



University of West Bohemia in Pilsen
Department of Computer Science and Engineering
Univerzitní 8
30614 Pilsen
Czech Republic

Metacognition in Intelligent Systems

PhD Study Report

Vilém Beneš

April, 2004

ABSTRACT

This thesis proposal delineates issues about developing artificial intelligent artifacts. Two main trends in AI are overviewed and their fusion is proposed. Then there is presented a general definition of intelligence. The main part of this document is concerned with abstract mechanisms that are designed to provide a substrate for emerging of intelligence. I suggest using a set of virtual environments where an intelligent agent learns universal abilities which working together provide effective behavior. Finally, one mechanism of perception is examined in greater detail. It is the mechanism of situation definition, attempt to solve transduction and meaning problem via bridging the real world and the inner world of symbols using unsupervised graph clustering together with simple reasoning.

keywords: situated cognition, evolution, developmental approach, action selection, behavior, intelligence, situation, uninterpreted sensor, clustering, meaning, perception, category learning

CONTENTS

ABSTRACT	2
CONTENTS	3
PREFACE	5
INTRODUCTION	6
(1) AI	8
1.1. Classical AI	9
1.2. Embodied AI	9
1.3. My Developmental Approach	11
(2) INTELLIGENCE	13
2.1. Ideas About Intelligence in Literature	13
2.1.1. Piaget	13
2.1.2. Sternberg	14
2.1.3. Turing	14
2.1.4. Brooks	14
2.2. Definition of Intelligence	15
2.2.1. Intelligence is Abstract Tool for Achieving Success	15
2.2.2. Only Outcome Matters	16
2.2.3. Ability to Develop and Perform Behavior	17
(3) AGENTS AND ARCHITECTURES	18
3.1. Architectures	18
3.1.1. Subsumption	18
3.1.2. ACR-T	20
3.1.3. Soar	21
3.1.4. 3T	22
3.1.5. Cog	23
3.1.6. General Cognitive Architecture – Development	27
3.1.7. My Architecture	29
3.2. Development and Mechanisms	31
3.3. Environments	34

(4) PERCEPTION	36
4.1. Defining a Situation	36
4.1.1. 2D Environment	36
4.1.2. Regions in State Spaces and New Defined Situation	37
4.1.3. Issues Related to Situation Definition	39
(5) FUTURE WORK	41
5.1. Aims of Doctoral Thesis	42
BIBLIOGRAPHY	43

PREFACE

For centuries we have studied nature and ourselves in order to answer the question: Who are we?

We are humans – animals most likely, but we differ from other species by extensive creating and using tools, we are transforming our surroundings and us to a large degree, we live in a society and have culture, we reflect on ourselves, we have feelings and are conscious – and at top of all – we think, we are intelligent.

Our development is accelerating. Our tools are more and more powerful. Our technology is processing information and helps us to acquire more knowledge. Knowledge is replacement for all other resources. Positive feedback loop is formed. Possessing more knowledge we can design more sophisticated technology and acquire more information and process it better.

We live in the most interesting era, maybe. Our own creations, machines, grow more intelligent and able every year. Time has come to think how the machines are going to redefine us – beating us in thinking and outperforming us in literally everything.

INTRODUCTION

Computer is the greatest invention of all times. It can be used for writing a letter or for calculating the price you will pay for purchasing all things on your shopping list. Computer can play music or movies or you can use it to read a book. Using networked computers you can chat with your friend who is far away, find out how much money you have on your bank account or watch weather satellite images in real time. Computers can be also used to monitor and control all kinds of machines. Computer can be used to design and produce better computers. Computers store and process data and information. You can program computer to do what you want.

Using computers you can simulate what is happening in a real world. You can create new artificial environments and give people an opportunity to experience various kinds of situations. You can program computer to create a world.

Because the computer can do many things, soon after first computers were constructed, researchers began to ask whether you can make computer to do things that are observed on humans, things that are useful, things that are considered to be demonstration of intelligence. This question grew into the whole field of research called artificial intelligence. Many problems and methods of solving them were studied within this field. One example of them is the chess game. To play chess on good level, you need abstract thinking, ability to learn in various ways, ability to anticipate advantages and disadvantages of your choices and many other abilities which smart people manifest. Computer beat the human world champion in a regular match in 1997.

Computer technology gave birth to artificial intelligence and brought platforms that we can use to create both intelligent creatures and rich worlds. Using virtual worlds occupied by acting and sensing creatures to study intelligence is a new promising way. Once again we can see better what is at the end of the tunnel.

This thesis provides broad overview of relevant research fields and presents analysis of directions of next possible research leading to creation of intelligent artifacts and to better understanding the thing called mind. After many years of thinking about thinking, consciousness, mind and intelligence – it seems we are still at the beginning. Our words cannot be easily converted into technical terms and we are still missing something vital, something that will allow us to build universal mind in machine – able to learn, able to be creative, able to help us to accelerate our own development.

In following chapters I will bring to light anything I believe could help to create thinking artifacts. Thesis is structured as follows: Chapter 1 describes the field of Artificial Intelligence and my approach that could lead to developing of universal intelligent agent. Chapter 2 presents definition of intelligence and discusses related issues. Chapter 3 is about agents, architectures and mechanisms of intelligent minds. Chapter 4 provides closer look on low level perception mechanisms.

(1) AI

Artificial Intelligence is endeavor of creating intelligent machines. According to [1], there are two principal goals of artificial intelligence: explaining human mind and creating intelligent artifacts as intelligent as possible - hopefully more intelligent than humans. Both these achievements could have tremendous impact on our society. Neither of them is easy to accomplish.

Research in many branches of science can contribute to progress in artificial intelligence – among others it is ethology, cognitive science, psychology, robotics, computer science, social science, biology, economics, mathematics or philosophy. Artificial intelligence as a field of research is already fragmented into many sub-fields. Reductionist approaches can give us some answers, but there is also need to join efforts to study mind and intelligence as a whole.

Look at the story of study of human body. At the beginning there were opinions that our body is uniform and not decomposable, miraculous thing powered by some sort of unknown energy. Then there were many distinct organs and systems recognized in a human body. And even then we started to ask how the body and many intertwined systems in it evolved through time. And now at the end we are closing the circle – we are realizing that our body is not fully decomposable to its parts – organs or even cells, that there are emerging properties resulting from interplay of simpler systems. This is what we need to repeat in study of mind. We realized that we can find distinct organs in our brains. I think the same could be done for abstract concept of our minds – we can find various simpler mechanisms that working together form a mind. Next we have to fully realize that our thinking was not created, it is evolved instead. It is much more easy to evolve a mind that to design its fully working form. The last step for a moment will be thinking what qualitatively new properties emerged via interplay of recognized mechanisms or organs.

Number of AI applications is growing along with their commercial success. Potentially any machine can benefit from AI mechanisms that are adapting and that allow more efficient work in the environment. AI helps to allocate resources better and to make better decisions, to create better user interfaces, to design better products. AI can improve medical care. AI systems will become our tutors. Possible applications are almost infinite.

In next sections I will review two paradigms of AI – classical and embodied. Then I will argue that successful fusion of both leads to best results, and I will show how to improve and accelerate development of intelligent machines.

1.1. Classical AI

Classical Artificial Intelligence, also known as Good Old Fashioned AI (GOF AI) was developing mainly since 1950 to 1980s. GOF AI was concerned with symbol manipulation techniques operating with representations. There was an underlining idea that pure formal symbol manipulation is both necessary and sufficient mechanism to produce general intelligent behavior [2].

Many problem solving methods were designed. Search, planning, pattern recognition, rule induction, genetic algorithms are few examples of those. GOF AI provides a rich source of control ideas – when interpreted via the medium of human. Problems arose when these control paradigms were applied to robotics, in particular to control autonomous mobile robots. There was unclear how to transform reality into inner world of symbols. Noise in sensors and maintaining model validity are serious issues here. Another serious issue is difficulty of reasoning about the effects of low-level actions. Additionally, accurate plans of classical AI are often both unnecessary and unachievable due to inaccurate real world sensors and motors [4].

Pure symbolic manipulation was proven not to be sufficient for general intelligence and there became apparent that understanding system-environment interaction is fundamental.

1.2. Embodied AI

Embodied AI, Nouvelle AI, New AI are newly coined terms for AI studying embodied cognitive systems. One of main characteristics of Embodied AI (EAI) is its investigation of system-environment interaction. It came to light that GOF AI is cursed with meaning of symbols. GOF AI systems were not designed to retrieve new symbols from the world they lived in, to define or redefine their meaning. Creating of symbols and assigning meaning to them was left to human designers. This leads to bad performance of classical AI systems in real world applications.

Embodied cognitive science is right about the fact that meaning of things comes out of our abilities – our possible actions in the environment. Consider a meaning of the word “river”. River can be an obstacle, source of water or fishes, transport route, temperature regulator, hiding place, orientation and navigation landmark, place of recreation or something that bring the danger of flooding or drowning. The way you cogitate river depends on our body, on the way you can act in the environment.

There was coined a term ‘agent’ for anything that is situated in the environment – anything that is acting and sensing in the world. Agents are embodied [3] and are sometimes said to be ‘structurally coupled’ (via perturbatory channels) with the environment. Chapter (3) of this thesis is concerned with agents and agent architectures. There are various kinds of agents – software agents, mobile robots or humans.

