

Situated Semantics is a Side-Effect of the Computational Complexity of Abduction

Tim Menzies

Artificial Intelligence Laboratory,
School of Computer Science and Engineering, University of New South Wales
PO Box 1, Kensington, NSW, Australia, 2033
timm@cse.unsw.edu.au

ABSTRACT: We develop a general abductive description of testing models. We find that this testing process is fundamentally slow and cannot be conducted exhaustively. Consequently, we argue that the usual case for model testing is non-exhaustive testing; i.e. some subset of the possible tests are chosen and executed. Note that if the tests result in model refinement, then different tests can result in different models. This leads to the hypothesis that different individuals form different "opinions" (i.e. models) about the world as a result of the different examples they push through their models. We prefer this symbolic explanation for situated semantics to non-symbolic proposals (e.g. neural).

1. Introduction

We endorse and continue an argument begun elsewhere; i.e.

...there is nothing wrong with classical logics in representing commonsense knowledge; there is, however, a problem with the assumption that to use logic we have to do deduction. David Poole [30].

We agree with Poole and Wang [41] that many of the arguments against logical/symbolic AI (e.g. Birnbaum [2]) are actually arguments against a more specific target; i.e. classical deduction. We disagree with (e.g.) Birnbaum that the obvious alternative to logical AI is some type of situated/functional cognition, the nature of which is yet to be explicated (but is explored in [5, 8]). Further, methodologically, we believe that is better to explore AI using a small number of general, well-understood and reproducible symbolic mechanisms than an indeterminate (but large) number of poorly understood domain-specific mechanisms [24].

In order to support our AI-logicist position, we seek explanations for human cognition in terms of the behaviour of non-deductive theorem provers. Wang's theorem prover, for example, selectively forgets unused theorems. Our work, like Poole's, focuses on *abductive* logics¹. We

¹ Consider a system with two facts a , b and a rule R_I :
 $If\ a \Rightarrow b$. *Deduction* is the inference from a to b . *Induction* is the process of learning R_I given examples of a and b occurring together. *Abduction* is inferring a , given b [19]. Abduction is a not a certain inference and its results must be checked by an inference assessment operator (see **BEST**, below). For more on

present here an abductive model of *situated semantics*²; i.e.

*The conclusions drawn from a model may vary according to the history of that model, who wrote the model, and the **TASK** at hand (**TASK** defined below).*

Such relativist knowledge is difficult to understand in terms of classical deduction which seeks context-free theorems. to represent knowledge³.

Using our abductive model, we describe a general computational device for testing models. We find that this test process is fundamentally slow and cannot be conducted exhaustively. Consequently, we argue that the usual case for model testing is non-exhaustive testing; i.e. some subset of the possible tests are chosen and executed. Note that test results can lead to model revisions (e.g. when a model fails a test, it is repaired). Different tests can result in different model revisions. This leads us to conjecture that individual form different opinions (i.e. models) about the world as a result of the different examples they push through their models. We offer this hypothesis as a symbolic explanation for situated semantics.

The structure of this article is as follows. Section 2 explores situated semantics. Sections 3 & 4 of develops and applies our abductive framework for testing models. Section 5 discusses the computational complexity of this framework. Section 6 discusses cognitive implications and section 7 discusses related work.

2. Situated Semantics

The concept of situation/context has become a key issue in contemporary models of scientific development, decision making, and knowledge acquisition (KA). Human knowledge appears in

abduction, see [21, 26, 30].

² Note that we avoid the similar term *situated cognition*. By situated semantics, we do NOT mean that symbolic representations are a post-hoc report of the physical interactions of the of the inferencing entity with its internal and external environment [7, 8]. Functional/situated cognition as discussed by Clancey and Birnbaum is a non-symbolic explanation for situated semantics. This paper explores symbolic explanations.

³ Birnbaum offers this as a diminutive summary of Nilsson's declarative knowledge proposal [25]. Our reading is that Nilsson seeks knowledge that is as re-usable as possible and not (as Birnbaum seems to believe) that is always re-usable in all contexts.

some social situation and that context seems to effect the generated knowledge:

- Kuhn notes that data is not interpreted neutrally, but (in the usual case) processed in terms of some dominant intellectual "paradigm" (which, if we represented it computationally, would be a model) [17].
- Phillips [27] and Bradshaw *et. al.* [3] describe model construction as a communal process that generates structure that explicate a community's understand of a problem. If the community changes then the explicit record of the communities shared understanding also changes; i.e. "truth" is socially constructed.
- Silverman cautions that systematic biases in expert preferences may result in incorrect/incomplete knowledge bases (KBs) [39].
- Compton [11] argues that the symbolic representations found in our knowledge bases are not records of structures inside the head of an expert. Rather, this "knowledge" is a situated report customised to the specific problem, the specific justification, the expert, and the audience. Like Phillips and Bradshaw *et.al.*, Compton argues that "truth" as expressed by human experts varies according to who says it.
- Agnew, Ford & Hayes summarises contemporary thinking in this area as:

...expert-knowledge is comprised of context-dependent, personally constructed, highly functional but fallible abstractions [1].

Two experiments demonstrate situated semantics. Shaw [38] took a group of geology experts and had them construct knowledge bases for the same problem. The experts then reviewed each other's knowledge base and, after 12 weeks, their own. Table 1 shows that experts strongly disagreed with each other. For example, experts only agreed with each other, at best one-third of the time.

