Introduction

Scientists are constantly engaged in various forms of reasoning, arguing that because *this* is the case, *that* must be the case. Some of these forms of reasoning are from what may broadly be called *data* to what may broadly be called *theory*. The data are things like observations, survey statistics, and experimental results. A theory is typically a more ambitious type of claim that often generalizes, expands, or otherwise "goes beyond" the data, such as by specifying what causes some type of event. For example, by the early twentieth century there was already a great deal of observational data suggesting that lung cancer is more frequent among tobacco smokers than among non-smokers. From this data most scientists eventually inferred that smoking *causes* lung cancer, and so that one may reduce one's chances of getting lung cancer by refraining from smoking.

The term "abductive reasoning" refers, at least for the purposes of this *Element*, to a specific way of engaging in data-to-theory reasoning. In particular, it refers to reasoning in which theories are evaluated at least partly on the basis of how well they would, if true, *explain* the available data. To see how this is supposed to work, consider how one might conclude that smoking causes lung cancer in the above example. The theory that smoking causes lung cancer seems to provide a good explanation, especially compared to rival explanations, of the observed difference in lung cancer frequency among smokers and nonsmokers. In particular, the theory that smoking causes lung cancer arguably provides a much better explanation of this data than various other theories one might think of, such as that the correlation between smoking and lung cancer is a mere coincidence, or that having lung cancer somehow causes smoking.[1] On these grounds, it seems reasonable to conclude that smoking causes lung cancer.

Abductive reasoning is arguably not only commonplace in the sciences, but also widespread in other situations in which we make inferences about the underlying explanations, such as the causes or grounds, of the things in our immediate environment. Some philosophers even claim that all cogent data-to-theory reasoning is abductive reasoning – that is, that reasoning from data to theory should always involve evaluating how well various theories would explain the data (e.g., Lycan, 1988). According to this view, even the most basic generalizations and predictions from past experience – such as inferring that the sun will rise tomorrow morning because it has risen every morning thus far – involve abductive reasoning as well, albeit in an implicit and indirect

[1] This latter type of explanation was seriously proposed by R.A. Fisher (1959), who suggested that having lung cancer might cause an unconscious irritation or pain, which in turn causes people to smoke.

way. Furthermore, several philosophers have argued that abductive reasoning is essential to philosophy itself – that philosophical theories should be evaluated on the basis of how well they explain some philosophical "data," such as our pre-theoretic judgments about hypothetical cases (e.g., Williamson, 2016).

Given the apparent significance of abductive reasoning, it should come as no surprise that philosophers of science have studied the nature of abductive reasoning intensely. In the past few decades, a number of subtly different accounts of abductive reasoning have emerged – many, though not all, of which have been inspired by Gilbert Harman's (1965) slogan "Inference to the Best Explanation." Roughly, Harman's idea was that one may infer a theory from some collection of data just in case the theory provides a better explanation of the data than any competing theory that one has considered, where inference involves coming to accept or believe that the theory is true. However, the popularity of Harman's slogan obscures how much disagreement there is about exactly how to understand it. A number of very serious, if not devastating, objections have prompted various philosophers to reconsider key elements of the slogan. Indeed, there are now prominent accounts on which Inference to the Best Explanation is not viewed as a form of *inference*, some accounts on which it does not involve inferring to the *best* explanation, and yet others on which one need not infer to an *explanation* at all.[2]

This Element has two main aims. The first is to give a systematic and opinionated overview of the current state of philosophical thinking about abductive reasoning. This involves not just discussing the various accounts of abductive reasoning that have been proposed, but also the many objections to previous accounts which have motivated philosophers to develop them. As this indicates, I will approach the topic in a problem-based manner, in the sense that the various accounts of abductive reasoning will be presented as responses to specific problems. However, some problems and accounts will not be discussed in detail, or indeed at all. This is partly for reasons of space and partly to keep the discussion accessible, since some important contributions to the topic are rather technical and require familiarity with various formal methods that would need to be introduced in an Element of their own.[3]

The second aim of this Element is to gradually construct, by drawing lessons from the various problems and accounts to be discussed, a systematic view of

[2] This curious situation evokes Voltaire's (1759, ch. 70) quip that the Holy Roman Empire was "in no way holy, nor Roman, nor an empire."