Embodied AI was designated while studying of one kind of agents – mobile robots. One of the developers of mobile robots is Rodney Brooks, leader of EAI research. He develops mobile robots based on his *subsumption* architecture – agent architecture that implements loosely-coupled layers each producing behavior from most simple to more advanced ones. Brooks advocates the behaviorist approach and he is arguing that mobile robots need no symbolic manipulation mechanisms at all.

1.3. My Developmental Approach

Classical AI as researched till 1980 suffers from inability to successfully assigning meaning to symbols and has problems with maintaining useful model of outer environment. Embodied AI is an attempt to show how intelligence originates in sensorimotor interactions with environment, but explaining high level mental functions in terms of low level interaction with environment may prove to be very difficult. Although I found Brooks' no-symbol approach useful for various reasons, I feel that the holy grail of AI lays in fusion of classical and embodied paradigms.

My approach to create of intelligent agents is characterized by three principles :

- generality
- evolution; self-organization; various mind-body(-environment) loops
- using (sequences of) virtual environment(s)

One of the attributes of intelligence is its general use. Intelligent agent should be able to adapt on various environments, to develop new useful methods and tweak old ones. In order to be able to manifest these abilities, the agent needs to have as general inner workings as possible.

It is unfeasible for a human to design an intelligent agent at once. This is why AI failed to produce intelligent agents. What we need here is a collection of general enough methods, that working together will exhibit self-organization process forming a thinking mind. How to facilitate this is unclear yet, but there is clear evidence in the nature that evolution can construct advanced things using only simple mechanisms. Evolving and emergence of mind is so strange for us due to unfamiliarity of general enough and rich enough control mechanisms and because we do not understand properties of the low level processes of perception. Mechanisms that should be incorporated into mind-evolution are proposed in chapter (3) of this thesis. Low level perception and emergence of meaning will be discussed in chapter (4).

So, hypothesize we have a bunch of general mechanisms that could, working together, eventually evolve into an intelligent mind. How this evolution should look like? What are the best conditions for fastest development of (most) intelligent agent? My way of answering this question is that virtual environments are most promising. The reason for this is that virtual environments are fully comparable to real ones and have additional advantages. Any relevant phenomena from real world can be seen in virtual environments too. Among these phenomena we count: only partial observability, noise in sensors, frequent changes, unpredictability, inexact motors and actuators, nonlinearity and chaos. Virtual environments have additional advantages – as a designer you can observe the virtual world completely. You can repeat the “simulation” and observe effects of small initial changes. You can design the environment precisely to study concrete aspects of agent evolution [5] [6].

Virtual environments will benefit from exponential computer speed growth. Unlike mobile robots operating in reality, virtual environment do not need any expensive equipment and run on any standard computer. And last but not least – there is fast growing industry concerned with creating complex virtual environments. Industry that makes more money than movie industry and that attracts many young people – industry of computer games [7] [8] [9] [10] [11].

Current state of the art artificial intelligence will soon appear in computer games and it will make the difference between the bad and the good ones. On the other hand science could benefit from money earned in gaming industry and new ideas relevant to AI research coming from people playing computer games.

(2) INTELLIGENCE

In this chapter I want to discuss the notion of intelligence. Every one of us feels in some way that the thing called intelligence makes a difference. But what it really means to be intelligent? What abilities you need to have to be considered intelligent. How you should differ from others to be considered more intelligent? Is being intelligent useful? Are animals intelligent and is it reasonable to speak about intelligent machines? I would like now to present ideas of other authors and then conclude my definition of intelligence, answers to mentioned questions and further thoughts about intelligence.

2.1. Ideas About Intelligence in Literature

2.1.1. Piaget

Jean Piaget was a Swiss developmental psychologist who dedicated considerable portion of his life to study of mental development of children. He is best known for organizing intellectual development into series of stages – levels of development roughly corresponding to infancy, childhood and adolescence [12].

The sensorimotor stage begins at birth, and lasts until the child is approximately two years old. At this stage, the child cannot form mental representations of objects that are outside his immediate view, so his intelligence develops through his motor interactions with his environment.

The preoperational stage typically lasts until the child is 6 or 7. According to Piaget, this is the stage where true “thought” emerges. Preoperational children are able to make mental representations of unseen objects, but they cannot use deductive reasoning.

The concrete operations stage follows, and lasts until the child is 11 or 12. Concrete operational children are able to use deductive reasoning, demonstrate conservation of number, and can differentiate their perspective from that of other people.

‘Formal operations’ is the final stage. Its most salient feature is the ability to think abstractly.

A central tenet of Piaget's theory is that increasingly complex intellectual processes are built on the primitive foundations laid in earlier stages of development. An infant's physical explorations of his environment form the basis for the mental representations he develops as a preoperational child, and so on.

2.1.2. Sternberg

Robert J. Sternberg is a professor of psychology at Yale University. In [13] Sternberg asks, *What is successful intelligence?* and provides the following definitions:

- The ability to achieve success in life in terms of one's personal standards, within one's socio-cultural context
- One's ability to achieve success depends on capitalizing on one's strengths and correcting or compensating for one's weaknesses
- Balancing of abilities is achieved in order to adapt to, shape and select environments
- Success is attained through a balance of analytical, creative and practical abilities

2.1.3. Turing

Alan Turing was mathematician, philosopher, codebreaker and the father of computer science. He replaced the question 'Can machines think?' with question whether the machine can win an imitation game [14]. In other words – if machine can fool an interrogator and successfully pretend to be a human being. Turing's imitation game is now widely known as Turing's test of intelligence.

Turing's test shows us that the only important thing is the outcome. We do not need to be concerned with the inner workings of the artifact or organism being under examination. If it displays outcomes interchangeable with outcomes produced by intelligent artifacts or organisms – it is intelligent.

2.1.4. Brooks

Rodney Brooks is professor of computer science, director of the MIT Computer Science and Artificial Intelligence Laboratory. He is leading person in behavior-based robotics research.

In [15] Brooks shows a timeline of the chief steps of historical human development. He points out that evolution took 3.5 billion years to create man and that man was developing for another 2.5 million years of which only the few hundred years are the era of "expert" knowledge and continues:

"This suggests that problem solving behavior, language, expert knowledge and application, and reason, are all pretty simple once the essence of being and reacting are available. That essence is the ability to move around in a dynamic environment, sensing the surroundings to a degree sufficient to achieve the necessary maintenance of life and reproduction.

This part of intelligence is where evolution has concentrated its time—it is much harder.” [15]

Brooks regards defining intelligence as slippery concept, but claims that “Intelligence is determined by the dynamics of interaction with the world.” [16]

2.2. Definition of Intelligence

My definition of intelligence:

“Intelligence is the ability to develop and perform effective behavior.”

Next sections explain meaning of this definition. Section 2.2.1 shows direct link between collecting high reward (achieving success/being effective/utility/value) and intelligence. Section 2.2.2 explains “only outcome matters” notion of intelligence. References to human ways of doing or thinking or references to knowledge, learning etc. are not needed to be stated explicitly in the definition of intelligence. Section 2.2.3 discusses developing and producing behavior.

2.2.1. Intelligence is Abstract Tool for Achieving Success

Sternberg coined *successful intelligence* – he speaks about achieving success in life in terms of ones personal standards, within ones socio-cultural context. I think that success or effectiveness is the fundamental attribute of intelligence – intelligence is an abstract tool for achieving of success. Agent’s success is defined by reward. Reward is a component of environment that is passed as a value to the agent. It is difficult to say what this reward is in case of human being. Humans like other advanced agents construct inner reward policies – methods of rewarding self. There is a reason for developing these inner reward policies. Positive reward from environment comes often in very rare but high-valued chunks, and during the rest of time agent is usually rewarded with small values or is not rewarded at all. Inner rewarding policies are ensuring that the agent rapidly develop useful ways of behavior that lead to mentioned high rewards.

Consider a chess game as an example. When you are playing chess game you need to make tens of moves until you receive first (and last) reward. Reward in chess is the final verdict at the end of the game – you win, you loss or you draw. This reward is relatively rare – considering work you need to carry out during the game. We can assign values to final results, but the question is how your behavior during the game influenced these results. What your ways of playing you should

consider to be good? What aspects of your playing you should consider beneficial¹? After you play some games you realize that one of wise things is to preserve your pieces. There are interesting counterexamples, but generally it is good to keep strength of your pieces bigger than strength of pieces of your opponent. So you transform the final win/loss/draw reward into more concrete reward policy that rewards you positively for taking opponents pieces and negatively for losing yours. This reward policy advises you in more concrete way what you should do.

To make things more knotty – you have a reward policy that is rewarding you if you find an interesting counter example against the first mentioned reward policy – in chess there is occasionally very commendable to sacrifice your most valued pieces to gain other advantages in the game. Good chess players are evaluating many aspects in the game at once – strength of pieces, mobility and influence of these pieces, safety of precious pieces and others.

Humans have many intertwined reward policies – some of them reward you for learning or exploring, some of them reward you for your reproductive behavior, there could be some that leads you eventually to sacrificing your life. Plentiful inner reward policies prevent us from seeing the primal reward from environment, but I claim there is such a thing.

2.2.2. Only Outcome Matters

According to Turing inner processes of intelligent agents are unimportant, what matters is their behavior – ways agents are using to change environment they lives in.

You take two agents and compare their results – their behavior in the environment and obtained reward. If this reward (on average) equals for both they are equally intelligent.

