ELSEVIER

Artificial Intelligence 126 (2001) 1–4

Artificial Intelligence

www.elsevier.com/locate/artint

Editorial

Computational tradeoffs under bounded resources

Eric Horvitz [a,*], Shlomo Zilberstein [b]

[a] *Microsoft Research, Redmond, WA 98052, USA*

[b] *Computer Science Department, University of Amherst, Amherst, MA 01002, USA*

Over the nearly fifty years of research in Artificial Intelligence, investigators have continued to highlight the computational hardness of implementing core competencies associated with intelligence. Key pillars of AI, including search, constraint propagation, belief updating, learning, decision making, and the associated real-world challenges of planning, perception, natural language understanding, speech recognition, and automated conversation continue to make salient the omnipresent wall of computational hardness. Early pioneers in AI research, including Allen Newell and Herbert Simon, established a long tradition of battling obvious intractabilities by resorting to approximations that relied on heuristic procedures—informal policies that appeared to perform acceptably on subsets of real-world problems. *Bounded rationality* was conceived and popularized in the context of sample applications that relied on such heuristic procedures to struggle through overwhelming complexity.

In the mid-1980s, several researchers began to pursue a line of research aimed at better understanding and formalizing tradeoffs under bounded representational and computational resources. During this time, a palpable shift in perspective occurred with regard to tackling resource limitations. Rather than viewing scarce resources as an unfortunate impediment, foiling at every turn attempts to perform automated problem solving on realistic challenges, investigators began to consider tradeoffs under scarce resources as a rich arena for focused AI research. Passionate researchers suggested that elusive principles of intelligence might actually be founded in developing a deeper understanding of how systems might grapple, in an implicit or explicit *resource aware* manner, with scarce, varying, or uncertain time and memory resources. Beyond computational resources, the interaction of limited resources and constraints associated with fixed problem-solving architectures were explored. Older, informal notions of bounded rationality soon gave way to richer, more comprehensive approaches to rational computational and real-world actions that incorporate considerations of resource costs and constraints.

* Corresponding author.
E-mail address: horvitz@microsoft.com (E. Horvitz).

0004-3702/01/$ – see front matter
PII: S0004-3702(01)00051-0

Over the last fifteen years, questions, definitions, and results on computational tradeoffs under bounded resources have flowed from the AI community at varying rates of progress. The research has been motivated by a set of difficult questions including: What is rationality under limited computational resources? How might limited agents compute appropriate beliefs and ideal actions under scarce time and memory? What are ideal architectures for problem solving and learning under uncertain resources? Can we construct procedures that perform in a provably optimal way under limited or varying resources? How can agents enhance the value of their actions by metareasoning about problem solving? What is the best partition of resources between metareasoning and reasoning? How can we best employ memory in compilation of action? What are ideal mixes of off-line compilation and real-time deliberation? What useful abstractions might be manipulated to enhance performance under scarce resources?

Several themes and perspectives have evolved. On the whole, researchers have leaned toward exploring principles and mechanisms for deliberating *about* problem solving, and have focused frequently on the use of metalevel representations and procedures in the design and operation of resource-limited systems. In addition, approaches and solutions to reasoning under bounded resources have underscored the critical role of *uncertainty* and procedures for handling uncertainties about resource availability and the outcomes of computation. Also, there has been an increasing reliance on the principles provided by probability and utility theory, both for providing a normative gold standard for evaluating action under limited information and resources, and for offering a conceptual framework for considering key tradeoffs. As such, the *Principle of Maximum Expected Utility*, derived as a fundamental theorem of decision theory, formulated by Morgenstern and von Neumann in the late 1940s, has been frequently invoked to justify taking actions that maximize measures of the expected value of outcomes of computation. Decision-theoretic formulations underlay several key concepts, including the *expected value of computation*, first described in the mid-1980s. The expected value of computation has served as a formal conceptual tool for guiding the design of algorithms and architectures, and for mediating the real-time allocation of resources in problem solving. Several other themes and approaches appear in work among researchers exploring computational tradeoffs. A number of investigators have explored time–space tradeoffs, alluding to analogous results developed in the Computational Theory community on the relationship between time and space in algorithmic complexity. Researchers have also explored the use of flexible computational procedures, or anytime problem solving—algorithms that seek to elucidate, justify and leverage models of performance that exhibit a relatively smooth surface for making resource allocation decisions.