Expert pairs	% Understands	% Agrees
E_1, E_2	62.5	33.3
E_2, E_1	61.1	26.7
E_1, E_3	31.2	8.3
E_3, E_1	42.9	33.3
E_2, E_3	44.4	20.0
E_3, E_2	71.4	33.3

Table 1: Expert E_x reviewing E_y 's rules.


Table 2 shows the expert's assessment of their own knowledge base, 12 weeks after they wrote it. Overtime, as an expert's situation changes so does their view on "correct" knowledge. This change may be very dramatic. For example, expert 1 could only understand 62.5% of what he'd written 12 weeks before. All experts disagreed (to some extent) with their own ideas

from the past (as shown in the *Agrees* column of Table 2).

Expert	Understands (max = 100)	Agrees (max = 100)
E_1	62.5	81.2
E_2	77.8	94.4
E_3	85.7	78.6

Table 2 Self-review of a specification, 12 weeks after it was written.

Our second demonstration of situated semantics is a small thought experiment. Consider a one-line mathematical model of exponential population growth:

$$EQ_1 : dN/dT = rN.$$


where T is time, N is the population and r is negative or positive in hostile or benign environments respectively. This model is wrong since population growth must taper off as it approaches C the maximum carrying capacity of the environment; i.e.

$$EQ_2 : dN/dT = rN(1-(N/C)).$$


In the case of a hostile environment and over-population, our intuition is that the population will fall. However in such circumstances, $N > C$, $r < 0$, and $rN(1-(N/C)) > 0$; i.e. the maths says that population will increase [34]. EQ_2 is therefore also incorrect.

We ask the reader, when did you become aware of the errors in EQ_1 and EQ_2 ; *before* or *after* we presented our examples of population growth tapering off as N approaches C (EG_1) and over-population in hostile environments (EG_2)? If *after*, then we have anecdotal evidence for situated semantics; i.e. models are situated in the history of their development (more precisely, situated in the examples used for their development). As to our own experience, we studied EQ_1 and EQ_2 extensively without detecting any errors. Yet when presented with EG_1 and EG_2 , the errors were obvious. EQ_1 and EQ_2 are not universal truths. Rather, they true in restricted contexts (i.e. not EG_1 , not EG_2 and not every other context where these equations fail). Elsewhere, we have argued that as soon as a model moves "out-of-context", it may generate inappropriate results. Sadly, models are rarely labelled with their contextual boundaries. When all available knowledge and examples are used to generate a model, out-of-context is never indicated since it represents the area(s) unexplored during development [22].

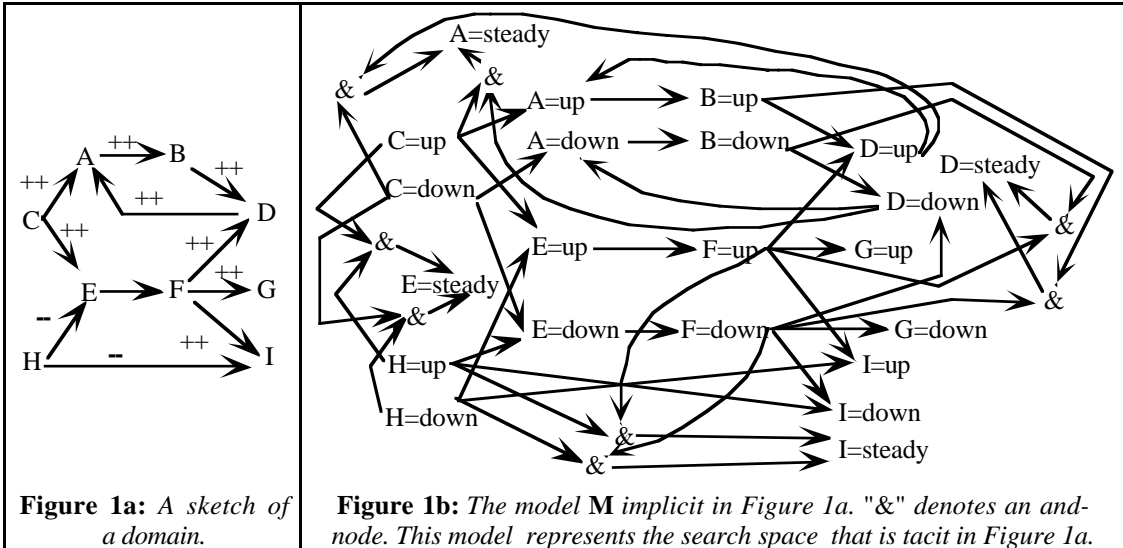


Figure 1a: A sketch of a domain.

Figure 1b: The model M implicit in Figure 1a. "&" denotes an and-node. This model represents the search space that is tacit in Figure 1a.

3. Abductive Models

3.1. Optional Inferences & Explanations

We view a model as the source of optional inferences which we could make, if it proved useful and possible to do so. This view is very different to the classical deductive interpretation of a model. Deductive rules such as *if a then b* are in no sense optional; in all worlds where *a* is true, *b* is also true. Our view of models as the space of possible inferences has much in common with the *RSpace* of Clark & Matwin (i.e. a specification of the space of rules from which ideal domain rules can be learnt [10]); or the *scenarios* of Poole:

The user gives true facts and a pool of possible hypothesis they are prepared to accept as part of an explanation to predict the expected behaviour [29].

