Proceedings of ICCM-5, Universitäts-Verlag Bamberg, 2003, ISBN 3-933463-15-7

The MicroPsi Agent Architecture

Joscha Bach (bach@informatik.hu-berlin.de)
Institut für Informatik, Humboldt-Universität zu Berlin
Unter den Linden 6, 10099 Berlin, Germany

Abstract

The MicroPsi agent architecture describes the interaction of emotion, motivation and cognition of situated agents based on the Psi theory of Dietrich Dörner. This theory touches on a number of questions, particularly about perception, representation and bounded rationality, in very interesting ways, but being formulated within psychology, has had relatively little impact on the discussion of agents within computer science. MicroPsi is an attempt to address this by formulating the original theory in a more abstract and formal way, at the same time enhancing it with additional concepts for building of ontological categories and attention. This paper gives an introduction into MicroPsi components and representation.

Introduction

The design of an agent architecture that is capable of mimicking the feats of human or animal cognition requires psychological theories with regard to motivation, perception, emotion and memory that are detailed and formal at the same time. Unfortunately, there are few theories in existence that meet this demand, partly because current psychological methodology does not encourage the formulation of comprehensive functional theories. (Newell 1990)

One of the most interesting and inspiring aspirants, however, may be the *Psi theory* by psychologist Dietrich Dörner (1999, 2002). The theory makes attempts at addressing issues like perception, imagination, emotion, planning and memory in often very original ways, and has even been implemented by its author as a software agent. But since most aspects of the theory have never been published within the context of AI and the formulation is sometimes loose, informal and incomplete, its impact on computer science has so far been somewhat limited.

The MicroPsi architecture, which is the subject of the following pages, is an agent architecture that incorporates many ideas from the Psi theory; in some way, it is meant to be a structured implementation of the original theory and a step towards its formulation within the context of computer science. The result is the description of a cognitive framework for autonomous, situated agents. Unlike in the successful (and more mature) architectures ACT (Anderson & Lebière 1998) and SOAR (Laird, Newell & Rosenbloom 1987), which concentrate mainly on cognition, a special focus is laid on motivation, emotion and flexible acquisition of object and action ontologies through interaction with the environment. Because the verbal abilities of the agent are fairly limited – so far it does not possess grammatical language – many aspects of human problem solving can not be modeled yet, which marks a current line of research of Dörner's group. Another limitation consists in the lack of social behavior, partly because social motivation and the perception of social aspects of the environment (i.e. mental states of other agents) are still missing.

Many concepts of MicroPsi are more abstract and formal than the Psi theory. In its formulation as an architecture, it makes use of ideas from BDI architectures (Bratman 1987, Rao & Georgeff 1998), but unlike common formulations of the BDI paradigm like PRS (Georgett & Ingrand 1988) and dMARS (d'Iverno et. al. 1998), it concentrates on modeling perception, emotion and interaction and incorporates notions from Sloman's theory of Cognition and Affect (1992, 1994). MicroPsi is not meant to be a toolkit for cognitive or social modeling, but sets out to serve as an experimental platform that allows for the discussion of phenomena of human and animal cognition, similar to the Conscious Agents of S. Franklin (1999).

Extensions of MicroPsi, compared to the original theory, took place on two levels. On the agent design level, MicroPsi possesses a more structured, partly parallel process layout, a divided memory and more general internal representations. On the level of the theory, methods for dynamic category building and for attention based processing have been added, even though implementation is still in its prototypic stages. The original low level perception mechanism of the Psi agents, which is based on a simple image recognition method, is currently not a part of MicroPsi.

In the following section, a brief and informal introduction into the main points of the MicroPsi architecture is given. The rest of the paper is concerned with a short description the MicroPsi node nets, which are the fundamental building blocks of representation and cognition of the agents. This is followed by some concluding remarks.

