1 ABDUCTION: BETWEEN CONCEPTUAL RICHNESS AND COMPUTATIONAL COMPLEXITY

Stathis Psillos

1.1 INTRODUCTION

The aim of this chapter is two-fold: first, to explore the relationship between abduction and induction from a philosophical point of view; and second, to examine critically some recent attempts to provide computational models of abduction. Induction is typically conceived as the mode of reasoning which produces generalisations over domains of individuals based on samples. Abduction, on the other hand, is typically seen as the mode of reasoning which produces hypotheses such that, if true, they would explain certain phenomena or evidence. Recently there has been some increasing interest in the issue of how exactly, if at all, they are related. Two seem to be the main problems: first, whether or not induction and abduction are conceptually distinct modes of reasoning; second, whether or not they can be modelled computationally in the same, or similar, ways. The second issue is explored in some detail by several chapters in this collection (e.g. the contributions by Aliseda, Mooney and Poole). The first issue is what the present chapter will concentrate on. My suggestion will be that abduction is the basic type of ampliative reasoning. It comprises as special case both Induction and what the American philosopher Charles Peirce called "the Method of Hypothesis".

In order to motivate and defend my thesis, I proceed as follows. Section 1.2 describes the basic logical features of ampliative reasoning. Section 1.3 takes its cue from Peirce's distinction between Induction and Hypothesis and raises the following question: should the fact that Induction and Hypothesis admit different logical forms

be taken to indicate that they are conceptually distinct modes of ampliative reasoning? I answer this question negatively and defend the view that Induction and Hypothesis are very similar in nature: they are instances of what can be called "explanatory reasoning", where explanatory considerations govern the acceptance of the conclusion. So, I suggest that explanatory reasoning is a basic type of ampliative reasoning, irrespective of the specific logical forms it may admit. In Section 1.4, I describe abduction as the basic type of explanatory reasoning. I suggest that it should be best understood as Inference to the Best Explanation. In particular, I deal with three problems. First, how abduction can acquire an eliminative-evaluative dimension; second, how abduction can produce likely hypotheses; and third, what the nature of explanation is. These are still open issues and what this chapter aims to do is motivate some ways to address them. Finally, Section 1.5 discusses some recent computational models of abduction and notes that there seems to be an inherent tension in the project of modelling abduction. Simple models of abduction are computationally tractable, but fail to capture the rich conceptual structure of abductive reasoning. And conversely, conceptually rich models of abduction become computationally intractable.

1.2 AMPLIATIVE REASONING

It was Charles Peirce who, following Kant's distinction between analytic and synthetic reasoning, called "ampliative" the kind of reasoning in which the conclusion of the argument goes beyond what is already stated in its premises (2.623).[1] A typical case of ampliative reasoning is the following more-of-the-same type of inference: 'All observed individuals who have the property *A* also have the property *A*; therefore, (probably) All individuals who have the property *A* also have the property *B*'. This is what is known as the rule of induction, where the conclusion of the argument is a generalisation over the individuals referred-to in its premises.

Ampliative reasoning is to be contrasted to what Peirce called "explicative reasoning". The conclusion of an explicative inference is already included in its premises, and hence contains no information which is not already, albeit implicitly, in them: the reasoning process itself merely unpacks the premises and shows what follows logically from them. Deductive inferences are explicative inferences. In contrast to this, ampliative reasoning is logically invalid: the conclusions of an ampliative argument can be false although all of its premises may be true. Consequently, the rules involved in ampliative reasoning do not guarantee that whenever the premises of an argument are true the conclusion will also be true. But this is as it should be: the conclusion of an ampliative argument is adopted on the basis that the premises offer *some* reason to accept it as plausible. Were it not for the premises, the conclusion would be unwarranted.

If ampliative reasoning is to be possible at all, one should be reasonable in accepting the conclusions of ampliative arguments, although further information might render them wrong. This feature of ampliative reasoning is called defeasibility. It is its

[1] All references to Peirce's work are given in the standard form and refer to the relevant volume and paragraph of his collected papers.

constitutive difference from explicative reasoning. The latter is not defeasible, since the addition of further information in the premises of a logically valid argument would not affect the derivation of the original conclusion. When it comes to ampliative reasoning, further evidence, which does not affect the truth of the premises, can render the conclusion false. Take for instance the simple inductive argument: 'All hitherto observed swans have been white; so, all swans are white'. The observation of a black swan falsifies its conclusion, without contradicting its premises. Given its defeasibility, one may wonder why ampliative reasoning should be accepted as a legitimate type of reasoning in the first place. The reason for this is that explicative reasoning is not concerned with one of the basic aspects of reasoning, *viz.*, how it is reasonable for someone to form and change their system of beliefs, or the information they hold true. All that explicative reasoning dictates is that since a certain conclusion logically follows from a set of premises, its likelihood of being true is at least as great as the likelihood of the premises being true, and it will remain so when further premises are added. But this is too thin. Judgements as to whether the conclusion, or the premises, are probable enough, or even plausible at all, to be accepted fall outside the province of explicative reasoning. When, for instance, the conclusion of an explicative argument is not acceptable to a reasoner, at least one of the premises should have to go (or the integrity of the derivation may be challenged). But explicative reasoning on its own cannot tell us which premise should go. This requires reasoning based on some considerations of plausibility, and only ampliative reasoning can tell the reasoner what to count as plausible and what not, given the information available. In order, however, to avoid a possible misunderstanding, the following should be stressed. There is nothing wrong with the claim that it is reasonable to accept a statement which logically follows from other premises accepted by a reasoner. Rather, what needs to be emphasised is that a) what makes premises acceptable in the first place is some sort of ampliative reasoning which renders them plausible, or reasonable, given the evidence available; and b) if the conclusion of a deductive argument is not acceptable, explicative reasoning alone cannot tell the reasoner where to revise.