This leads us to discovery that can be little difficult to understand – when examining intelligence, mind and body can never be considered isolated. Think about this setup: environment is real world – say jungle, rewarded is who prevails in this environment longer. Here you manage to be eaten by a bear – so this bear has to be considered more intelligent than you. These thoughts originated in thinking about virtual agents – body of virtual agent is more easily understandable as being one with mental processes than in the case of non-virtual agents.

Implication for our reality is this: intelligence of beings is “caused” or “developed” partly by evolution of their bodies. If you have better body and

¹ In other words - what situations you should try to create in the game? What are your subgoals? What is worth to think about? What is worth to perceive in the current game state. What are wise things to do considering particular state of the game? How you should behave in certain situations?

“comparable” other “components of intelligence” there is a good chance that you are more intelligent. Two “things” we denote as “mental processes” and “body” forms one system, where the first from the latter cannot be separated.

Next fact revealed using “only outcome matters” is that intelligence depends on environment (including ways of reward) where we are examining it. In jungle bears can be more intelligent than you, in environment where you have access to machines like guns or vehicles you will be probably more intelligent.

What is the real frontier of our body? Should not we consider our machines to be part of our body too, we control them the same way as parts of our own body. Mind, body and environment have to be considered together when thinking about intelligence.

2.2.3. Ability to Develop and Perform Behavior

Prerequisite for intelligence is an ability to change environment you live in by performing actions. Sequence of actions along with their overall effect can be abstracted and handled as a behavior. Behaviors are thence interesting sequences of actions, and their use typically depends on perceived facts.

Agent’s effectiveness in any environment depends on its ability to develop sophisticated behaviors. There are some general features of sophisticated behaviors:

- lead to immediate reward from environment
- lead to exploring of environment
- lead to exploring possibilities of an agent
- save resources, utilize resources rationally
- lead to development of other effective behaviors

Usually used abstract parts of intelligence: creativity, awareness, experience, skills, adaptation, global optimization, can be explained as abilities in terms and features of developing and performing of effective behaviors.

Various mechanisms that are cooperating on development of behaviors will be outlined in section 3.2 along with description of gradual development of the whole agent from tabula rasa state.

(3) AGENTS AND ARCHITECTURES

In this chapter I will be concerned with intelligent agent architectures. The questions are: *What parts working together can form an intelligent whole?*, *How these parts work?* and *How to connect them?*. Intelligent agent architectures are also known as cognitive architectures or robot control architectures.

I will present a brief overview of Subsumption, ACR-T, SOAR, 3T, and Cog cognitive architectures in section 3.1. Then I will present general view of development of an intelligent agent in previously unknown environment, I will outline various *mechanisms* that are believed to appear in a working architecture of an intelligent agent (section 3.2).

Look at intelligence depends on chosen environment. There are various kinds of environments that differ fundamentally in their qualities, resulting in diverse levels of complexity. Section 3.3 presents taxonomy of environments and speaks about the relation of environment kind and inner machinery that an agent needs to successfully cope with this kind of environment.

3.1. Architectures

3.1.1. Subsumption

Classical control scheme for industrial robots is a sense-model-plan-act scheme. Data from sensors are processed in several sequential steps (see Fig. 1).

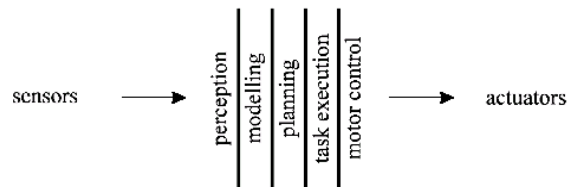


Fig. 1: Classical sense-plan-act control scheme.

This classical sense-plan-act scheme was tried in a domain of autonomous mobile robots with little success. It seems that the major drawback of sense-plan-act architecture is its inability to produce swiftly simple reactive behaviors.

Rodney Brooks follows reactive behavior paradigm, he is designing altogether successful mobile robots based on his ‘subsumption’ architecture (see Fig. 2, below) [17]. In this approach, components of the architecture are composed in layers, from bottom to top. Low layers are performing basic behaviors like object avoiding and

wandering, the lowest layer is grounded in the robot’s sensors and effectors. High level layers are performing high level behaviors while using lower layers for simple task handling. Newly added components and layers exploit the existing ones hence the subsumption architecture. For choosing the appropriate behavior producing layer there is a priority based arbitration performed. Data from sensors are processed concurrently. Each layer is designed not to be strongly coupled with other layers. This provides modularity and testability of the architecture and robustness to failure.



Fig. 2: Subsumption architecture (Brooks).

Brooks tell us that using properly designed architecture based on sensor-triggered behaviors, you can avoid placing explicit planning and modeling components into your architecture. He says: “World is its own best model”. If the world can provide the information directly (through sensing), it is best to get it this way, than to store it internally in a representation (which may be inadequate – large, slow, expensive and outdated).

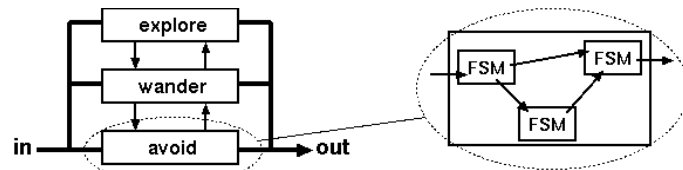


Fig. 3: Subsumption architecture. Individual behaviors are composed of finite state machines.

3.1.2. ACR-T

ACT-R is cognitive architecture which resulted from researching of various human cognitive functions (see Fig. 4). These including perception and attention (visual search, eye movements, task switching, etc.), learning and memory (list memory, skill acquisition, category learning, updating memory etc.), problem solving and decision making (tower of Hanoi, spatial reasoning, game playing, etc.), language processing (parsing, analogy and metaphor, sentence memory) and other (cognitive development, emotion).

The ACT-R theory admits three basic binary distinctions. First, there is a distinction between two types of knowledge – declarative knowledge of facts and procedural knowledge of how to do various cognitive tasks. Second, there is the distinction between the performance assumptions about how ACT-R deploys what it knows to solve a task and the learning assumptions about how it acquires new knowledge. Third, there is a distinction between the symbolic level in ACT-R that involves activation-based processes that determine the availability of these symbolic structures [18].

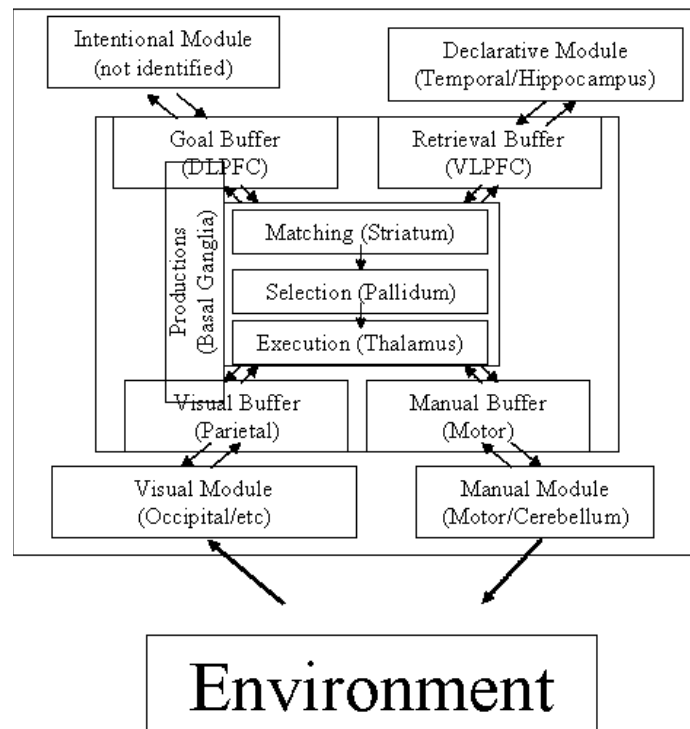


Fig. 4: Cognitive architecture ACR-T 5.0.

Declarative knowledge stores facts and is organized as a network of small units called chunks. Procedural knowledge captures various cognitive skills, such as

mathematical problem-solving skill. Procedural knowledge is stored in the form of production rules – condition-action units which respond to various problem-solving conditions with specific cognitive actions. The steps of thought in a production system correspond to a sequence of such condition-action rules which execute one after another. Declarative and procedural knowledge is based on symbolic representations, however in ACT-R there is an intricate set of subsymbolic computations determining which chunks and productions come to mind (if at all).

ACT-R assumes there are real categories of things in the world based on sets of features. Rational categorization is a task of predicting the probability that an object will display a specific feature given that it displays certain other features. There is no architectural primitive in ACT-R that is performing categorization. There are two ways how to try to solve problem of categorization. First, ACT-R can try to retrieve an example similar to the test stimulus and categorize the test stimulus with whatever category is stored with the example. This is the example of using declarative knowledge. Alternatively, ACT-R can also take a rule-based approach and use production rules to process the values on the dimensions as voting for one category or another. This will rely on procedural memory. There is also evidence that category learning may result from mixture of different methods of categorization. The Exemplar-Based Random Walk (EBRW) and Rule-plus-Exception (RULEX) models of categorization and their possible combinations are described in [19].

