Popper and computer induction

Donald A. Gillies

Introduction

In a recent article in *BioEssays*, John F. Allen uses Popper's philosophy of science to argue on p. 107 that "There is no induction machine", and that looking for one will result in "waste of resources spent on the search itself." This argument is of great interest to me because I started my research life as a graduate student in philosophy of science in Popper's department at the London School of Economics in the years 1966–68. Popper had just published his book *Conjectures and Refutations*, 1963 in which he says on p. 53: "Induction, i.e. inference based on many observations, is a myth. It is neither a psychological fact, nor a fact of ordinary life, nor one of scientific procedure." Indeed I remember Popper making the statement: "induction is a myth" in one of his lectures in 1967, to which he added: "and those who claim that there is induction do not know what they are talking about." I need hardly add that I was completely convinced by Popper's very forceful arguments for this thesis which I believed for many years.

But can computers perform induction?

Sometime in the late 1980s I heard for the first time that groups of computer scientists working in a new field called machine learning were trying to program computers to carry out induction. My first reaction to this news was exactly the same as that expressed by John Allen in his article. I thought that it would be impossible for these computer scientists to carry out their research programme, and that they were wasting their time attempting an impossible task. I was, however, sufficiently intrigued to examine the field of machine learning more closely to see what had been achieved, and I discovered to my great surprise that researchers in machine learning had done what I had thought to be impossible, and had produced programs which genuinely enabled computers to carry out induction. Thus induction was not a myth after all, but a reality, and it became clear that Popper's philosophy would need to be modified. I have tried to carry out what I think are the necessary modifications to Popper's philosophy in my 1996 book: *Artificial Intelligence and Scientific Method*, and will now indicate a few of the results which I present there in more detail.

Golem and protein secondary structure

In chapter 2 of my 1996 book, I give a number of examples of machine learning programs which seem to me to have genuinely carried out computer induction.[2] Perhaps the most striking of these is the program "Golem" produced by Stephen Muggleton and his collaborators who follow an approach known as *inductive logic programming*. Allen refers to this group rather sarcastically as (2001, 107) authors who "use the I-word explicitly", and expresses scepticism as to whether they can achieve their objective: "to find rules to assist in the prediction of protein structure from sequence" (Ref. 1, p. 107). In reality, however, Golem has already succeeded in doing just that in some cases. In 1996, 50–55, I give a description of Golem and of one of the laws that it discovered by induction. This was a previously unknown rule gives a set of conditions that cause a residue to form part of an α-helix. The work was originally described by Muggleton, King and Sternberg.[3]

But many of Popper's ideas remain valid

The Golem work and many similar achievements in the field of machine learning show that induction is a reality rather than a myth, and this in turn shows that Popper's philosophy of science needs to be modified in some respects. The new results do not, however, show that Popper's ideas about science are to be totally rejected. On the contrary it turns out, perhaps surprisingly, that some of Popper's ideas are actually useful for creating computer programs that carry out automated induction. This leads to a number of points where I agree with what Allen says.[1] Allen rightly stresses the central importance of falsifiability in Popper's philosophy, and falsifiability is actually used in the program Golem just mentioned. Golem works by generating hypotheses from data according to a logical algorithm. The hypotheses generated are then tested out against further data, and those that fail these tests are eliminated. This process continues until eventually a hypothesis is found that passes all the tests carried out. In effect Golem works by a process of conjectures and refutations. The difference from Popper's original scheme is that the conjectures are generated mechanically rather than by the use of human intuition. Popper also emphasises the need for background knowledge in scientific investigations, and this is also borne out by Golem and other machine learning programs. The naïve scheme of induction is that a hypothesis (h) is induced from observational evidence (e), or, in symbols, $e \rightarrow h$. In all the machine learning programs that I have studied, the real inductive scheme is that a hypothesis (h) is induced from observational evidence (e) together with background knowledge (k), or in symbols $e\&k \rightarrow h$. In fact background knowledge is coded into Golem and other similar programs. This puts Golem in the same position as a human scientist who has learnt the background knowledge in his or her branch of science and uses it, together with observational

evidence, to form a new hypothesis. It should be stressed that the novel hypotheses discovered by Golem are in no sense implicit in, or contained in, the background knowledge. They constitute genuine additions to this background knowledge.

A comment on Allen's challenge

The point about background knowledge is relevant to the challenge that Allen makes in the following sentence (Ref. 1, p. 106): ''Everything I have said above would be disproved by a single demonstration that a purely logical (i.e. hypothesis-free) process, when applied to data alone, is sufficient to produce a gain in understanding.'' I do not think this challenge can be met at the moment, but this does not show that computer induction is impossible. Computer induction proceeds not from ''data alone'' as Allen requires, but from ''data and background knowledge''.