Such opening remarks lead us directly to the *problem of justification* of ampliative reasoning: given that ampliative reasoning is not necessarily truth-preserving, how can it be justified? This is *Hume's problem*, for although David Hume first raised it for induction, his challenge concerns ampliative reasoning in general. His point hinges on the fact that ampliative reasoning is defeasible. Since, the Humean challenge goes, the premises of an ampliative argument do not logically entail the conclusion, there are possible worlds in which the premises are true and the conclusion false. How then, the challenge goes on, can we show that the actual world is one of the possible worlds in which whenever the premises of an ampliative argument are true, its conclusion is also true? Or even, how can we show that in the actual world most of the times in which the premises of an ampliative argument are true, the conclusion is also true? The Humean challenge is precisely that the only way to do this is bound to presuppose that ampliative reasoning is rational and reliable; hence, its bound to beg the question.

What the Humean challenge is taken to suggest is that the premises of an ampliative argument cannot confer warrant or rational support on its conclusion. This is the central philosophical issue concerning ampliative reasoning. Any substantial de-

fence of the rationality of ampliative reasoning should either solve or dissolve Hume's problem. Yet, this is not the place to deal with this philosophical problem. Instead, this chapter will concentrate on another problem, which needs to be dealt with independently of the problem of justification. It is the *descriptive problem*: what is the structure of ampliative reasoning? No-one, including Hume himself, but save Popper, denies that humans are engaged in ampliative inferential practices. What exactly these inferential practices involve, and whether or not they admit specific logical forms, are issues worth looking into. One may call the descriptive problem *Peirce's problem*, since Peirce was, arguably, the first who tried to address it systematically.

1.3 EXPLANATORY REASONING: INDUCTION AND HYPOTHESIS

That ampliative reasoning admits specific logical forms goes back to Peirce's early work on logic and inference. As is well-known (and further explained in the introduction of the present book), early Peirce (2.372-388) attempted to model ampliative reasoning on the logical form of explicative reasoning. Take, for instance, the following typical case of explicative reasoning:

D: All *A*'s are *B*; *a* is *A*; therefore, *a* is *B*.

An obvious re-organisation of *D* is

I: *a* is *A*; *a* is *B*; therefore All *A*'s are *B*.

I (what Peirce originally called "Induction") moves from some observations about a set of individuals (i.e., that the individuals in the sample are both *A* and *B*) and returns a generalisation over *all* individuals of a certain domain. But the deductive rule *D* above admits of yet another re-organisation:

H: *a* is *B*; All *A*'s are *B*; therefore *a* is *A*,

where the premises are a particular known fact (*a* is *B*) and a generalisation (All *A*'s are *B*), while the conclusion is a particular hypothesis (that *a* is *A*).

The fact that argument-patterns *I* and *H* have different logical forms suggests that there may well be two different and distinct types of ampliative reasoning. While the argument-pattern *I* clearly characterises the logical form of the intuitive more-of-the-same rule of induction, the argument-pattern *H* is more difficult to characterise. Peirce called "Hypothesis" (or the "method of hypothesis") the mode of reasoning which corresponds to *H*. It can be illustrated by using Peirce's own example: given the premises "All the beans from this bag are white" and "These beans are white", one can draw the hypothetical conclusion that "These beans are from this bag" (2.623). Peirce seems to have thought that the argument-patterns *H* and *I* correspond to two distinct modes of ampliative reasoning, since he noted that "induction classifies, whereas hypothesis explains" (2.636). As he put it: "Induction is where we generalise from a number of cases of which something is true, and infer that the same thing is true of a whole class. (...) Hypothesis is where we find some very curious circumstance, which would be explained by the supposition that it was a case of a certain general rule, and thereupon adopt that supposition" (2.636). However, scholars of his work, most notably (Fann, 1970, p.22-23), suggest that he was not prepared to separate sharply the two forms of inference, but that he conceived of induction and hypothesis as occupying opposite

ends of the continuum of ampliative inference. Be that as it may, I want to focus on and defend the following thesis. The different logical forms of Induction and Hypothesis should not obscure the fact that they are very similar in nature: they are instances of what one may call explanatory reasoning, where explanatory considerations govern the acceptance of the conclusion. So, I want to suggest that explanatory reasoning is a basic type of ampliative reasoning, irrespective of the specific logical forms it may admit. In order to defend this thesis I shall first discuss the case of Induction.

In order to see how a nomological generalisation of the form "All *A*'s are *B*" is explanatory we need to consider the following *contrastive* explanation-seeking question: 'Why is this sample of individuals which are *A* also *B*, rather than not-*B*'? (e.g., 'Why is this sample of ravens black, rather than white'?). When this question is asked, what is looked for is a relevant difference between an actual case (e.g. the sample containing only black ravens) and an unactualised, but possible, case (e.g., the sample containing white ravens, or both white and black ravens) (cf. (Lewis, 1986)). The relevant difference is that by virtue of a law, the contrastive class of *A*'s which are not *B* is empty. In other words, the relevant difference is that there is a nomological connection between being *A* and being *B* which makes it the case that *all A*'s are *B*. Therefore, the nomological generalisation "All *A*'s are *B*" explains why the sample has failed to contain an individual which is *A* but not *B*. This can be suitably extended to statistical generalisations. The nomological generalisation that *x*% of *A*'s are *B* explains why the random sample of individuals has displayed the observed frequency of *A*'s which are *B*.[2]

What needs to be stressed is that good inductive reasoning involves comparison of alternative explanatory hypotheses. In a typical case, where the reasoning starts from the premise that 'All *A*'s in the sample are *B*', there are two possible conclusions that can be drawn. The first is that the observed correlation in the sample is due to the fact that the sample is biased. The second is that the observed correlation is due to the fact that there is a nomological connection between being *A* and being *B* such that All *A*'s are *B*. Which hypothesis should be chosen as the appropriate conclusion will depend on explanatory considerations. Insofar as the conclusion "All *A*'s are *B*" is accepted, it is accepted on the basis it offers a better explanation of the observed frequencies of *A*'s which are *B* in the sample *in contrast to* the (alternative potential) explanation that someone has biased the sample in order to make us think that all *A*'s are *B*.[3]

In order to see how hypothetical reasoning of the form *H* above is explanatory, lets take a toy-example. Suppose that we observe a black bird (*a* is *B*) and that, by instantiating schema *H*: {*a* is *B*; All *A*'s are *B*; therefore *a* is *A*}, we infer that, given that All ravens are black, this bird is a raven (*a* is *A*). We have thereby answered the explanation-seeking question "Why is individual *a B*?" by hypothesising that *a* is *A* and by appealing to some sort of nomological connection between being *A* and being *B*. The nomological connection between property *A* and property *B* is part of the information contained in the premises of the explanatory argument *H*. The conclusion of the argument, that *a* is *A*, does not explain (how could it?) this nomological connection, but it is itself a potential explanation of the observation that *a* is *B* only in

[2] A similar point is made by John Josephson in his chapter in the present volume.
[3] Gilbert Harman (Harman, 1965) has also emphasised this point.

virtue of this nomological connection. This observation seems to be the essence of the Hempelian Deductive-Nomological account of explanation (cf. (Hempel, 1965)). On this account, explanation amounts to nomic expectability. A singular event e (the *explanandum*) is explained iff a description of *e* is the conclusion of a valid deductive argument, whose premises, the *explanans*, involve essentially a law-like statement L, reporting a law of nature, and a set C of initial, or antecedent, conditions. So, the event *e* is explained by showing how this event should have been expected, if the relevant laws and certain initial conditions were taken into account. For instance, on this account, we offer a potential explanation of the fact that the beer keg exploded in the basement by citing the law which connects the pressure of a liquid with its temperature and by appealing to a certain antecedent condition, *viz.*, that the temperature of the beer in the keg rose rapidly. We therefore explain the explanandum by subsuming it under a law. It then appears that schema *H* is nothing but the Hempelian Deductive-Nomological account of explanation. We ask: why did *e* happen? And we answer the question by constructing an argument of the type *H* above, whose premises are law-like statements and statements of initial conditions.

My suggestion then is that one should simply see both *I* and *H* as species of one and the same genus of reasoning: explanatory reasoning, where hypotheses are being generated and accepted on an explanatory basis, irrespective of the logical form that these hypotheses might take. This is not to suggest that there are no differences between *I* and *H*. The reasoning process behind *I* produces generalisations. In the case of *H*, the reasoning begins with the observation that *a* is *B*, and then, in one and the same act, it asserts that this observation would be explained if one hypothesised that there was a nomological connection between *A* and *B* and that *a* was *A*. So, the reasoning process behind *H* produces both a general hypothesis (asserting the nomological connection between *A* and *B*) and a particular hypothesis (asserting that the individual *a* is *A*). In any case, we shouldn't lose sight of the fact that the generation and acceptance of general hypotheses is the product of explanatory reasoning no less than the generation and acceptance of particular hypotheses. The fact that *I* can lead only to general hypotheses, whereas *H* can lead to both general and particular hypotheses does not amount to a fundamental category difference between the two. In either case, it is *explanatory reasoning* which leads to the generation and acceptance of hypotheses, be they general or particular.[4]

1.4 ABDUCTION

In his famous characterisation of abduction, Peirce described "abduction" as the reasoning process which proceeds as follows: "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" (5.189). We can easily see that this process can underlie both argument-patterns *I* and *H*. Suppose that the surprising fact is the particular fact that *a* is *B*. This is ex-

[4]With one possible exception, *viz.*, the predictive inference, as in the case of next-instance induction. There, we move from *n* observed *A*'s being *B* to conclude that the next A is going to be *B*. The conclusion is a singular statement and is clearly non-explanatory. But one may think of predictive inference as parasitic on the implicit generalisation "All *A*'s are *B*".

plained by saying that if *a* is *A* and All *A*'s are *B*, then *a* is expected to be *B*. This piece of reasoning is nothing but an instance of the argument-pattern *H*. Suppose now that we allow the surprising fact to be the correlation of two properties *A* and *B* in a sample of individuals. Then, we can explain this by saying that the sample is the way it is because All *A*'s are *B*. This is an instance of the argument-pattern *I*. So, abduction can incorporate both *I* and *H*, and therefore can lead to the generation of generalisations no less than to the generation of hypotheses stating particular facts. Accordingly, I shall reserve the term abduction for explanatory reasoning in general and suggest that abduction comprises both argument-patterns *I* and *H*. This may tally with what Peirce himself thought of abduction, when in his later years introduced abduction as a distinct type of reasoning. But I refrain from engaging in interpretative work.[5] Instead, I will try to characterise more precisely what exactly is involved in abduction as a reasoning process, drawing on some of Peirce's thoughts and suggestions and pointing out some open problems.

Three, I think, are the big problems that any precise characterisation of abduction faces. The first is what I will call the "multiple explanations problem". The second concerns the connection between the connection between reasoning process behind abduction and the likelihood of the hypotheses that it generates. The third is the nature of explanation itself. Let me consider them in turn.

1.4.1 Abduction as inference to the best explanation

There is a clear sense in which the reasoning process that Peirce's quotation captures is inadequate. For there are, typically, more than one mutually incompatible hypotheses T_i,...,T_n such that, if true, they would make the explanandum-event *e* a matter of course. If the mere fact that the explanatory hypothesis made the explanandum "a matter of course" were enough to render this hypothesis plausible, then an unlimited number of mutually incompatible hypotheses would be equally plausible. I shall call this problem the "multiple explanations problem". As a result of this, one has either to resolve for the view that abduction is impotent to impose any restriction on the acceptance of hypotheses, or to beef up the reasoning process behind abduction so that it acquires an evaluative-eliminative component.

The search for (types of) explanatory hypotheses should be preferential. The search should aim to create, as Peirce nicely put it (Peirce, 1957, p.254), "good hypotheses". Consequently, this search should produce an evaluation of hypotheses which ranks them in an order of preference, reflecting a distinction of hypotheses into better and worse. Those hypotheses are ranked higher which a) explain *all* the facts that led to the search for hypotheses; b) are licensed by the existing background beliefs;[6] c) are, as

[5]The interested reader should look at the Introduction of this book for a description of Peirce's account of abduction. More relevant literature includes (Burks, 1946; Hanson, 1965; Fann, 1970; Thagard, 1981; Flach, 1996).

[6]Using background beliefs to give a hypothesis a certain place in the order of preference is going to influence the likelihood of the hypothesis, and hence its acceptability, since background beliefs themselves are, typically, supported by evidence to some degree.

far as possible, simple; d) have unifying power [7] e) are more testable, and especially, are such that entail novel predictions.[8] These factors are not algorithmic in character, but this does not mean that one cannot decide, on their basis, which hypothesis should be ranked highest. In fact, in most typical cases, these factors will lead to a definite conclusion, be it about medical diagnosis or car mechanics, or what have you. So, for instance, a diagnostician, pretty much like a good car mechanic, will look for a hypothesis about the cause of symptoms such that: it accounts, if possible, for all the symptoms; it is consonant with background knowledge as what types of causes produce these symptoms; it avoids, in the first instance, attributing the symptoms to multiple causes; it can yield further predictions that can be tested (e.g., that the patient will recover if they take a certain medicine which acts on the cause of the symptoms). It's not implausible to think that although virtually never do we go through all such factors explicitly when we are engaged in abductive reasoning, all these factors have nonetheless been internalised by a good reasoner who, then, applies them implicitly in the case at hand. The internalisation of these factor may well be what Peirce called "good sense" (7.220). In most typical cases, an explicit reconstruction of the reasoning process will reveal the implicit reliance on such factors. Similarly, in most typical cases, the product of the reasoning will be just one hypothesis which is ranked as most plausible. But when there is more than one (e.g., the light does not come on; is it because the light-bulb is gone; because the fuse is blown; or because of a power-cut?), the reasoning process itself contains obvious resources which will lead to adjudication. To the extent that the application of these evaluative-ranking criteria mark the degree of goodness of a hypothesis, it is reasonable to say that abduction is nothing but what (Harman, 1965) has called "Inference to the Best Explanation". According to this mode of reasoning, a hypothesis H is accepted on the basis that a) it explains the evidence and b) no other hypothesis explains the evidence as well as H does. So, not only is there a reasoning process which underlies abduction, but also this reasoning process has a certain logical, though not algorithmic, structure.

1.4.2 Abduction and confirmation

It should be clear that the product of abductive reasoning – the explanatory hypothesis – is not guaranteed to be true. This is not surprising, given that abductive reasoning is defeasible. But, surely, one may think, what is at issue here is not the obvious fact of defeasibility. Instead, the objection may be that abductive reasoning cannot return an explanatory hypothesis which might be reasonably said to be (likely to be) true. For, one might ask, what reasons would govern such judgements of likelihood? Yet, in the end of the day, a good reasoner should want to adopt hypotheses that are likely to be true, or that she has reasons to think that they are likely to be true. Peirce was surely aware of the problem: "A hypothesis then has to be adopted which is likely in itself and renders the facts likely. This process of adopting a hypothesis as being suggested

[7] Or, breadth, as Peirce put it (7.220-1 &7.410).
[8] cf. Peirce (7.220 & 7.115).

by the facts is what I call abduction" (7.202). The question is: how can abduction be this process? How, that is, can abduction render the chosen hypothesis likely?

For a plausible solution to this problem we may take our cue from late Peirce's suggestion that abduction should be seen as part and parcel of the method of enquiry (cf. 7.202ff.). So, the reasoning process that underlies abduction should be embedded in a more general framework of inquiry so that the hypotheses generated and evaluated by abduction can be further tested. The result of this testing is the confirmation or disconfirmation of the hypothesis which, naturally, affects its likelihood to be true. We should therefore conceive of abduction as the *first stage* of the reasoner's attempt to add reasonable beliefs into his belief-corpus in the light of new phenomena or observations. The process of generation and ranking of hypotheses in terms of plausibility (abduction) is followed by the derivation of further predictions from the abduced hypotheses. Insofar as these predictions are fulfilled, the abduced hypothesis gets confirmed. Peirce himself thought that the process of generating predictions is deductive and came to call "Induction" the testing these predictions, and hence the process of confirming the abduced hypothesis (cf. 7.202ff).[9] Leaving once again aside some important interpretative issues, I make the following use of Peirce's idea: although a hypothesis might be reasonably accepted as the most plausible hypothesis based on explanatory considerations (abduction), the *degree of confidence* in this hypothesis is tied to its degree of subsequent confirmation. The latter has an antecedent input, i.e., it depends on how good the hypothesis is (i.e., how thorough the search for other potential explanations was, how plausible a potential explanation is the one at hand etc.), but it also crucially depends on how well-confirmed the hypothesis becomes in light of further evidence. So, abduction can return likely hypotheses, but only insofar as it is seen as an integral part of the method of inquiry, whereby hypotheses are further evaluated and tested.

1.4.3 Explanation

As we have already seen, in his famous characterisation of abduction, Peirce noted that the abduced hypothesis makes the surprising fact to be explained a matter of course. This reference to matter of course is not accidental. It suggests that the explanatory hypothesis should be such that it removes the surprise from the occurrence of the explanandum. But, although it is certainly part of an explanation that it renders the explanandum non-surprising, what needs to be added is in exactly what ways the explanandum is rendered non-surprising. And although it is intuitively clear that to explain an explanandum-event is to provide information about its causal history, there is substantive disagreement over *how exactly* we should understand this last claim. Explanation is effected by pointing to some causal-nomological connections between the explanandum and the fact that is called upon to do the explaining. But the nature of these causal-nomological connections are under heavy dispute.

[9] "But if [abduction is] to be understood to be a process antecedent to the application of induction, not intending to test the hypothesis, but intended to aid in perfecting that hypothesis and making it more definite, this proceeding is an essential part of a well-conducted inquiry" (7.114) And "Induction is a process for testing hypotheses already in hand. The induction adds nothing" (7.217).

Two are the important points of dispute. The first centres around how exactly the explanatory connection is to be understood. Some philosophers (e.g. (Hempel, 1965; Kitcher, 1981)) argue that explanation proceeds via derivation. They claim that explanations are, essentially, arguments such that an event-type P explains an event-type Q iff (a description of) the explanandum-event logically follows from a set of premises which essentially involve (a description of) P. The well-known Deductive-Nomological account of explanation is an instance of this approach. What's typical of this approach is that causal order follows from (instead of being presupposed by) explanatory-derivational order: what causes what is settled after we have settled the question what explains (in a derivational sense) what. Opposite to the above approach is the view (advocated, among others, by (Salmon, 1984)) that explanation are not arguments. Instead, they should characterise the causal mechanisms that bring about the explanandum-event, irrespective of whether (descriptions of) these mechanisms can be captured in the premises of an argument whose conclusion is(a description of) the explanandum-event.[10] The second (related) dispute focuses on the role of laws in explanation. On one approach, laws and reference to nomological connections are essential part of an explanation, whereas on another view, causal stories can be complete even though they make no reference to laws, or even though there may be no relevant laws to refer to. These issues are in the forefront of the current philosophical debate. So, here I will not try to examine them further (but cf. (Salmon, 1990)). It should be enough to keep in mind two things. First, it is still an open issue what exactly an explanation is. Second, whatever the explanatory relation is taken to be in its details, its connection with causal and/or nomological information about the explanandum-event and its function as a surprise-remover should be pretty uncontroversial. I think the best way to capture the latter function is to point out that explanation is typically linked with improving our understanding of why an event happened and that improvement of understanding occurs when we succeed in showing how an event can be made to fit in the causal-nomological nexus of things we accept. We remove the surprise of the occurrence of an event if we show that the acceptance of certain explanatory hypotheses, and their incorporation into our belief-corpus, helps to include the explanandum-event into this corpus. Schematically, if BK is this belief corpus, e is the explanandum-event and T is a potentially explanatory hypothesis, then T should be accepted as a potential explanation of e if BK alone cannot explain e, but $BK \cup T$ explains e.

To sum up, abduction, conceived as Inference to the Best Explanation has a rather definite logical structure: it is the reasoning process in which the reasoner generates and evaluates a number of potentially explanatory hypotheses, in the light of background knowledge. Judging the plausibility of each of them, and ranking them accordingly, is precisely the respect in which abduction is evaluative. The degree of

[10] Well-known counter-examples to the Deductive-Nomological account of explanation suggest that there is more to causal explanation than can be captured by the DN-pattern. The DN-patterns is symmetric, but causation is not. For instance, one can explain the length of the shadow of a flagpole in a DN-fashion by constructing an argument whose premises are general laws about the propagation of light and particular conditions about the height of the flagpole. Yet, one can use the length of the shadow as the initial condition and DN-explain the height of the flagpole by reversing the above DN-argument. The latter DN-derivation cannot count as a genuine explanation because it does not respect the relation of cause and effect.

confidence in the chosen hypothesis, however, is a matter of how well the hypothesis will stand up further testing.

1.5 ABDUCTION AND COMPUTATION

Having outlined a conceptual model of abductive reasoning, I shall now turn my attention to two major attempts to provide computational models of abductive reasoning. The aim of this section is to motivate (but not prove) the point that, because of its rich structure, abduction resists an adequate and computationally tractable formal model. Before I embark on this task, I should note that the following points are meant only to be part of a general philosophical critique of (some aspects of) the computational approach to abduction which does not aim to minimise or bypass the important technical achievements related to the use of abduction in Artificial Intelligence.[11]

In Logic Programming (LP) (Kakas *et al.*, 1992; Console *et al.*, 1991), abduction operates in the context of a logic program. The aim of an abductive problem is to assimilate a new datum O into a knowledge-base (KB). So, KB is suitably extended by a certain hypothesis H into KB' such that KB' incorporates the datum O. Abduction is the process through which an H is chosen. The logical form of an abductive problem in LP is the following (call it F): given a KB and a datum O, search for a hypothesis H such that i) $KB \cup H \models O$ and ii) $KB \cup H$ is consistent. In a typical LP, abduction is used to detach (and affirm) the antecedent (the body) of a conditional (rule) which is part of the domain theory (KB) in order to show how its consequent (its head) can be proved. Take, for instance, the following well-known toy-example. The domain theory consists of the following two statements: KB: {grass is wet, if rained last night; grass is wet, if sprinkler was on}. In order to explain the observation O that the grass is wet we may abduce the hypothesis H that it rained last night (or, alternatively that the sprinkler was on). Given H, and given that it is consistent with KB, we can then run the program to prove that $KB \cup H \models O$.

As it stands, F cannot adequately capture the structure of an abductive problem. Here are some of the reasons why. *First*, if the only task is to suitably augment KB so that (i) and (ii) above are satisfied, then the reasoner (program) might trivially incorporate O straight into KB, without bothering for an H. *Second*, if (i) and (ii) are the only elements of an abductive problem, then, as the toy-example shows, there can be more than one hypothesis that satisfy them. (i) and (ii) above cannot, on their own, distinguish between the many H's that satisfy them. From a syntactic-computational point of view, all H's which satisfy (i) and (ii) are the same. *Third*, searching for hypotheses H that satisfy non-trivially (i) and (ii) above requires a conceptual space of hypotheses from which H's can be drawn. Consistency with KB is too permissive a criterion because the reasoner (program) might well end up examining all kinds of irrelevant, but consistent with KB, hypotheses before starting investigating the relevant ones. So, other constraints should be incorporated which guide the search and order the hypotheses to be examined in a preferential way. Typically, the generation of H's is constrained by the existing KB. Hence, the selected H's should not be merely

[11] For a recent survey of the role of abduction in AI, see (Konolige, 1996).

consistent with *KB*, but they may stand in a stronger relation to it (e.g., they are made likely by *KB*.) Notice also that the required relation cannot be entailment of *H*'s by *KB*, unless one is willing to accept only mutually consistent hypotheses as potential explanations of *O*. If all potential *H*'s are entailed by *KB*, then they have to be mutually consistent. However, it is essential for abduction to be able to deal with mutually inconsistent explanatory hypotheses. *Fourth*, knowledge assimilation is typically abductive, but it is a much more complicated process than the one characterised by (i) and (ii) above. Here, let me only stress that requirement (ii) above can be too restrictive. It may well be the case that the assimilation of datum *O* requires extensive modification of the existing *KB* in such a way that the adopted *H* is inconsistent with the existing *KB*, although, of course, the new KB' which includes *H* should be internally consistent. *Fifth*, the explanation of the datum *O* need not be deductive. It may well be the case that *O* does not logically follow from *H* and *KB*, but that still $KB \cup H$ explain *O*, by showing how *O* was to be expected (e.g., by showing how $KB \cup H$ make *O* likely, or more likely than not-*O*.).

An adequate computational model of abduction should be able to deal with such problems. That is, it should incorporate these features into the computation. Naturally, there should be a trade-off between the need to set-up a conceptually adequate model and the need for the model to be computationally tractable. But at least computational modelling should aim to characterise as adequately as possible the rich conceptual structure of an abductive problem.

LP-theorists have attempted to improve on *F* above. There are three main ways in which *F* has been improved. *First*, the logical space from which *H*'s are drawn comprises a set *A* of domain-specific hypotheses, called *abducibles*. In the toy-example above the abducibles are {rain last night; sprinkler was on}. In the event calculus the set of abducibles is a set of events (or event-predicates of the form *happens (E)*) which are abduced to hold at a time t_1 in order to explain how a property holds at t_2 ($t_2 > t_1$) (cf. (Shanahan, 1989)). *Second*, the updates of *KB* are subjected to a set of *integrity constraints* (IC), i.e., a set of meta-rules which specify which changes of the *KB* are not allowed and, therefore, specify which abducibles are not acceptable. In the event calculus, an *IC* is such that a property cannot hold at time t_2 even though it held at time t_1 ($t_2 > t_1$) if there was an event that terminated P at a time t_{12}, where $t_2 > t_{12} > t_1$. *Third*, the abduced hypothesis must be minimal, i.e., it must not be decomposable into two others each of which could on its own explain the datum. So, *F* above gives way to the following schema (F'): an abductive problem is characterised by the triple $< KB, A, IC >$ and it consist in searching for a minimal *H* such that i') $KB \cup H \models O$ and ii'') $KB \cup H$ satisfies *IC*.

There is no doubt that F' is on the right track. But its own limitations suggest some very general problems with the computational approach. Two such problems stick out. *First*, although it should be clear that the specification of a set of abducibles is necessary for any computational model of abduction, its doubtful that this specification can only be achieved by syntactic-computational resources. An abductive problem is not merely the search of an explanation *H* of a datum *O* such that $KB \cup H \models O$. Rather, it is the search of an explanation of a *particular type*, one that the background information suggests that is relevant to the understanding why *O* occurred. When,

for instance, the computer does not come on we look for blown fuses, power-cuts, or internal failures, but we do not look for astral influence, or for who switched it off last time etc. Some hypotheses are relevant while others are not. The first should be properly called abducible. Yet, these judgements of relevance – and hence the specification of the appropriate type of abducibles – are not-syntactic, although once in place, they may admit a certain logical-computational form.[12] The *second* problem with F' is that it does not yet have built into it the required preferential structure. As it stands it simply seems to lack the resources to rank abducibles in some order of preference. The requirement of minimality says that of two mutually consistent abducibles the minimal should be preferred, but as it stands it applies only to abducibles with a definite logical structure, e.g., p and $p\&q$. It does not say, for instance, among two or more mutually inconsistent abducibles which one should be preferred. Nor does it say among two equally simple, but mutually consistent hypotheses, which should be chosen. In a nutshell, F' does not yet capture the rich structure of abductive reasoning.

LP-theorists have developed several techniques to deal especially with the multiple explanations problem.[13] At this stage, however, they are sets of heuristics which are not fully incorporated into the computational framework of abductive reasoning. According to Michalski in (Michalski, 1993, p.120), however, the computational characterisation of abduction need *not* capture its preferential structure. He suggests that what needs to be formalised is the process of generating (creating) *an* explanation, not the evaluation of which explanation is the best. Michalski's abstract model conceives of abduction as "reversed deduction", or as he puts it, as tracing backwards an implication rule: where in deduction the reasoner looks for conclusions of the premises she already accepts, in abduction she looks for the premises that, if true, would entail a certain conclusion she already accepts. (Hence, Michalski's abduction looks very much like early Peirce's Hypothesis.)

Michalski is quite right in stressing that the determination of the best among a set of alternative explanations is not always easy. Abductive reasoning will not always rank hypotheses in such a way that one, and only one, comes out the best. But his objection cuts much deeper than that. He thinks that the logical properties of abductive reasoning do *not* depend on any measure of the goodness of an explanation. It should be clear, however, that if abduction was just what Michalski thinks, then abduction would generate an infinity of crazy explanations. Michalski's own example is instructive. Envisage a case in which one wants to explain why one's pencil is green. Abduction could easily trace backwards the following implication: {My pencil is grass; Everything that is grass is green; Therefore, my pencil is green}. The result

[12](Console *et al.*, 1991, p.668) give a syntactic characterisation of abducible as follows: the abducible symbols are exactly those not occurring in the head of any clause in the theory. This characterisation works, however, only in the limited case in which the explanatory hypothesis is already included in the background theory. If the explanation is to be sought outside the theory and is such that, together with the theory, it explains the datum, then Console *et al.* need to explicitly introduce a set of abducibles (cf. p.676). It should then be clear that the specification of this set cannot be made syntactically.

[13]See (Kakas *et al.*, 1997) and (Evans and Kakas, 1992). Evans and Kakas use the notion of corroboration to select explanations. But it should clear that the notion of corroboration is not related to the search for explanations but rather to the degree of confidence in the chosen explanation. Corroboration is more akin to Peirce's later use of induction rather than to his abduction.

would be that the reasoner might consider as an explanation of the fact that the pencil is green that it is grass. It is precisely because the computational characterisation of abduction should avoid such trivialities, that some measure of goodness of the abduced hypotheses should be incorporated in it.

Michalski does, after all, build into his abstract model some measure of goodness of potential explanation. He makes abduction dependent on some estimation of the likelihood of what he calls a "mutual implication". According to his suggestion, whether or not a hypothesis of the form "All *A*'s are *B*" is a good explanation depends on the backward strength of the converse implication: "All *B*'s are *A*". If it is likely that 'If something is a *B*, then it is also an *A*' then, upon finding a *B* we may conclude that it is also an *A*. So, on this suggestion, inferring from "*a* is *B*" and "All *A*'s are *B*" that probably "*a* is *A*" depends on how likely it is that "All *B*'s are *A*". If it is very likely, then the reasoner may accept that *a* is *A*, but not otherwise. In the example above, it would be silly to infer that my pencil is grass because the reversed implication "If something is green, then it is grass" is not at all likely. Since there is no much space at present to evaluate properly Michalski's theory, the only point I will stress is that his suggested measure of goodness does not depend on explanatory considerations. His suggestion amounts to the claim that the likeliest hypothesis should be chosen. This is a sound piece of advice, if we already know which is the likeliest hypothesis. But if we do know that, then there is no reason to generate any other than the likeliest hypothesis.

Bylander and his collaborators in (Bylander *et al.*, 1991) have aimed to offer a computational model of abduction which captures its evaluative element. According to them, an adduction problem is a tuple $< D_{all}, H_{all}, e, pl >$ where: D_{all} is a finite set of all the data to de explained; H_{all} is a finite set of all the individual hypotheses; e is a map from all subsets of H_{all} to subsets of D_{all}; pl is a map from subsets of H_{all} to a partially ordered set representing the plausibility of various hypotheses. In this model, an explanation is a set of hypotheses H such that H is complete and parsimonious, i.e., such that $e(H) = D_{all}$ and there is no proper subset H' of H such that $e(H') = D_{all}$. The *best* explanation is the H with the highest place in the plausibility ordering. Lets call this model J.

There are clear senses in which J is an improvement over F' and over Michalski's model. Its most distinctive improvements are a) that it is not built into the model that an explanation should be a deductive argument and b) that potential explanations are ordered in terms of plausibilities. Allowing for an initial plausibility ordering takes account of the way in which background information and explanatory considerations may affect the trustworthiness of a hypothesis. In the case of medical diagnosis, where J has been applied, the plausibility ordering suggests, for instance, that not all hypotheses concerning the causes of a set of symptoms are equally licensed by background information. The plausibility ordering is also helpful from a technical point of view. Given that J conceives of explanation as a function from subsets of H_{all} to subsets of D_{all}, it should be clear that there will, normally, be a large number of such potential explanations. If they are ranked in terms of plausibility, then some of them will be deemed implausible and will not be further entertained.

The *J* model, however, has some weaknesses, too. Some computational difficulties have been noted by the authors of *J* themselves. They point out that *J* makes abductive problems computationally tractable only if it assumed – as a rule, implausibly – that there are no incompatibility relationships between the competing hypotheses. Besides, if it is required that there always should be one most plausible (best) explanation, then there intractability is guaranteed. Some conceptual in nature problems with *J* have been noted by (Thagard and Shelley, 1997). What one may add is that plausibility in *J* is taken as a primitive notion. Although it is right to say that the details of the plausibility ordering will be domain-specific, *J* needs to say more about its general structure in order to accommodate explanatory factors into abductive reasoning. To be sure, (Josephson and Josephson, 1994, ch.9) offer a weaker model of abductive reasoning which is computationally tractable. In this model the task of an abductive problem is to explain as much as possible of the data with acceptable levels of confidence. Completeness and maximal plausibility have to be sacrificed in favour of the weaker aim of maximising explanatory coverage. There are algorithms for this model which compute the result in polynomial time. This is clearly an improvement in respect of computation, but some of the element which make their original model *J* conceptually rich have to go.

1.6 CONCLUSIONS

To recapitulate, I have argued that abduction has a rich conceptual structure which comprises induction as a special case. Abduction is the mode of reasoning in which a hypothesis *H* is accepted on the basis that a) it explains the evidence and b) no other hypothesis explains the evidence as well as *H* does. So, the reasoning process which underlies abduction has a certain logical, though not algorithmic, structure. Induction produces generalisations (be they universal or statistical), but these are explanatory and their acceptance is governed by explanatory considerations. So, although induction may be taken to be superficially distinct from abduction, it is an instance of explanatory reasoning.

As for the second theme of this chapter, i.e., the critical discussion of the recent computational modelling of abduction, I wish to sum it up with a conjecture: the more conceptually adequate a model of abduction becomes, the less computationally tractable it is. This may leave us with a dilemma: either we may have to go for computational tractability at the expense of conceptual richness, or we may have to resolve for the view that a rich conceptual model of abduction cannot be adequately programmed. The solution, if any, lies with future research.

Acknowledgments

Many thanks to Peter Lipton whose book “Inference to the Best Explanation” (Lipton, 1991) has been a great source of inspiration; to John Josephson for many helpful comments on an earlier draft; to Bob Kowalski for many hours of discussions about the role of abduction in Artificial Intelligence; and to Francesca Toni for her patient defence of the Logic Programming approach to abduction. Research for this chapter was conducted under a British Academy Postdoctoral Fellowship. I am grateful to the Academy for all the help. Many ideas of this chapter have been

stimulated by, and have found a great companion in, the views expressed in the book "Abductive Inference" by John Josephson and his collaborators (Josephson and Josephson, 1994), whom I wish to thank.

¶

References

Burks, A. (1946). Peirce's theory of abduction. *Philosophy of Science*, 13:301–306.

Bylander, T., Allemang, D., Tanner, M., and Josephson, J. R. (1991). The computational complexity of abduction. *Artificial Intelligence*, 49:25–60.

Console, L., Theseider Dupré, D., and Torasso, P. (1991). On the relationship between abduction and deduction. *Journal of Logic and Computation*, 1(5):661–690.

Evans, C. A. and Kakas, A. C. (1992). Hypothetico-deductive reasoning. In *Proceedings of the International Conference on Fifth Generation Computer Systems*, pages 546–554. ICOT.

Fann, K. T. (1970). *Peirce's Theory of Abduction*. Martinus Nijhoff, The Hague.

Flach, P. A. (1996). Abduction and induction: syllogistic and inferential perspectives. In Flach, P. A. and Kakas, A. C., editors, *Proceedings of the ECAI'96 Workshop on Abductive and Inductive Reasoning*, pages 31–35. Available on-line at `http://www.cs.bris.ac.uk/~flach/ECAI96/`.

Hanson, N. R. (1965). Notes towards a logic of discovery. In Bernstein, R., editor, *Critical Essays on C.S. Peirce*. Yale University Press.

Harman, G. (1965). The inference to the best explanation. *Philosophical Review*, 74:88–95.

Hempel, C. G. (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. Free Press, New York.

Josephson, J. R. and Josephson, S. G. (1994). *Abductive Inference: Computation, Philosophy, Technology*. Cambridge University Press, New York.

Kakas, A. C., Kowalski, R. A., and Toni, F. (1992). Abductive logic programming. *Journal of Logic and Computation*, 2(6):719–770.

Kakas, A. C., Kowalski, R. A., and Toni, F. (1997). The role of abduction in logic programming. In Gabbay, D. M., Hogger, C. J., and Robinson, J. A., editors, *Handbook of Logic in Artificial Intelligence and Logic Programming*, volume 5, pages 233–306. Oxford University Press.

Kitcher, P. (1981). Explanatory unification. *Philosophy of Science*, 48:251–281.

Konolige, K. (1996). Abductive theories in artificial intelligence. In G.Brewka, editor, *Principles of Knowledge Representation*. CSLI Publications.

Lewis, D. (1986). Causal explanation. In *Philosophical Papers*, volume 2. Oxford University Press.

Lipton, P. (1991). *Inference to the Best Explanation*. Routledge & Kegan Paul, London.

Michalski, R. S. (1993). Inferential theory of learning as a conceptual basis for multistrategy learning. *Machine Learning*, 11:111–151.

Peirce, C. S. (1957). *Essays in the Philosophy of Science*. Liberal Arts Press.

Salmon, W. C. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton University Press, Princeton.

Salmon, W. C. (1990). *Four Decades of Scientific Explanation*. University of Minnesota Press, Minneapolis.

Shanahan, M. (1989). Prediction is deduction, but explanation is abduction. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 1055–1060, Detroit, MI.

Thagard, P. R. (1981). Peirce on hypothesis and abduction. In *C.S. Peirce Bicentennial International Congress*. Texas University Press.

Thagard, P. R. and Shelley, C. (1997). Abductive reasoning: Logic, visual thinking and coherence. In Chiara, M. D., editor, *Logic and Scientific Methods*, pages 413–427. Kluwer.