tripleC 4(2): 328-332, 2006
ISSN 1726-670X
http://tripleC.uti.at

Testing Reasoning Software. A Bayesian Way

Bertil Rolf

School of Management, Blekinge Institute of Technology
SE-372 25 Ronneby, Sweden
bro@bth.se

Abstract: Is it possible to supply strong empirical evidence for or against the efficacy of reasoning software? There is a paradox concerning tests of reasoning software. On the one hand, acceptance of such software is slow although overwhelming arguments speak for the use of such software packages. There seems to be room for skepticism among decision makers and stakeholders concerning its efficacy. On the other hand, teachers-developers of such software (the present author being one of them) think the effects of such software are obvious. In this paper, I will show that both positions – skepticism vs. belief in efficacy – can be compatible with evidence. This is the case if (1) the testing methods differ, (2) the facilities of observation differ and (3) tests rely on contextual assumptions. In particular, I will show that developers of reasoning software can, in principle, know the efficacy of certain design solutions (cf. van Gelder, 2000b, Suthers et al., 2003). Other decision makers may, however, be unable to establish evidence for efficacy.

Keywords: software, reasoning, education, test, efficacy

1 Clarification.

By “empirical evidence”, I refer to observations and measurements, outcome of tests and of experiments, where the evidence is elaborated by inductive methods. Such methods include Bayesian inductive reasoning. By “reasoning software”, I refer to graphically based, general purpose reasoning supporting software to be described below.

Here, the question is not whether we actually can show such impact. The question is whether possible impact would be detectable by empirical methods, provided that such impact was present.

I will focus on a type of general-purpose reasoning supporting software directed at professionals or students in higher education. There are five or six such software packages. Some of them are purely experimental and only two or three have the kind of finish that makes them usable in real courses where they have been tried out. Two well known such packages are shown below (van Gelder, 2000a, Rolf and Magnusson, 2002):

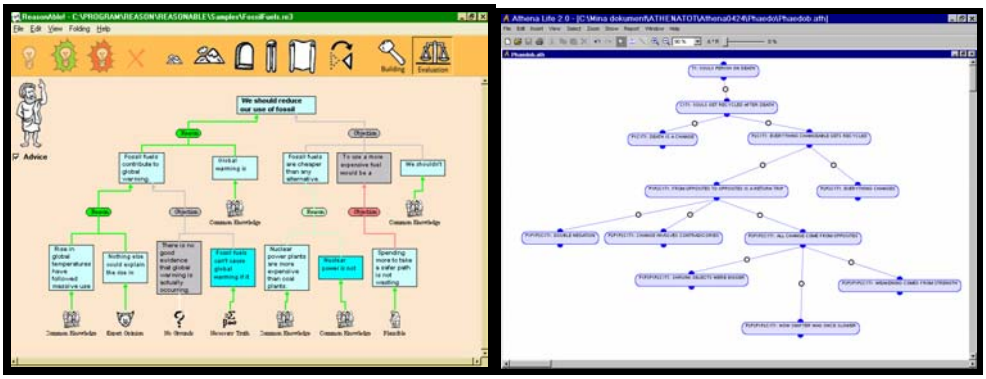

Figure 1: Two software packages for general purpose reasoning support. Tim van Gelder's Reason!Able (left) the author's Athena Standard (right).

Such software packages are typically based on a kind of "soft" (i.e. non-axiomatized) theory of argumentation, they lay claims to applicability in a wide range of argumentation and they would, in principle, be usable in courses in critical thinking. The possible market for such packages is very large.

2 Using Bayesian Testing: Converging from Subjective to Intersubjective Probabilities

Bayesian inductive reasoning satisfies the Kolmogorov axioms of probability, just like the Fisher-Neyman-Pearson (FNP) methods do. A major difference, however, is in the interpretation of the axioms. The FNP-interpretation assigns probabilities only to classes of events. Therefore, the FNP-methods cannot make sense of probabilities of hypotheses. Bayesians claim to make sense of such assignments representing degrees of personal beliefs (Howson and Urbach, 1993).