References

1. Allen JF. ''Bioinformatics and discovery: induction beckons again'', Bioessays 2001; 23:104–107.
2. Gillies DA. *Artificial Intelligence and Scientific Method*, 1996. Oxford University Press.
3. Muggleton S, King RD, Sternberg MJE. 'Protein Secondary Structure Prediction using Logic-Based Machine Learning', Protein Engineering, 1992;5/7:647–657.
4. Popper KR. Conjectures and Refutations: The Growth of Scientific Knowledge, 1963 Routledge & Kegan Paul.

Donald Gillies
Professor of Philosophy of Science and Mathematics
King's College London. UK.
E-mail: Donald.Gillies@kcl.ac.uk

On John Allen's critique of induction

Lawrence A. Kelley and Michael Scott

In a recent paper in *BioEssays*, John F. Allen expresses concern over whether current scientific techniques used in genomics/bioinformatics can be reconciled with a well-founded scientific methodology. He is particularly critical of ''data mining'' and computational approaches used to identify patterns in large quantities of data, such as the relationship between amino acid sequences and protein structures, supposedly relieving scientists of the need to think up and test hypotheses. While Allen identifies some important questions, we believe that his concerns about bioinformatics are misplaced and his arguments conflate some distinct issues.

Allen offers two objections. The first draws on falsificationist theory in philosophy of science, according to which science should proceed through thinking up conjectures and attempting to falsify them. Successful theories are those that survive elimination through falsification. In support of this, Allen cites Hume's scepticism about the rationality of inductive reasoning (the inference from, e.g., observed As are Bs to ''All As are Bs'') and endorses the dramatic claim ''verification is impossible'', which he attributes to Popper. If correct, the computational techniques in question would indeed be in difficulty. For, despite their alleged irrationality and impossibility, they do use induction and take evidence to confirm theories.(1) However, a great deal of science besides bioinformatics will fall foul of the approach Allen suggests — specifically, the use of any experimental outcome to count as positive evidence for the hypothesis that explains it. Presumably, the Darwinian theory of evolution, which is taken as confirmed by fossil evidence but not falsified by gaps in the fossil record, would not count as scientific on this basis. Nor, more relevantly, would statistical hypotheses. A statistical hypothesis is typically an (inductive) generalisation from a relationship observed in the sample data; moreover, such a hypothesis is in principle unfalsifiable. The problems Allen poses should, if anything, raise doubts about the correct order of his analysis. That is, whether scientific procedures should be dictated to by a priori considerations about scientific method, or whether our account of scientific method should rather take its cue from successful scientific practice.

Allen's second objection takes an entirely different direction. Apparently relaxing the falsificationist strictures implied by his first objection, Allen points out that even the computational approach must involve hypothesis testing: ''Every search for a sequence feature embodies an hypothesis.'' This, of course, is quite true but only trivially so. Any investigative procedure can be construed as testing some assumption, in this case the search algorithm contained in the data analysis programme. The departure from traditional hypothesis testing involved in the computational approach is not that these techniques lack any assumptions whatsoever (nobody claims the programmes appear by magic). The point is that the assumptions involved do not present an explanation of the phenomena in question, but only a strategy for identifying relationships between strings of data.

Allen concludes his paper with a plea for understanding and some polemical observations on the need for current research

BioEssays 23:860–861, © 2001 John Wiley & Sons, Inc.

to be informed by ideas. This suggests that his real complaint lies not with induction or even the role of hypotheses, but rather the kind of understanding of the phenomena afforded by the computational approach. This concern is entirely understandable but not to the point. In principle, any approach that works is of course welcome, but in the case of the protein folding problem thirty years of research has failed to yield a solution. As the problem stands, the computational approach seems to be the most useful direction of investigation. It makes little sense to insist on positing underlying mechanisms before analysing the huge quantities of complex genomic and structural data now available.

References

1. ILP S. Muggleton. Inductive Logic Programming. New Generation Computing 1991;8(4):295–318.

Lawrence A. Kelley
Biomolecular Modelling,
Imperial Cancer Research Fund
44 Lincoln's Inn Fields,
P.O. Box 123, London, WC2A 3PX, U.K.
E-mail: L.Kelley@icrf.icnet.uk

Michael Scott
Department of Philosophy,
University of Edinburg, George Square,
Edinburgh, EH8 9JX, UK.
E-mail: Michael.Scott@ed.au.uk

Hypothesis, induction and background knowledge. Data do not speak for themselves. Replies to Donald A. Gillies, Lawrence A. Kelley and Michael Scott

John F. Allen

Reply to Donald A. Gillies

The intention of my article in BioEssays[1] was to draw attention to what I believe is an absurd proposition, namely, that analysis of data can enlarge human understanding in the absence of any hypothesis or preconceived idea. Donald A.Gillies, I am sure, understands my position perfectly.[2] I think that our only point of disagreement stems from different senses in which he and I use the word ''induction'', and I hope that we will be in agreement if we resolve this ambiguity. By ''induction'' I mean ''logical induction''. Gillies uses ''induction'', and even ''inductive logic'', in a different sense. His usage corresponds to one of the definitions given in the Shorter Oxford English Dictionary: ''The adducing of a number of separate facts, particulars, etc., especially for the purpose of proving a general statement''. If we take ''prove'' to mean ''test'' (as in ''The exception proves the rule'') rather than ''to establish the truth of'', then I think there is no difference between Gillies's position and mine. I would then readily agree both with Gillies and with Kelly and Scott that computers can, and do, help us with testing (''proving'') general statements. However, the problem of where general statements come from in the first place remains.

A different definition of ''induction'' in the Shorter Oxford English Dictionary corresponds to the sense in which I intended to use the word in my article.[1] It is as follows: ''Logic. Of the process of inferring a general law or principle from the observation of particular instances (opp. DEDUCTION, q.v.)''. The reference to ''deduction'' is important, because deduction is the transfer of truth from a premise to a conclusion. Thus ''induction'', in this second sense, strongly implies the transfer of truth from a number of observations to a general principle. In this sense Popper was right to say ''induction is a myth...and those who claim that there is induction do not know what they are talking about''.

Although Gillies's position is very much a case of ''been there, done that'', I am pleased to acknowledge that he is correct, and he has obviously studied these issues in greater depth than I. Gillies's recollection of Popper's lectures is of interest to me. My own contact with Popper's ideas derives from reading his books, but my original decision to do so was based on once losing an argument about whether science produces increasingly probable explanations of the world. I thought it did, but it doesn't. The person who won the argument had already read, and understood, Popper, and knew him personally. That person was Colin McClare, a protein biochemist in the Department of Biophysics at King's College London. The Department was then in Drury Lane, just a short walk from the London School of Economics. At the LSE

Popper was Professor of Philosophy, and Gillies had been a research student just a few years before. McClare, was an excellent lecturer, and my undergraduate tutor for one half-year, I think in 1970–71. He was passionate about his science and about the philosophy of science. He would have understood, and I hope enjoyed, the whole of this discussion.

Reply to Lawrence A. Kelley and Michael Scott

There seem to be serious philosophical differences between Kelley, Scott[3] and myself. For example, Kelley and Scott argue that "...a great deal of science besides bioinformatics will fall foul of the approach Allen suggests". They suggest that, if I'm correct, then no experimental outcome can serve as positive evidence for the hypothesis that explains it. I think this is a mistake, and the same as that made by Holliday in an earlier contribution to *BioEssays* on this topic.[4] My position on "evidence" is as follows. When an observation is consistent with, and can be explained by, an hypothesis, then we regard it as evidence for that hypothesis, particularly if no alternative hypothesis can explain the same observation, and especially if there is an alternative hypothesis that predicts something quite different. This, in fact, is the only rational basis that we have for preferring one hypothesis to another, according to Popper. It is also his solution to Hume's psychological problem of induction.[5] A new and better hypothesis is one that can explain everything accounted for previously, but one that also makes predictions about which earlier, or competing, hypotheses are either mistaken or silent. These same predictions are precisely the potential falsifications, or tests, of the new hypothesis,[5] and the agreement between these predictions and observation is what we count as evidence in its favour.

I think that what many people expect from Popper is that he provides an explanation of why better hypotheses are more probably true than worse ones (as I wished to do in the argument with McClare). Popper himself is quite clear that he is unable to provide this explanation. More importantly, his rather startling conclusion is that we cannot know about the probability of something being true at all (see Fig. 2 of Ref. 1). This is a point that has not been widely understood, and has had little impact, unfortunately, on most people's expectations of science. I am sure it will require further discussion. Even if we ever stumbled on the complete truth we could not know that that is what it was. Restated, if ever we found a statement whose "real" probability of truth, p, had the value $p = 1$, we still could not know the value for p.

Kelly and Scott state "It makes little sense to insist on positing underlying mechanisms before analysing the huge quantities of complex genomic and structural data now available". My true position is even more radical than they seem to suspect. It can be summarised as follows. *It makes little sense to insist on collecting genomic and structural data before you, or someone else, has posited an underlying mechanism*. Without having an underlying mechanism — in essence an explanatory, tentative hypothesis — you have no basis on which to decide which data to collect.[6] Data do not, and cannot, "speak for themselves".[1]

References

1. Allen JF. Bioinformatics and discovery: induction beckons again. Bioessays 2001;23:104–107
2. Gillies DA. Popper and computer induction. Bioessays 2001;23:859–860.
3. Kelley LA, Scott M. Bioessays 2001;23:860–861.
4. Holliday R. The incompatibility of Popper's philosophy of science with genetics and molecular biology. Bioessays 1999;21:890–891.
5. Popper KR. Objective knowledge. An evolutionary approach. Oxford: Oxford University Press. 1972.
6. Allen JF. In silico veritas. Data-mining and automated discovery: the truth is in there. EMBO Reports 2001;2:1–3

John F. Allen
Plant Biochemistry
Lund University
SE-221 00 Lund
Sweden
E-mail: john.allen@plantbio.lu.se