3.1.3. Soar

Soar is agent architecture based on production system proposed by Newell and his colleagues. It is based similarly to ACT-R on production system. Unlike ACT-R, Soar attempts to model cognitive function with relatively simple universal mechanisms of using production rules, subgoaling and learning by *chunking*. All behavior is seen as occurring in a problem space, made up Goals, Problem spaces, States and Operators.

```

sp {waterjug*propose*fill
  (state <s> ^name waterjug
    ^jug <j>)
  (<j> ^free > 0)
  -->
  (<s> ^operator <o> +)
  (<o> ^name fill
    ^jug <j>)}

```

Fig. 5: Soar – production rule.

At each step during execution (see Fig. 6), productions rules are *fired* as long as there are rules with satisfied conditions. During this phase are proposed operators

that will be eventually used as agent's actions. Arbitration mechanism is then used to choose one operator to carry out.

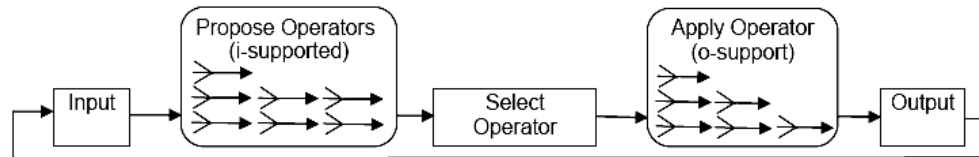


Fig. 6: Soar – Execution cycle.

3.1.4. 3T

3T architecture separates the general robot intelligence problem into three interacting layers or tiers and is thus known as 3T [21]:

- A set of hardware-specific situated skills that represent the architecture's connection with the world. The term "situated skills" is intended to denote a capability that will achieve or maintain a particular state in the world.
- A sequencing capability that can activate the situated skills in order to direct changes in the state of the world and accomplish specific tasks. This tier of architecture is implemented using Reactive Action Packages [22].
- A deliberative planning capability which reasons in depth about goals, resources and timing constraints.

The architecture works as follows: the deliberative layer takes a high-level goal and synthesizes it into a partially ordered list of operators (tasks). Each of these operators corresponds to one or more Reactive Action Packages (RAP). RAP is in [22] proposed as the basic building block of situation-driven execution system. RAP is representation that groups together and describes all of the known ways to carry out a task in different situations. Each task in situation-driven execution system has condition that indicates end of the task – until the success is reached, methods included in given RAP are continually used to try to transform world into desired state. This approach handles possibility of method failures caused by various real world reasons including uncertainty.

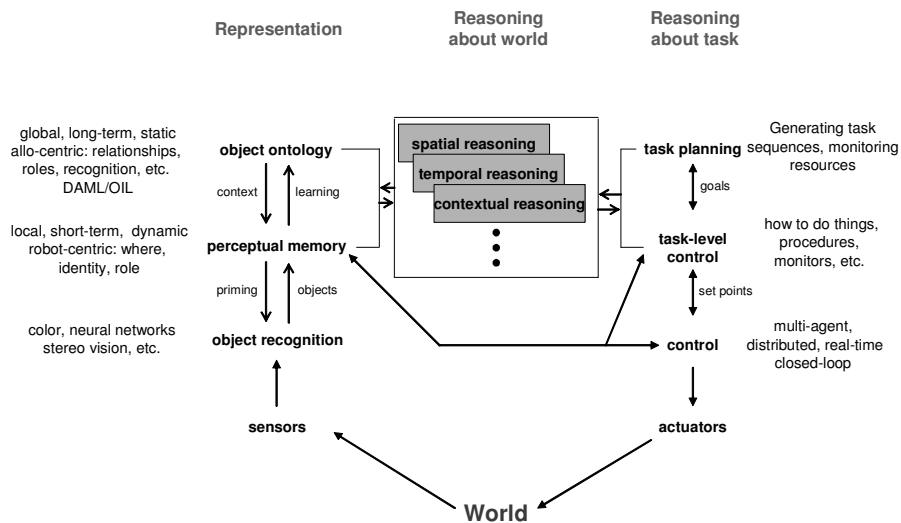


Fig. 7: 3T architecture.

Situation-driven execution basis:

- tasks contain sense methods
- methods in tasks are chosen appropriately to current situation
- tasks have their own success measures

3.1.5. Cog

Cog is a real humanoid robot, physically interacting with people, designed on MIT. There is an underlining idea to create a learning system in which actions and states are learned/learnable entities, not hard-coded primitives.

There will be overviewed a toy-finger robot example from [23]. This robot architecture consists of *meso* – virtual musculature, *pamet* – controlling mechanisms, *tactile* – tactile sensor module and *emotion* – module connected to tactile sensor, emotion is source of reward (see Fig. 8 Cog architecture and Fig. 9 toy-finger robot).

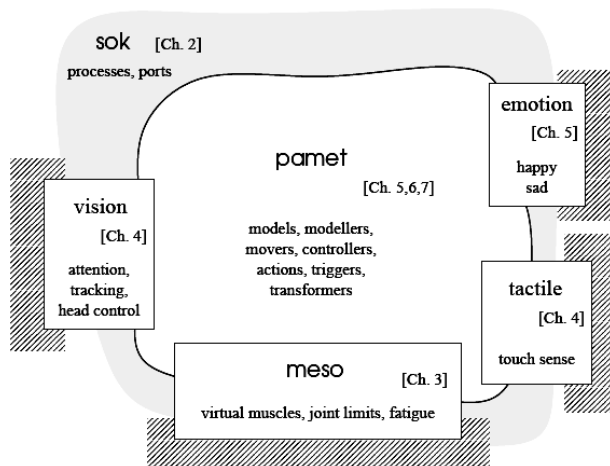


Fig. 8: Cog architecture.

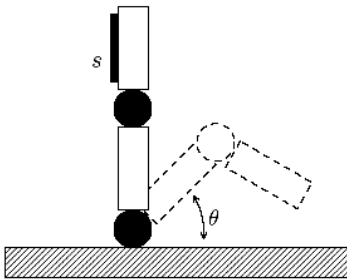


Fig. 9: Cog – A toy robot example. It is just a finger with a single actuator and a single tactile sensor. Its position is fully determined by the single value θ , the sensor yields a single value s .

Figures Fig. 10, Fig. 11 and Fig. 12, show a sequence that is illustrating how the robot explores its capabilities and outer environment. The sequence has three stages:

- stage one: creating a model of effects of two movers – curl and extend; new controller is derived and a new action is defined (Fig. 10)
- stage two: an *actor* is created – potential of the new controller derived in stage one is explored and its link to reward is investigated (Fig. 11)
- stage three: a *trigger* is created – trigger is activating curl-extend actor according to tactile sensor, when there is a reward expected (Fig. 12)

In other words, stage one can be describe as learning to move, stage two as learning what to do, and stage three as learning when to do it. We see the birth of new reactive behavior – from exploring actuation possibilities and investigating their influence on reward, to defining triggering mechanisms that initiates new actions according to perceived situation.

The actuators are not actually completely subservient to external activation; they also have internal random activation. At random times, a mover will activate itself for a random duration and with a random activation level. The rate of random activation is initialized to some maximum value when the mover is created. As the mover is activated more and more by external sources, this rate of internal activation decreases.

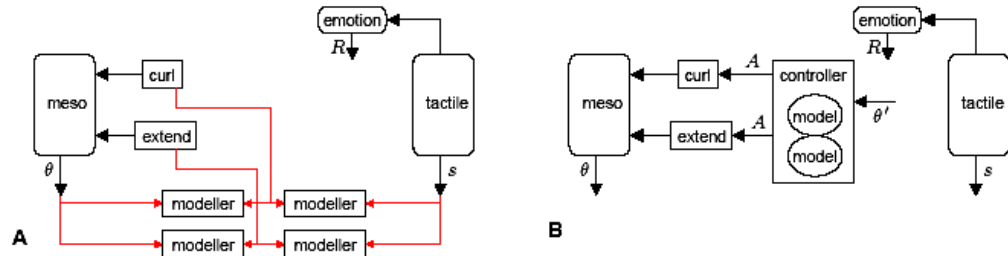


Fig. 10: Cog – The robot has two movers, one for curling and one for extension. (A) Each is observed by mover modellers, to see what effect, if any, they have on θ and s . (B) Two models relating the movers to θ are learned, and a controller for the velocity θ' is created.

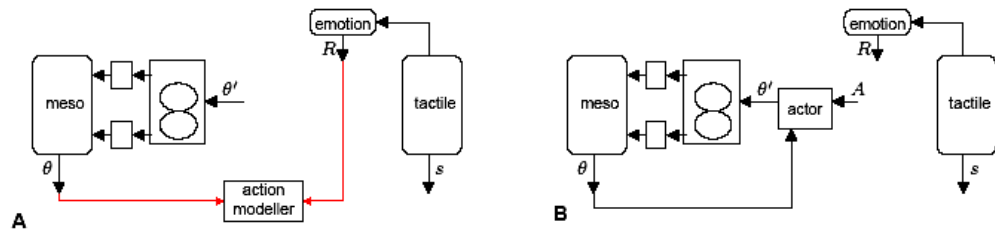


Fig. 11: Cog – (A) An action modeller is generated to find rewarding functions of θ . (B) After training, an actor module is created. When activated, it drives θ to a prototypical value θ_0 via the controller.

