Alan Bundy
December 3, 1997

A Scientific Checklist

1 Introduction

In this note I give a first draft of a checklist of questions that one should ask about a piece of AI research. This is intended as an aid to researchers when presenting their research, evaluating the research of others or even when proposing a piece of research. I am considering putting this in the Researchers Bible, so I would particularly value constructive feedback.

2 Testing Hypotheses

Any piece of scientific research, including AI research, should state a hypothesis, provide evidence for (and/or against) that hypothesis and conclude whether that hypothesis is supported or refuted. This is true even of a piece of purely implementational AI. The hypothesis, in this case, might be:

> *System* P *is good for task* X.
> or
> *System* P *is better than rival systems* Q *and* R *for task* X.

Similar hypothesis can be formulated by replacing the word "system" above by "theory", "technique" or "parameter".

In this context *"good for"* and *"better than"* are usually with respect to some properties on one or more of the following three dimensions:

Behaviour: the correctness or quality of the solutions to task X which are produced by the systems;

Coverage: the range of examples of task X which the systems can be applied to or can solve; or

Efficiency: the resources consumed by the systems in doing task X.

The evidence provided to support (or refute) the hypothesis can be theoretical, experimental or both. Theoretical evidence often involves some mathematics and the proving of some theorem. Experimental evidence requires testing; usually a program is tested on a range of examples.

3 Questions to Ask

This analysis suggests various questions that should be asked of any proposed or completed piece of research.

3.1 The Hypothesis

(1) *What is the hypothesis and is it explicitly stated?*

Often the hypothesis of a piece of AI research is only implicit. It may be that, in the research area concerned, the standard hypotheses is so obvious that it does not require explicit statement. However, it frequently aids presentation and methodological clarity if it is, nevertheless, explicitly stated.

(2) *Is the hypothesis falsifiable?*

According to Popper, a hypothesis is not scientific unless you can imagine circumstances under which it would be refuted. Of course, you need faith in the truth of your hypothesis to see you through those dark days of despair. But you should also be able to put yourself in the shoes of your potential critics: referees, examiners, seminar questioners, and see what tough questions they would ask. Asking these will help in formulating the programme of analysis and testing you will need to subject your hypothesis to.

*Notes in this series are for ϵ-baked ideas, for $1 \geq \epsilon \geq 0$. Only exceptionally should they be cited or distributed outwith the Mathematical Reasoning Group.

(3) *If some theory, technique, system, parameter is being hypothesised to be good, or better than some rival, at some task:*

  (a) *what is the task?*

  (b) *are the properties that are being measured or compared: behavioural; coverage; efficiency or some combination of these?*

  (c) *is the evidence being advanced theoretical; experimental; or some combination?*

3.2 Measuring and Comparing Properties

(4) *If a behavioural property is being measured, how is the scale of correctness or quality determined?*

Sometimes some theoretical measure of correctness or quality of solutions is available. Sometimes it is necessary to appeal to some authority, *e.g.* a human expert at task X.

(5) *If coverage is being measured:*

  (a) *can the full range of examples of task* X *be characterised?*

  (b) *can the range of examples to which* P *is applicable be characterised?*

  (c) *can the range of examples on which* P *is successful be characterised?*

Sometimes a theoretical characterisation is available. Sometimes it is only possible to explore a range experimentally. In this case, characterisations are usually vague.

(6) *If efficiency is being measured, are we concerned with usage of space, time or of some other resource?*

3.3 Types of Evidence

(7) *If the evidence provided is theoretical, are any simplifying assumptions justified?*

Theoretical evidence almost always requires some simplifying assumptions to make the analysis tractable. Some variable may be considered negligible and ignored. Another variable may be assumed to vary in a predictable way. If these simplifying assumptions are invalid then the theoretical analysis may be worthless. Validating them may require experimental testing.

(8) *If the evidence provided is experimental, are the test examples representative?*

Usually, the full range of examples is infinite and we can only test finitely many. Some kind of characterisation of the full range is required to determine that the test range is representative of it. In particular, we must guard against the criticism that the test examples especially favour P.

3.4 Hypothesis Confirmation

(9) *Does the evidence really support (or refute) the hypothesis?*

Apparently good evidence may not be due to the cause claimed. For instance, one system may perform better than another, but this may not be due to the features we are ostensibly comparing. They may differ in some other feature, and this may provide the real explanation. If rival systems only differ in one feature then this must be the cause. Unfortunately, it is sometimes difficult to arrange for systems to differ only in the compared feature. Micro-analysis of system behaviour may be necessary to confirm the real cause of behavioural differences.

Differences in experimental evidence may only be due to chance. Statistical analysis may be required to show that these differences are significant.

4 Conclusion

In this note I have listed 9 questions to ask of any piece of research, including AI research. I propose that this list be used to check your research. For instance, you might run through it with the first draft of a research paper. Better still, you might use it to check a project proposal *before* you ask for support or start working on it. You could also use it when refereeing the research proposals and papers of others.

I expect there are a lot more than 9 questions of this kind to ask and would welcome suggestions of additional questions — the more generally applicable the better. I would also welcome any suggestions for improvement of the questions proposed.

5 Self Reflection

Practice what you preach. So here goes for the note above.

1. **Hypothesis:** AI research can be assessed with a small list of generic questions.
2. **Falsifiability:** This hypothesis would be refuted if a piece or kind of AI research was discovered to which these questions were inapplicable or irrelevant.
3. **Nature:** the task is AI research; the properties are behavioural and coverage; the evidence is experimental.
4. **Behavioural Scale:** the scale of correctness and quality is determined by appeal to experts in assessing AI research.
5. **Characterising Coverage:** in each case the range is all AI research.
6. **Type of Efficiency:** trivially true, since efficiency will not be measured.
7. **Simplifying Assumptions:** trivially true, since no theoretical evidence will be provided.
8. **Representative Test Set:** not applicable yet, since this is only at proposal stage, but I hope these questions will be used in a wide area of AI, *e.g.* if it were put in Researchers Bible and used by all DAI PhDs.
9. **Support for Hypothesis:** The non-applicability of any question to any piece of AI research would require reformulation of the hypothesis, so there is no possibility of false comfort.