Distinguishing Abduction and Induction under Intensional Complexity

José Hernández-Orallo[1] *and* **Ismael García-Varea**[2]

Abstract: This paper presents a theoretical and general differentiation among descriptional induction, explanatory induction and abduction. Descriptional induction is based on the idea of compression (justified by mean- or cross-validation). Explanatory induction is characterised by a 'balanced' compression (exception-free validation). Finally, abduction is the more elusive notion, where the validation comes from a background theory. Since this background theory can also be used in both kinds of induction, we must distinguish between an auxiliary use and a necessary or 'consilient' use of the background knowledge.
We introduce many new concepts and formalisations for this goal, mainly the idea of 'intrinsic exception or anomaly', consilience and an operative measure of reinforcement for logic programs. Finally, the difference between induction and abduction is seen in the context of growth of knowledge and theory revision.
Keywords: Abduction, Induction, Explanation, Compression, Reinforcement, Kolmogorov Complexity, Intensional Complexity, Consilient / Coherent Theories, Philosophy of Science, ALP, ILP, EBL.

1 Introduction

Abduction (Sherlock Holmes' intelligence [Josephson & Josephson 1994]) is a kind of hypothetical inference introduced by Sanders Peirce (1839-1914) because, in his opinion, neither deduction nor induction, alone or combined, could unveil the internal structure of *meaning* [Yu 1994].

Although we will get back on the problem of meaning in the last section, nowadays, abduction is usually considered as a special kind of induction or, at most, both are seen as different kinds of hypothetical inferences, as [Michalski 1987] points out: "*inductive inference was defined as a process of generating descriptions that imply original facts in the context of background knowledge. Such a general definition includes inductive generalisation and abduction as special cases*".

More concretely, it is usually accepted that *abduction* is a mechanism for *completing* knowledge about a certain individual (generally inventing a fact to fit with a theory that is given), thus *explaining why the given observations* were not predicted by the initial knowledge. On the contrary, *induction* tries to *extend* knowledge (or to make a new theory) for *predicting future observations*.

In our view, the difference may be more of nature than of purpose: induction works without constraint (although an auxiliary background theory can be used) whereas abduction tries to find a hypothesis that is 'compliant' with some higher law that constrains how hypotheses can be. In this way, abduction may be seen as induction in a fixed context, closer to Peirce's original postulate [Flach 1996]:

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.

This "matter of course" is usually represented as a background theory or common-sense theory T (known as paradigm in philosophy of science or a constraint bias in inductive learning). Accordingly, abduction can be represented as usually:

$$A \cup T \models C$$

but with the additional condition that A cannot be an anomaly in the context of T and it cannot be an invention either (a fantastic but possible assumption). Since many A's could be found, some selection criteria must be chosen in order to find the most appropriate one. Two criteria are generally used: simplicity and consilience. Simplicity means that a very large assumption does not seem a good explanation. On the other hand, consilience means that A and C have to be *consilient* with T (or in misspelled words, T and C must be 'consiliated' by A), for turning the anomaly into a matter of course.

Still, there is another kind of induction that is known as "explanatory induction". Moreover, abduction often has been characterised as the inference to the best explanation [Harman 1965], where an explanation distinguishes from an enumerative induction [Ernis 1968] using some coherence criteria (or metrics [Ng & Mooney 1990]). Given an observation, in the absence of noise, an explanation must give the causes for the whole observation. For instance, if we have seen smoke and fire co-occurring 999 times over 1000 times, we can describe this observation as "P(smoke-fire) = 0.999" and we have a reliable prediction. However, no explanation is given, mainly because there is not any underlying mechanism justifying the co-occurrence nor the anomaly.

It is clear that an explanation must have some degree of plausibility to avoid fantastic hypotheses, but in many applications, like scientific discovery or abduction, we must regard an explanation as an investment, even a "risky bet" that could be soon falsified. This is merely Popper's criterion of falsifiability. He argued [Popper 1962] that one does not always want the most likely explanation, because it is also the less informative.

This dilemma between informative and probable hypotheses pervades most of discussion about the nature of abduction, too. The usual equation $P(h) = 2^{-I(h)}$, being P the probability of h and $I(h)$ the information of h makes the idea of information rather counterintuitive, because any redundancy (even useful) makes a theory less probable. This paper shows that it is possible to look for a good compromise between information and probability of an explanation.

Finally, there is another trait of abduction related with causation. Different frameworks for formalising causation soon appeared with the early expert diagnosis systems, exemplified by causal networks, especially Peng and Reggia's *causal abductive network* [Peng & Reggia 1987], along with other probabilistic or possibilistic frameworks like Pearl's Causal Theories, using a Bayesian belief network [Pearl 1988], [Pearl 1993]. In this context, an explanation consisting of an only cause for all the data is frequently preferable over separate causes for co-occurring phenomena, following Reichenbach's *principle of common cause* [Reichenbach 1956]. We will introduce different variants of the notion of consilience to show how and when this common cause may be ficticious.

[1] Universitat Politècnica de València, Departament de Sistemes Informàtics i Computació, Camí de Vera 14, Aptat. 22.012 E-46071, València, Spain. E-mail: jorallo@dsic.upv.es. On-line papers: http://www.dsic.upv.es/~jorallo/escrits/escritsa.htm.
[2] Universitat Politècnica de València, Institut Tecnològic d'Informàtica, Camí de Vera 14, E-46071, València, Spain. E-mail: ivarea@iti.upv.es

The paper is organised as follows. Section 2 presents intensional complexity and explanatory complexity, two information-theoretic concepts derived from Kolmogorov Complexity, which avoid 'intrinsic anomalies' in the hypotheses. These will be used in section 3 to characterise descriptive induction and explanatory induction as compression and 'comprehension', seen the latter in terms of explanatory complexity.

Sections 4 translates these ideas to logical theories and compares them with other criteria in explanatory induction. Section 5 addresses abduction, comparing consilience with simplicity and generality. Once established what abduction is, section 6 introduces the necessity of weights, costs or probabilities to select from all the possible abductions. Section 7 integrates induction and abduction in the context of growth of knowledge and theory revision under the notion of reinforcement. Section 8 relates abduction with meaning and intelligence. Section 9 closes the paper discussing the results and the open questions.

2 Intensional Complexity

We have mentioned that the purpose of an explanation is to turn anomalies into "matter of course". So we need first to discern what an anomaly (or exception) is. There is only a theory where we can characterise objectively and generically what an anomaly is: algorithmic complexity.