[3] Happily, there is another Element, *Bayesianism and Scientific Reasoning* (Schupbach, 2022), that covers much of the ground I have in mind here, especially recent discussions of formal measures of explanatory power and how they could be leveraged in an account of abductive reasoning.

the nature and purpose of abductive reasoning. This view is difficult to summarize briefly at this stage, but at a very general level it holds that abductive reasoning is a collection of inferential strategies that serves to approximate different forms of probabilistic reasoning. Depending on the exact nature of the probabilistic reasoning that is being approximated, the inferential strategy may be more or less demanding. In particular, I will suggest that some of the probabilistic conclusions we wish to reach are quite modest, for example, when determining which theory to investigate further; in those cases, abductive reasoning is not very demanding. In other cases, we may want abductive reasoning to warrant a reasonably high level of probabilistic confidence that a theory is true; in those cases, abductive reasoning is an evidentially demanding and temporally extended process that may not deliver the desired conclusion at all.

The rest of this Element is structured as follows. Section 1 briefly summarizes the history of philosophical thought about abductive reasoning from the advent of modern science to the middle of the twentieth century. Section 2 surveys contemporary accounts of abductive reasoning, based on a three-fold distinction between accounts that construe abductive reasoning as (i) a form of *inference*, (ii) a *probabilistic* process, or (iii) *both* of the above. Section 3 focuses on the fact that in abductive reasoning, one is told to infer or prefer the *best* explanation. But what reason, if any, is there for scientists to prefer "better" explanations in this way? As we shall see, there are several quite different types of answers to this question, leading to different ideas about the role of abductive reasoning in science. Section 4 then discusses a different set of problems for accounts of abductive reasoning, having to do with whether abductive reasoning is somehow irrational or incoherent in some cases. In particular, it has been suggested that some common accounts of abductive reasoning imply that one should sometimes infer theories that are, by one's own lights, very likely to be false; or that one should assign probabilities to theories in ways that are, by one's own lights, demonstrably irrational. Finally, Section 5 weaves together various threads from the previous sections to briefly present a holistic view of abductive reasoning that, I hope, avoids the various problems for abductive reasoning discussed in this Element while retaining the core insight that much of scientific reasoning is governed by explanatory considerations.

1 A Brief History of Abductive Reasoning

This section introduces the topic of this Element by way of a brief historical overview of philosophical thinking about abductive reasoning. In particular, we will look at examples of scientists and philosophers who deployed or implicitly endorsed forms of abductive reasoning, such as Charles Darwin and René

Descartes (§1.1); discuss Charles S. Peirce's pioneering work on the form of reasoning he dubbed "Abduction" (§1.2); consider the extent to which the "hypothetico-deductive model" is a forerunner to abductive reasoning (§1.3); and, finally, examine Gilbert Harman's seminal notion of "Inference to the Best Explanation" (§1.4). This overview sets the stage for the next section, in which more recent (and arguably more sophisticated) accounts of abductive reasoning are surveyed.

1.1 The Historical Roots of Abductive Reasoning

As is so often the case with methodological novelties, abductive reasoning seems to have emerged first as an implicit scientific practice rather than an explicit philosophical theory. This is perhaps clearest in the writings of Francis Bacon (1561–1626), often regarded as the father of "the scientific method." Rebelling against the Aristotelian idea that natural philosophy (i.e., science) can discover the essences of things, Bacon explicitly advocated an austere form of "inductivism" in his influential *Novum Organum* (Bacon, 1620). According to Bacon, scientists should proceed by first collecting data, for example, by observing that this or that pot of water boils at 100°C. Having collected such data, they should then generalize from observed correlations in that data, for example, by concluding that water *always* boils at 100°C. In short, Bacon's official view identified scientific reasoning with extrapolation from data.

In practice, however, Bacon seems to have allowed for a different type of reasoning to play an important role in science (McMullin, 1992, 175–179). Bacon was an early advocate of what was later dubbed the kinetic theory of heat, which holds that heat can be identified with the motion of unobservably small parts of the heated body (i.e., what we would now call *molecules*). But how could Bacon establish that these unobservably small parts move around within the heated body in the first place, or indeed that they exist at all? Baconian generalization from a correlation among observations cannot do the trick, since there was never a correlation to generalize; there is no correlation between observations of hot bodies and observations of bodies consisting of small parts in motion, simply because those parts are hypothesized to be too small to see. So, in his scientific practice, Bacon seems to have been relying on some additional form of reasoning in which we are given license to postulate the existence of unobservable entities to explain observable phenomena, such as heat.

