Australasian Journal of Philosophy,
Vol. 74, no. iv, (December 1996), 699-700.

REVIEW OF "ABDUCTIVE INFERENCE", JOHN R. AND SUSAN G. JOSEPHSON (CAMBRIDGE, CAMBRIDGE UNIVERSITY PRESS, 1994). AUS$125.00

Many introductory reasoning courses focus largely on deduction (inference of necessary consequences) and induction (generalising from a set of examples that share a certain feature to further unobserved cases). Charles Peirce, as in many philosophical inquiries, saw a third term lacking in this account of human reasoning. This third term he called *abduction*, and in 1901 he defined it 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.

Abduction is therefore a species of inference to the best explanation (inference to the best explanation of something surprising). Only abductive reasoning, he believed, could introduce an "original suggestion" or hypothesis into our thinking, without which science (or any other progress in human inquiry) would be inconceivable.

A prime example of abductive inference offered by the Josephsons is medical diagnosis. When a patient presents a doctor with a (surprising and unpleasant) symptom, such as enlargement of the liver, the doctor will frame hypotheses from which the symptom follows as a matter of course, such as the existence of a liver tumour.

This form of reasoning has been picked up by researchers in Artificial Intelligence who see deduction and induction as insufficient for modelling human reasoning. This book is an account of recent work by the Laboratory for Artificial Intelligence Research (LAIR) based at Ohio State University. LAIR researchers aim to make abduction a fundamental building block of modelled

reason. They believe that it can be formalised in a way that is new, more rigorous than ever before yet, interestingly, does not map onto (deductive) classical mathematical logic or (inductive) probability theory. Rather, abductive inference is a proposed new approach for AI.

The book defines abduction as the authors see it, sketches their proposed formalisation, and then reports the results of a series of computer programs designed to implement it in various real-life reasoning situations in which, they argue, it performed well. Thus the book is more than a theoretical claim for the importance of abduction, but offers sound empirical argument as well.

The Josephsons' account of abduction, offered in Chapter 1, is broader in scope than Peirce's 1901 definition. It is:

> "D is a collection of data (facts, observations, givens)
> H explains D (would, if true, explain D)
> No other hypothesis can explain D as well as H does
> Therefore, H is probably true." (p.5)

Where the Peircean formulation cited above seems aimed at the cutting edge of human inquiry, which moves forward by searching for explanations for epistemological surprises received by scientists, the Josephsons concentrate on abduction as a common-or-garden reasoning tool, used continually in everyday life. They therefore allow that abduction can provide positive confidence to a hypothesis, rather than mere enthusiasm for putting it to the (inductive) test, as Peirce claimed in 1901 (though it is arguable whether he maintained this view of abduction throughout his writings). Also noteworthy in this formulation is the LAIR researchers' emphasis on the importance of ruling out alternative explanations as a means of arriving at positive confidence when reasoning abductively.

Chapter 2 deals with a philosophical question that has bedevilled the AI field - what exactly is AI *for*? For example, are researchers merely trying to simulate reason, or are they aiming to breathe computational life into new thinking beings? Should AI research remain faithful to human reason

in every detail, or is it allowed to "surpass" it? And what would that mean? Susan Josephson explores this question through a fourfold distinction of possible justifications for doing AI.

First comes "AI as Engineering", where scientists design technology to solve practical problems, and the success of that technology is measured purely by its contribution to solving those problems. Second is "AI as Traditional Science", where writing and running AI programs is an experiment to test a particular theory about the mind, and the success or failure of the programs is measured purely by the scientists' success or failure at testing their theory. The last two categories fall under what Searle has called "Strong AI". "AI as Art" sees scientists trying to create something that will be a mind in its own right, and to the external observer as much like a human mind as possible, whereas in "AI as Design Science" the aim is to discover and reproduce the abstract principles behind human cognition, of which human minds are just a special case, and thereby possibly extend cognition in ways never before "known".

The final approach is the one the authors of this book favour. They claim that this design science approach to AI represents "a new paradigm of science" (and scientific explanation). This is because it seeks not merely to represent but also to order the world in new ways through studying recreating and creating *function*, rather than merely mapping events and their (efficient) causes. It is therefore imbued with an ineliminable teleology. I found this gestalt-shift in the philosophy of science intriguing.

The main structure of the book describes a series of abduction machines of ever-increasing sophistication built and tested by the LAIR. The first two were designed merely to show how an example of abduction may be formalised and programmed on a digital computer, by using programmable strategies that are as general as possible. These machines were written for a specific task - examining blood and offering "decision support" for identifying red-cell antibodies in the blood of someone about to undergo a transfusion. The machines are described in detail and the results they arrived at given.

Later stages in the project acknowledged that these first machines, though they tested hypotheses, did so in a way that was fairly algorithmic, with the hypothesis assembly strategy hard-wired during system programming. The next machine, which they called PEIRCE in honour of that still relatively unrecognised American philosophical pioneer, was designed to "opportunistically" improve a working hypothesis until it is satisfactory, during the course of problem solving.

In the final chapter and two appendices the Josephsons step back from the immediate details of software programming to consider some of the fascinating philosophical issues raised by their research. First of all, John Josephson makes the bold claim that all perception is "abduction in layers" (p. 238). What might this mean? He claims that perception is plausibly regarded as the progressive extraction of high-level information from low-level stimuli, and that we should think of each level as an inference to the best explanation of the stimuli on the level below. (This would seem to fit visual depth perception quite well, where the positions of three-dimensional objects are inferred from disparities between the two-dimensional inputs of two eyes. An interesting question, however, is how such an account might cope with colour.) They also discuss language use, which they claim is merely a special case of perception, and its abductive structure.

The book closes with a discussion of the concept of "plausibility" (of hypotheses) which the authors lean heavily on throughout the book. They explore the relationship of plausibility to the traditional formally elaborated "probability", and also the features which make this concept unique, such as the strong role which analogy may play in the generation of plausible hypotheses. A further philosophical theme in the background of this book, and worth further thought, is a new pragmatic account of seeking truth by seeking a confident explanation, provided in context, for a set of given facts.

This is a bold, programmatic book, which puts forward a strong, new analysis of abduction, demonstrates a new empirical approach to epistemology and outlines a possible new direction for AI, while raising important and interesting issues in philosophy of mind and philosophy of science.

Australian National University / Sydney University .