

---

This space is reserved for the EPiC Series header, do not use it

---

# Boldly Going Where No Prover Has Gone Before

Giles Reger

University of Manchester, Manchester, UK

## Abstract

I argue that the most interesting goal facing researchers in automated reasoning is being able to solve problems that cannot currently be solved by existing tools and methods. This may appear obvious, but focusing on this as a primary goal allows us to examine other goals in a new light. Many successful theorem provers employ a portfolio of different methods for solving problems. This changes the landscape on which we perform our research: solving problems that can already be solved may not improve the state of the art and a method that can solve a handful of problems unsolvable by current methods, but generally performs poorly on most problems, can be very useful. We acknowledge that forcing new methods to compete against portfolio solvers can stifle innovation. However, this is only the case when comparisons are made at the level of total problems solved. We propose a movement towards focussing on unique solutions in evaluation and competitions i.e. measuring the potential contribution to a portfolio solver. This state of affairs is particularly prominent in first-order logic, which is undecidable. When reasoning in a decidable logic there can be a focus on optimising a decision procedure and measuring average solving times. But in a setting where solutions are difficult to find, average solving times lose meaning, and whilst improving the efficiency of a technique can move potential solutions within acceptable time limits, in general, complementary strategies may be more successful.

## 1 Setting the Scene

Before I begin, let me note that the contents of this paper are the result of working very closely with Andrei Voronkov on Vampire. However, whilst he influenced these views heavily, they remain my own. The paper will use the royal ‘we’ throughout as I am reflecting on the general philosophy of the group to which I belong. The aim of the discussion is to provide a starting point for the discussion of the role of portfolio solvers in research in general, and evaluation in particular.

In this discussion we consider the reasoning problem of taking a formula and answering whether the formula is satisfiable or not<sup>1</sup>. In some cases, this problem may be decidable (e.g. for propositional logic) and in more interesting cases (at least for this paper) it will not be (e.g. for first-order logic, or theories of arithmetic). This represents a large and diverse set of research areas and realistically our focus is on theorem proving in first-order logic (our main area of experience), however we attempt to argue more generally.

Solutions to the above problem take the form of *proof search* where we apply the term liberally to describe, for example, both the approach of searching a finite set of possible solutions (as in SAT) or the approach of searching an infinite clause space (as in saturation-based proving for first-order logic).

---

<sup>1</sup>In the case of first-order theorem proving this is usually phrased as checking validity via unsatisfiability of the negation.

Even in the decidable cases, this proof search is exploring a huge space that is rarely structured in such a way that a single search strategy will lead to a solution (within the given resources). This leads to the introduction of *heuristics* to guide proof search in directions likely to contain a solution. Such heuristics may prefer different parts of the search space and could give up qualities of proof search usually considered important, such as *completeness*, with the aim of finding a solution more quickly.

To handle the above issue, a common solution is the use of *portfolio solvers* that combine different strategies for exploring the solution space by trying a sequence of different approaches (we do not conflate this approach with *collaborative* approaches, which we discuss shortly). Note that this description does not differentiate between differing proof search strategies implemented within a single solver and the combination of distinct solvers implementing fundamentally different approaches.

The typical portfolio approach is to perform *time slicing* e.g. to slice up the available solving time between different strategies. This approach is based on the hypothesis (which we revisit later) that **if a solution exists then a quick solution exists**, which implies that it is better, for example, to try 100 *complementary* strategies for one second rather than one strategy for 100 seconds. We highlight the term *complementary* as this approach relies on the identification of strategies that are likely to explore different parts of the search space. Finally, we note that this time slicing approach is straightforwardly parallelisable or at least made concurrent where different strategies are interleaved.

In addition to portfolio approaches, solvers can be combined to achieve a synergistic effect. Collaboration can have two forms. A problem can be split into smaller parts and different approaches collaborate by focussing on different parts, or each different approach considers the whole problem and there is collaboration at the level of things learned that are, in general, helpful for proof search. An example of a collaborative approach of the first kind would be the typical Nelson-Oppen SMT solver where decision procedures are combined to solve different parts of the problem based on their signature. Examples of the second kind include the historic *Team-Work* method for clause sharing (among many others), or our more recent work utilising the AVATAR architecture in Vampire to share knowledge about the search space between concurrent proof attempts.

A key advantage of both portfolio and collaborative approaches is that strategies particularly suited for a specialised set of difficult problems can be combined to give a more general approach. This relies on techniques for deciding which strategies to apply based either on (syntactic or semantic) characteristics of a problem, or on some user guidance (e.g. if the user thinks they know the answer they can already restrict search to strategies optimised for satisfiability or unsatisfiability).

## 2 The Challenge: Solving The Unsolved

Our position is that the main and overriding goal of research on this (very general) problem is, or at least should be, to be able to **solve as many problems that people care about as possible**. This statement has two parameters that need addressing. Firstly, how do we characterise those problems that people care about and, secondly, what does it mean to solve a problem e.g. under what resource constraints? We argue that important problems should be motivated by some application and what counts as a solution should be tied to this application. For example, many program analysis tools that pass reasoning tasks to solvers give very short time limits and solutions should be found within these limits to ‘count’.

The above goal appears self-evident but acknowledging that this should be an overriding goal (a matter of debate perhaps) might change how we approach research in the field. Indeed, it suggests the following activities are important:

- Extending existing techniques to promote complementary proof search
- Identifying characterisable sets of problems that are currently difficult to solve

- Developing specialised solutions to such sets
- Exploring methods for identifying and combining diverse strategies
- Developing methods for collaboration between approaches

These are already common activities within the field. However, our position also suggests that some activities are less desirable. For example, it does little to help a portfolio approach to improve an existing method that is not already the best method for a particular set of problems, unless the improvements demonstrably make it such. In addition, improving average solving times is, in itself, not helpful to this goal as it tells us little about the contribution to portfolio solving. Although we acknowledge that there are important research activities complementary to the above stated goal (e.g. advancing deeper theoretical understanding) that require different forms of research.

### 3 (Re)thinking about Evaluation

We acknowledge that taking the above approach can have potential shortcomings that stifle innovation as comparing a new technique against a portfolio approach makes it unlikely that the new approach can be shown to be a direct improvement. However, we believe that this reflects on poor evaluation criteria, rather than the portfolio approach itself. Therefore, we suggest some evaluation activities that could help promote working towards the above identified goal:

- It is important that the benchmark problems used for evaluation reflect the problems that people care about, particularly the hard ones that are difficult to solve
- Evaluation should focus on the contribution of a new technique to the union of possible techniques. This means that:
  - Experiments should focus on short solving times, based on the hypothesis (see below) that if a solution exists a quick solution exists and hence this quick solution should be preferred for a schedule of strategies
  - Results should identify the (number of) unique problems solved by each approach (or subset of approaches)
  - Results should consider how to build schedules of strategies from the experimental results to obtain the best (resource constrained) results overall
  - When comparing sets of proof search parameters it is useful to consider the contribution of a single parameter value to the ‘best’ schedule of strategies
- Where possible, the problems for which an approach is particularly suited, or not suited, should be characterised
- Ideally, competitions should focus on the sets of problems that are unsolved, creating challenges to the community to solve those problems, rather than focussing on how well solvers can tackle problems known to be solvable.

Finally, the natural conclusion of this argument might appear to be that we should have a single prover capable of handling everything and representing the union of the state of the art. This is not our intent. We believe that competition is still an important component in driving research forward. In particular, diverse approaches are those that are most likely to be complementary in the end.

## 4 Quick Solutions Exist (Most of the Time)

The previous discussion refers to a hypothesis that if a solution exists a quick solution exists. We would caveat this with ‘most of the time’ – there are clearly difficult problems that are just difficult however you approach them but it is our experience that the majority of problems (notably theorems) can be solved quickly once they can be solved at all.

We present some evidence for this hypothesis beyond our appeal to folklore but do not claim that this evidence is substantially rigorous. We have inspected a large data set of results kindly provided by Martin Suda<sup>2</sup> running our Vampire theorem prover (for 60 seconds) on all relevant problems in TPTP using 801 different strategies employed in competitions (CASC and SMT-COMP) over the last few years (this experiment consists of significantly more than 10 million runs of Vampire).

Out of 14,722 solved problems, 13,756 (96%) have a solution taking less than 10 seconds, 12,390 (87%) have a solution taking less than 1 second, and 9690 (68%) have a solution taking less than 0.1 seconds. But perhaps these problems are generally easy. However, we also compared the maximum and mean solution times in the following table. Each cell can be read as *When the fastest solution is at most x seconds the mean/max solution takes more than y seconds*. Let us take one cell as an example. There are 8106 problems where some strategy takes more than 30 seconds but the fastest solution takes less than a second. For those problems (more than half of those we solve) picking the ‘wrong’ strategy can make a substantial difference to how long it takes to solve the problem. For around 10% of the problems the mean strategy took over 10 seconds whilst the best strategy took less than a second i.e. on average we would pick the ‘wrong’ strategy.

		Faster solution in at most (seconds)			
		0.1	1	10	
Mean/Max solution takes more than	Mean	1	4224 (29%)	6727 (47%)	-
		10	292 (0%)	1442 (10%)	2476 (17%)
		30	1 (0%)	7 (0%)	137 (0%)
	Max	1	7421 (52%)	9992 (70%)	-
		10	6499 (49%)	9036 (63%)	10228 (72%)
		30	5682 (40%)	8106 (57%)	9158 (64%)

We suggest that this data supports our hypothesis that quick solutions exist and can act as evidence for prioritising the search for complementary approaches. Although we should not ignore those problems unsolved within 60 seconds or those only solved ‘slowly’ (e.g. the 4% or 13% of problems taking more than 10 or 1 second respectively).

## 5 Concluding Remark

In this short paper we have argued for the portfolio approach as it supports what we see as the main goal of automated reasoning research. This argument might seem oversimplified but we hope that it encourages lively discussion at ARCADE and potentially leads to changes in how we evaluate our research.

<sup>2</sup>Please contact Martin for this data.