The Interplay of Emotion and Cognition," pp. 181–5, in A. S. Fox, R. C. Lapate, A. J. Shackman, and R. D. Davidson (eds.) *The Nature of Emotion: Fundamental Questions*. 2nd ed. Oxford: Oxford University Press.

Oppenheimer, D. M. and Frank, M. C. (2007) "A Rose in Any Other Font Would Not Smell as Sweet: Effects of Perceptual Fluency on Categorization," *Cognition,* 106, 3, pp. 1178–94.

Opdyke, N. D. (2003) "The Birth of Plate Tectonics," pp. 95–107, in N. Oreskes (ed.) *Plate Tectonics: An Insider's History of the Modern Theory of the Earth*. Colorado: Westview Press.

Oreskes, N. (1999) *The Rejection of Continental Drift: Theory and Method in American Earth Science*. New York and Oxford: Oxford University Press.

Oreskes, N. (ed.) (2003) *Plate Tectonics: An Insider's History of the Modern Theory of*

the Earth. Colorado: Westview Press.

Osborne, H. (1984) "Mathematical Beauty and Physical Science," *British Journal of Aesthetics*, 24, 4, pp. 291–300.

Pellegrini, P. A. (2019) "Styles of Thought on the Continental Drift Debate," *Journal for General Philosophy of Science*, 50, pp. 85–102.

Pellegrini, P. A. (2022) "About the Reaction to Styles of Thought on the Continental Drift Debate," *Journal for General Philosophy of Science*, 53, pp. 573–82.

Pessoa, L. (2018) "The Cognitive-Emotional Brain," pp. 202–6, in A. S. Fox, R. C. Lapate, A. J. Shackman, and R. D. Davidson (eds.) *The Nature of Emotion: Fundamental Questions*. 2nd ed. Oxford: Oxford University Press.

Pitman, W. (2003) "On Board the *Eltanin-19*," pp. 86–94, in N. Oreskes (ed.) *Plate Tectonics: An Insider's History of the Modern Theory of the Earth*. Colorado: Westview Press.

Plakias, A. (2021) "The Aesthetics of Food," *Philosophy Compass*, 16:e12781. https://doi.org/10.1111/phc3.12781

Popper, K. ([1960] 1987) "The Growth of Scientific Knowledge," pp.171–80, in. D. Miller (ed.) *A Pocket Popper*. Glasgow: Fontana Press.

Potochnik, A. (2017) *Idealization and the Aims of Science*. Chicago and London: The University of Chicago Press.

Proust, J. (2013) *The Philosophy of Metacognition: Mental Agency and Self-Awareness*. Oxford: Oxford University Press.

Psillos, S. (1999) *Scientific Realism: How Science Tracks Truth*. London and New York: Routledge.

Putnam, H. (1979) "What Is Mathematical Truth?" pp. 60–78, in *Philosophical Papers: Volume 1, Mathematics, Matter and Method*. New York: Cambridge University Press.

Ramachandran, V. S. and Seckel, E. (2012) "Neurology of Visual Aesthetics: Indian Nymphs, Modern Art, and Sexy Beaks," pp. 375–89, in A. Shimamura and S. E. Palmer (eds.) *Aesthetic Science: Connecting Minds, Brains, and Experience*. Oxford: Oxford University Press.

Reber, R. (2018) "Beauty and Truth in Mathematics: Evidence from Cognitive Psychology," pp. 252–67, in S. Bangu (ed.) *Naturalizing Logico-*

Mathematical Knowledge: Approaches from Philosophy, Psychology and Cognitive Science. New York and Abingdon: Routledge.

Reber, R., Winkielman, P., and Schwarz, N. (1998) "Effects of Perceptual Fluency on Affective Judgments," *Psychological Science*, 9, 1, pp. 45–8.

Reber, R. and Schwarz, N. (1999) "Effects of Perceptual Fluency on Judgments of Truth," *Consciousness and Cognition*, 8, 338–42.