Algorithmic Complexity, Descriptional Complexity or, commonly, Kolmogorov Complexity are different names for a simple concept. The algorithmic complexity of a string x for a given machine ϕ is defined as the length in bits of the shortest program p in ϕ which outputs x. By the Church-Turing thesis, a Turing machine can emulate any computable function. This, this length depends only on the string x to be described and not on the descriptional machine ϕ, up to a constant denoted by $O(1)$. If an arbitrary Turing machine or other computable mechanism is fixed, algorithmic complexity is simply denoted by $C(x)$. In the following, p_x denotes any program for x. If we use an arbitrary prefix-free Turing machine, Kolmogorov Complexity is denoted by $K(x)$ and it has some additional interesting properties. See [Li & Vitányi 1997] for accurate and detailed definitions.

In [Hernández-Orallo & García-Varea 1998] the formal notion of the "exception degree" of a description has been formalised for any descriptional mechanism, but, at the same time, it has been shown that, in general, it is only possible if space considerations are taken into account. In addition, it has been shown that it is only effective if time is considered.

Informally, "*an exception is something we can take apart from a description so leaving it much simpler with respect to the magnitude of the length of the elements removed*" [Hernández-Orallo & Minaya-Collado 1998]. More concretely, a description is exception-free if it does not exist a subdescription that produces almost all the data, i.e., there is not a reduction in the description that could be greater that the corresponding reduction in the described data.

The description p_x for the data x is c-exception-free (denoted $\Delta_c(p_x) = 0$) iff there does not exist a subprogram p_y of p_x such that $K(p_x) - K(p_y) \geq [K(x) - K(y)] / c$. Note that in the case it exists, $p_x - p_y$ is the exception. The parameter c can be tuned depending on the deductive framework and the approximation for computing K, usually computing the length instead. In the following, c is assumed to be 1.

Obviously, a formalisation of subprogram is necessary in the deductive framework which would be chosen. As we will see, in the case of logical theories, this question is trivial but, in other cases, it can be very arduous.

EXAMPLE 2.1:

Consider the facts $F = \{ f_1, f_2, \ldots, f_{10} \}$ and a theory $Ta = \{ t_1, t_2 \}$ that covers these facts in the following way: t_1 covers f_1 to f_9 and, separately, t_2 covers f_{10}. Since t_1 and t_2 are separable, we can check the condition simply as $K(t_1) \geq K(f_{10})$. If it is the case, we say that f_{10} is an exception wrt. to Ta. In contrast, we may find a theory $T_b = \{t_1, t_2, t_3\}$ longer than T_a that covers the facts in the following way t_1 covers f_1 to f_4, t_2 covers f_5, f_6, f_7 and t_3 covers f_8, f_9, f_{10}. It is said that this theory is 'balanced' if $K(t_1) \approx K(t_2) \approx K(t_3)$. Finally, we can consider another theory $Tc = \{ t_1 \}$ longer than Tb which is not only balanced, but t_1 cannot be split up to cover separately subsets of F. That is to say, Tc consiliates F.

Later we will give formal characterisations of a consilient logic program and exception-free logic program. For the moment, we can give a general measure of the quality of descriptions, avoiding extensional parts, forcing all the theory to describe the data in an *intensional* way:

DEFINITION 2.1

The *Intensional Complexity* of a string x on a bias β, denoted $E\beta(x)$, is defined as follows:

$$E\beta(x) = \min \{ l\beta(p_x) : \Delta(p_x) = 0 \}$$

i.e., the shortest program for x without intrinsic exceptions. $l\beta(p_x)$ denotes the length of p_x in β.

There can be short intensional descriptions whose computational cost would be so high that they are of little use as theories. In addition, definition 2.1 turns out to be non-computable (like $K(x)$). In [Hernández-Orallo & Minaya-Collado 1998] is introduced an explanatory variant of intensional complexity which is defined in the following way:

DEFINITION 2.2

The *Explanatory Complexity* of a string x on a bias β, denoted $Et\beta(x)$, is defined as follows:

$$Et\beta(x) = \min \{ LT\beta(p) : \Delta(p) = 0 \}$$

$LT\beta(p)$ is chosen $l\beta(p)$ + log cost $\beta(p)$ —the same weighing as Levin's Kt— because it provides a good compromise between space and computational time [Levin 1973], but another parameterised relation could be tuned.

There are good reasons to choose a time-weighted definition of the best explanation. The intuitive view of explanation entails that the hypothesis can be *explained* to others. At the moment a system has to *tell* or communicate the explanation to other system (or internally work with it), there are two important topics: the space of the discourse and the time the system will need to relate it. Moreover, people and Science expect that nature has underlying mechanisms that emerge 'quickly' in our observations, simply because nature is not a reliable computer for executing long programs.

3 Descriptive vs. Explanatory Induction

The principle of simplicity, represented by Occam's razor, selects the shortest hypothesis as the most plausible one. This principle was rejected by Karl Popper because, in his opinion (and at that moment) there *was* no objective criterion for simplicity. However, Kolmogorov complexity *K(x)* is an objective criterion for simplicity. This is precisely what R.J.Solomonoff proposed as a 'perfect' theory of induction [Li & Vitányi 1997]. Algorithmic Complexity inspired J. Rissanen in 1978 to use it as a general modelling method, giving the popular MDL principle [Rissanen 1978], recently revised as a one-part code [Rissanen 1996] instead of two-part codes.

It is remarkable (and often forgotten) than Kolmogorov Complexity just gives consistency to this theory of induction,

but Occam's razor is *assumed*[3] but not proven. Nonetheless, some justifications have been given in the context of physics, reliability and entropy, but, in our opinion, it is the notion of *reinforcement* (or cross validation) which justifies the MDL principle more naturally. In general, it *seems* that the higher the mean compression ratio the higher the mean reinforcement ratio.

The problem of the MDL principle for explanation is that for the sake of maximum mean compression, some part of the hypothesis can be not compressed at all, resulting in a very compressed part plus some additional extensional cases. This extensional part is not validated, makin the whole theory weak.

Summing up, the MDL principle says that, in absence of any other knowledge about the hypotheses distribution, we should select the prior $P(h) = 2^{-K(h)}$. For explanatory induction we propose to use $P(h) = 2^{-Ex(h)}$ instead. This principle is known as the shortest explanatory description (SED). In this way, we give priority to the avoidance of extensionality over simplicity.

There are other approaches to finding intensional theories. Wexler claimed that the subset principle was an intensional principle [Wexler 1992], for the case of positive data only. The subset principle (also known as Least General Generalisation (lgg) by Plotkin [Plotkin 1970]) means that if two theories explain some positive data, we should select the more specific one, because it is the more informative (and the more falsifiable). The problem of the subset principle is that it must be combined with some simplicity criterion, because, if not, the more specific hypothesis is the data themselves, which is completely extensional.