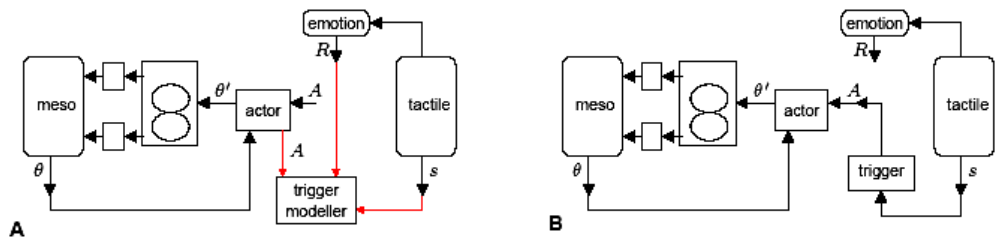


Fig. 12: Cog – (A) A trigger modeller is created to discover rewarding conditions (values of s) in which to activate the action. (B) After training, a trigger module is created, which activates the action when the sensor value s is close enough to a prototype s_0 .

3.1.6. General Cognitive Architecture – Development

Franklin in [24] derives general cognitive architecture (see Fig. 13) from works of Kanerva (sparse distributed memory), Drescher (schema mechanism), Maes (behavior networks), Jackson (pandemonium theory), Hofstadter and Mitchell (copycat architecture).

Here are questions and answers that Franklin presents as overview of latest opinions related to embodied cognitive architectures:

Q1) "Is it necessary for an intelligent system to possess a body...?"

Q2) "What are necessary elements of embodied architectures?"

Q3) "What drives these systems?"

Q4) "How are we to proceed in a science of embodied systems?"

Q5) "How is [meaning] related to real objects?"

Q6) "What sort of ontology is necessary for describing and constructing knowledge about systems?"

Q7) "Which ontologies are created within the systems...?"

A1) Software systems with no body in the usual physical sense can be intelligent. But, they must be embodied in the situated sense of being autonomous agents structurally coupled with their environment.

A2) An embodied architecture must have at least the primary elements of an autonomous agent, sensors, actions, drives, and an action selection mechanism. Intelligent systems typically must have much more.

A3) These systems are driven by built-in or evolved-in drives and the goals generated from them.

A4) We pursue a science of embodied systems by developing theories of how mechanisms of mind can work, making predictions from the theories, designing autonomous agent architectures that supposedly embody these theories, implementing these agents in hardware or software, experimenting with the agents to check our predictions, modifying our theories and architectures, and looping ad infinitum.

A5) Real objects exist, as objects, only in the "minds" of autonomous agents. Their meanings are grounded in the agent's perceptions, both external and internal.

A6) An ontology for knowledge about autonomous agents will include sensors, actions, drives, action selection mechanisms, and perhaps representations, goals and subgoals, beliefs, desires, intentions, emotions, attitudes, moods, memories, concepts, workspaces, plans, schedules, various mechanisms for generating some of the above, etc. This list does not even begin to be exhaustive.

A7) Each autonomous agent uses its own ontology which is typically partly built-in or evolved-in and partly constructed by the agent.

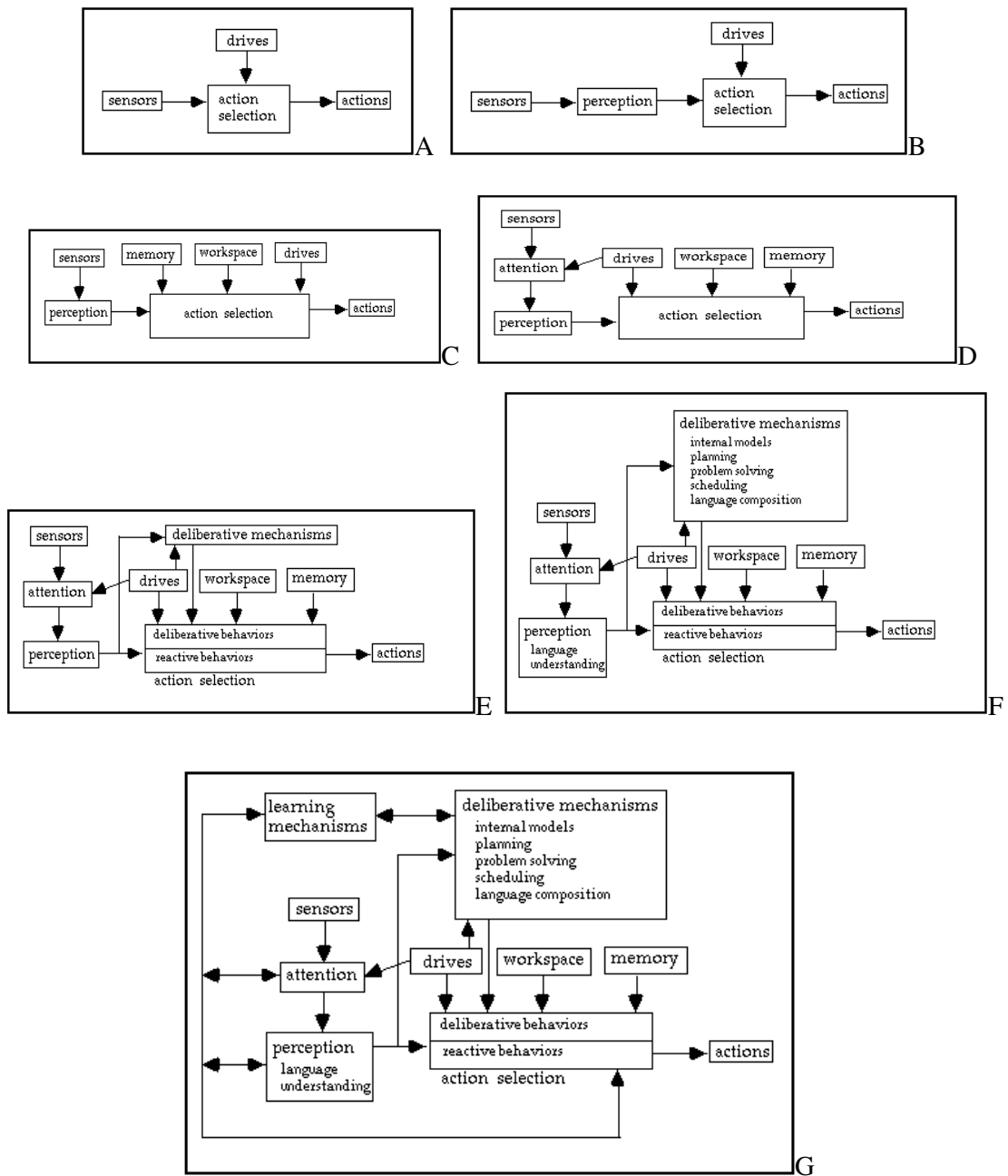


Fig. 13: General cognitive architecture – development.

3.1.7. My Architecture

In this section I will present my general agent architecture. Emphasis is given on ability to support various mechanisms that working together provide substrate for emerging of intelligence. These mechanisms are abstract way of describing inner working of agent, aspects of cognitive functions we observe in human performance, too. Among these mechanisms I count various tricks, heuristics and skills that contribute to effective function of perception, categorization and world modeling, behavior selecting and others. More detailed view on mechanisms is presented in section 3.2.

My architecture presented on figure Fig. 14 incorporates Brook's ideas about creating reactive behavior without much world modeling and symbolic processing. Elaborated behaviors can be learned, optimized and then automatized and so they can be triggered by perceived circumstances and performed without intervention of deliberative mechanisms.

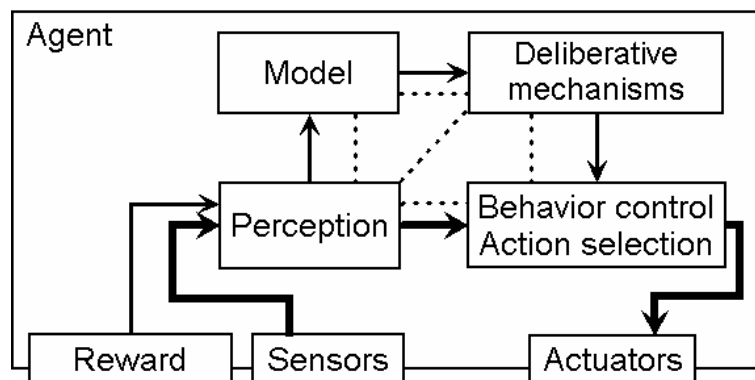


Fig. 14: My architecture.

On the other hand, this architecture incorporates symbol manipulation abilities. High level concepts used in perception, world model and deliberation parts of architecture are represented via symbols. Symbols are useful abstract and compact descriptions of circumstances, but they have to be grounded somehow in agent's actuator capabilities.

Very important functional part of the architecture is hence transduction from non-symbolic sensory data to symbols and then from symbols to actions. This very transduction is described via a set of mechanisms, one of them is unsupervised graph clustering of sensory data. Interesting patterns and features are detected in sensory data, their correlations with received rewards and with effects of previous and future actions are investigated and stored as world model. There should be various densely interlinked ways how to detect and represent things.

Next important sign of the structure of intelligent agent is omnipresent self-organization. Among other things self-organization apply to allocating inner/outer

resources; to choosing, verifying and refinement of inner methods; to developing of inner functions based on previous evidence about processes in environment. It is necessary to maintain history of past performance of the system and of things that influenced self development of the agent. When something goes wrong, the agent can do self-reflection and retrospection to find out where the flaw is and where it originated.