The underlying logic of Bayesian testing will focus on the quotient between probabilities of hypotheses. The Bayesian interpretation of probability is focused on supplying evidence that will make posterior odds converge. The underlying logical relation is illustrated with a formula:

Posterior odds	= The likelihood ratio	*	Prior odds
P(H1/e1…en)/P(H2/e1…en)	= P(e1…. en/H1)/P(e1…en/H2)	*	P(H1)/P(H2)

Assignment of prior odds is largely a matter of subjective guessing. In order to reach an intersubjective conclusion, we need to focus on such evidence that might make the posterior odds converge.

Typically, the assignment of Bayesian probabilities starts with a subjective judgment or guess about probability. Different guesses about probabilities can converge towards an interpersonal assignment of revised probability on the basis of rather few observations. A table with fictional values will illustrate how rapidly convergence occurs:

Quotient initial probabilities. Odds against effect.	Initial probability of effect. From col 1.	Quotient posterior probabilities 1 case. Assuming LR = 10.	Posterior probability of effect 1 case. From col 3.	Quotient posterior probabilities of 2 cases. Assuming LR = (10) ^2.	Posterior probability of effect 2 cases.
1	0.50	0.10	0.91	0.01	0.99
2	0.33	0.20	0.83	0.02	0.98
3	0.25	0.30	0.77	0.03	0.97
4	0.20	0.40	0.71	0.04	0.96
5	0.17	0.50	0.67	0.05	0.95

6	0.14	0.60	0.63	0.06	0.94
7	0.13	0.70	0.59	0.07	0.93
8	0.11	0.80	0.56	0.08	0.93
9	0.10	0.90	0.53	0.09	0.92
10	0.09	1.00	0.50	0.1	0.91

Figure 2: Table showing convergence from initial probabilities to posterior probabilities after two cases.

Above, "LR" stands for the likelihood ratio between the two hypotheses for and against effect. A Bayesian inductivist starts with a guess about the odds against an effect – in the table ranging between 1 and 10. If we chose H2 as the negation of H1, we can solve the probability of H1 via the formula:

$$\frac{X}{1-X} = \frac{P(H1, \text{given the evidence})}{P(H2, \text{given the evidence})}$$

This gives us probabilities for effect ranging from 0.09 to 0.50. After one well-chosen test case, the Bayesian's probabilities for effect will vary from 0.50 to 0.91. After two well-chosen tests, her probabilities will vary between 0.91 and 0.99. In this way, Bayesians can reach intersubjectively valid inductive conclusions on the basis of few, well chosen observations or measurements (*Stanford Encyclopedia of Philosophy*).

3 Using Strongly Discriminating Observations

It is hard to use mainly qualitative evidence to estimate convergence unless the ratio converges towards 0 or towards very large numbers. We therefore wish to look for evidence e1…. en that strongly discriminates between H1 and H2, that is to say, evidence that takes the likelihood ratio either towards 0 or towards very large values. Such evidence would be highly unexpected without the use of software and rather expected with the use of the software.

What might be highly unexpected evidence in a course, using reasoning software? Athena based courses and software have been designed to encourage students to improve on four key factors to good reasoning:

- Robustness, i.e. that all evidence has been supplied that might change a rationally based conclusion.
- Structure, emphasizing the tree structure of arguments, sorting pros and cons at various levels of argument, creating branches in the argument tree.
- Relevance, in the sense that subordinate arguments should support superior conclusions.
- Acceptability, in the sense that the arguments supplied should be as likely (if factual) or normatively acceptable as possible.

It is very unusual that these features occur in an inquiry even among professional experts. Both the course and the use of Athena software imply that these features should be present in student inquiries at a level far above what is common among students or young professionals.

Observational competence will have to be assumed in all inductive testing. In order to supply inductive arguments based on the pieces of evidence e1…en, one needs competence to establish e1… en.

The more importance we attach to strongly discriminating evidence, the more we demand of observational competence to discriminate the intended unusual and specific effects. Such

observational competence can be presumed among qualified teachers of reasoning and designers of software who know which features that will discriminate.

Not all educational decision makers will exercise such observational competence. They may want to improve students' capacities for critical thinking in general, not merely in those respects supported by a particular software package. It would be natural for them to design a broad test of critical ability. Such tests may not discriminate features related to a specific software package. Reasoning software captures only some aspects of critical thinking. The broader tests that are applied, the less chance there is of discovering highly discriminating evidence.