More precisely, we define an abductive model to be a directed, possibly cyclic graph of edges E , vertices V , and invariants I . V is either (i) literals referring to entities that the model author is familiar with; or (ii) and-nodes (described below). I is an invariant predicate that is satisfied iff some set of vertices cannot be believed simultaneously without violating some constraint; e.g. $I(p, \neg p)$, $I(\text{day=monday}, \text{day=tuesday})$, $I(A=\text{up}, A=\text{down})$. E is the union of the explanations that are acceptable to the model author(s); e.g. $e(\text{rain=yes}, \text{grass=wet})$ means that rain is an acceptable explanation for wet grass.

The library of known behaviour B of M is a set of pairs $\langle \text{IN}_i, \text{OUT}_i \rangle$ where IN_i and OUT_i are subsets of V . The *TASK* of our model M is, roughly speaking, to explain OUT_i in terms of IN_i . Explanations (also known as worlds W) are maximal unions (with respect to size) of acyclic proof trees P :

- Whose roots are from IN_i and whose leaves are members of OUT_i
- Which have the right number of parents for each vertex in a proof. And-nodes must share a proof with all their parents. In general, all other vertices must share a proof with 1 parent. Exceptions: IN_i vertices and **DEFAULT** vertices. A **DEFAULT** vertex has no parents and is in IN_i - **FACTS** (**FACTS** = $\text{IN}_i \cup \text{OUT}_i$). These vertices need no parents in a proof.
- That contain no vertex that contradicts the **FACTS** or any other vertex in that world.

M may not be parsimonious, complete, deterministic, or consistent. In the general case we can generate N consistent explanations ($0 \leq N \leq \infty$) for some subset of OUT_i using some subset of E (see example, below).

3.2. An Example

Suppose that an expert sketches Figure 1a using the notation that -- denotes *discourages* and ++ denotes *encourages*. Let us assume that (i) the vertices $\{A, \dots, I\}$ in Figure 1a have the states $\{\text{up}, \text{down}, \text{steady}\}$; (ii) a conjunction of an *up* and a *down* can explain a *steady*; and (iii) a *steady* cannot explain anything else. Figure 1a is like a graphical macro language that we can expand into the model of Figure 1b. In the case of $\langle \text{IN}_1, \text{OUT}_1 \rangle = \langle \{C=\text{up}, H=\text{up}\}, \{B=\text{up}, D=\text{up}, G=\text{up}, I=\text{down}\} \rangle$ and the invariant that the states $\{\text{up}, \text{down}, \text{steady}\}$ are mutually exclusive, then we can generate the following proofs:

- $P_1: H=up \rightarrow E=down \rightarrow F=down \rightarrow I=down$
 $P_2: H=up \rightarrow I=down$
 $P_3: C=up \rightarrow E=up \rightarrow F=up \rightarrow G=up$
 $P_4: C=up \rightarrow A=up \rightarrow B=up \rightarrow D=up$
 $P_5: C=up \rightarrow E=up \rightarrow F=up \rightarrow D=up$
 $P_6: C=up \rightarrow A=up \rightarrow B=up$
 $P_7: C=up \rightarrow E=up \rightarrow F=up \rightarrow D=up \rightarrow A=up \rightarrow B=up$

An assumption A_i is a vertex in P that is not in $FACTS$. In our example, P contains assumptions A for vertices $\{A,B,E,F\}$. An interesting subset of A are the controversial assumptions A_c that are *base* (i.e. are dependant on no other A_c). The exclusions X of the base controversial assumptions A_b are the assumptions that contradict the maximal consistent subsets of A_b . A proof belongs in a world if it does not use the base controversial assumptions excluded from that world. For our example, $A = \{A=up, B=up, E=up, E=down, F=up, F=down\}$; $A_c = \{E=up, E=down, F=up, F=down\}$; and $A_b = \{E=up, E=down\}$ (since F is fully dependant on E). Our exclusions are $X_1 = \{E=down\}$, and $X_2 = \{E=up\}$. Each exclusion X_i defines one world W_i (see Figure 2). We also define W_0 to be the world that excludes all A_b ; i.e. $X_0 = \{E=up, E=down, F=up, F=down\}$ (see Figure 2a).

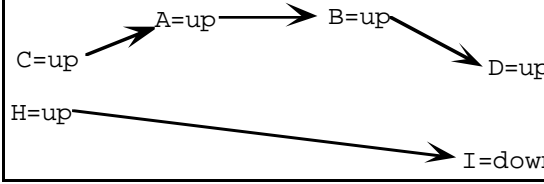


Figure 2a: W_0 : All P_i that avoid X_0 .

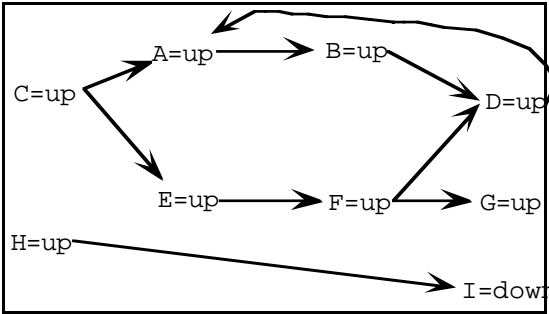


Figure 2b: W_1 : All P_i that avoid X_1 .