Something similar can be said of René Descartes (1596–1650). In contrast to Bacon the empiricist, Descartes the rationalist held that scientific knowledge (*scientia*) is grounded in the "simple natures" of objects, which we can come

to know through direct apprehension, or "intuition." In his influential methodological essay, *Rules for the Direction of the Mind* (1985/1628), Descartes repeatedly warns against settling for "merely probable cognition," instead urging us to "resolve to believe only what is perfectly known and incapable of being doubted" (Descartes, 1985/1628, 10). This may seem to leave little room for abductive reasoning – which, after all, uses empirical data rather than direct apprehensions into simple natures, and delivers theories that are very much capable of being doubted.

However, a closer look at Descartes's own scientific writings, especially in his later work *Principles of Philosophy* (1985/1644), paints a more nuanced picture. Accompanying Descartes's official rationalist theory of scientific reasoning, scholars have found an implicit scientific methodology that resembles abductive reasoning in some important respects (Clarke, 1992; Dellsén, 2017b). In more than 300 separate sections, Descartes posits various novel and ingenious mechanisms to explain numerous natural phenomena, such as why bodies fall toward the earth, how magnets work, and why glass is transparent. Descartes prefaces the discussion by telling us that he wishes to "put forward everything that I am about to write simply as a hypothesis," adding in the French edition that it "is perhaps far from the truth" (Descartes, 1985/1644, 255). Clearly, then, Descartes felt the need to employ some other form of reasoning – in which hypotheses are fallibly posited to explain known phenomena – in addition to his official rationalist and infallibilist methodology.

The methodological necessity of some form of abductive reasoning is also apparent in the writings of various prominent scientists of the early modern period (Thagard, 1978). For example, Antoine Lavoisier's (1743–1794) work on chemical phenomena such as combustion and calcination led him to posit the existence of oxygen, because with it "all the phenomena were explained with an astonishing simplicity" (Lavoisier, 1862, 623). Similarly, Charles Darwin ends his famous discussion of a vast range of empirical facts about biological species that support his theory of evolution by writing: "It can hardly be supposed that a false theory could explain, in so satisfactory a manner as does the theory of natural selection, the several large classes of facts above specified" (Darwin, 1962, 476). Darwin explicitly defended his use of this "method of arguing" by pointing out that "it is a method used in judgment of the common events of life, and has often been used by the greatest natural philosophers" (Darwin, 1962, 476).

In sum, then, it appears that something like abductive reasoning – in which theories are posited to explain known phenomena – emerged during the advent of modern science amongst scientific luminaries such as Bacon, Descartes, Lavoisier, and Darwin. However, as noted above, this form of reasoning

appears to have been largely implicit amongst working scientists of this period, rather than being based on an explicit account of how reasoning of this kind ought to proceed.

1.2 Peirce's Notion of "Abduction"

This began to change with the work of the American pragmatist Charles S. Peirce (1839–1914), from whom the term "abduction" and its cognates seem to originate. Peirce wrote a number of works touching on the topic over his long career, often contrasting "Abduction" with both "Deduction" and "Induction." In one frequently quoted passage, Peirce (1958, 5.145) writes that Abduction follows the following schema:

> The surprising fact C is observed.
> But if A were true, C would be a matter of course.
> Hence, there is reason to suspect that A is true.

For example, one may notice the surprising fact that a burning object placed in a vacuum immediately stops burning. If, as Lavoisier claimed, combustion is a process in which a burning substance combines with oxygen, then this surprising fact would be a matter of course. Hence, according to Peirce's Abduction schema, there is reason to suspect that Lavoisier's theory is true.

It is worth noting that Peirce was arguably not entirely consistent over time about how he defined 'Abduction' – or, indeed, regarding which term he used for it, preferring "Hypothesis" and "Retroduction" in his earlier work. Moreover, most contemporary readers of Peirce agree that his use of the term "Abduction" differs in important ways from how the term tends to be used and understood today. In particular, several scholars (Hanson, 1958; Kapitan, 1992; Minnameier, 2004; Campos, 2011) have argued that in his most influential works, Peirce uses "Abduction" to refer to a psychological process of generating or suggesting new hypotheses. Put differently, the standard interpretation of Peirce's work is that his notion of Abduction primarily describes the process by which we can or should come to think of novel theories, namely by considering what type of theory would potentially explain the facts before us, regardless of whether those theories can be considered true or plausible.