So teacher-designers of software will, in general, possess evidence more suited to draw inferences about the efficacy of their own software packages.

4 Context Dependent Testing

In all induction, probabilities are assigned relative to some unspecified background knowledge. It is assumed that there is no unknown factor of influence, systematically influencing the outcome. In classical statistics, randomization takes care of some such factors that are known to us. Unknown factors are supposed to exercise no systematic influence. We assume in our background knowledge that there is no systematic influence from sunspots or planetary positions.

In testing effects of software use, however, important background factors contribute to effects. Software in itself has no learning effects whatsoever. All effects arise from the usage of software by teachers, classes and individual students.

Software is no different from textbooks or other educational facilities in this respect. Whatever learning effects they have, depend on the way they are used. The effects of educational facilities are always blended with effects from tasks, teaching methods, examination, institutional frame factors etc.

A teacher-designer can modify both software and its uses in order to produce desired effects. In developing software and using it in education, one will normally plan and observe such uses and modify software or modify teaching on the basis of inferences from such observations. If desired effects do not arise, a teacher-designer can either modify the software package, the instructions, the tasks or the teaching methods (Rolf, 2003, Rolf, 2004).

When the teacher-designer uses Bayesian induction, it is presupposed that the context is fixed or that the interesting software effects arise from software together with background factors. Such a presupposition makes sense if you are acquainted with the context of use.

Other decision makers are less privileged with respect to background knowledge. Background assumptions are not transparent. The background of use if seldom described, classified, codified or controlled for in testing.

5 Conclusion

The efficacy of reasoning software is in principle intersubjectively testable on the basis of a few observations. If (1) Bayesian induction is used, (2) highly discriminating evidence is established on the basis of observational competence and (3) local context of usage can either be presumed constant or be explicitly controlled for, then it is possible to know the efficacy of reasoning software on the basis of few observations.

But in practice, observations and contextual background will often be sufficiently transparent only for teachers-designers involved in the processes of education and software design. For decision makers outside the design-education-test process, a rational acceptance of effects would need large scale testing across systematically varied contexts of use. (cf Hitchcock, forthcoming) It is an open matter whether it is realistic to demand such tests.

References

Howson, C. & Urbach, P. (1993, Second Edition). *Scientific Reasoning: The Bayesian Approach*, Chicago: Open Court.

van Gelder, T. (2000a). Learning to Reason: A Reason-Able Approach, in C. Davis and T. van Gelder et al. (eds.), *Cognitive Science in Australia: Proceedings of the Fifth Australasian Cognitive Science Society Conference*, Adelaide: Causal.

van Gelder, T. (2000b). *The Efficacy of Undergraduate Critical Thinking Courses. A Survey in Progress*, http://www.philosophy.unimelb.edu.au/reason/papers/efficacy.html.

Hitchcock, D. (forthcoming). *The Effectiveness of Computer-Assisted Instruction in Critical Thinking*, in Informal Logic.

Joyce, J. (2003). Bayes' Theorem. *The Stanford Encyclopedia of Philosophy* (Winter 2003 Edition), E. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2003/entries/bayes-theorem/>.

Rolf, B. & Magnusson, C. (2002). Developing the Art of Argumentation, in F. van Eemeren et al. (eds.), *Proceedings of the Fifth Conference of the International Society for the Study of Argumentation,* Amsterdam: Sic Sat. 2003.

Rolf, B. (2003). Educating Reason, in C. Constantinou and Z. Zacharia (eds.), *Computer Based Learning in Science: Conference Proceedings,* Nicosia: Dept. of Educational Sciences. University of Cyprus.

Rolf, B. (2004). Cognitive And Social Strategies In Teaching Reasoning, in K. Fernstrom (ed.), *Proceedings of the 5th International Conference on Information Communication Technologies in Education*, Athens: National and Kapodistrian University of Athens.

Suthers, D. & Hundhausen, C. et al. (2003). An Exploratory Comparison of the Roles of Representations in Face to Face and Online Collaborative Learning, in *Proceedings of the 36th Hawai`i International Conference on the System Sciences (HICSS)*.