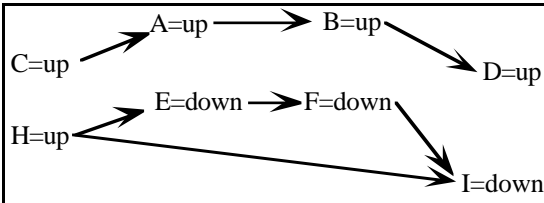


Figure 2c: W_2 : All P_i that avoid X_2

Formally, this process is abduction [21]. Given a theory M , some assumables A , a goal OUT_i and

invariant knowledge I , abduction is the inference to explanations that cover OUT_i without generating inconsistencies; i.e.

$$\begin{aligned}
 & A' \subseteq A, \text{ CAUSES} \subseteq IN_i, \text{ COVERED} \subseteq \\
 & \quad \text{OUT}_i \\
 & M \cup \text{CAUSES} \cup A' \vdash \text{COVERED} \\
 & M \cup \text{CAUSES} \cup A' \not\vdash \text{false (i.e. } \neg I \text{)}.
 \end{aligned}$$

Abduction is not a certain inference. Our example above generated three worlds and we must choose between them using some **BEST** assessment operator. Example **BESTs** include **BEST₁**: returning all W_i with fewest assumptions ; **BEST₂**: with the smallest number of inputs $|CAUSES|$; **BEST₃**: with shortest proof size ($\forall P_i, j \in W_i, \Sigma[P_j]$); **BEST₄**: with the largest number of outputs $|COVERED|$; or **BEST₅** which avoids edges with low likelihood (assuming that such meta-knowledge about edges is available; e.g. some edges were proposed as part of a theory you wish to fault)⁴.

3.3. Plausible

In this section, we argue for the plausibility of our abductive model as a psychological theory of expert reasoning.

We have characterised inference over a model as the *extraction* of consistent beliefs (W_i) from some background knowledge (M) that is relevant to some **TASK** (which we define to be the tuple $\langle\langle IN_i, OUT_i \rangle, BEST_j \rangle$). Numerous commonly-used representations can be mapped into our definition of a model; e.g. the dependency graph between literals in a propositional theory, a rule-based expert system⁵, a unfolded first-order system⁶, and declarative frame-based systems⁷ satisfy our M definition. Elsewhere, we have argued that this extraction process is the core inference underlying prediction, classification, explanation, diagnosis, qualitative/causal reasoning, design, recognition, case-based reasoning, and natural language understanding [22]. For example, single-fault set-covering diagnosis is just a specialisation of **BEST₂** (i.e. $|CAUSES| = 1$). More generally, we find that abduction directly operationalises the model extraction process which Clancey [9] and Breuker [4] argue is at the core of expert system inference [21].

⁴ Bylander *et. al.* call **BEST** the plausibility operator *pl* [6].

⁵ Less its conflict resolution strategy.

⁶ That is, unfolded until it is ground (i.e. all variables bound).

⁷ In partial-match systems, the disjunction of slots can lead to a frame. In total-match frame systems, the conjunction of slots can lead to a frame. In both cases, inferring a subclass can lead to inferring the superclass (e.g. *if emu then bird*).

We find that this inference model offers symbolic explanations for certain idiosyncracies noted in human reasoning. Consider how certain special cases of this abductive inference would appear to an outside observer:

- i) In certain cases, world generation is indeterminate. Consider Figure 2 in the case where (somehow) W_1 did not occur. In this case, $BEST_4$ cannot decide between W_0 and W_2 since they both cover the same number of OUT_i . Yet these worlds condone different beliefs (W_2 condones a belief in $\{E=down, F=down\}$ while W_0 does not). When inference assessment operators are inconclusive, an *abductive model-based reasoning system* (hereafter, AMBRS) could not decide if certain literals are believed or not. An outside observer would note that the AMBRS was undecided about certain issues, even after due deliberation.
- ii) Checking that inferences deduced from an abductive models is hard (to be precise, it is NP-hard [6, 37]). An AMBRS may not allocate sufficient resources to complete that inference. An outside observer would note that such an AMBRS held certain inconsistent beliefs.
- iii) Assuming that sufficient resources are allocated to world generation, it is still a slow process. If some inference is required during that generation time, then an AMBRS would produce a conclusion that they may retract very soon afterwards. An outside observer would note that such an AMBRS is making mistaken decisions that it tries to patch afterwards

Like AMBRS, human beings (i) appear undecided about certain issues, even after due deliberation; (ii) hold inconsistent beliefs; and/or (iii) make mistaken decisions that they patch afterwards. Hence, we conjecture that certain aspects of human cognition can be modelled by our symbolic formalism.

We view this as a plausible, but not conclusive, argument. An alternative symbolic explanation for these behaviours is simply that the human theorem-prover is non-abductive (e.g. deductive) somehow resource limited (e.g. due to short-term memory limitations). For a more convincing argument for humans-as-AMBRS, see our subsequent discussion regarding situated semantics (see section 6).

4. How to Test a Model

In this section we discuss methods of testing a model M .

