Biologically Inspired Cognitive Architectures (2014) 9, 105–122

Available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/bica

INVITED ARTICLE

Goal reasoning as a general form of metacognition in BICA

Alexei V. Samsonovich

Krasnow Institute for Advanced Study, George Mason University, 4400 University Drive MS 2A1, Fairfax, VA 22030-4444, USA
Exelis, Inc., 2560 Huntington Ave., Alexandria, VA 22303, USA

Received 15 March 2014; received in revised form 14 July 2014; accepted 14 July 2014

KEYWORDS
Cognitive architectures;
Goal reasoning;
Autonomy;
Intelligent agents

Abstract
A key capability that distinguishes humans from intelligent agents is the ability to generate and select new goals in an unexpected situation, while re-prioritizing existing goals. This level of cognitive autonomy becomes practically vital for actors working as team members with humans or in unpredictable environments. The work is focused on a new perspective on this problem based on biologically inspired cognitive architectures (BICA). Aspects of goal reasoning and goal management capabilities that are currently available in selected BICA are reviewed. A generalizing model of goal reasoning as a form of metacognition is defined, that integrates multiple mechanisms. On top of the traditional cognitive cycle, the model includes a goal-reasoning cycle triggered by notable phenomena. This goal reasoning cycle integrates mechanisms like agitation of desires by drives and their instantiation in goals, or metacognitive reasoning about goals. The model is illustrated by consideration of specific scenarios at micro- and macro-cognitive scales to evaluate the use and potential capabilities of this goal reasoning model for BICA. New capabilities that will be enabled based on this approach are expected to increase the human compatibility and cognitive autonomy of future intelligent agents. The general goal reasoning capability is a key component included in the BICA Challenge and is required for integration of intelligent agents into human teams.

Introduction

To integrate into human teams as teammates, future intelligent agents will need certain general capabilities that will make them human-compatible, and therefore believable and acceptable to humans as reliable partners (Samsonovich, 2012). For example, these artificial actors[1]

E-mail address: asamsono@gmu.edu

http://dx.doi.org/10.1016/j.bica.2014.07.003
2212-683X/

[1] Following a modern trend (e.g., Ghallab, Nau, & Traverso, 2013), the term ''actor'' is used to refer to virtual agents and autonomous robots, as well as to human participants when implied.

will need to create and manage their goals, model and reason about their own and others' beliefs, reason about human emotions, and acquire knowledge about new tasks and environments. This work focuses on goal reasoning, while keeping in mind that all aforementioned components of higher human-level intelligence are mutually dependent and need to be integrated in their study.

Like most modern paradigms in artificial intelligence (AI), goal reasoning does not refer to something radically new in terms of theoretical approaches and methodology, but rather offers an original perspective on familiar approaches. However, this perspective challenges traditional ways of thinking about AI at their foundation. In a traditional understanding, AI tools are developed to aid humans in solving given tasks, specifically, tasks that require human-level intelligence and rationality (Russell & Norvig, 2003). In goal reasoning, this tacit basic assumption is turned upside down. When a goal-reasoning actor is given a certain goal, it may start investigating questions like ''Why should I pursue this goal?'' or ''Should I pursue another goal instead?'' As a result, this actor may abandon the given goal, formulate a different task for itself, and start performing it.

This paradigm may seem strange, useless, and potentially dangerous. At first, it is not obvious that this goal-reasoning actor can assist the user. But perhaps a bigger concern is that uncontrolled autonomous goal reasoning could be harmful. In any case, the task of designing a useful goal-reasoning actor does not seem well-formulated or well-understood, because its definition requires metrics for success. On the one hand, it is difficult to define general metrics for goal reasoning, because it is difficult to construct a fully fleshed-out mapping of actor's states to goals that would guarantee correct behavior. It might be impossible to provide a complete mapping of this sort for a class of real-world situations. On the other hand, if a precise mapping were given, then there would be no goal-reasoning task to solve. And without having precise and objective metrics, it seems impossible to study the problem scientifically.

Yet humans frequently encounter and successfully respond to most practically meaningful and vital goal-reasoning challenges encountered in real life without a priori associated goals or rigorously defined metrics for success. The spectrum of such real-life situations can be illustrated with examples given below.

(1) Imagine that you are entering a grocery store with the goal to buy some food. As you are opening the door (suppose it is not automatic), you see a person approaching from the other side who is carrying groceries in both hands. Will you hold the door for her — or continue immediately toward your goal? Although typical goal-changing behaviors like the one normally expected in this situation could be learned through education (in humans) or pre-programmed (in robots), it would be useful to have an artificial actor that can discover and solve such goal-reasoning situations ''on the fly''.
(2) Imagine that you are a student solving exam problems in a classroom. Suddenly an earthquake starts, and an expensive computer starts falling from the table. What will you do: try to save the computer, run for safety, or continue solving problems? In many cases like this example, the rational analysis and evaluation of the choice of behavior comes afterwards, yet humans can decide what to do immediately, in most cases correctly, and on their own, being driven by common sense.
(3) An important aspect of goal reasoning is metacognition that operates on the goal states represented in the actor's mind. This process may not be triggered by a particular exogenous event. For example, imagine that you try to navigate in the forest in heavy fog, and have no compass. You set the goal to move north, yet keep unexpectedly returning to previously visited landmarks. As a result, you re-evaluate your ability to navigate in the fog and change your current goal accordingly.
(4) In many cases goal reasoning involves learning: not only learning of what goal should be pursued in a given state, but also learning in a broader sense. Imagine that you enter an unknown videogame environment, where you need to discover objects, rules of the game, possible actions and their consequences, and rewards through trial and error, before you can learn a system of values in this environment and dependencies among possible states and various conditions, together resulting in a *hierarchy of goals*, from immediate and concrete targets to long-term and more abstract goals. An artificial actor in a similar situation may need to infer and learn values and goals, moving step-by-step through many levels.

To constructively address these and related general questions, the rest of the paper is organized as follows. Section 'Background on goal reasoning and cognitive architectures' provides relevant background knowledge on cognitive architectures and on goal reasoning, ending with a brief summary of specific goal-reasoning approaches and capabilities implemented in cognitive architecture. Section 'Integrated model of goal reasoning in BICA' attempts to construct a generic, integrated and unifying framework for design and implementation of goal reasoning capabilities in BICA. The general framework is further tested using example scenarios. Finally, Section 'Comparative analysis of goal reasoning capabilities offered by selected cognitive architectures' performs comparative analysis of goal reasoning capabilities available in several selected popular cognitive architectures, and summarizes results in a table. Future promising directions of research in this area and connections to other components of the BICA Challenge are discussed in the last section.

Background on goal reasoning and cognitive architectures

Cognitive architectures

A *cognitive architecture* is a computational framework for designing intelligent agents (Gray, 2007). An agent, or *actor*, here is a cognitive system embedded (not necessarily embodied) in a physical or virtual environment, such that it can perceive information and perform actions to satisfy its needs. A *cognitive system* is a dynamic information-processing system whose elements are functionally related to the semantics of the processed information. Biologically inspired cognitive architectures (BICA) capture principles and mech-