Reber, R., Schwarz, N.., and Winkielman, P. (2004) "Processing Fluency and Aesthetic Pleasure: is Beauty in the Perceiver's Processing Experience?" *Personality and Social Psychology Review*, 8, pp. 364–82.

Reber, R., Brun, M., Mitterndorfer, K., (2008) "The Use of Heuristics in Intuitive Mathematical Judgment," *Psychonomic Bulletin & Review*, 15, pp. 1174–8.

Reber, R., Unkelbach, C., (2010) "The Epistemic Status of Processing Fluency as Source for Judgments of Truth," *Review of Philosophy and Psychology*, 1, pp. 563–81.

Reder, L. M. (1987) "Strategy Selection in Question Answering," *Cognitive Psychology*, 19, pp. 90–138.

Reder, L. M. (1996) *Implicit Memory and Metacognition*. Mahwah, N.J.: Erlbaum.

Reder, L. M., and Ritter, F. E. (1992) "What Determines Initial Feeling of Knowing? Familiarity With Question Terms, Not With the Answer," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 3, pp. 435–45.

Ritson, S. (2023) "Something from Nothing: 'Non-Discovery' and Transformations in High-Energy Experimental Physics at the Large Hadron Collider," pp. 201–14, in M. Ivanova and A. Murphy (eds.) *The Aesthetics of Scientific Experiments*. New York: Routledge.

Rozenblit, L. and Keil, F. (2002) "The Misunderstood Limits of Folk Science: an Illusion of Explanatory Depth," *Cognitive Science*, 26, pp. 521–62.

Saito, Y. (2007) *Everyday Aesthetics*. Oxford: Oxford University Press.

Shimamura, A. and Palmer, S. E. (eds.) (2012) *Aesthetic Science: Connecting Minds, Brains, and Experience*. Oxford: Oxford University Press.

Schindler, S. (2007) "Rehabilitating Theory: Refusal of the 'Bottom-up' Construction of Scientific Phenomena," *Studies in History and Philosophy of*

Science, 38, 1, pp. 160–84.

Schindler, S. (2014) "A Matter of Kuhnian Theory Choice? The GWS Model and the Neutral Current," *Perspectives on Science*, 22, 4, pp. 491–522.

Schindler, S. (2017a) "Theoretical Fertility McMullin-style," *European Journal for the Philosophy of Science*, 7, 1, pp. 151–73.

Schindler, S. (2017b) "Kuhnian Theory-Choice and Virtue Convergence: Facing the Base Rate Fallacy," *Studies in History and Philosophy of Science*, 64, pp. 30–7.

Schindler, S. (2018) *Theoretical Virtues in Science: Uncovering Reality Through Theory*. Cambridge: Cambridge University Press.

Schindler, S. (2020) "Theoretical Virtues in Science," in *Oxford Bibliographies*. Oxford University Press. https://doi.org/10.1093/OBO/9780195396577-0409

Schindler, S. (2022) "Theoretical Virtues: Do Scientists Think What Philosophers Think They Ought to Think?" *Philosophy of Science*, 2022, 89, 3, 542–64.

Schindler, S. (2025) "Explanatory Stabilization in Situations of Evidential Uncertainty," *Philosophy of Science*, 92, 1, pp. 204–22.

Schwartz, B. (1994) "Sources of Information in Meta-Memory: Judgements of Learning and Feelings of Knowing," *Psychonomic Bulletin & Review*, 1, 3, pp. 357–75.

Schwartz, B. and Metcalfe, J. (2011) "Tip-of-the-tongue (TOT) States: Retrieval, Behavior, and Experience," *Memory & cognition*, 39, 5, 737–49. https://doi.org/10.3758/s13421-010-0066-8

Schwartz, B. and Díaz, F. (2014) "Quantifying Human Metacognition for The Neurosciences," pp. 9–23, in S. M. Fleming and C. D. Frith (eds.) *The Cognitive Neuroscience of Metacognition*. Berlin and Heidelberg: Springer.

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., and Simons, A. (1991) "Ease of Retrieval as Information: Another Look at the Availability Heuristic," *Journal of Personality and Social Psychology*, 61, 2, 195–202.

Schwarz, N. (1998) "Accessible Content and Accessibility Experiences: The Interplay of Declarative and Experiential Information in Judgment,"

Personality and Social Psychology Review, 2, 2, pp. 87–99.

Schwarz, N. (2018) "Of Fluency, Beauty, and Truth: Inferences from Metacognitive Experiences," pp. 25–46, in J. Proust and M. Fortier (eds.) *Metacognitive Diversity: An Interdisciplinary Approach*. Oxford: Oxford University Press.

Šešelja, D. and Weber, E. (2012) "Rationality and Irrationality in the History of Continental Drift: was the Hypothesis of Continental Drift Worthy of Pursuit?" *Studies in History and Philosophy of Science*, 43, pp. 147–59.

Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., and Frith, C. D. (2014) "Supra-Personal Cognitive Control and Metacognition," *Trends in Cognitive Sciences*, 18, 4, pp. 186–93.

Shea, N. (2024) "Metacognition of Inferential Transitions," *The Journal of Philosophy*, 121, 11, pp. 597-627.

Silva, R. R., Garcia-Marques, T., and Reber, R. (2017) "The Informative Value of Type of Repetition: Perceptual and Conceptual Fluency Influences on Judgments of Truth," *Consciousness and Cognition*, 51, pp. 53–67.

Silvia, P. J. (2012) "Human Emotions and Aesthetic Experience: An Overview of Empirical Aesthetics," pp. 250–75, in A. Shimamura and S. E. Palmer (eds.) *Aesthetic Science: Connecting Minds, Brains, and Experience*. Oxford: Oxford University Press.

Sliwa, P. (2015) "Understanding and Knowing," *Proceedings of the Aristotelian Society, CXV, part 1*, pp. 57–74.

Sloman, S., and Fernbach, P. (2017) *The Knowledge Illusion*. Pan Macmillian.

Smith, B. (2012) "Complexities of Flavour," *Nature*, 486, p. 6. https://doi.org/10.1038/486S6a

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

Stolnitz, J. (1960) *Aesthetics and Philosophy of Art Criticism: A Critical Introduction*. Houghton Mifflin.

Stoltz, J. (2007) "Gettier and Factivity in Indo-Tibetan Epistemology," *The Philosophical Quarterly*, 57, 228, pp. 394–415. https://doi.org/10.1111/j.1467-9213.2007.493.x

Stuart, M. T. (2023) "The Future Won't Be Pretty: The Nature and Value of

Ugly, AI-Designed Experiments," pp. 215–33, in M. Ivanova and A. Murphy (eds.) *The Aesthetics of Scientific Experiments*. New York: Routledge.

Suess, E. (1904-24) *The Face of the Earth*. Vol 1-5. Transl. H. B. C. Sollas. Oxford: Clarendon Press.

Swinburne, R. (1999) *Simplicity as Evidence of Truth: The Aquinas Lecture*. Milwaukee: Marquette University Press.

Thagard, P. (2006) *Hot Thought: Mechanisms and Applications of Emotional Cognition*. Cambridge and London: The MIT Press.

Thagard, P. (1978) "The Best Explanation: Criteria for Theory Choice," *The Journal of Philosophy*, 75, 2, pp. 76–92.

Thompson, V. (2009) "Dual-process Theories: A Metacognitive Perspective," pp. 17–96, in J. Evans and K. Frankish (eds.) *In Two Minds: Dual Processes and Beyond*. Oxford: Oxford University Press.

Thompson, V. A., Turner, J. A. P., and Pennycook, G. (2011) "Intuition, Reason, and Metacognition," *Cognitive Psychology*, 63, pp. 107–40.

Thompson, V., Turner, J. A. P., Pennycook, G., Ball, L. J., Brack, H., Opir, Y. and Ackerman, R. (2013) "The Role of Answer Fluency and Perceptual Fluency as Metacognitive Cues for Initiating Analytic Thinking," *Cognition*, 128, pp. 237–51.

Thompson, V. A. and Johnson, S. C. (2014) "Conflict, Metacognition, and Analytic Thinking," *Thinking & Reasoning*, 20, 2, pp. 215-44.

Thompson, V. A. and Markovits, H. (2025) "Fast Reasoning and Metacognition," *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-025-02662-0

Topolinski, S. and Strack, F. (2009) "The Analysis of Intuition: Processing Fluency and Affect in Judgements of Semantic Coherence," *Cognition and Emotion*, 23, 8, pp. 1465–503.

Todd, C. (2008) "Unmasking the Truth Beneath the Beauty: Why the Supposed Aesthetic Judgements Made in Science May Not Be Aesthetic at All," *International Studies in the Philosophy of Science*, 22, 1, pp. 61–79.

Todd, C. (2017) "Fitting Feelings and Elegant Proofs: On the Psychology of Aesthetic Evaluation in Mathematics," *Philosophia Mathematica*, 26, 2, pp. 211–33.

Tulodziecki, D. (2025) *Underdetermination and Theoretical Virtues*. Cambridge: Cambridge University Press.

Turner, D. D. (2019) *Paleoaesthetics and the Practice of Palaeontology*. Cambridge: Cambridge University Press.

Turner, D. D. and AboHamad, A. (2023) "Narrative Explanation and Non-Epistemic Value," *Journal of the Philosophy of History*, 17, pp. 53–76.

Tversky, A. and Kahneman, D. (1974) "Judgment under Uncertainty: Heuristics and Biases," *Science*, 185, 4157, pp. 1124–31.

Tversky, A. and Kahneman, D. (1983) "Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment," *Psychological Review*, 90, 4, pp. 293-315.

Undorf, M., Söllner, A., and Bröder, A. (2018) "Simultaneous Utilization of Multiple Cues in Judgments of Learning," *Memory and Cognition,* 46, pp. 507–19.

Undorf, M., & Bröder, A. (2020) "Cue Integration in Metamemory Judgements is Strategic," *The Quarterly Journal of Experimental Psychology,* 73, 4, pp. 629–42. https://doi.org/10.1177/1747021819882308

Undorf, M., Navarro-Baez, S., and Zimdahl, M. F. (2022) "Metacognitive Illusions," pp. 307–23, in R. Pohl (ed.) *Cognitive Illusions: Intriguing Phenomena in Thinking, Judgement, and Memory*. 3rd ed. London and New York: Routledge.

Vaidyanathan, B., Haraburda, B., and Jacobi, C. J. (2023) "Beauty in Biology: An Empirical Assessment," *Journal of Biosciences*, 48, 15, pp. 1–9.

Van Fraassen, B. C. (1980). *The Scientific Image*. New York: Oxford University Press.

van der Gracht, W. A. J. M. v. W. (ed.) (1928) *Theory of Continental Drift: A Symposium on the Origin and Movement of Landmasses Both Inter-Continental and Intra-Continental, as Proposed by Alfred Wegener*. Tulsa, Oklahoma: The American Association of Petroleum Geologists.

van Geert, E. and Wagemans, J. (2020) "Order, Complexity, and Aesthetic Appreciation," *Psychology of Aesthetics, Creativity, and the Arts*, 14, 2, pp. 135–54.

Vickers, P. (2013) *Understanding Inconsistent Science*. Oxford: Oxford University

Press.

Vine, F. J. (2003) "Reversals of Fortune," pp. 46–66, in N. Oreskes (ed.) *Plate Tectonics: An Insider's History of the Modern Theory of the Earth*. Colorado: Westview Press.

Vine, F. J. and Matthews, D. H. (1963) "Magnetic Anomalies Over Ocean Ridges," *Nature*, 199, pp. 947–9.

Vrij, A., Mann, S. A., Fisher, R. P., Leal, S., Milne, R., and Bull, R. (2008) "Increasing Cognitive Load to Facilitate Lie Detection: The Benefit of Recalling an Event in Reverse Order," *Law and Human Behavior*, 32, 3, pp. 253–65. https://doi.org/10.1007/s10979-007-9103-y

Warren, R. P. (1980) "Youthful Truth-Seeker, Half-Naked, at Night, Running Down Beach South of San Francisco," pp. 25–6, in *Being There: Poetry 1977–1980*. Random House, Inc.

Weinberg, S. (1993) *Dreams of a Final Theory: The Search for the Fundamental Laws of Nature*. London: Vintage.

Weber, E. and Šešelja, D. (2020) "In Defence of Rationalist Accounts of the Continental Drift Debate: A Response to Pellegrini," *Journal for General Philosophy of Science*, 51, pp. 481–490.

Wechsler, J. (ed.) (1978) *On Aesthetics in Science*. Cambridge, MA. and London: the MIT press.

Wegener, A. (1966) *The Origin of Continents and Oceans*. Translated from the 4th revised German ed. by J. Biram. New York: Dover Publications, inc.

Williams, B. (1985) *Ethics and the Limits of Philosophy*. Fontana Press.

Williams, E. F., Duke, K. E. and Dunning, D. (2020) "Consistency Just Feels Right," *Journal of Experimental Psychology: General*, 149, 12, pp. 2395-405.

Williamson, T. (2000) *Knowledge and Its Limits*. Oxford: Oxford University Press.

Williamson, T. (2024) *Overfitting and Heuristics in Philosophy*. New York: Oxford University Press.

Wilson, E. O. (1978) *On Human Nature*. Cambridge, MA and London: Harvard University Press.

Wilson, T. (1965) "A New Class of Faults and Their Bearing on Continental Drift," *Nature*, 207, pp. 343–7.

Wolpert, L. And Richards, A. (1997) *Passionate Minds: The Inner World of Scientists*.

Oxford, New York, and Tokyo: Oxford University Press.

Worrall, J. (1989a) "Fresnel, Poisson and the White Spot: The Role of Successful Predictions in the Acceptance of Scientific Theories," pp. 135–57, in D. Gooding, T. Pinch, and S. Schaffer (eds.) *The Uses of Experiment: Studies of Experimentation in Natural Science*. Cambridge: Cambridge University Press.

Worrall, J. (1989b) "Structural Realism: The Best of Both Worlds?" *Dialectica*, 43, 1–2, pp. 99–124.

Wragge-Morley, A. (2020) *Aesthetic Science: Representing Nature in the Royal Society of London, 1650–1720*. Chicago and London: The University of Chicago Press.

Wragge-Morley, A. (2023) "Robert Boyle (1627–1691) and the Aesthetics of Chemical Experimentation," pp. 59–79, in M. Ivanova and A. Murphy (eds.) *The Aesthetics of Scientific Experiments*. New York: Routledge.

Wylie, C. D. (2021) *Preparing Dinosaurs: The Work Behind the Scenes*. Cambridge and London: MIT Press.

Yaniv, I. and Foster, D. (1995) "Graininess of Judgement Under Uncertainty: An Accuracy-Informativeness Trade-Off," *Journal of Experimental Psychology: General*, 124, 4, pp. 424–32.

Yaniv, I. and Foster, D. (1997) "Precision and Accuracy of Judgemental Estimation," *Journal of Behavioural Decision Making*, 10, pp. 21–32.

Yeung, N. and Summerfield, C. (2012) "Metacognition in Human Decision-Making: Confidence and Error Monitoring," *Philosophical Transactions of the Royal Society: B*, 367, pp. 1310–21.

Zajonc, R. B. (1968). "Attitudinal Effects of Mere Exposure," *Journal of Personality and Social Psychology: Monograph Supplement*, 9, pp. 1–27.

Zangwill, N. (1995) "The Beautiful, the Dainty and the Dumpy," *British Journal of Aesthetics*, 35, 4, pp. 317–29.

Zangwill, N. (1999) "Feasible Aesthetic Formalism," *Noûs*, 33, 4, pp. 610–2.