The MicroPsi architecture

Overview

MicroPsi consists of several main components, which are connected to their environment by a set of somatic parameters (like 'intactness' and 'hunger') and external sensors. From these, somatic desires ('urges'), immediate percepts and modulators are automatically derived. The activity of the agent consists of a number of internal and external behavior modules. While the former are 'mental' actions, the latter send sequences of actuator commands to the environment.

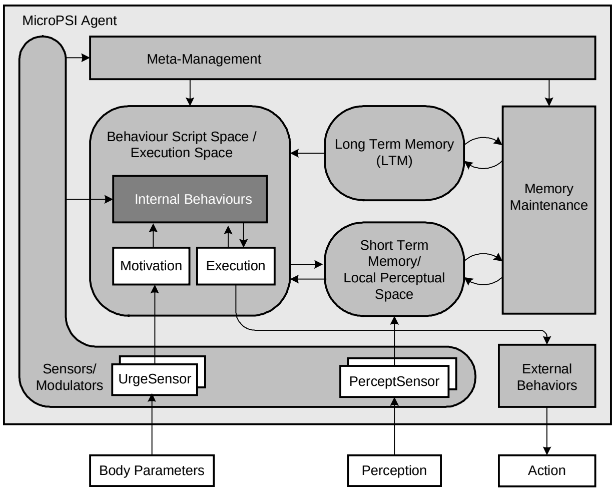


Figure 1: Overview of Architecture.

The representations that can be derived from external percepts (in conjunction with knowledge that has been acquired earlier) are stored in the agent's access memory. The agent also possesses a long term memory that holds its history and concepts that have been derived from interaction with the environment. The exchange between long term and access memory takes place through a set of autonomous memory maintenance processes, which also handle memory decay, concept generation and so on. The agent's internal behaviors are meant to handle its higher cognitive processes. They are triggered by motivations and modified by a set of modulators. The management of processing resources between the internal behaviors and the memory maintenance mechanisms is handled by the meta-management module.

Main concepts

A full description of the MicroPsi implementation of the Psi theory is beyond the scope of this paper; the most important concepts shall be briefly mentioned here:

Representation: Objects, situations, categories, actions, episodes and plans are all represented as hierarchical networks of nodes. Nodes may be expanded into weighted conjunctions or disjunctions of subordinated node nets, and ultimately 'bottom out' in references to sensors and actuators. Thus, the semantics of all acquired representations result from interaction with the environment or from somatic responses of the agent to external or internal situations. For communicating agents, they may potentially be derived from explanations, where the interaction partner (another software agent or a human teacher) refers to such experiences or previously acquired concepts.

Emotion: The Psi theory does not aim for merely simulating mental processes like emotions, imagination or learning. Rather, it boldly attempts to define these processes formal and robust enough as to allow for the implementation of these concepts. This results in agents that are meant not to simulate but to actually undergo these mental processes. That claim, however, has to rest on the comprehensiveness of the understanding of mental processes underlying the definitions. Emotion, in Dörner's framework, is a set of configurations of the cognitive system of an individual. These configurations influence how an agent perceives, plans, memorizes, selects intentions, acts etc. in a certain way, and they are defined by a set of *modulators*, like *arousal*, *resolution level* and *selection threshold*. Thus, they depend both on the modulation and on the embedded cognitive system that is being modulated. The modulation is designed to allocate mental resources in way that is suitable to a given situation and reduce the computational complexity of the tasks at hand. (Dörner & Schaub 1998)

Motivation: The agent possesses a number of innate desires (urges) that are the source of its motives. Events that raise these desires are interpreted as negative reinforcement signals, whereas a satisfaction of a desire creates a positive signal. Currently, there are urges for intactness, energy (food and water), affiliation, competence and reduction of uncertainty. The levels of energy and social satisfaction (affiliation) are self-depleting and need to be raised through interaction with the environment. The cognitive urges (competence and reduction of uncertainty) lead the agent into exploration strategies, but limit these into directions, where the interaction with the environment proves to be successful. The agent may establish and pursue sub-goals that are not directly connected to its urges, but these are parts of plans that ultimately end in the satisfaction of its urges.

The execution of internal behaviors and the evaluation of the uncertainty of externally perceivable events create a feedback on the modulators and the cognitive urges of the agent.

Perception: External perceptions are typically derived from hypotheses about the environment which are then tested against immediate external percepts (this is called 'hypothesis based perception', or *hypercept*). Only if the expectations of the agent fail, and no theory about the perceived external phenomena can be found in memory ('*assimilation*'), a new object schema is acquired by a scanning process ('*accommodation*') that leaves the agent with a hierarchical node net. Abstract concepts that may not be directly observed (for instance classes of transactions – like 'giving' – or object categories – like 'mammals') are defined by referencing multiple schemas in such a way that their commonalities or differences become the focus of attention.

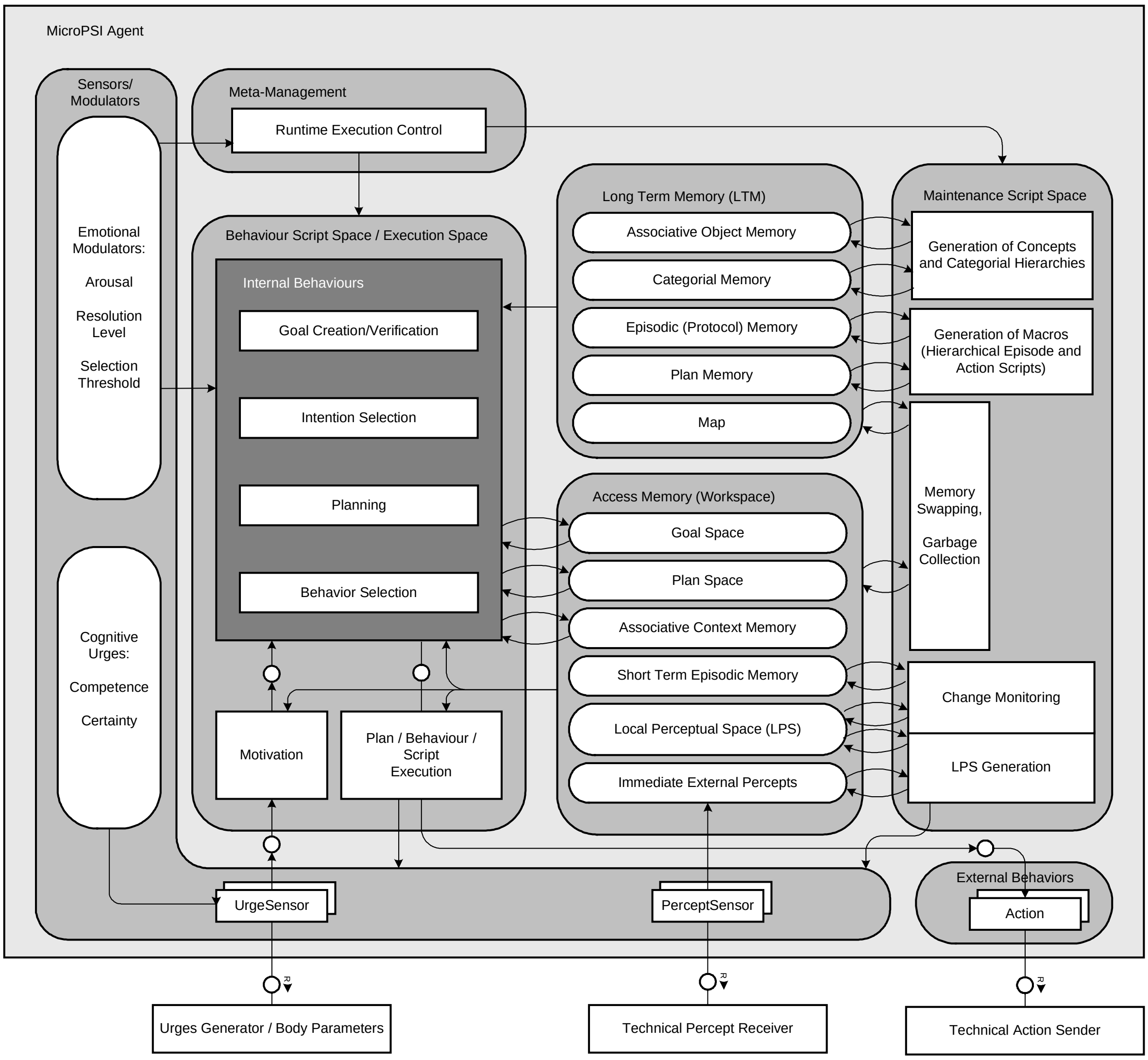

MicroPSI Agent
Sensors/ Modulators
Emotional Modulators:
Arousal
Resolution Level
Selection Threshold
Cognitive Urges:
Competence
Certainty
Meta-Management
Runtime Execution Control
Behaviour Script Space / Execution Space
Internal Behaviours
Goal Creation/Verification
Intention Selection
Planning
Behavior Selection
Motivation
Plan / Behaviour / Script Execution
Long Term Memory (LTM)
Associative Object Memory
Categorial Memory
Episodic (Protocol) Memory
Plan Memory
Map
Access Memory (Workspace)
Goal Space
Plan Space
Associative Context Memory
Short Term Episodic Memory
Local Perceptual Space (LPS)
Immediate External Percepts
Maintenance Script Space
Generation of Concepts and Categorial Hierarchies
Generation of Macros (Hierarchical Episode and Action Scripts)
Memory Swapping, Garbage Collection
Change Monitoring
LPS Generation
External Behaviors
Action
UrgeSensor
PerceptSensor
Urges Generator / Body Parameters
Technical Percept Receiver
Technical Action Sender

Figure 2: Main Components.

External percepts are mapped into a space of sensors ('immediate external percepts'), from which a representation of the agent environment is created ('local perceptual space'). Changes in the environment are recorded into the agent's short term episodic memory. The mechanisms responsible for this form the autonomous external perception of the agent.

Action: The agents represent actions as triplets of nodes, where the first references the elements of a situation that form the pre-condition of an action, the second the actuator that leads to the change in the environment, and the last the changes that form the post-condition. The actuator often refers to other chains of actions ('macros' or 'scripts'), which makes long plans feasible by packing sub-plans into chunks. Since all internal behaviors – perception, goal identification, planning, meta-management etc. – may be formulated as node-chains and can be subject to the evaluation and planning of the agent, it has the tools to re-program its own strategies. Language may become an important structuring aid for planning, especially when referring to internal behavior plans.

Planning: MicroPsi agents possess a simple set of planning algorithms. Given a goal situation (which is derived from the motivational process), agents try to find a chain of actions that leads from the current situation to the goal (*automatism*). If no such chain is remembered, its construction is attempted by combining actions (see above). This may happen by different search algorithms (forward, backward, A* etc.), where depth and width of the search are controlled by the modulators.

Meta-management: Although there is no singular control structure (central execution), the different processes forming the internal behaviors and the memory maintenance are being allocated processing resources according to the given situation. This may happen by calling them with varying frequencies or by using algorithms that consume different amounts of memory and processing time. Thus, different layers of reactivity within the agent can be realized. Note that this does not happen by distinguishing behaviors based on their level of reactivity, but by promoting a cognitive process if its successful execution needs more attention, and by demoting it if it runs smoothly. The evaluation of the performance of such processes is the task of the meta-management. The meta-management is not to be confused with awareness or some form of consciousness of the agent; rather, it is a cognitive behavior like others and can also be subject to different levels of processing.[1]

Alarms: The MicroPsi agent is not guaranteed to execute the meta-management in short intervals or with high attention, which can prevent it from reacting quickly to environmental changes. Dörner has proposed a 'securing behavior' that should be executed by the agent in regular intervals, while Sloman describes a system which he terms 'alarms', with the same purpose: to quickly disrupt current cognitive processes if the need arises. In MicroPsi, an orientation behavior would follow if unexpected rapid changes in the low level perception or urge detection were encountered. This is not part of the current MicroPsi, because its environment is not very hostile so far.

Memory: Node nets act as universal data structures for perception, memory and planning. (See fig. 3 for simple sensoric schema, and fig. 4 for a chain of action schemas.) Even though the Psi theory does not distinguish between different types of memory, we have found that splitting it into different areas (node spaces) helps to clarify the different stages of cognitive processing. The main distinction that has been introduced into MicroPsi is the split into long term memory and workspace. This enables agents to represent and manipulate data quickly according to a given context, and to establish and test new hypothesises without compromising established long term memory.

Main kinds of information in the short term memory include the actual situation, the current course of events, a contextual background for objects in the current situation, as well as currently active goals and plans.

The long term memory stores information about individual objects, categories derived from these objects, a biography of the agent (protocol memory), a library of plans and plan-components, and a map of the environment.

Both long term and short term memory face a decay of links between nodes (as long as the strength of the links does not exceed a certain level that guarantees not to forget vital information). The decay is much stronger in short term memory, and is counterbalanced by two mechanisms:

– usage strengthens the links, and
– events that are strongly connected to a positive or negative influence on the urges of the agent (such as the discovery of an energy source or the suffering of an accident) lead to a retro gradient connection increase of the preceding situations.

If a link deteriorates completely, individual isolated nodes become obsolete and are removed. If gaps are the result of such an incision, an attempt is made to bridge it by extending the links of its neighbors. This process may lead to the exclusion of meaningless elements from object descriptions and protocol chains.

Hierarchical categories: the similarity of node schemas can be established by a complete or a partial match. If the resolution level of an agent is low, the comparison of spatial and temporal features and the restriction to fewer features may allow for greater tolerances. If the depth of the comparison is limited too, the agent may notice *structural similarity*, for instance between a human face and a cartoon face. However, the key to structural similarity is the organization of node schemas into hierarchies (where an abstract face schema may consist of eye, nose and mouth schemas in a certain arrangement, and can thus be similar to a 'smiley'). Furthermore, many objects can *only* be classified using abstract hierarchies.[2] It seems that humans tend to establish not more than 5–9 elements in each level of hierarchy, so that these elements can be assessed in parallel. (Olson & Jiang 2002)

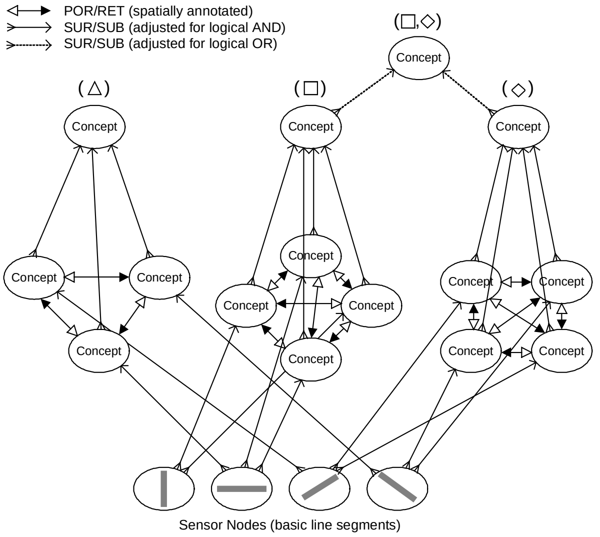
Such hierarchies can be derived mainly in three ways: by identifying prominent elements of objects (that is, structures that are easy to recognize by interaction or perception and also good predictors for the object category), by guessing, and by communication. (Bach 2002)

Shortcomings of the agent include a very limited potential for social behavior, mainly because means for the assessment of social situations (such as representations about mental states of other agents) are missing, and MicroPsi does not possess urges that would result in dominant, accepting, socially possessive or altruistic behaviors.