4 Exception-Free Logic Programs

The preceding definitions and ideas are general enough to be adapted to any inductive framework, in order to distinguish between descriptive induction and explanatory induction. Nevertheless, this generality renders difficult the comparison with other works and it cannot be made operative easily. Particularly, we will adapt the previous notions to first order logic, providing a means to identify explanatory hypothesis in Inductive Logic Programming (ILP) [Muggleton 1991]. Furthermore, this concretion for logic theories will make possible to address the problem of understanding what exactly abduction is, which has been more generally studied in a logical framework.

Although the notion of subprogram is easy, we will try to refine the notion of partition before.

DEFINITION 4.1

Consider a program P as a set of Horn clauses with its minimal Herbrand model $M^+(P)$ equal to the set of ground literals L_i such that $P \models L_i$.

P is n-separable in the partition of *different* programs $\Pi = \{ P_1, P_2, \dots , P_n \}$ iff

$$M^+(P) = \bigcup_{i=1..n} M^+(P_i) \text{ and}$$

$$\forall_{i=1..n} \left(M^+(P_i) \neq \varnothing \right)$$

DEFINITION 4.2

P is *non-subset* n-separable in the partition $\Pi = \{ P_1, P_2, \dots , P_n \}$ iff it is n-separable into Π and

$$\forall_{i, j=1..n} \left(P_i \subseteq P_j \text{ implies } i = j \right).$$

[3] Furthermore, in the case the universal distribution $2^{-K(x)}$ is assumed, giving a priori predilection of short programs, the a posteriori optimality of the MDL principle is proven, supposing the *randomness of the hypothesis to the data* [Vitányi & Li 1997]. But precisely in explanatory prediction, if the hypothesis is random to the data, it cannot be the cause!

The existence of a non-subset 2-separation can be regarded as a condition to avoid exceptions. However, this exception-free condition would be so strict that it would ban any *modularity* in the program.

DEFINITION 4.3

P is *disjoint* n-separable in the partition $\Pi = \{ P_1, P_2, \dots , P_n \}$ iff it is n-separable into Π and

$$\forall_{i, j=1..n} \left(P_i \cap P_j = \varnothing \right)$$

DEFINITION 4.4

P is *non-subset model n-separable* in the partition $\Pi = \{ P_1, P_2, \dots , P_n \}$ iff it is n-separable into Π and

$$\forall_{i, j=1..n} \left(M^+(P_i) \subseteq M^+(P_j) \text{ implies } i = j \right).$$

DEFINITION 4.5

P is *disjoint model n-separable* in the partition $\Pi = \{ P_1, P_2, \dots , P_n \}$ iff it is n-separable into Π and

$$\forall_{i, j=1..n} \left(M^+(P_i) \cap M^+(P_j) = \varnothing \right)$$

To show they differ, we give some examples:

EXAMPLE 4.1

Given the following program P_1= { p(a). q(X) :- r(X). r(a). } it is separable for all the definitions we have given in the partition Π = {{p(a)} , {q(X) :- r(X). r(a)}}.

The program P_2= { q(X) :- r(X). r(a). } is not separable for any of the definitions we have given.

The program P_3= { q(X) :- r(X). p(X) :- r(X). r(a). } is non-subset (model) separable into Π = {{ q(X) :- r(X). r(a)}, {p(X) :- r(X). r(a). }} but it is not disjoint (model) separable.

The program P_4= { q(a). p(X) :- q(X). p(a) } is non-subset (model) and disjoint separable into Π = {{ q(a). p(X) :- q(X). }, {p(a)}} but it is not disjoint model separable.

The program P_5= { s(X):- p(X), q(b). p(X) :- q(X). t(X):-p(X),q(a) } is non-subset (model) and disjoint separable model into Π = {{ s(X) :- p(X), q(b). p(X) :- q(X) }, { p(X) :- q(X), t(X) :- p(X), q(a) } but it is not disjoint separable.

Moreover, it is trivial to show the following theorems:

THEOREM 4.1

If a program P is *disjoint separable* then it is *non-subset separable*.

THEOREM 4.2

If a program P is *disjoint model separable* then it is *non-subset model separable*.

At this point, different notions of exception can be given by using definitions 4.1 (single partition), 4.2 (non-subset partition), 4.3 (disjoint partition), 4.4 (non-subset model partition), 4.5 (disjoint model partition) that we dub *modes*.

We translate the informal definition we have given: "an exception is something we can take apart from a program so leaving the program much simpler with respect to the magnitude of the length of the elements removed" to Horn logic programs.

DEFINITION 4.6

A program P has $e = \text{card}(M^+(P_E))$ c-exceptions, denoted $\Delta_c(P) = e$, generated from P_E, iff there is a partition $P = \{ P_R, P_E \}$ such that:

$$l(P) - l(P_R) \geq \left[l(M^+(P)) - l(M^+(P_R)) \right] / c$$

Definition 4.6 means that what is reduced in the length (l) of the program is greater than what is reduced in the consequences, but it would be slightly different depending on which of definitions 4.1-4.5 is used.

The greatest value of c that still makes a program exception-free (i.e., $\Delta_c(P) = 0$) is known as its *consilience* level. On the other hand, when not indicated it is assumed to be 1, and we say that a program is exception-free. Finally, there are many ways to estimate the length of logic programs $l(P)$, but, customarily, a syntactical measure is used.

Let us illustrate the difference between explanatory induction and descriptional induction in an example:

EXAMPLE 4.2

Given the facts F = { even(0). even(s(s(0)). even(s(s(s(0)))), ¬even(s(0)) } the following programs can be induced:

P_1 = { even(0). even(s(s(X)) }, which is the shortest one but it is separable in all cases and *even*(0) is an exception.

P_2 = { even(0). even(s(s(X)) :- even(X) }, which is not separable in any case and logically it has not any exceptions.

P'_1 = { even(0) :- *fant*. even(s(s(X)) :- *fant*. *fant*. }, which is non-subset (model) separable, but it is not disjoint (model) separable and it has exceptions for the two first modes.

The last program from example 4.2 shows that a 'fantastic' concept can make a program non-separable for some modes, *hiding* exceptions. It is easy to prove that any separable program in the disjoint modes can be extended to a non-separable program using a fantastic concept. We say that the concept is not fantastic (it is really consilient) when it must reduce the size of the consiliated part. This implies that it is impossible to make every program exception-free, i.e., intensional.

Although the non-subset mode alone is too strict and the disjoint mode easy to cheat, the *non-subset* mode *combined* with the value of c=1 for exceptions are appropriate to distinguish a consilient program for most applications. Indeed, different modes and values for c can be combined for various degrees of desired explanatory induction.

The main problem of the definition of exception-free is that it must be computed w.r.t. to the given data (facts), because all the possible consequences can be infinite. In the next section, we will use these definitions to address the problem of abduction.

It is outside of this paper to take into account the presence of noise, but a degree or ratio of exceptions could be fitted to the expected ratio ε making $\Delta_c(p) = \varepsilon$.

In conclusion, Whewell [Whewell 1847] coined the term *consilience* to comprise the relevant basics in scientific theories: prediction, explanation and unification of fields. In this sense, the previous notions around the idea of intensional/explanatory complexity present a formal view of Thagard's notion of "explanatory coherence" [Thagard 1978]. In his view, a hypothesis exhibits explanatory coherence with another if it is explained by the other, explains the other, is used with the other in explaining other propositions, or if both participate in analogous explanations[4].

5 Consilience and Abduction

Abduction in Logic Programming sometimes has been identified with different approaches of non-monotonic reasoning, where new hypothetical facts (assumptions) are introduced in the way that the 'anomalies' and 'novelties' are explained as consequences of the revised or extended model of the program. The purest approach in this line is known as ALP (Abductive Logic Programming) [Kakas et al. 1993].

[4] It is important to note that this notion differs with Thagard's current works on coherence (see [Thagard 1998]), seen just as constraint satisfaction or maximization.

When abduction is seen in a qualitative manner (without costs, weights or probabilities), different selection criteria are advocated, depending on the dilemma between informative explanation and probable explanation.

One criterion, which is generally accepted, is subset minimality [Konolige 1991], [Bylander et al. 1991], [Poole 1985]: given two explanations E_1 and E_2 which explain the same fact, where $E_1 \subset E_2$, we should select the shortest one.

Things do not stand so clear in the case of specificity vs. generality. The Most Specific Abduction [Stickel 1990] is the interpretation into abduction of Popper's criterion of falsifiability. It has been used particularly well in diagnostic tasks, because it restricts the possible worlds (we look for an informative reason for the failure). Probability is not so crucial here because it can be checked revising the piece that is supposedly malfunctioning.

Stickel also presents the contrary one, the least specific abduction, useful in natural language interpretation. Stickel argues that many times the interpretation of an observation is just the observation itself, without any further intention.

Very related with the previous one, the Least Presumptive Explanation [Poole 1989] selects the hypothesis that makes less assumptions, and since it leaves more worlds open, model-theoretically, it is the most likely explanation. If not restricted, the least presumptive explanation coincides with the least specific abduction, completely extensional.

In other cases, the least presumptive explanation is estimated by the length or number of assumptions instead of the model simplicity [Ng & Mooney 1990]. In this case, we have the MDL principle applied to the abducibles (or possible hypotheses). Let us study the use of the MDL principle to the abducibles and to the whole model.

Given a theory T and a fact C, if we are looking for an A such that it explains C under T, i.e. $A \cup T \models C$, in many cases, the abducible A which minimises $K(A \cup T \cup C)$ or $K(A)$ is just A=C, which is extremely probable but explains nothing.

In response of this, we formulate the central argument for our position: *consilience is more important than whole simplicity for abduction*.

Formally, the *optimal* A is the one that minimises $Et(A \cup T \cup C)$, i.e., the shortest hypothesis without exceptions. For those cases where the theory has intrinsic exceptions previously, we will select P_R of definition 4.6 as T.

Consider simply the following example:

EXAMPLE 5.1

Given the program T = { p. lawn-wet :- rain. lawn-wet :-sprinkler-on } and the observation C = { lawn-wet } we have that A_1= C , A_2 = { rain }, A_3 = { sprinkler-on } are the shortest ones. But A_1 is an exception because $l(A_1 \cup T) - l(\mathrm{T}) = l(A_1) \geq l(M^+(T) + C)) - l(M^+(T)) = l(C)$. Contrarily, for A_2 (and A_3) we have $l(A_2 \cup T) - l(T) = l(A_2) < l(\mathrm{M}^+(T) + C + A_2) - l(M^+(T)) = l(C + A_2)$.

Even more, for an A_4 like { lawn-net :- p } we have that $l(A_4) \geq l(M^+(T) + C)) - l(M^+(T)) = l(C)$, so it is not a valid explanation.

It is frequently assumed in the abduction literature (see e.g. [Aliseda 1996]) that the additional condition $A \not\models C$ for explanation, i.e., the anomaly must not imply the observation *alone*, is sufficient to characterise abduction. Example A_4 shows that this idea has important flaws: A_1 is discarded as a valid abduction, but A_4 is not. The usual solution to this problem is to characterise abduction in a modal way or to restrict syntactically the abducibles to facts only.

These all definitions, conveniently adapted, would allow a different way for integration abduction and explanation in ILP

(in a different manner from the so-dubbed AILP (Abductive Inductive Logic Programming) [Adé & Denecker 1994]).

6 Quantitative Abduction

In the preceding section we have dealt with a characterisation of abduction, but, among all possible assumptions, we must still select the 'best one' without falling into the most probable nor the most informative.

The following example from [Poole 89] shows that we cannot do much about selection when two assumptions have isomorphic derivations.

EXAMPLE 6.1

T = { intd-in-hardware → intd-in-logic ∧ intd-in-CS.
intd-in-formal-AI → intd-in-logic ∧ intd-in-CS.
intd-in-logic → borrows-logic-books.
intd-in-CS → writes-computer-programs. }

Given the observation C = { borrows-logic-books ∧ writes-computer-programs } we have the following hypotheses satisfying $A \cup T \models C$:

$A_1 = C$ is the least specific hypothesis but it is an exception.

A_2= { intd-in-logic ∧ intd-in-CS } is the least presumptive explanation, it is intensional but it is long[5].

A_3= { intd-in-hardware } and A_4= { intd-in-formal-AI } are intensional and short and both are the most specific.

A_5= { intd-in-hardware ∨ intd-in-formal-AI } is the most general but it is more presumptive than A_2. Besides, it is intensional.

In this example, our criterion would select either A_3 or A_4, but we have not any ground to select the most plausible one from them.

Commonly, as [Leake 1995] remarks, the "best" explanation is based on 'probabilities' or 'costs' of the assumptions. These costs can be assigned to a selected list of abducibles (what Stickel calls predicate specific abduction [Stickel 1990]), to the rules of the theory or to both.

There are two main approaches, based on weights or probabilities. The advantage of using weights is that there is more freedom of how to distribute these weights for both abducibles and theory. In addition, there are many different ways of how to operate with the weights (see e.g. Stickel's chained specific abduction [Stickel 1990]). On the other hand, Poole's approach based on probabilities [Poole 1989] must establish too many restrictions about independence of the abducibles to make consistent the computation of probabilities.