Apart from textual evidence supporting this interpretation, there are philosophical reasons for taking Peircean Abduction to be something other than a rule of inference – or, at most, to be a very weak rule of inference. After all, it should be clear that the same set of facts may lead, via a Peircean Abduction, to quite different, indeed incompatible, theories. Put in terms of the above schema, for each C there will arguably be several incompatible theories $A_1, \ldots, A_n$ such

that if each A_i were true, then C would be "a matter of course." For example, note that Lavoisier's oxygen theory of combustion is not the only theory on which we should expect an object to stop burning once placed in a vacuum. Consider instead the theory that burning involves the transfer of a specific substance, phlogiston, from the object to the surrounding air. This theory also explains why nothing burns in a vacuum, because in a vacuum there is no air to receive the phlogiston that would otherwise be transferred from the object. So which theory, Lavoisier's oxygen-based theory or this phlogiston-based theory, should be inferred? (We cannot infer both, since the two theories contradict each other.) Peircean Abduction, by itself, does not answer these questions, which in turn suggests that Peirce did not intend it to be a rule of inference at all.

A note on terminology is appropriate at this point. As I have intimated, contemporary authors usually use the term "abduction" to refer to an epistemic process of providing support for explanatory hypotheses (see, e.g., Douven, 2021). This is a process that is meant to make certain theories plausible or believable, as opposed to merely helping us come up with those theories. In order to prevent confusion between Peirce's notion of Abduction and the contemporary notion of abduction, I have chosen to use the term "abductive reasoning" when referring to the latter; and, on those occasions I refer to the former, I will use "generation of explanatory hypotheses." Keeping these notions clearly distinct from one another is important for a number of reasons. For example, some accounts of abductive reasoning (e.g., Lipton, 2004) take it to involve, as one part of the process, the generation of explanatory hypotheses (see §2.2).

1.3 The Hypothetico-Deductive Model

Peirce's notion of Abduction is an important early precursor to contemporary accounts of abductive reasoning. Another idea that is arguably just as important a precursor to such accounts is the so-called *hypothetico-deductive model* (the HD model; also known as the hypothetico-deductive *method*), often associated with William Whewell, Hans Reichenbach, and Carl G. Hempel, among others.[4]

[4] See, for example, Sankey (2008, 251) and Okasha and Thébault (2020, 774). With that said, as far as I know, none of the authors mentioned above advocate the simple version of the HD model described below. Of the three, Hempel is perhaps the one that comes closest to doing so in his textbook *The Philosophy of Natural Science* (Hempel, 1966, 196–199). However, a discussion in a textbook can hardly be assumed to accurately reflect Hempel's own considered views on the topic. Indeed, Hempel (1945) proposes a much more nuanced theory of confirmation that conflicts in important ways with the HD model (on this, see Crupi, 2021, §2.1).

The HD model can be thought of as a combination of two ideas. The first idea is about the temporal priority of theory over data. The HD model says, in direct opposition to the inductivism of Francis Bacon, that one should formulate one's theory before one starts gathering data (e.g., by making observations and doing experiments). In other words, one should start by "hypothesizing." At this point, the theory is merely a guess, a hypothesis; it is not something the theorist must take to be true, probably true, or even particularly plausible. According to Hempel (1966, 201–207) there are no rules of rationality for how one should go about coming up with such hypotheses – one may simply let one's imagination roam free in search of some guess that might work. Indeed, it would be impossible to formulate such rules, according to Hempel, because oftentimes the correct guess will be completely different from one's earlier way of approaching the issue, and also very different from the empirical data one has gathered so far. In particular, the guess might well postulate the existence of some new type of entity that cannot be directly observed at all, such as subatomic particles or electromagnetic fields.

The other part of the HD model concerns how this guess – this hypothesis – is evaluated. According to the HD model, the hypothesis is evaluated by testing its empirical consequences. An empirical consequence of a hypothesis is something that can be deduced from it, given background assumptions, and that can be directly verified in some way, such as by an observation or experiment. If these empirical consequences are shown to be correct, the theory from which they have been deduced is confirmed or supported according to the HD model. So the logical structure of scientific confirmation, according to the HD model, is as follows:

The HD model (scientific confirmation): A theory T is confirmed (to some extent), given some background assumptions $\mathcal{A}$, just in case:

(i) T, together with $\mathcal{A}$, deductively implies an empirical consequence E; and
(ii) E is indeed correct, as shown by empirical data.

We are now in a position to see why the HD model has the word "deductive" in it. It's because, in order for the theory to be supported by the observations or experimental results, the empirical consequences which serve as evidence for the theory must be *deducible* from the theory. However, note that what is being deduced is not the theory itself; rather, it is the empirical consequences of the theory. And yet it is the theory that is being supported or confirmed, not (just) its empirical consequences.

There is a caveat to the HD model as presented above that will prove to be important as we contrast it below with prominent accounts of abductive