Theoretical and empirical studies guided by new forms of resource awareness have continued, increasing the AI community's understanding of core problems of tradeoffs under bounded resources, and yielding new principles, architectures, and real-world applications. The papers collected in this special issue reflect a cross-section of current research. Each paper submitted for the special issue was carefully reviewed by two to four expert reviewers. Several papers underwent one or more cycles of revision in response to critical comments provided by experts. The review of manuscripts that included a co-editor as an author was managed exclusively in a discrete manner by the uninvolved co-editor.

In the special issue, Adnan Darwiche describes work on a solution paradigm that provides a smooth tradeoff between time and space for probabilistic inference. Exact and approximate probabilistic inference have long been known to be NP-hard. Indeed, the complexity of probabilistic inference motivated some of the earliest approaches to the formal control of tradeoffs under bounded resources, and led early on to the introduction of notions of decision-theoretic control, expected value of computation, and flexible inference procedures. Darwiche's elegant work on trading space for time for inference introduces new insights into probabilistic inference, and highlights the potential for understanding analogous time-space tradeoffs for a variety of problem classes beyond probabilistic inference. Carla Gomes and Bart Selman explore critical issues and tradeoffs with the use of portfolios of solution procedures to minimize the overall run time of problem solving. Work on the constitution and control of ensembles of solution strategies executed in parallel or in a time-shared manner shows great promise for allowing problem solvers to manage uncertainty and hedge bets about the performance of algorithms in an ideal manner. Gomes and Selman demonstrate the value of algorithm portfolios for grappling with the high variance in performance of randomized search procedures in the context of constraint satisfaction and mixed integer programming. Lev Finkelstein and Shaul Markovitch explore the tradeoff between the time spent on monitoring a computational process and time spent on the base-level computation itself. They present results on optimal schedules for monitoring flexible procedures in the context of several applications. The work is interesting in its primary focus on monitoring policies, but also by analogy to related problems on the ideal partition of resources among multiple stages of computational problem solving. Weixiong Zhang formulates search problems into flexible approximations by introducing methods for reducing the number of nodes under consideration through an iterative node-pruning procedure. He combines the heuristic pruning procedure with a branch-and-bound strategy and tests the methods on three intractable combinatorial optimization problems. The methods trade solution accuracy for tractability, and converge on the optimal analysis when sufficient resources are available. Eric Hansen and Shlomo Zilberstein describe the promise of harnessing dynamic programming methods to monitor and control problem solving of flexible procedures. The research extends work over a decade ago on decision-theoretic control of computation that relied largely on the application of approximations of the expected value of computation in myopic and semi-myopic metalevel deliberation. The dynamic-programming approach focuses attention on the potential for additional study of means for increasing the horizon of consideration in metareasoning. Finally, Eric Horvitz explores an extension of his earlier work on the use of expected value of computation and flexible procedures in metareasoning to problems of *continual computation*—reasoning incessantly about potential future problems, in addition to current challenges that face a computational system. The work pursues the identification of tractable policies and scenarios against the background of an intractable combinatorial optimization problem. As highlighted by applications presented to illustrate key principles, continual computation shows promise for endowing computing systems with the ability to harness all resources all of the time.

We hope that this collection of articles will serve to highlight current research and stimulate new research efforts. We are grateful for the support provided by the editorial board of the journal *Artificial Intelligence* and for the assistance provided by Jennet Batten of the AI Journal staff.