In general, there are two categories for tests of a model: we can look at its internal structure or we can assess it on some external criteria. KB

verification tools detect anomalies with internal syntactic structures such as contradiction, tautologies, circularities and logical subsumption [32]. KB validation tools apply some external semantic criteria such as test suite assessment. The test suite may be naturally occurring or may be generated via an analysis of the dependencies within the KB [16, 42].

We argue that external semantic criteria takes precedence over internal syntactic criteria. The model *if a then b* contains no syntactic anomalies yet may be irrelevant to the task of inferencing over important domain entities (e.g. $\{c,d,\dots\}$). Further, it is known that fielded expert systems may contain internal syntactic anomalies, yet perform adequately. Preece & Shinghal document fielded expert systems that contain numerous logical anomalies such as unused inputs, unsatisfiable conditions and unusable consequences [33]. These expert systems still work, apparently because in the context of their day-to-day use, this erroneous logic is never exercised.

Our preferred external semantic test is test suite assessment. If a model cannot reproduce behaviour some required behaviour, and/or the behaviour of the thing that it is modelling, then it is definitely wrong. Hence, we propose $BEST_4$ (maximal B -coverage) as the definitive test for a model⁸. $BEST_4$ is an *exhaustive, relative* measure suitable for *under-specified models*:

Exhaustive: Given a single model, we can fault it iff after generating all possible consistent explanations, we still cannot cover known behaviour. Note that this search cannot be culled at the local propagation level. The utility of using each local inference (some edge in E) has to be assessed by a meta-interpreter using the global criteria: "will it eventually lead to maximal coverage?". This global criteria cannot be applied till after all possible paths are collected. We will return to this point below.

Relative: Given a selection of possible models, we can rank them according to $BEST_4$. The best models cover the highest percentage of OUT_i . Note that we do not demand total B -coverage. We know of cases in neuroendocrinology where existing models are known to be faulty, yet they are still being used since (i) it takes too long to remove all bugs from the models and (ii) these models represent the current high-water mark in that domain.

Under-specified models: Given an under-specified indeterminate model, we can still assess it using

⁸ This definition of test was originally inspired by Popper's view that ideas are never "true" in some absolute sense. Rather, the models we currently believe are the ones that have survived active attempts to refute them [31].

BEST₄. Inference over indeterminate models requires making assumptions and maintaining those assumptions in mutually exclusive worlds; i.e. the abductive process described above. The ability to process indeterminate models is an important feature of real-world model assessment. Many KB domains are poorly-specified and, hence, indeterminate. If they it were otherwise, then algorithmic approaches would suffice for KB problems.

5. Complexity of BEST₄

This section discusses the computational complexity of applying **BEST₄**. We will find that both theoretically and experimentally that this process has significant computational limits.

The complexity of testing a model using **BEST₄** is at least the complexity of the abductive process described in Section 3. Most known abductive inference engines exhibit exponential runtimes for real-world inputs, even for sophisticated algorithms. Hence, many of the articles in [26] are concerned with heuristic optimisations of abduction. Eshghi report a class of polynomial-time abductive inference problems, but this class of problems require a non-cyclic background theory [15]. Bylander reports techniques for tractable abduction [6], but many of these techniques (e.g. rule-out knowledge to cull much of the search space) are not applicable to arbitrary models developed in poorly-measured domains (e.g. our test domain of neuroendocrinology).

Selman & Levesque show that even when only *one* explanation is required, and **M** is restricted to acyclic theories, then abduction is NP-hard [37]⁹; i.e. very likely to be computational intractable in the worst-case. Recall that our search is for inexplicable behaviours; i.e. we must search for *all* explanations because only then can we decide what behaviours are inexplicable. This exhaustive search is hence even slower than standard abduction.

Theoretical discussions aside, experiments have shown our abductive **B**-coverage semantics is a practical model validation tool for models with $|V| < 850$ and $|E|/|V| < 7$ [21]. For these experiments, we used our HT4 abductive inference engine. For reasons of efficiency, HT4 makes two assumptions about **M**: (i) the model is generated and cached prior to inferencing; (ii) **I** is restricted to symmetric invariants of arity of 2. Assumption (i) lets us use fixed-sized bitstrings to represent much of the sets processing. Assumption (ii) lets us quickly pre-compute and cache with each vertex a list of other **FORBIDDEN** vertices.

Earlier versions of HT4 [23] used a basic chronological backtracking approach (i.e. no

memoing) that only terminated for very small models. Basic chronological backtracking has the disadvantage that any feature of the space learnt by the search algorithm is forgotten when backtracking on failure [13, 20]. HT4 learns and caches as much as it can about the search space as it executes.

- *Forward sweep*: **A_c** is inferred as a side-effect of forward chaining from **IN_i** (ignoring **I**). Once this sweep terminates, a linear-time post-processor can find all **A_c**.
- *Backward sweep*: **A_b** is inferred as a side-effect of growing proofs back from **OUT_i** through the space found by the forward sweep towards **IN_i**. Each **P_i** stores (i) its **ROUTE** (the set of vertices it uses); its own **FORBIDDEN** set (i.e. the vertices that, with the **ROUTE**, would violate **I**); and the upper-most **A_c** vertices found during proof generation¹⁰. Candidate vertices **V_j** for inclusion in **P_i** must satisfy $V_j \notin P_i.ROUTE$ (loop detection) and $V_j \notin P_i.FORBIDDEN$ (consistency check). After all proofs are generated, the union of all **P_i.A_c** is **A_b**.

- *Worlds sweep*: Once **A_b** is known then the worlds **W** can be calculated by looping over all proofs and all exclusions as follows:

```

i ← 0;
for Xi ∈ exclusions(Ab) do begin
  i ← i + 1; Wi ← ∅;
  for Pi ∈ P do
    if Xi ∩ Pi = ∅ then Wi ← Wi + Pi
  ;
end;
```

Figure 4 shows the results of executing HT4 with **BEST₄** for 94 models and 1991 **<IN_i, OUT_i>** pairs. Note the abort time shown in Figure 3. Exhaustive abduction is slow and, in a resource-bounded environment, this implies some "give-up" time. None of the models over $|V| = 850$ terminated within this time frame and so the average runtime curve lies somewhere into the grey area to the right of Figure 4. Our reading of Figure 4 is that (i) runtimes are exponential on model size and (ii) the knee of an exponential curve for HT4 is found around $|V| = 800$.

⁹ A result endorsed by Bylander *et al* [6].

¹⁰ Note that at and-nodes, the proof generation is more intricate since all consistent combinations of all proofs for all parents must be computed. Combining proofs also implies combining the **P_i.A_c** sets. See [21] for details.

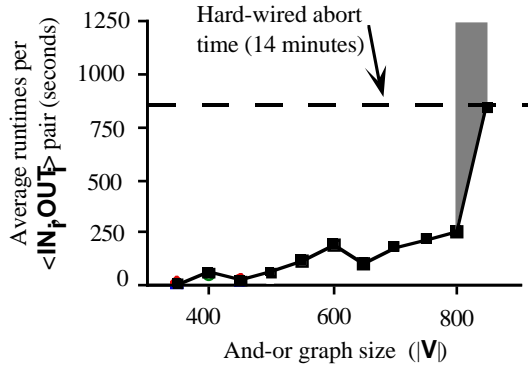


Figure 4: Average runtimes using **BEST₄** for models of varying size.

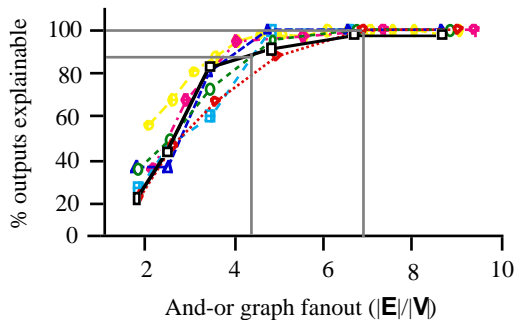


Figure 5: Average **B-coverage** for seven models of varying fanout.

Figure 4 is experimental confirmation of the theoretical prediction that testing-as-abduction is intractable for large models. Figure 5 shows another limit-to-testing. In the Figure 5 experiment, edges were randomly added to 7 models to increase their fanout ($|E|/|V|$) from 2 to 10. These models were then run using randomly generated $\langle \text{IN}_i, \text{OUT}_i \rangle$ pairs. Note that after a fanout of 4.4, 89% of all behaviours were explicable. Further, after a fanout of 6.9, nearly all behaviours were explicable. Note that a test that cannot distinguish between different models is a useless test; i.e. our generalised test procedure is only applicable for models with $|E|/|V| < 7$.

In order to place the $|V|$ and $|E|/|V|$ limit in perspective, we include here figures for model $|V|$ and $|E|/|V|$ in some fielded expert systems (see Table 3).

Application	$ V $	$ E / V $
<i>mmu</i>	65	7
<i>tape</i>	80	4
<i>neuron</i>	155	4
<i>displan</i>	55	2
<i>DMS-1</i>	510	6

Table 3: Model size and average fanout from real-world expert systems. From [33].

Table 3 is telling us that we are already using models near our limits. Systems larger than Table 3 (e.g. [18, 40]) are clearer well beyond our limits to testing.

6. Cognitive Implications

An AMBRs that uses a model bigger than our limits cannot rigorously explore all the test cases available to it. Some subset must be chosen. We know of no general principles for a priori selecting significant test cases.

Lacking such guidance, an AMBRs will select a subset test cases that *appear* to be more important (possibly using some selfish criteria that satisfy some local agenda). Given a community of AMBRs working on the same problem, they may all select a different subset of test cases.

Testing a model can lead to a modification of a model. Testing with different cases can hence lead to different models. Our AMBRs community could all develop different models of the same thing.

If we can find evidence for communities of humans that evolve different models for the same thing, then we have some evidence that humans are AMBRs. We believe that situated semantics is such evidence. For example, we conjecture that Shaw's geology experts learnt their knowledge base tool in their own way and used their different experiences (instantiated as their own personal models) to build their knowledge bases. 12 weeks later, their local models had changed. Perhaps due to their professional experiences since the initial study, their personal models had again changed. Hence, Shaw's reported intra- and inter-expert disagreements.

7. Related Work

To our knowledge, the cognitive implications of the computational complexity of abductive logic has not been previously discussed in the literature. While Poole explores non-deductive logics, he as (to date anyway) refrained from exploring the cognitive implications of his abductive frameworks.