Self-organization is manifested apart from other things by agent’s tendencies to automatize its inner processes and thus save computational resources. Deliberative mechanisms are examining themselves while working on problems in environment and parts of thinking that can be performed automatically are transformed into reactive behaviors.

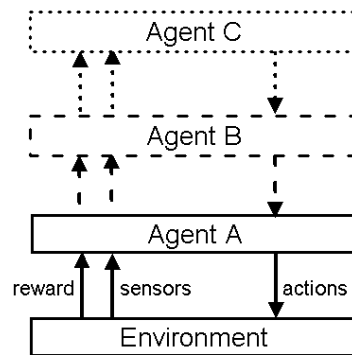


Fig. 15: Recurrent control hierarchy.

Fig. 15 shows control hierarchy. Abilities that agent uses directly (translated into actions in environment) can be also used to improve agent’s overall performance. It is done by creating recurrent hierarchy. Agent A is said to be new environment where operates new copy of the same agent – agent B. B observes A’s inner working and has no direct connection to outer environment. B’s actions are actions that change A’s inner working – like choosing methods and their parameters or allocating A’s resources. When passed reward from outer environment, B can tune A’s performance, prevent it from doing dull loops, or perform high-level learning. Consider similarity between A’s investigating of effect of an action in environment and B’s investigating of effects of alternating A’s methods – very similar process, different level.

Recurrent hierarchy is supposed to produce advanced behavior while using relatively simple architecture. We can expect that new skills acquired by A while working in environment can probably improve B’s performance and vice versa.

Recurrent hierarchy is based on Minsky’s A-brain, B-brain idea [26]. Similar ideas are expressed by SASE (Self-Aware Self-Acting) paradigm for agents.

3.2. Development and Mechanisms

We should take a closer look at development in unknown environment. Intelligent agent is an adaptive entity. To be intelligent, it is needed to be able to observe reward coming from environment and to conclude how the reward is linked to features that can be influenced and sensed by the agent. Questions about reward, influencing features and sensing features are fundamental. I am going to outline a set of abstract mechanisms, which could be identified in intelligent agent mental machinery.

Table 1. Inner mechanisms of an intelligent agent.

<ol style="list-style-type: none">1. performing random actions2. observing utility of actions (reward they resulted in)3. performing better action more often4. searching for goals (states in sensor data with high reward)5. composing actions into macroactions (abstract actions, behaviors)6. discovering effects of actions7. situation definition (perception – abstract sensors)8. deriving rules that describe effects of actions and situations (model, category learning)9. planning (constructing useful behaviors, optimization of actions being performed in terms of reward, anticipating)10. definition of inner actions and inner sensors of the agent => new environment for new copy of the agent (building control hierarchy)11. correction mechanisms – dull loop detection, model verification, model consistency checking, improving predicting abilities12. learning and adaptation aids – abstraction, model simplifying, lower complexity examples discovery (dimensionality reduction, sandbox, fishpool), hypotheses proposing, experiment conducting, investigate best/worst scenarios

Table 1 summarizes abstract mechanisms that are believed to take part in thinking machinery of intelligent agents. They are sorted from most simple to more advanced. All mechanisms are working simultaneously. However, advanced mechanisms need to use outcome of simpler mechanisms – so, after appearing in completely unknown environment, the agent will start to employ these mechanisms approximately in order how they are listed in the table.

Let's suppose that an agent appears in a new and completely unknown environment and that this agent has no experiences from other environments. The only thing that the agent is aware of, is a list of actions that can be performed in the environment and actual sensory data. At this moment, the agent is fully capable of sensing, but the sensory data gives absolutely no sense. There are no colors, shapes or patterns of any kind that could be identified and used now.

First phase of agent's behavior in environment is not dependent on sensory data, agent performs random actions.

Next step is tracing the influence of performed actions on reward. Some actions are better in terms of reward and it is commendable to use them more often (but not to use them all the time). While observing which actions get good reward,

the agent can also conduct discovery of correlations between good rewards and certain forms of sensory data or certain patterns in sensory data.

If interesting correlations between sensory data and reward are found it is then useful to investigate effects of actions (the way they change observed circumstances in environment). After understanding how certain actions influence environment and hence sensory data, this insight can be used for attempt to guide the environment into states with high reward. This process is called *defining situation* – the agent defines a state space partition. There is bounded a region in state space that is interesting somehow. For this region and for the rest of the sensory space are then derived two different behaviors. This mechanism of defining situation gives ability to recognize objects, features and processes and opportunity to construct more effective behavior. Note that dividing sensory state space into region and the rest leaves no uncovered places in state space. Behavior is designed separately for the interesting region and then for the whole rest of the sensory state space. This brings robustness to agent's behavior. Defining a situation is a fundamental mechanism for intelligence, which allows meaning to be assigned to patterns in sensory data. More detailed view on this mechanism is presented in chapter (4)).

To this point, mechanisms 1 to 8 from Table 1 were described. As I have mentioned, all these mechanisms work simultaneously. For example when new situation is defined, identified region in sensory state space could represent so different state of environment from average considered previously, that the best thing to do here is to perform random actions again (mechanism 1). Consider an agent learning to walk in the city. After successfully learning to travel by foot, agent may accidentally appear in a car. Being in a car could² be identified as a new situation. Trying to use behavior learned before (walking) is not a good idea in a car. So the agent has to perform random actions and try to deduct from newly collected information how to use a car.

Mechanisms 9 to 12 are more advanced ones. After discovering what the effects of actions (macroactions, behaviors) are and after discovering the most rewarding states in sensory state space, the agent may try to prepare a sequence of decisions and associated behaviors that could eventually lead to state with high reward (planning). Mechanism 10 is concerned with building control hierarchy (see Fig. 15 in section 3.1.7). Inner sensors and actions are defined and another copy of agent architecture is instantiated to work with these, resulting in high-level control of original agent via the same principles used by the original agent to cope with the outer environment. Under number 11 there is a set of correction mechanisms. An agent disposes of rules and subsymbolic representations that describe environment

² This is not that easy as it seems to us. An agent can recognize situations (as a patterns in sensory state space), but this is done thanks to significantly alternated actuator possibilities in state space region that forms a new action. This means that to recognize being in car, this being in car has to have significant effect on next agent actions and reward.

(world model). World model must be maintained – constantly checked for its usefulness. All deviations that cause flaws in agent’s behavior in environment have to be tracked down and corrected. Among flaws in behavior we can count all phenomena causing losses in reward (when compared to usual behavior, or when compared to “common sense” behavior). It is very easy for example for a computer program to get stuck in endless loop. This is a catastrophic scenario for an agent working in an environment. So, these cases have to be inevitably prevented. Apart of this, all kinds of bad decisions should be recorded, examined, and their causes tracked down and corrected if possible. Last set of mechanisms mentioned in the table are learning and adaptation aids. These mechanisms are concerned with suggesting behavior that could lead to rapid exploration and building and verifying of world model. These mechanisms are seeking for ways of reducing ongoing experience, to direct cognitive resources to relevant dimensions, features and cases.

All described mechanisms can be seen to be optimizing agent’s mental processes – with ultimate goal of acquiring of highest reward(s) possible and thus making agent as intelligent as possible.

3.3. Environments

I have argued that environment cannot be omitted when thinking about intelligence because intelligence depends on the point of the view we have – on environment we are considering.

Environment is also a living place for agents. The way the agent develops depends on properties of its environment. So, I think, we need to study how various environments are related to development of intelligent agent. What are the fundamental qualities, how intelligent agents need to cope with various aspects of environments they live in. Table 2 summarizes well known characteristics of environments along with features of environment I found important.

Table 2. Factors that determine complexity of given environment.

<ul style="list-style-type: none"> ▪ progress is repeatable, restart is possible, there are actions that restore state of environment ▪ number and degree of influence of possible actions, range of action parameters, degree of nonlinearity in action effects, mutual influence between actions ▪ dimensionality of sensory data, quality of sensors ▪ degree of influence of hidden features on reward ▪ degree of continuity of reward in state space ▪ environment is real time demanding? (is there time for off-line processing?) ▪ observability – are sensors detecting all aspects relevant for action selection?

- deterministic / stochastic
- episodic / sequential
- static / dynamic
- discrete / continuous
- single / multi agent – is the environment friendly/neutral/hostile? are there other agents maximizing similar reward?

(4) PERCEPTION

AI theorists are speaking about symbol grounding problem, transduction problem, anchoring problem and frame problem. Major issue of these problems is how to represent outer environment inside the agent, how to define things, situations, relations and processes, how to detect them in outer environment and how to keep inner representation of outer circumstances useful.

In following section I am going to present my method of assigning meaning to sensory data – defining situation(s). The way of defining new situation is described on an example of agent working in a simple environment.

4.1. Defining a Situation

In this section I am going to present my method of defining a situation. Term situation is coined here as a (sensory) state space partition into (certain interesting) region and the rest of the state space. So, *situation* is a very abstract notion that is incorporating all kinds of sensory experiences of an agent and is consistent with the common notion of occurring or appearing of things, objects or features. We can define a situation *being in car*, so we have a state space region that represent being in car and the rest of the sensory state space representing not being in car. Situations can be composed into one – so we can speak about a situation *being in car when having flu* or *being in expensive car*.

As I already mentioned, an agent can detect and recognize only circumstances that have a meaning for the agent. This means that there must be a possibility that these circumstances (will eventually) have influence on agent's next behavior.

4.1.1. 2D Environment

Defining a situation (section 4.1.2) will be described using an example of agent behaving in a simple 2D environment called *Region World* (see Fig. 16).

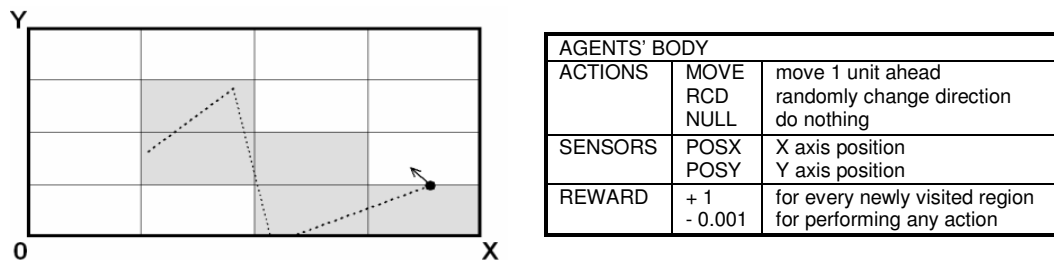


Fig. 16: Region World. Left: Rectangular 2D environment. Agent moves and color visited regions. Right: Agent's body – available actions, available sensors and reward function.

Put yourself into agent's shoes in Region World. Take closer look onto Fig. 16 and guess what you will be able to discover in the environment having this body. Note that the agent has no sensor for direction it is heading, so it will be complicated to find out that the environment and regions in it are rectangular shaped. Next, there is nearly no clue what for the reward is received – it is very hard to discover it is for newly visited regions when the agent has no idea about how the environment looks like from above (Fig. 16-Left).

After agent appears in unknown environment, it commences employing of mechanisms sketched in section 3.2, Table 1. To define new situation, the agent has to collect some statistical material by randomly choosing and performing of actions.

Next step in exploring the world is finding regions of sensory state space, where the agent acquires high reward. Analyzing collected statistical material, no such interesting regions with high reward are found. It is due to the fact that the agent is rewarded for visiting regions that were not visited before. Places with reward are therefore distributed uniformly on average through the whole sensory state space.

4.1.2. Regions in State Spaces and New Defined Situation

Carrying out broad analysis of collected material it is discovered, that action MOVE is tied together with receiving reward, but performing this action in endless loop leads inevitably to significant drop in effectiveness (agent get stuck at the border and cant visit new regions). These circumstances lead the agent to performing analysis of effects of the MOVE action (see Fig. 17).

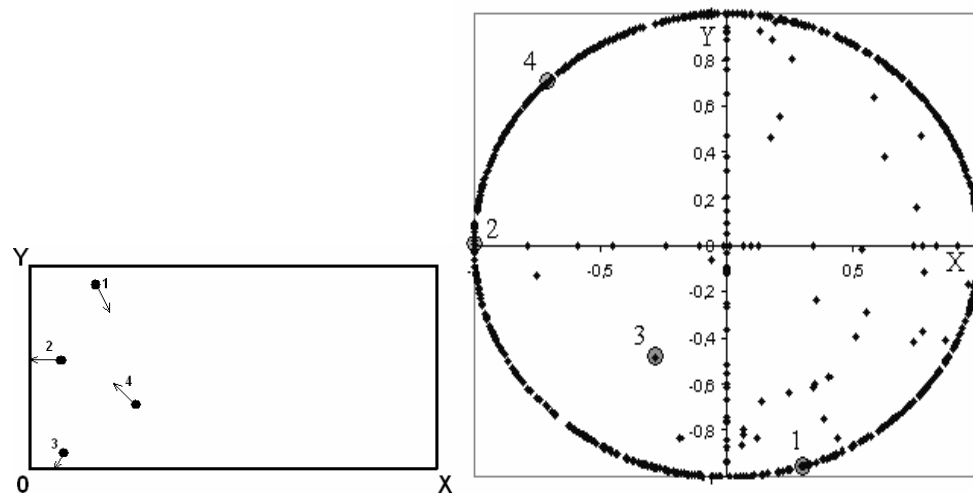


Fig. 17: Effects of MOVE action. Left: Four examples of performing MOVE action. Right: Effect of action move – differences in sensors POSX and POSY caused by performing MOVE action (5000 cases). Four examples from left are highlighted.

So, what are the effects of action MOVE? MOVE is precisely defined in the environment (moving 1 unit ahead, in direction agent is oriented OR moving less than 1 unit when hitting the wall). Because the agent has no sensor for direction, it is impossible to precisely discover the very effects of MOVE.

But, the agent will leastways try to catch some interesting patterns in effect of MOVE action. The question is: *What is the usual effect of MOVE?* Agent can answer this question. Simply enough, usual effect of MOVE action is the region in sensory state space with high frequency of occurring cases after performing of MOVE action³.

To find interesting regions, there is used a graph clustering method. This hierarchical unsupervised clustering method finds arbitrary shaped region, in which there is high probability (comparing to the size of the region) that point representing actual effect of certain action will be located here.

Clustering method finds high frequency region and thus defines a state space partition: found high-frequency region and the rest of the sensory state space.

³ Note that Fig. 17-Right shows differences in sensory values caused by MOVE, not the absolute values. Effect of MOVE on absolute sensory values is not so interesting for the agent as effect on sensory value differences that's way effect on absolute sensory values is not showed here. This doesn't mean that the agent is not performing action effect analysis in absolute sensory values. Possible action TELEPORT-TO-BASE would have had interesting effects on absolute sensory values and less interesting on differences in sensory values.

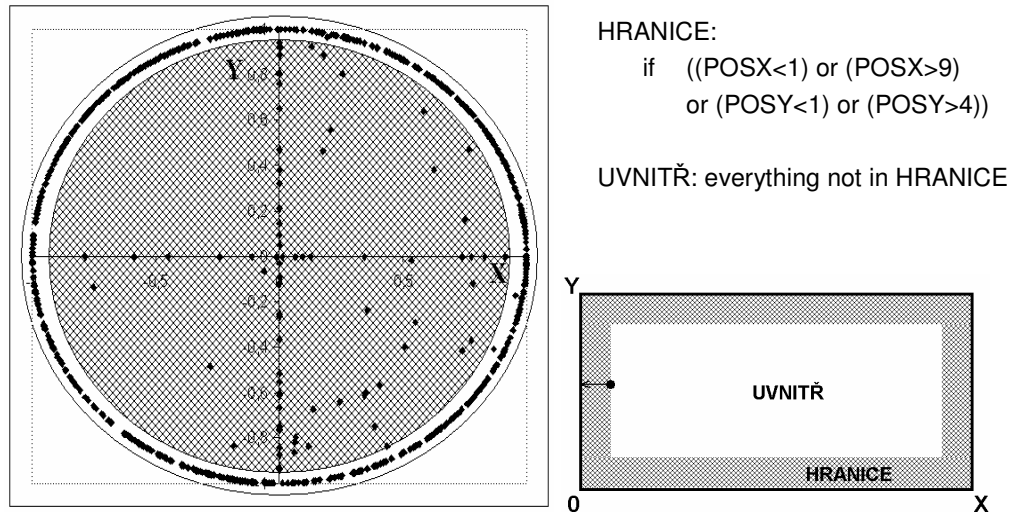


Fig. 18: New situation definition. Left: Usual and unusual effect of MOVE action. Right below: Demonstration of places (in POSX, POSY sensory state space) where occurs unusual effect of MOVE⁴. This is new situation – its meaning is being/not being near the border. Arbitrary symbols are chosen for places where unusual effect of MOVE occurs (*HRANICE*) and for the rest (*UVNITŘ*). Right up: Rules that describe new situation.

Fig. 18 shows definition of new situation. Situation is new partition of sensory state space, based in this case on unusual effect of action MOVE. For us, humans with common sense, this automatically defined situation has a meaning – being in vicinity of the border. Agent is now able to detect this situation (using similar procedure to the defining; or using rules that were derived (Fig. 18-up)) and it can derive two different behaviors according to this situation. We intuitively see that near the border it is advisable to try change direction and avoid getting stuck in wall, while there is no need to turning being far from border.

Newly defined situation helps the agent to significantly improve its effectivity in the environment.

4.1.3. Issues Related to Situation Definition

Situation defining is the bridge between non-symbolic continuous noisy outer environment and inner symbolic representation. For frequency clustering I have used graph clustering algorithm based on Delaunay triangulation. Because

⁴ Usual effect occurs everywhere in POSX, POSY sensory state space, unusual only near the border – this is what the agent has to find out while performing additional analysis.

triangulation is computationally expensive process, it is needed to let an agent learn to find equilibrium between utility of outcome and cost of computation. There are various ways how to reduce demandingness of situation definition process. Dimensionality of data can be reduced to obtain space where interesting patterns will most likely arise. Data set can be reduced via sampling. Previous graph creating work can be reused. There should be analysis telling the agent how much data is needed for discovering new facts of some sort.

Situation definition works as follows: collected statistical material (performed actions, effects of actions, detected features) is converted into points in state space – triangulation is performed to represent the data using a graph. This graph is then processed to find interesting region. Dividing state space into this region and the rest, new situation is formed. Then there is question how to recognize whether sensory data will fall into interesting region. In previous section I presented rules describing new situation – but deriving these rules is no (computationally) easy process. The same steps which are performed during defining of new situation are performed when detecting whether the actual sensory readings fall into interesting situation-region or not. After enough trials and enough collected cases of sensory data falling into and outside new-situation-region rule deriving analysis could be performed. How this analysis should look like? In previous section (see Fig. 18) derived rules catch the fact that no unusual effect of MOVE happens behind some sensory value. So this is one way how to derive rules – find a certain value of sensor which divides sensory space into two – one where something (unusual effect of MOVE) appeared and the rest.

Agent's behavior is influenced by defined situations and derived behavior in these situations. Outcome of situation definition process depends on collected statistical material. This statistical material depends among other things⁵ on agent's behavior. This loop can potentially result into badly flawed picture of circumstances of the environment.

Similar problem to defining situations is learning classes of agent's actions where there are actions with continuous parameters available. This process is described in [27] on example of grouping actions of a mobile robot into categories that produce similar outcome.

⁵ Fig. 18 would be for example different for another size of environment.

(5) CONCLUSIONS AND FUTURE WORK

In this document I was presenting theoretical background and my own ideas related to creating intelligent agents.

I have presented new definition of intelligence. Although it seems strange a little bit, it captures the essence of intelligence and reflects the notion discovered by embodied cognitive science that mind and body are not separable. Additionally, this definition can be easily translated into technical terms. As we can conclude from works of Brooks and Piaget, basic elements of intelligence are physical interactions with environment. Turing tells us that when defining intelligence, we have to look upon it only from outside – examining its demonstrations, instead of describing it in terms of mental processes. Strinberg is speaking about *successive intelligence*. But success (reward) is fundamental for intelligence in general.

Artificial intelligence – as the endeavor of creating intelligent machines and examining (human) mind – is growing since 1950. Until 1980 it was concerned mainly with formal problem solving methods (GOFAI). Thanks to Brooks and others, around 1980 there emerged new branch of AI called embodied AI. EAI constitutes a new view on the problem of creating intelligent agents. It has become clear that intelligence is not property of isolated entities. To be intelligent, an organism or a machine has to be embodied in the environment. It means that this organism or machine (agent) needs to have a body that is structurally coupled with the environment via perturbatory channels. Agent is influencing the environment, environment is influencing the agent. Intelligence is the quality of the agent that could emerge only when agent and environment are mutually influencing each other. I am proposing in this document a fusion of GOFAI and EAI – to use sensing and acting and reacting body, together with abstract symbolic methods. To combine this two approaches successfully, there is needed a transduction between the sensors and actuators of the body and inner symbols. I am developing methods of bridging body and inner world of symbols. One of these mechanisms that deal with meaning of symbols is described in this work as *defining situations*.

Intelligent agent must display robust sophisticated behavior. Designing the full agent (for any of non-trivial domains) from scratch is an intractable task. Instead of this I am proposing an approach of designing an agent that is universal – able to evolve and adapt to arbitrary conditions. To identify fundamental mechanisms that are providing adapting and evolving capabilities of an agent, I am proposing using sequence of virtual environments. In this thesis proposal I am arguing that virtual environments are most suitable for evolution of intelligent agents.

Our mind and our mental processes are still defying full understanding. Among other things, it is because our retrospection is far from complete. Our thinking machinery was evolving for tenths of thousands of years and many fundamental processes are used by automatic, but they are still hidden to us. This does not mean that these fundamental processes need to be too intricate to

understand. I think many of them are pretty simple, but we do not have the right point of view, yet. I suppose that examination of mechanisms that are employed at the low level of perception (defining situations, automatic category learning, rule induction using uninterpreted sensory data) and at the low actuator level (examining effects of actions, action grouping, action parameter space investigation) may bring significant insight.

5.1. Aims of Doctoral Thesis

Aims of my doctoral thesis, based on ideas mentioned before, are:

- design mechanisms of low level perception (category learning, rule induction) as the bridge between environment and inner symbolic processing of intelligent agent
- suggest general agent architecture, capable of evolution and displaying emergence of intelligence
- design virtual environments, uncover properties of these environments in relation to evolution of agent's intelligence

BIBLIOGRAPHY

- [1] Winston, P. H. (1984). *Artificial Intelligence*. Second Edition, Addison-Wesley, Reading, MA.
- [2] Simon, H. (1957). *Administrative Behavior: A Study of Decision-Organization, (2nd ed.)*. New York: Macmillan.
- [3] Quick, T., Dautenhahn, K., Nehaniv, C. L., Roberts, G. (1999). *On bots and bacteria: Ontology independent embodiment*. Proc. European Conference on Artificial Life (ECAL'99). <http://citeseer.ist.psu.edu/quick99bots.html>
- [4] Duffy, B.R., Joue, G.. (2000). *Intelligent Robots: The Question of Embodiment*. Brain-Machine 2000 December 20-22, 2000, Ankara, Turkey.
- [5] Etzioni, O. (1993): *Intelligence without robots: A reply to brooks*. AI Magazine, vol. 14, no. 4, pp. 7 -- 13.
- [6] Lueg, Ch., Salomon R. (1997). *A New AI Perspective on Software Agents: Preliminary Report*. Proceedings of the Second German Workshop on Artificial Life (GWAL 97) pp. 59-60. Dortmund, Germany, April 17 – 18.
- [7] Laird, J. E. (2000). *It knows what you're going to do: adding anticipation to a quakebot*. Proceedings of the AAAI Spring Symposium Technical Report, 2000.
- [8] Laird, J., van Lent, M. (2001). *Human --Level AI's Killer Application, Interactive Computer Games*. Artificial Intelligence Magazine, v.22, n.2, Summer, pp. 15-25.
- [9] Curtis, P. (1992). *Mudding: Social Phenomena in Text-Based Virtual Realities*. Proceedings of the 1992 Conference on the Directions and Implications of Advanced Computing, Berkeley, May 1992.
- [10] Atkin, M., Westbrook, D. L., Cohen, P. R. (1999). *Capture the Flag: Military Simulation Meets Computer Games*. Presented at the AAAI Spring Symposium on AI and Computer Games.
- [11] Amir, E., Doyle, P. (2002). *Adventure games: A challenge for cognitive robotics (full version)*. AAAI'02 workshop on Cognitive Robotics. Also, available at the author's website (<http://www.cs.uiuc.edu/eyal/papers>).
- [12] Piaget, J. (1963, 2001). *The psychology of intelligence*. New York: Routledge.
- [13] Sternberg, R. J. (2003). *A broad view of intelligence: A theory of successful intelligence*. Consulting Psychology Journal: Practice and Research, 55, 139-154.
- [14] Turing, A.M. (1950). *Computing Machinery and Intelligence*. Mind, 1950, vol. 59, no.236, pp. 433 – 460.
- [15] Brooks, R. A. (1991). *Intelligence without Representation*. Artificial Intelligence, Vol.47, 1991, pp.139-159.

- <http://citeseer.ist.psu.edu/brooks91intelligence.html>
- [16] Brooks, R. A. (1991). *Intelligence Without Reason*. Proceedings, IJCAI-91, Sydney, Australia.
- <http://citeseer.ist.psu.edu/article/brooks91intelligence.html>
- [17] Brooks, R. A. (1986). *A Robust Layered Control System for a Mobile Robot*. IEEE Journal of Robotics and Automation, Vol. 2, No. 1, March 1986, pp. 14–23.
- [18] Anderson, J. R., Matessa, M. (1998). *The rational analysis of categorization and the ACT-R architecture*. M. Oaksford & N. Chater (Eds.) Rational models of cognition, pp. 197-217. Oxford: Oxford University Press.
- [19] Anderson, J. R., Betz, J. (2001). *A hybrid model of categorization*. Psychonomic Bulletin and Review, 8, 629-647.
- [20] ACR-T homepage: <http://act-r.psy.cmu.edu/>
- [21] Leon, V.J., Kortenkamp, D., Schreckenghost, D. (1997). *A Planning, Scheduling and Control Architecture for Advanced Life Support Systems*. NASA Workshop on Planning and Scheduling for Space, October, 1997.
- [22] Firby, R. J. (1989). *Adaptive execution in complex dynamic worlds*. Doctoral Thesis, Yale University, May 1989.
- [23] Marjanović, M. J. (2003). *Teaching an old robot new tricks: Learning novel tasks via interaction with people and things*. Doctoral Thesis, Massachusetts Institute of Technology, June 2003.
- [24] Franklin, S. (1997). *Autonomous Agents as Embodied AI*. Cybernetics and Systems, special issue on Epistemological Issues in Embodied AI, 28:6 499-520.
- [25] Selman, B., Brooks, R., Dean, T., Horvitz, E., Mitchell, T., Nilsson, N. (1996). *Challenge Problems for Artificial Intelligence*. Proceedings of AAAI-96, Thirteenth National Conference on Artificial Intelligence, Portland, Oregon, August 1996. AAAI Press, Menlo Park, California, pp. 1340-1345.
- [26] Minsky, M. (1988). *The Society of Mind*. Simon & Schuster, New York.
- [27] King, G., Oates, T. (2001). *The Importance of Being Discrete: Learning Classes of Actions and Outcomes through Interaction*. Lecture Notes In Computer Science; Vol. 2056. Proceedings of the 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence.