An Architecture for Rationality – Statement of Interest

Piotr J. Gmytrasiewicz
Department of Computer Science
University of California
Riverside, California 92521–0304
piotr@cs.ucr.edu

January 31, 1994

The notion of rationality, defined as a behavior that maximizes the agent's expected utility, is usually taken to pertain to the courses of physical action available to the agent. And while there is no reason why it should not be useful in providing normative theories of activities that are not purely physical like planning, learning, inference, and gathering information, the formalization of the latter types of activities in decision-theoretic terms is not straightforward. The basic difficulty is that these activities are not directly aimed at changing the physical environment, based on which the utility gain could be conveniently expressed, but rather they change the state of the agent itself. Thus, their purpose is more of an indirect effect of *putting the agent in a better position for subsequent physical interaction* with its environment, and it is the general and robust formalization of this effect that remains a challenge.

In our work we have noticed that it is very difficult to even begin to address these issues on a fundamental level without postulating a particular, although possibly quite abstract, architecture for an agent in question. In the framework we have investigated an agent is represented as a following tuple: (IS, K, P, BA, I, R). Below, we briefly describe the components of this representation and the relations among them.

The first element, IS, is the agent's *information structure*, which is the representation of the explicit information the agent has about the past, present and the possible future evolution of the physical environment. IS can be thought of as consisting of a set of branching time lines each of which corresponds to a possible state of the world and its future development. Additionally to possibilistic information, the information structure contains the probabilities that represent the agent's estimate of the likelihoods of the various possibilities actually being realized. Formally, therefore, the information structure is the temporal generalization of the Bayesian Kripke structure. It supports the calculations of the expected utility, as presented in our "Elements of a Utilitarian Theory of Knowledge and Action" (IJCAI-93), including the notion of the value of time, or urgency. As a complete representation of the explicit information an agent has about the outside world, IS plays a central role in the agent's rational choice of the appropriate behavior.

The element K is intended as the agent's knowledge base, which we will take as consisting of a body of implicit, general information about the world. It consists of relations among classes of objects (in the story of Dudley saving Nell from being mashed by a train, popularized by McDermott, these could include the facts that trains are heavy, and that human bodies are fragile), events, and of the causal laws of various kinds (for example, that being mashed by a train causes death). This implicit information can be made to modify and enrich the explicit information about the world contained in the information structure IS during inference.

The preferences, P, represent the assignment of utility to the states of the world. For Dudley, for example, the states of the world in which Nell is dead are not desirable, and less preferable to the states in which Nell is only scared, which are in turn less preferable to her not being negatively affected at all. It is frequently convenient to represent preferences as a hierarchical structure depicting the more general preferences being composed of the more specific ones. For example, Dudley's concern for Nell's well-being could be composed of physical and mental factors, while his preference not associated with Nell may include material factor measured in US dollars. As we mentioned, the preference hierarchy is used to assign desirability to states of the world, and further, to time lines, whole information structures and finally, actions. The explicit inclusion of the preferences in the formal make-up of the agent gives us the ability to talk of the purposeful nature of an agent. The agent's purpose and the driving factor behind all of its undertakings is the maximization of the expected utility specified by the attributes contained in the preference hierarchy.

BA is the set of basic actions that the agent is capable of performing, that is the actions for which no additional planning is required to make them executable. It seems desirable to postulate that all of the computational, sensory, as well as physical actions belong to this set and be treated on an equal footing, although their character and effects differ substantially. Physical actions influence the agent's physical environment, and sensing actions change the agent's knowledge of the physical environment, as represented by the information structure IS. The computational actions are themselves diverse and influence different elements of the agent's state. For example, planning actions maintain and elaborate the agent's intentions I. Inference actions can be roughly described as aimed at transforming the general implicit knowledge contained in the knowledge base K into the explicit representation of the environment in the information structure.

I is the agent's current intention structure, which consists of a planning hierarchy supplemented with the expected utility assignments to plans at the leaves of the hierarchy. The leaf that has the highest expected utility is defined to be the agent's current intention. The agent's intention is, therefore, closely related to the process of planning, to its beliefs about the environment residing in the information structure IS, and to its preferences. In other words, we are postulating that the agent's intentions are to be viewed as rational consequences of the agent's explicit beliefs about the world, contained in IS, and its preferences P. Each of the plans in the intention structure is formed as a sequence of one or more actions, which are behavioral specifications on any level of abstraction, lowest of which are the basic actions of the agent. The plans grow progressively more detailed as they become elaborated during the planning process.

R is the set of reactive rules, application of which relieves the agent from rational deliberation. We take reactions to be compilations of, or short-cuts through, the general but tedious rational decisionmaking processes. The construction and usage of such compilations was investigated before by Heckerman and by Russell, and we will build on this work to integrate reactive behaviors into our framework.

Our basic approach to assessing the utilities of actions and plans can be found in our "Elements of a Utilitarian Theory of Knowledge and Action" (IJCAI-93). In this approach actions are treated as transformations of the agent's information structure IS, and the expected utilities of physical and sensing actions can be computed as a difference between the utilities of IS's after and before a given action. Since inference can be thought of as enriching the explicit information in the information structure using the explicit knowledge in K, we expect that a similar approach can be used to compute the expected utilities of inferencing actions. A remaining challenge is to formalize the utility of the planning actions themselves, which is the focus of our present research.