We prefer our simple abductive inference model to Wang's more elaborate architecture. Our experiments have convinced us of the practical utility and insightfulness of our approach. To our knowledge, Wang has not entered the experimental state.

Operators for implementing preference criteria for assessing possible worlds has been widely discussed in the literature. Most researchers argue that the best worlds must at least cover all the known output. Some argue that the "best" explanation is the smallest one (e.g. [12, 35]). Poole [28] and Console *et. al.* [12] have proposed the additional criteria that the "best" explanation also uses the most specific terms from a taxonomic hierarchy; e.g. they prefer explanations in terms of *emu* rather than in terms of the more general term *bird*. We prefer not to hard-wire

world assessment into our formalism. World assessment knowledge is still domain-specific knowledge and should be customisable.

We assume that all worlds will be generated. An alternative approach is to generate a single world and only move from that single world if its base assumptions somehow break down [14]. Our all-worlds approach is closer to DeKleer's ATMS [13] and default logic extensions [36]. However, we differ from ATMS/ default logic in that our worlds only contain relevant literals (i.e. only those literals that exist on proofs between inputs and outputs). This means that, unlike default logic extensions, not all consequences of a literal exist in a world containing that literal. For example, consider the following example:

```

model:      b if a; c if a;
            e if (b or c);
            d if b; f if c; z if y.
contradicts: {d,f}
<IN2,OUT2> = <{a}, {e}>

```

Our approach would generate two proofs which could exist in the one world, i.e. $W_1 = \{P_1, P_2\}$, $P_1 = \{a, b, e\}$, $P_2 = \{a, c, e\}$. Standard ATMS/ default logic would analyse all literals in the model to generate two extensions (one with d and the other with f). Both of these extensions would contain the same proofs of e in terms of a . We view these two worlds as irrelevant and wasted computation.

For further notes on related work, see [21, 22].

8. Conclusion

We have argued that **BEST₄** (test-suite coverage) is the definitive test for a model and have explored its theoretical and experimental limits. We have found those limits to be $|E|/|V| < 7$ and $|V| < 850$. While the $|E|/|V|$ limit seems fundamental to the problem of generalised test, the $|V|$ limit could be increased by using faster platforms¹¹. However, given the fundamentally exponential nature of the process¹², we do not expect significant increases to the $|V|$ limit to be achieved in this manner.

We note that testing models bigger than these limits necessitates exploring a subset only of the possible test cases. Communities of AMBRs exploring different subsets can lead to different models being generated for the same problem. Such a community would exhibit situated semantics. Hence, we offer this symbolic test-based explanation for this phenomena.

Finally, we return to the claim at the start of the paper that is nothing wrong with classical logics

in representing commonsense reasoning. We find that if we refrain from simple deductive logic, we can explain certain observed cognitive phenomena (situated semantics) without recourse to non-symbolic approaches (e.g. connectionism). We prefer our symbolic model of human cognition, if only because it permits a precise description of its theory, implementation, and limitations in this short paper. We would be more convinced by the non-symbolic school if they could deliver a similar description.

9. Acknowledgements

Hugh Clapin, ANU, was kind enough to critic early drafts of this paper.

10. References

1. Agnew, N.M., K.M. Ford, and P.J. Hayes, *Expertise in Context: Personally Constructed, Socially elected, and Reality-Relevant?* **International Journal of Expert Systems**, 1993. **7**(1).
2. Birnbaum, L., *Rigor Mortis: A Response to Nilsson's "Logic and Artificial Intelligence"* **Artificial Intelligence**, 1991. **47**: p. 57-77.
3. Bradshaw, J.M., K.M. Ford, and J. Adams-Webber. *Knowledge Representation of Knowledge Acquisition: A Three-Schemata Approach* in **6th AAAI-Sponsored Banff Knowledge Acquisition for Knowledge-Based Systems Workshop, October 6-11 1991**. 1991. Banff, Canada:
4. Breuker, J. *Components of Problem Solving and Types of Problems* in **8th European Knowledge Acquisition Workshop, EKAW '94**. 1994.
5. Brooks, R.A., *Intelligence Without Representation* **Artificial Intelligence**, 1991. **47**: p. 139-159.
6. Bylander, T., D. Allemang, M.C. Tanner, and J.R. Josephson, *The Computational Complexity of Abduction* **Artificial Intelligence**, 1991. **49**: p. 25-60.
7. Clancey, W., *Book Review of Winograd & Flores, Understanding Computers and Cognition: A New Foundation for Design* **Artificial Intelligence**, 1987. **31**: p. 233-250.
8. Clancey, W., *Book Review of Israel Rosenfield, The Invention of Memory: A New View of the Brain* **Artificial Intelligence**, 1991. **50**: p. 241-284.
9. Clancey, W.J., *Model Construction Operators* **Artificial Intelligence**, 1992. **53**: p. 1-115.
10. Clark, P. and S. Matwin. *Using Qualitative Models to Guide Inductive Learning* in **Proceedings of the Tenth International Machine Learning Conference, ML-93**. 1993. Department of Computer Science, Ottawa University, Canada.
11. Compton, P.J. and R. Jansen, *A philosophical basis for knowledge acquisition*. **Knowledge Acquisition**, 1990. **2**: p. 241-257.

¹¹ HT4 was built using Smalltalk on a Macintosh Powerbook 170.

¹² Recall that **BEST₄** is exhaustive abduction and Selman & Levesque have shown that non-exhaustive abduction is NP-hard.

12. Console, L., D.T. Dupre, and P. Torasso, *On the Relationship Between Abduction and Deduction* **Journal of Logic Programming**, 1991. **1**(5): p. 661-690.
13. DeKleer, J., *An Assumption-Based TMS Artificial Intelligence*, 1986. **28**: p. 163-196.
14. Doyle, J., *A Truth Maintenance System* **Artificial Intelligence**, 1979. **12**: p. 231-272.
15. Eshghi, K. *A Tractable Class of Abductive Problems* in **IJCAI '93**. 1993. Chambéry, France:
16. Ginsberg, A. *Theory Reduction, Theory Revision, and Retranslation* in **AAAI '90**. 1990.
17. Kuhn, T., **The Structure of Scientific Revolutions**. 1962, New York: Cambridge Press.
18. Lenat, D.B. and R.V. Gutha, *CYC: A Midterm Report* **AI Magazine**, 1990. (Fall): p. 32-59.
19. Levesque, H. *A Knowledge-Level Account of Abduction (preliminary version)* in **IJCAI '89**. 1989. Detroit, Michigan, USA:
20. Mackworth, A.K., *Consistency in Networks of Relations* **Artificial Intelligence**, 1977. **8**: p. 99-118.
21. Menzies, T., *Principles for Generalised Testing of Knowledge Bases*. 1995, University of New South Wales:
22. Menzies, T. and P. Compton. *The (Extensive) Implications of Evaluation on the Development of Knowledge-Based Systems* in **Proceedings of the 9th AAAI-Sponsored Banff Knowledge Acquisition for Knowledge-Based Systems Workshop**. 1995.
23. Menzies, T., A. Mahidadia, and P. Compton. *Using Causality as a Generic Knowledge Representation, or Why and How Centralised Knowledge Servers Can Use Causality* in **Proceedings of the 7th AAAI-Sponsored Banff Knowledge Acquisition for Knowledge-Based Systems Workshop**. 1992. Banff, Canada, October 11-16,:
24. Menzies, T.J. *A Precise Semantics for Vague Diagrams* in **AI'94**. 1994. Armidale, Australia:
25. Nilsson, N.J., *Logic and Artificial Intelligence* **Artificial Intelligence**, 1991. **47**: p. 31-56.
26. O'Rourke, P., *Working Notes of the 1990 Spring Symposium on Automated Abduction*. 1990, University of California, Irvine, CA.:
27. Phillips, L.D., *A Theory of Requisite Decision Models* **Acta Psychologica**, 1984. **56**: p. 29-48.
28. Poole, D. *On the Comparison of Theories: Preferring the Most Specific Explanation* in **IJCAI '85**. 1985.
29. Poole, D., *A Logical Framework for Default Reasoning* **Artificial Intelligence**, 1988. **36**: p. 27-47.
30. Poole, D., *A Methodology for Using a Default and Abductive Reasoning System* **International Journal of Intelligent Systems**, 1990. **5**: p. 521-548.
31. Popper, K.R., **Conjectures and Refutations**,. 1963, London: Routledge and Kegan Paul.
32. Preece, A.D., *Principles and Practice in Verifying Rule-based Systems* **The Knowledge Engineering Review**, 1992. **7**(2): p. 115-141.
33. Preece, A.D. and R. Shinghal. *Verifying Knowledge Bases by Anomaly Detection: An Experience Report* in **ECAI '92**. 1992. Vienna:
34. Puccia, C.J. and R. Levins, **Qualitative Modelling of Complex Systems: An Introduction to Loop Analysis and Time Averaging**. 1985, Cambridge, Mass.: Harvard University Press. 259.
35. Reggia, J., D.S. Nau, and P.Y. Wang, *Diagnostic expert systems based on a set covering model*. **Int. J. of Man-Machine Studies**, 1983. **19**(5): p. 437-460.
36. Reiter, R., *A Logic for Default Reasoning* **Artificial Intelligence**, 1980. **13**: p. 81-132.
37. Selman, B. and H.J. Levesque. *Abductive and Default Reasoning: a Computational Core* in **AAAI '90**. 1990.
38. Shaw, M.L.G. *Validation in a Knowledge Acquisition System with Multiple Experts* in **Proceedings of the International Conference on Fifth Generation Computer Systems**. 1988.
39. Silverman, B.G., *Survey of Expert Critiquing Systems: Practical and Theoretical Frontiers* **Communications of the ACM**, 1992. **35**(4): p. 106-127.
40. Soloway, E., J. Bachant, and K. Jensen. *Assessing the Maintainability of XCON-in-RIME: Coping with the Problems of a VERY Large Rule-Base* in **AAAI '87**. 1987.
41. Wang, P., *From Inheritance Relation to Non-Axiomatic Logic*. 1993, Center for Research on Concepts and Cognition, Indiana University, Bloomington, Indiana, 1993:
42. Zlatareva, N. *Distributed Verification and Automated Generation of Test Cases* in **IJCAI '93 workshop on Validation, Verification and Test of KBs**. 1993. Chambéry, France: