Philosophica 61 (1998, 1) pp. 9-17

ABDUCTION-PREDICTION MODEL OF SCIENTIFIC INFERENCE REFLECTED IN A PROTOTYPE SYSTEM FOR MODEL-BASED DIAGNOSIS

John R. Josephson

ABSTRACT

This paper describes in some detail a pattern of justification which seems to be part of common sense logic and also part of the logic of scientific investigations. Calling this pattern "abduction," the paper lays out an "abduction-prediction" model of scientific inference as an update to the traditional hypothetico-deductive model. According to this newer model, scientific theories receive their claims for acceptance and belief from the abductive arguments that support them, and the processes of scientific discovery aim to develop theories with strong abductive support. It is suggested that the study of diagnosis presents a good opportunity for studying abduction under somewhat simpler and more reproducible conditions than occur in scientific discovery. A computer-based diagnostic system is described which provides a small-scale validation of the abduction-prediction model by showing that a version of it can be made precise enough to be implemented and to perform correctly for diagnosis.

1. Introduction - abductive inference

Consider the pattern of reasoning given by:

D is a collection of data (facts, observations, givens),
Hypothesis H explains D (would, if true, explain D),
No other hypothesis explains D as well as H does.

Therefore, H is probably correct.

This pattern is very similar to what Lycan (1988) has called "the explana-

tory inference," what Harman (1965) and Lipton (1991) have called "inference to the best explanation," what Hanson (1958) and Peirce (1899) have called "retroduction," and what Peirce has called "abduction." I will use the term "abduction" in this paper, but whatever it is called, and however it is formalized, I hope my readers recognize it as distinctive and familiar, and as having a kind of intuitively recognizable evidential force. In fact, we can readily observe that people very commonly justify their conclusions by direct or barely disguised appeal to this pattern, which shows that speaker and hearer share a common understanding of it. Thus, abduction seems to be part of "commonsense logic" as an ordinary and presupposed structure for justifying conclusions, both to others, and, presumably, to ourselves.

It will be useful to distinguish abduction as a pattern of justification from abduction as a reasoning process. In a process of trying to explain some experience, or pattern of experiences, the object is to arrive at an explanation that can be confidently accepted. An explanation that can be confidently accepted is an explanation that can be justified as being "the best explanation" in consideration of various factors, and in contrast with alternative explanations. Thus, an explanation-seeking process – an "abductive reasoning process" – aims to arrive at a conclusion that has strong "abductive justification."

To reason abductively, to seek explanations, an agent must process information in various ways. It must: focus attention on data needing explanation, generate explanatory hypotheses, evaluate hypotheses, compare hypotheses, and decide whether to accept a hypothesis as being sufficiently justified. These can be thought of as the characteristic subgoals, subfunctions, and subprocesses of abductive reasoning, which must be accomplished successfully for abductive reasoning to be successful. The term "abduction" has often been used for the hypothesis-generation part alone, although we will find it convenient to use the term for the whole process of reasoning to the acceptance of explanations.

Our intuitions probably agree that the strength of an abductive justification depends on:

- how decisively H surpasses the alternatives.
- how good H is by itself, independently of considering the alternatives,
- confidence in the accuracy of the data, and
- how thorough was the search for alternative explanations.

As far as I know this list of considerations is complete. Let us consider

each one in a bit more detail.

How decisively H surpasses the alternatives. Negative evidence against a hypothesis becomes positive evidence for rival hypotheses. Thus, the rejection, or negative evaluation, of rival hypotheses is important for justifying an abductive conclusion. One way that a hypothesis may acquire a negative evaluation is by authorizing a prediction that fails to be true.

How good H is by itself, independently of considering the alternatives. Even the best explanation available is not strongly supported by the evidence if it has weaknesses in itself, for example if it authorizes failed predictions or is excessively complicated or inconsistent.

Confidence in the accuracy of the data. Support for H is weak if support for the propositions expressing the data is weak. This consideration does not need to be listed separately if we agree that hypotheses such as noisy data, error, accidental correlation, and the like, count as possible explanations, and thus they undercut confidence in other explanations.

How thorough was the search for alternative explanations. We can undercut an abductive argument by raising a plausible alternative explanation. This is very common. (In fact, all four criteria point to corresponding blocking moves. See our "Diagnosis and abductive justification" in Josephson & Josephson [1994] pp. 9-12.) This particular criterion shows most clearly why the process of discovery is a logical matter, and why logic cannot simply be confined to matters of justification. The strength of an abductive justification depends, in part, on evaluating the quality of the search for alternative explanations, and in so doing, on evaluating characteristics of the discovery process.

Besides the strength of an abductive argument, acceptance of the conclusion generally also depends on *pragmatic considerations*, such as the cost of error, and on how strong the need is to come to a conclusion at all, especially considering the possibility of gathering further evidence before deciding.

Abductive arguments are, of course, *fallible*, and a conclusion might be false, even if the premises are true, unlike deductions. Yet, sometimes abductive justification is strong, and its conclusion has a strong claim to acceptance and belief. A broad search for possible explanations for solid data might turn up a good explanation which decisively surpasses alternative explanations. In such a case, the conclusion is strongly supported, even thought there remain particles of doubt as to whether the search has

been broad enough, whether the leading hypothesis might yet authorize false predictions, etc.

2. Abduction-Prediction Model

What we may call "the abduction-prediction model" or "A-P model" of scientific inference holds that:

- Theory formation, evaluation, and acceptance in science are inferentially well characterized as abduction (inference to the best explanation). Scientific theories receive their claims for acceptance and belief from the abductive arguments that support them, and the processes of scientific discovery aim to develop theories with strong abductive support.
- Alternative explanatory hypotheses are evaluated, in part, based on the success of their predictions.
- A hypothesis gains strength from the weakness of rivals. A hypothesis may also have strengths and weaknesses independent of rivals, which come from its evaluation according to such criteria as: explanatory power, predictive success, consistency, simplicity, precision, coherence with background knowledge, etc. These criteria may also be applied comparatively.

The A-P model is thus very similar to the traditional hypothetico-deductive (H-D) model except that the A-P model:

- requires hypotheses to be explanatory,
- denies that predictions are always deductive, and
- emphasizes the rejection of rival hypotheses.

The claim that predictions are not necessarily deductive is not important for the purposes of this paper, and will not be argued here (but see my "Conceptual analysis of abduction" in Josephson & Josephson [1994]).

3. Diagnosis

Besides scientific discovery, abduction appears to be a useful way to characterize a variety of information-processing tasks including natural language understanding, judgments of guilt or innocence in law cases, inferring goals from behavior, and diagnosis. The study of diagnosis, in particular, presents a good opportunity for studying abduction under somewhat simpler and more reproducible conditions than occur in scientific discovery, and so diagnosis can serve as a kind of "laboratory model" of scientific discovery. Diagnosis is simpler than scientific discovery in being usually more concrete, working in a closed conceptual world, generating a theory of the individual case rather than general theory, and in requiring lower levels of creativity. Similarities between diagnosis and scientific discovery include: the need to select a best explanation from among alternatives, requiring the synthesis of a composite hypothesis (usually), and in evaluating hypotheses by the success and failure of predictions along with other traditional criteria such as simplicity and explanatory power. Diagnosis has been studied for several years from a computational perspective in artificial intelligence (AI), with many practical systems actually deployed.

4. Prototype system for model-based diagnosis

I will now describe a working prototype system for real-time model-based diagnosis. The system has been documented in detail by Wu (1997). The overall structure of the system and its inference strategies were designed to be as general and domain-independent as possible, while retaining specificity for the task of diagnosis. Thus, the details of the specific application domain (manufacturing operations) are not needed for understanding how the system works at the inference level. In short, it works as follows.

1. Devices in a *component library* are represented as having:

a. input and output *ports*,

b. one or more *modes* (states of a device that effect its behavior)

c. associated with each mode are one or more *functions* that map inputs to outputs. Functions may also change device modes. Functions will usually specify the time delays that will occur between inputs and cor-

responding outputs. Typically, functions can also be read backwards, mapping outputs to possible inputs.

2. The *target device* is represented as a system of components and connections using components from the library. The figure below depicts the specific target device represented in Wu's prototype. The device is represented as a flat system of components and connections, but extending the architecture to represent hierarchies of subdevices and multiple levels of organization does not appear to be problematical.

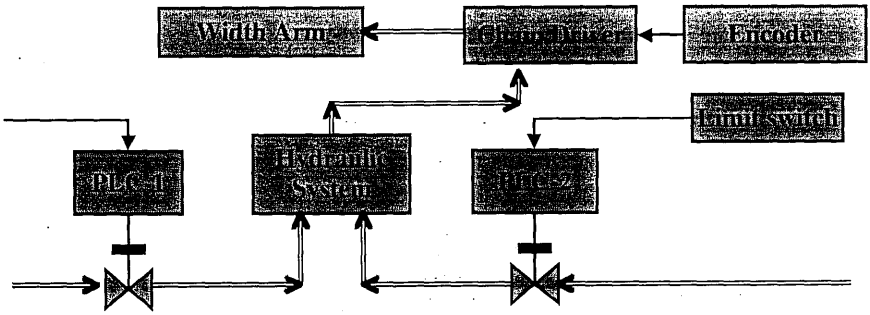

Figure 1

Cause-to-effect *forward simulation* for the target device is performed by propagating state descriptions from inputs to outputs across components, and across connections, starting with initial conditions and using functions and assumed modes. Similarly, effect-to-cause *backward simulation* is performed by mapping outputs to inputs, and across connections, considering all possible modes. Backward simulation generates branching alternative causal paths.

3. Fault detection, or more generally, *detection of deviations from expectations*, is done by forward simulation from normal or last-updated state and comparing the results in real time with data coming from the device. A significant mismatch implies a deviation from expectations. A deviation from expectations induces a *goal of explaining the deviation*, as in Peirce's view that inquiry begins with puzzlement.

4. If a deviation from expectations is detected, backward simulation is used to *generate hypotheses*, which are causal stories (causal paths) that

potentially explain the deviation from expectation. As hypotheses are being built by backward simulation, causal paths are cut off, whenever possible, by comparing with device data and eliminating those where mismatch occurs.

5. To *evaluate hypotheses*, forward simulation from the surviving hypotheses is used to produce predictions. Predicted observables are matched against data coming from the target device. Mismatching hypotheses are eliminated.

6. If more than one hypothesis remains, the *ambiguity* may be *resolved* by watching and waiting. Forward simulation from each of the alternatives is used to continually make predictions from the surviving set of hypotheses. Predictions are continually compared with new incoming data. Mismatching hypotheses are eliminated. The diagnostic system thus maintains a working set of plausible hypotheses that continually produce new predictions and are continually compared with new incoming data to eliminate hypotheses that are inconsistent with new observations.

5. Results

A small portion of a manufacturing plant was represented in the prototype system. A simulated malfunction was used to generate data. Detection of deviations of observable values from expected values was performed correctly. This triggered backward simulation from the points of deviation, which found three possible root causes after cutting off paths that included states that differed from observations. Forward simulation from the hypothesized faults and mismatching observations eliminated the two incorrect diagnoses, leaving only the correct (simulated) cause. The system got the right answer.

6. Discussion

The reasoning strategy has thus been demonstrated to work well, at least on one case. This constitutes a small-scale validation of the computational strategy for diagnosis. It also tends to validate the A-P model of scientific inference by showing that a version of it can be made precise enough to be implemented and to perform correctly for diagnosis.

The prototype system reflects the A-P model of scientific inference in that it implements a pattern of reasoning in which all of the major subfunctions of abductive reasoning, mentioned earlier, are reflected, at least in simplified forms:

- Deviations from expectation focus the system's attention on data needing explanation.
- Explanatory hypotheses are generated as alternative causal stories explaining the deviations from expectation. All possible causal paths (known to the system) are generated.
- Hypotheses are evaluated,
  - first, by direct comparison with observables, eliminating those inconsistent with observations,
  - second, by using cause-to-effect reasoning to generate predictions, with hypotheses that license failed predictions being eliminated,
  - third, by using continuing cause-to-effect reasoning to generate further predictions as time passes, with hypotheses that license failed predictions being eliminated.
- A hypothesis is accepted if it is the sole surviving explanation after hypotheses are evaluated.

Moreover, a conclusion that is accepted is the best explanation in contrast with its rivals, the rivals having been eliminated for inconsistency with observations or failures of prediction. The prototype system thus implements a pattern of reasoning which aims to arrive at a conclusion that has strong "abductive justification."

In the prototype system, a model is used to generate a causal story to explain the specific case. A new model is not generated, as would be needed to reflect theory formation in science. However, it is not difficult to imagine that the software technology for component libraries that supported the representation of the target device in the prototype system just described could be used under computer control to link components previously represented in the library and thus to compose novel device descriptions that could be used as hypothesized models to explain some classes of observational data.

Acknowledgements: I wish to thank Susan Josephson and B. Chandrasekaran for their helpful comments on an earlier draft. The work was supported in part by the DARPA RaDEO program and the Office of Naval Research under Grant No. N00014-96-1-0701.

Ohio State University

REFERENCES

Hanson N.R. (1958), *Patterns of Discovery*, Cambridge University Press, London.

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

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

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

Lycan W.G. (1988), *Judgement and Justification*, Cambridge University Press, Cambridge.

Peirce C.S. (1899), from the Collected papers of Charles Sanders Peirce, Vol. 1, paragraph 139, Edited by C. Hartshorne and P. Weiss, 1960, Harvard University Press.

Wu H. (1997), Use of multi-purpose knowledge database in simulation and diagnosis, Master's Thesis (Chemical Engineering), The Ohio State University, Columbus, Ohio, USA.