The problem of exactly establishing the abducibles and their costs as well as the theory results in an inversion of the problem of abduction into a non-monotonic deductive problem.

In all these cases, the following questions arises: who chooses the assumptions? [Poole 1997], and who assigns the probabilities?

To answer these questions, in our opinion, a coherent account for induction and abduction must be understood in the context of theory construction, growth and revision.

[5] Anyhow, we can convert T into T'
T' = { interested-in-hardware → p.
interested-in-formal-AI → p.
p → interested-in-logic ∧ interested-in-CS.
interested-in-logic → borrows-logic-books.
interested-in-CS → writes-computer-programs. }
and p would be selected along with A_3 and A_4.

7 Knowledge Acquisition and Revision

Abduction has been usually seen as belief revision [Boutilier & Becher 1995], usually combined with induction [Aliseda 1996]. A theory T is constructed as the data suggest and, when new observations C are received, we can have three possible situations:

- Prediction Hit. The observations are covered without more assumptions, i.e., $T \models C$. The theory is reinforced.
- Novelty [Aliseda 1996]. The observation is uncovered but consistent with the theory T , i.e., $T \not\models C$ and $T \cup C \not\models \square$. Here, the possible actions are: (1) T can be extended with a good explanation, (2) revised if a good explanation cannot be found, (3) left it as an exception or (4) ignored.
- Anomaly [Aliseda 1996]. The observation is inconsistent with the theory T, i.e., $T \not\models C$ and $T \cup C \models \square$. In this case, T can be revised if a good explanation cannot be found, left it as an exception or ignored.

A non-explanatory approach to theory formation is Kuhn's theory of changing paradigms. According to the MDL principle, when too many exceptions to the paradigm are found, they are difficult to quote and the whole paradigm (or part of it) must be changed.

In a different way, explanatory knowledge construction must minimise the exceptions, so the revisions are much more frequent. Even more, the goal is anticipating instead of preserving the current knowledge, as many approaches to minimal revisions aim for [Mooney 1997], supported by the obvious fact that a minimal revision is usually less costly than making the whole theory from scratch.

Whatever the approach to knowledge construction (lazy or eager), the revision of knowledge must come from a loss of *reinforcement* (or apportionment of credit [Holland et al. 1986]). We present a way to compute the reinforcement map for a given theory, depending on past observations.

DEFINITION 7.1

The pure reinforcement $\rho\rho(r)$ of a rule r of a theory T wrt. to some given observation $C = \{c_1, c_2, \ldots, c_n\}$ is computed as the number of proofs of c_i where r is used. If there are more than one proof for a given c_i, *all* of them are reckoned, but in the same proof, a rule is computed only once.

DEFINITION 7.2

The (normalised) reinforcement $\rho(r) = 1 - 2^{-\rho\rho(r)}$.

These definitions show that, in general, the most reinforced theory is not the shortest one. In addition, redundancy does not imply a loss of reinforcement ratio. However, the measure of reinforcement of the *theory* would present the same problems of fantastic concepts we have been discussing in section 4. Fortunately, the solution comes from measuring the validation *wrt. the data*.

DEFINITION 7.3

The *course* $\chi(f)$ of a given fact f wrt. to a theory is computed as the product of all the reinforcements $\rho(r)$ of all the rules r used in the proof of f. If a rule is used more than once, it is computed once. If f has more than one proof, we select the greatest course.

In this case, we can select the theory with the greatest *mean* of the courses of all the data presented so far. If we want a compensated theory, we can use a *geometric mean* instead. If we do not want exceptions we can discard theories where a fact has a course value less than the mean divided by a constant. The following example shows the use of these *new* criteria for knowledge construction:

EXAMPLE 7.1

Suppose we have an incremental learning session as follows:

♦ Given the following background theory B = { s(a,b), s(b,c), s(c,d) } we observe the evidence E = { e_1^+: r(a,b,c), e_2^+: r(b,c,d), e_3^+: r(a,c,d), e_1^-: ¬r(b,a,c), e_2^-: ¬r(c,a,c) } :

The following programs could be induced, with their corresponding reinforcements and courses:

P_1 = {r(X,Y,Z) :- s(Y,Z) : ρ = 0.875}
$\chi(e_1^+)= \chi(e_2^+)= \chi(e_3^+)= 0.875$

P_2 = {r(X,c,Z) : ρ = 0.75
r(a,Y,Z) : ρ = 0.75}
$\chi(e_1^+)= \chi(e_2^+)= \chi(e_3^+)= 0.75$

P_3 = {r(X,Y,Z) :- s(X,Y) : ρ = 0.75
r(X,Y,Z) :- s(Y,Z) : ρ = 0.875}
$\chi(e_1^+)= \chi(e_2^+)= \chi(e_3^+)= 0.875$

P_4 = {r(X,Y,Z) :- t(X,Y), t(Y,Z) : ρ = 0.875
t(X,Y) :- s(X,Y) : ρ = 0.875
t(X,Y) :- s(X,Z), t(Z,Y) : ρ = 0.5}
$\chi(e_1^+)= \chi(e_2^+)= 0.7656$, $\chi(e_3^+)= 0.3828$

P_5 = {r(X,Y,Z) :- t(X,Y) : ρ = 0.875
t(X,Y) :- s(X,Y) : ρ = 0.875
t(X,Y) :- s(X,Z), t(Z,Y) : ρ = 0.5}
$\chi(e_1^+)= \chi(e_2^+)= 0.7656$, $\chi(e_3^+)= 0.3828$

At this moment, P_1 and P_3 are the best options by far. For the moment, P_4 and P_5 seem fantastic theories according to the evidence

♦ e_4^+ = r(a,b,d) is observed.

P_1 does not cover e_4^+ and it is patched to:

P_{1a}' = {r(X,Y,Z) :- s(Y,Z) : ρ = 0.875
r(a,b,d) : ρ = 0.5}
$\chi(e_1^+)= \chi(e_2^+)= \chi(e_3^+) = 0.875$, $\chi(e_4^+) = 0.5$
Mean = 0.78, GeoMean = 0.76

P_{1b}' = {r(X,Y,Z) :- s(Y,Z) : ρ = 0.875
r(X,Y,d) : ρ = 0.875 }
$\chi(e_1^+)= \chi(e_2^+)= \chi(e_3^+)= \chi(e_4^+) = 0.875$

P_2' is reinforced = {r(X,c,Z) : ρ = 0.75.
r(a,Y,Z) : ρ = 0.875}
$\chi(e_1^+) = 0.875$, $\chi(e_2^+) = 0.75$, $\chi(e_3^+)= \chi(e_4^+) = 0.875$

P_3' is reinforced = {r(X,Y,Z) :- s(X,Y) : ρ = 0.875.
r(X,Y,Z) :- s(Y,Z) : ρ = 0.875}
$\chi(e_1^+)= \chi(e_2^+)= \chi(e_3^+)= \chi(e_4^+) = 0.875$

P_4' is reinforced.

P_4' = { r(X,Y,Z) :- t(X,Y), t(Y,Z) : ρ = 0.9375
t(X,Y) :- s(X,Y) : ρ = 0.9375
t(X,Y) :- s(X,Z), t(Z,Y) : ρ = 0.75}
$\chi(e_1^+)= \chi(e_2^+)= 0.8789$, $\chi(e_3^+)= \chi(e_4^+) = 0.6592$
Mean= 0.77, GeoMean = 0.76

P_5' is slightly reinforced

P_5' = { r(X,Y,Z) :- t(X,Y) : ρ = 0.9375.
t(X,Y) :- s(X,Y) : ρ = 0.875
t(X,Y) :- s(X,Z), t(Z,Y) : ρ = 0.5}

At this moment, P_{1b}' and P_3 are the best options. Now P_4 seems less fantastic.

♦ We add e_3^- = ¬r(a,d,d)

P_{1a}' remains the same and P_{1b}' and P_{2a}' are inconsistent. The following two theories could also be 'patches' for them:

P_{2a}' = {r(X,c,Z) : ρ = 0.75.
r(X,b,Z) : ρ = 0.75}
$\chi(e_1^+)= \chi(e_2^+)= \chi(e_3^+)= \chi(e_4^+) = 0.75$

P_{2b}' = {r(X,Y,Z) :- e(Y) : ρ = 0.9375.
e(b) : ρ = 0.75
e(c) : ρ = 0.75}
$\chi(e_1^+)= \chi(e_2^+)= \chi(e_3^+)= \chi(e_4^+) = 0.7031$

P_3' and P_4' remain the same and P_5' seem to be inconsistent.

♦ We add e_5^+ = r(a,d,e)

P_{1a}', P_{2a}', P_{2b}' can only be patched with e_5^+ as an exception and not abduction is possible.

P_3' has abduction as a better option.

P_3'' = {s(d,e) : ρ = 0.5
r(X,Y,Z) :- s(X,Y) : ρ = 0.875
r(X,Y,Z) :- s(Y,Z) : ρ = 0.9375}
$\chi(e_1^+)= \chi(e_2^+)= \chi(e_3^+)= 0.9375$,
$\chi(e_4^+) = 0.875$, $\chi(e_5^+) = 0.46875$
Mean= 0.831, GeoMean = 0.805

P_4' makes the same abduction

P_4'' = { s(d,e) : ρ = 0.5
r(X,Y,Z) :- t(X,Y), t(Y,Z) : ρ = 0.96875
t(X,Y) :- s(X,Y) : ρ = 0.96875
t(X,Y) :- s(X,Z), t(Z,Y) : ρ = 0.875}
$\chi(e_1^+)= \chi(e_2^+)= 0.9385$, $\chi(e_3^+)= \chi(e_4^+) = 0.8212$, $\chi(e_5^+) = 0.4106$
Mean= 0.786, GeoMean = 0.754

The example illustrates the advantages of explanatory induction (the shortest theories are not the best ones in general[6]). More importantly, it also shows that as soon as a theory gains some solidity, abduction can be applied.

The way reinforcements are reckoned makes that very complex programs are avoided, but redundancy is possible. In some way, computational complexity could also be restricted if a rule was computed more than once, when used more than once in a proof.

Furthermore, there is not any risk of fantastic concepts. Formally, for any program P composed of rules r_i of the form { h :- t_1, t_2, .. t_n }, which covers m examples E = { e_1, e_2, ... e_m } and their reinforcements ρ_i, a *fantastic* rule r_f can be added to the program and all the rules can be modified in the following way r_i = { h :- t_1, t_2, .. t_n , r_f}. The program *seems* consilient, but as the following theorem shows is not reinforced over the original one:

THEOREM 7.1

We cannot increase the course of any example by the use of *fantastic* concepts.

PROOF

Since the *fantastic* concept *rf* now appears in all the proofs of the m examples, the reinforcement of r_f is exactly $1 - 2^{-n}$ and the reinforcements of all the r_i remain the same. Hence, the course of all the m examples is modified to $\chi'(e_j) = \chi(e_j) \cdot r_f = \chi(e_j) - \chi(e_j) \cdot 2^{-n}$. From here, for all $e_j \in E$, $\chi'(e_j)$ never can be greater than $\chi(e_j)$.□

8 Abduction and Meaning

Briefly, meaning is definitely associated with intention, and the latter with an agent or source S which pursues a goal G with an action N. When one receives as perception some of the effects C of action N, one tries to discover the final cause G. This is also applicable to language and meaning. If the background theory T contains the description of the *mentality* of the source T_S (its intentions) and the 'usual' actions and environment of S (its behaviour T_{SB}), the discovering of the cause G can be translated into two chained abductions $N \cup T_{SB} \models C$ and then $G \cup T_S \models N$. Hence, abduction should be considered as the main mechanism for hermeneutics. In general, and in natural language interpretation, the chain of cause-effects must be stopped at an informative (or interesting) point but still with some reliability. This differenct in *interest* makes that [Stickel 1990] could consider the most specific and the least specific abductions for language interpretation.

Finally, and more rhetorically, this ability to *comprehend* (or looking for explanations) has been recognised as fundamental for intelligence and it is usually exploited by psychometrics when making IQ tests. There have been many propos-

[6] Although —in the limit [Gold 1967]— the MDL principle is an excellent principle for achieving reinforcement

als for an inherent characterisation of meaning directly related with intelligence [Hofstadter 1979]:

> *It would be nice if we could define intelligence in some other way than "that which gets the same meaning out of a sequence of symbols as we do". [...] This in turn would support the idea of meaning being an inherent property.*

In [Hernandez-Orallo & Minaya-Collado 1998] a "C-test" was developed from *k*-comprehensible strings using explanatory complexity. The correlation of abduction problems with IQ tests was 0.68 and the correlation with explanatory prediction was 0.77, whereas the correlation with compression was much lower. This suggests (cautiously since one psychometric study is not representative) that the ability of discovering explanations (either inductive or abductive) is at the core of the set of computational abilities that intelligence tests measure.

9 Discussion and Future Work

We have studied the nature of explanations under the view of intensional complexity, which is a formalisation of the idea of consilience or coherence of a theory, distinguishing three kinds of "non-deductive reasoning":

- Descriptional (or Enumerative [Ernis 1968]) Induction uses background knowledge as a help but it has no expectancy of the source to conciliate (and no restriction either), so a hypothesis is constructed as the data suggest, and in this case, the MDL principle can be used.
- Explanatory Induction sometimes looks for more informative theories instead of the most probable, regarding the process as an investment, something that is also very common in the case of analogical reasoning.
- Abduction can be defined as a biased explanatory induction, where the hypothesis (usually facts) must be a "matter of course" w.r.t. the background knowledge, which does not only ensure consistency but also consilience.

The distinction between the last two is subtle and sometimes only terminological. In fact, as long as our background theory grows and the language is more expressible[7], the assumptions to abduce are so small wrt. to the paradigm *T* that not only facts but complete theories could be abduced as cases from *T*. In other words, what makes a difference (and in our opinion a continuous one) between abduction and induction is the importance, solidity, constraint and size of the background theory *T* wrt. to the explanation *A* [Goebel 1997].

The most important critique to our approach is abduction in the presence of noise. We are working on induction under a known ratio of exceptions, intrinsically characterisable in the hypothesis using intensional complexity. In any case, for explanation, it is difficult to accept a very short hypothesis (with some minor unexplained anomalies) if we are able to find a much more complicated exception-free explanation. Think for instance that the noise surrounding Newton's physics was used by Einstein to develop a new theory.

Finally, Deduction must not be longer seen as a static process that does not bring any information and cannot be creative. Moreover, Induction and Abduction should not be seen as inverse processes of Deduction, *in terms of information gain*. Indeed, any computable induction and abduction must be done in a computer system, so it is deductive somehow[8]. Further work must be done to reconcile deduction, induction and abduction [Hernandez-Orallo 1998].

Summing up, the distinction between abduction and (explanatory) induction is more general than some syntactical, causal or modal approaches. We also think that the role of induction and abduction in knowledge acquisition is portable even from expert systems and diagnostic systems to neural networks (training = induction, recognition = abduction). However, a logical framework seems an extremely adequate tool to advance and combine different areas and applications: ILP, ALP, EBL, Analogical Reasoning, Reinforcement Learning and some kinds of non-monotonic reasoning. Our current work deals with evaluating in practice our intensional principles in inductive systems [Hernandez-Orallo & Ramirez-Quintana 1998].

As future work, we plan to relating reinforcement with explanatory induction in a more formal way (a very interesting work which is connected with this goal is [Dieterich & Flann 1997]. It would also be very attractive to include negative facts in the reinforcement measures and to study thoroughly the entanglements between the length of rules and reinforcements.

Acknowledgements

This work has been extremely benefited from the constructive comments of the anonymous referees, by suggesting the comparison with the works of Poole and Stickel. We also hope that the examples and the extended space had helped to clarify our position.

References

[Adé & Denecker 1994] "AILP: Abductive Inductive Logic Programming" Dep. of Computer Sci., K.U.Leuven, Belgium 1994.

[Aliseda 1996] Aliseda, A. "A Unified Framework for Abductive and Inductive Reasoning in Philosophy and AI" in M. Denecker, L. De Raedt, P. Flach and T. Kakas (eds) Working Notes of the ECAI'96 Workshop on *Abductive and Inductive Reasoning*, pp. 7-9, 1996.

[Barker 1957] Barker, S.F. *Induction and Hypothesis* Ithaca, 1957.

[Boutilier & Becher 1995] Botilier, C.; Becher, V. "Abduction as belief revision" Artificial Intelligence 77, 43-94, 1995.

[Bylander et al. 1991] Bylander, T.; Allemang, M.C.; Tanner, M.C.; Josephson, J.R. "The computational complexity of abduction" Artificial Intelligence, 49:25-60, 1991

[Dieterich & Flann 1997] Dietterich, T.G.; Flann, N.S. "Explanation-Based Learning and Reinforcement Learning: A Unified View" Machine Learning, 28, 169-210, 1997.

[Dimopoulous and Kakas 1996] Dimopoulos, Yannis; Kakas, Antoni Kakas "Learning Abductive Theories" W. Wahlster (ed.) *ECAI 96, 12th European Conference on AI.*, John Wiley & Sons Ltd. 1996.

[Ernis 1968] Ernis, R. "Enumerative Induction and Best Explanation" *The Journal of Philosophy*, LXV (18), 523-529, 1968.

[Flach 1996] Flach, Peter "Abduction and Induction: Syllogistic and Inferential Perspectives" in M. Denecker, L. De Raedt, P. Flach and T. Kakas (eds) Working Notes of the ECAI'96 Workshop on *Abductive and Inductive Reasoning*, pp. 7-9, 1996.

[Goebel 1997] Goebel, R.G. "Abduction and its relation to constrained induction" in Peter Flach and Antonis Kakas (eds), Proceedings of the IJCAI'97 Workshop on Abduction and Induction in AI, Nagoya, Japan 1997.

[Gold 1967] Gold, E.M. "Language Identification in the Limit" *Inform. and Control*., 10, pp. 447-474, 1967.

[Grünwald 1997] Grünwald, P. "The Minimum Description Length Principle and Non-Deductive Inference" in Peter Flach and Antonis Kakas (eds), Proceedings of the IJCAI'97 Workshop on Abduction and Induction in AI, Nagoya, Japan 1997.

[7] Think about higher-order logic, where a complete theory can be an abduction of another supertheory. Or even, complete theories can be deduced from 'megatheories', as it is usual in theoretical physics nowadays.

[8] The idea backs to the introduction in 1949 of the deductive-nomological model of explanation by Hempel and Oppenheim [Hempel 1965], which argued that abduction is just a selection of possible phenomena derived from very general laws (*nomos*)

[Harman 1965] Harman, G. “The inference to the best explanation” *Philosophical Review*, 74, 88-95, 1965.
[Hempel 1965] Hempel, C.G. “Aspects of Scientific Explanation” The Free Press, New York, N.Y. 1965.
[Hernández-Orallo 1998] Hernández-Orallo, J. “Explanatory Induction and Informative Deduction using Intensional Complexity” Thesis Dissertation, forthcoming, 1998.
[Hernández-Orallo & García-Varea 1998] Hernández-Orallo, J.; García-Varea, I. “On Autistic Interpretations of Occam’s Razor” TR, http://www.dsic.upv.es/~jorallo/escrits/autistic21.ps.gz, 1998.
[Hernández-Orallo & Minaya-Collado 1998] “A FDI based on an intensional variant of algorithmic complexity” Proc. of the Intnl. Symp. of Engin. of Intelligent Systems, EIS'98, ICSC Press 1998.
[Hernandez-Orallo & Ramirez-Quintana 1998] “Inductive Inference of Functional Logic Programs by Inverse Narrowing” J. Lloyd (ed) CompulogNet Area Meeting on Computational Logic and Machine Learning of the 1998 Joint International Conference and Symposium on Logic Programming (JICSLP’98), June 1998.
[Hofstadter 1979] Hofstadter, D.R. “Gödel, Escher, Bach: An eternal golden braid” New York: Basic Books, 1979.
[Holland et al. 1986] Holland, J.H.; Holyoak, K.J.; Nisbett, R.E.; Thagard, P.R. “INDUCTION, Processes of Inference, Learning and Discovery” The MIT Press 1986.
[Kakas et al. 1993] Kakas, A.C.; Kowalski, A.; Toni, F. “Abductive Logic Programming” *J. of Logic and Comp.*, 2 (6): 719-770, 1993.
[Kolmogorov 1965] Kolmogorov, A.N. “Three Approaches to the Quantitative Definition of Information” Problems Inform. Transmission, 1(1):1-7, 1965.
[Konolige 1991] Konolige, K. “Abduction versus closure in causal theories” Artificial Intelligence, 52:255-72, 1991.
[Kuhn 1970] Kuhn, T.S. “The Structure of Scientific Revolution”, University of Chigago 1970.
[Lavrac & Dzeroski 1997] Lavrac, Nada; Dzeroski, Saso (eds.) “Inductive Logic Programming. 7th International Workshop, ILP-97” Lecture Notes in Artificial Intelligence, Springer 1997
[Leake 1995] Leake, David B “Abduction, Experience, and Goals: A Model of Everyday Abductive Explanation” The Journal of Experimental and Theoretical Artificial Intelligence 1995.
[Li & Vitányi 1997] Li, M.; Vitányi, P. “An Introduction to Kolmogorov Complexity and its Applications” 2nd Ed. Springer-Verlag 1997.
[Michalski 1987] Michalski, R.S. "Concept Learning" in S.C. Shapiro (ed). "Encyclopedia of Artificial Intelligence" 185-194, John Wiley, Chicester, 1987.
[Mooney 1997] Mooney, R.J. “Integrating Abduction and Induction in Machine Learning” in Peter Flach and Antonis Kakas (eds), Proceedings of the IJCAI’97 Workshop on Abduction and Induction in AI, Nagoya, Japan 1997.
[Muggleton & De Raedt 1994] Muggleton, S. & De Raedt L. “Inductive Logic Programming — theory and methods” Journal of Logic Programming, 19-20:629-679, 1994.
[Ng & Mooney 1990] Ng.H.; Mooney, R. "On the role of coherence in abductive explanation" in *Proceedings of the Eighth National Conference on Artificial Intelligence*, pp. 337-342 Boston, MA. AAAI, 1990.
[O'Rorke 1989] O'Rorke, P. "Coherence and abduction" *The Behavioural and Brain Sciences*, 12 (3), 484, 1989.
[Pearl 1988] Pearl, J. "Probabilistic Reasoning in Intelligent Systems: Networks of plausible inference" San Mateo, CA, Morgan Kaufmann, 1988.
[Pearl 1993] Pearl, J. "Belief Networks Revisited" Artificial Intelligence (59), 45-56, 1993.
[Peirce 1867/1960] Peirce, C.S. "Collected papers of Charles Sanders Peirce" Cambridge. Harvard University Press 1960.
[Peng & Reggia 1987] Peng, Y.; Reggia, J.A. "Abductive Inference Models for Diagnostic Problem-Solving, Symbolic Computation Series. Springer-Verlag, 1987.
[Plotkin 1970] Plotkin G. “A note on inductive generalization” Machine Intelligence, Vol. 6, Edinburgh Univ. Press, 1970.
[Poole 1985] Poole, D. “On the Comparison of Theories: Preferring the Most Specific Explanation” in *IJCAI’85*, pages 144-147, 1985.
[Poole 1989] Poole, D. “Explanation and Prediction: An Architecutre for Default and Abductive Reasoning” Computational Intelligence 5(2), 97-110, 1989.
[Poole 1997] Poole, D. “Who chooses the assumptions?” in P. O’Rorke (ed) *Abduction*, AAAI/MIT Press, 1997.
[Popper 1962] Popper, K.R. “Conjectures and Refutations: The Growth of Scientific Knowledge” Basic Books, New York 1962.
[Reichenbach 1956] Reichenbach, H. "The Direction of Time" University of California Press, Berkeley and Los Angeles, 1956.
[Rissanen 1978] Rissanen, J. “Modelling by the shortest data description” Automatica-J.IFAC, 14:465-471, 1978.
[Rissanen 1996] Rissanen, J. “Fisher Information and Stochastic Complexity” IEEE Trans. Inf. Theory, 1(42): 40-47, 1996.
[Shapiro 1981] Shapiro, E. “Inductive Inference of Theories form Facts” Research Report 192, Department of Computer Science, Yale University, 1981.
[Solomonoff 1964] Solomonoff, R.J. “A formal theory of inductive inference” Inf. Control.. vol. 7, 1-22, Mar., 224-254, June 1964.
[Stickel 1990] Stickel, M. E. “Rationale and Methods for Abductive Reasoning in Natural-Language Interpretation” in R. Studer (ed.) “Natural Language and Logic” Lecture Notes in AI 459, pp. 233-252, Springer-Verlag 1990.
[Thagard 1978] Thagard, P. “The best explanation: Criteria for theory choice” Journal of Philosophy, 75, 76-92, 1978.
[Thagard 1989] Thagard, P. “Explanatory coherence” *The Behavioural and Brain Sciences*, 12 (3), 435-502, 1989.
[Thagard 1998] Thagard, P. “Probabilistic networks and explanatory coherence” in P. O’Rorke & J. Josephson (eds), Automated Abduction: Inference to the best explanation, Menlo Park, AAAI Press 1998.
[van den Bosch 1994] van den Bosh "Simplicity and Prediction" Master Thesis, department of Science, Logic & Epistemology of the Faculty of Philosophy at the University of Groningen, 1994.
[Vitányi & Li 1997] Vitányi, P.; Li, M. “On Prediction by Data Compression”, Proc. 9th European Conference on Machine Learning, Lecture Notes in Artificial Intelligence, Vol. 1224, Springer-Verlag, Heidelberg, 14-30, 1997.
[Wexler 1992] Wexler, K. “The Subset principle is an intensional principle” in *Knowledge and Language: Issues and Representation and Acquisition* (E. Reuland and W. Abrahamson, eds.), Kluwer Academic Publishers, 1992
[Whewell 1847] Whewell W. “The Philosophy of the Inductive Sciences” New York: Johnson Reprint Corp. 1847.
[Yu 1994] Yu, C. H. "Abduction? Deduction? Induction? Is there a Logic of Exploratory Data Analysis?" Annual Meeting of American Educational Research Associations, New Orleans, 1994.