While it is possible to associate concept with labels and thus exchange simple descriptions of situations, objects or plans with an agent, this is a far cry from having it learn grammatical language that would be helpful as a structuring aid for planning and reflection.

[1] Attention, in our understanding, is the focusing of processing resources, while awareness is an integration of active elements from different cognitive behaviors into a single process with high attention. Awareness is currently not a part of MicroPsi.

[2] Trees may be a good example: their similarity is not very apparent in their actual shape, rather, it is limited to being rooted in the ground, having a wooden stem which is connected to the root, and ends in some equally wooden branches on the opposite side, whereby the branches may or may not carry foliage. These features form an abstract object representation and need to be individually validated when an object that is being suspected to qualify as a tree is encountered.


POR/RET (spatially annotated)
SUR/SUB (adjusted for logical AND)
SUR/SUB (adjusted for logical OR)
(□,◇)
(△)
(□)
(◇)
Concept
Sensor Nodes (basic line segments)

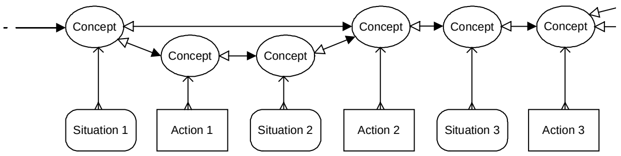


Figure 4: Example of an action chain.
Situations refer to sensoric schemas,
Actions eventually lead to actuator nodes.

Figure 3: 'Sensoric schema' for simple objects.

Representation with Nodes and Node Spaces

The central building block of internal representation within the MicroPsi agents are networks of nodes.[3] Here, **concept nodes** are the basic node type.

Generally, a node n_i consists of an input activation in_i and an array of gates $\mathbf{o}_i$, where each gate o_{ij} is characterized by an activation a_{ij}, a threshold t_{ij}, an amplification factor f_{ij} and an array $\mathbf{w}_{ij}$ of links to other nodes. For a link to a node n_k, this array contains weights w_{ijk}. (Additionally, links may be annotated with relative spatial and temporal coordinates, accuracy information concerning these coordinates, and with a certainty value.)

Concept nodes have nine different gates: general activation (GEN), links for causal relations forwards (POR) and backwards (RET), for *part-of* and *contains* relations (SUR, SUB), for membership (CAT, EXP), and for naming (SYM, REF). Of these, POR, RET, SUR and SUB are part of the original description of the Psi theory and derived from a theory of representation by Klix (1984); the others have been added to simplify the implementation and notation.

Nodes may be connected to special nodes, so-called **directional activators** *(act*$_{\text{POR}}$, *act*$_{\text{RET}}$, etc.) which refer to individual gates. There are activators for each gate type. Gates may only transmit an activation, of their corresponding activator is active, which allows for a *spreading activation* mechanism. A set of Nodes N which is connected to the same set of directional activators is called a *node space*.

The incoming activation of a node n_i in a given node space is then computed using

$$in_i = \sum w_{hji} \times a_{hji}$$

where w_{hji} is the weight of the link from a gate o_{hj} of a node n_h to the node n_i. The gate activations are given by

$$a_{ij} = \begin{cases} \min\left(act_{type\left(o_{ij}\right)} \times \left(in_i - t_{ij}\right) \times f_{ij}, 1\right), \text{ if } in_i \geq t_{ij} \\ 0, \text{ else} \end{cases}$$

By choosing appropriate weights and thresholds, links between nodes can express logical AND and OR terms. It can be demonstrated that this notation is suitable to express first order logic (Dörner 2002). On the other hand, information retrieval with node spaces is very similar to using hierarchical Case Retrieval Networks with directional activation (Burkhard & Lenz 1998).

Node spaces may contain other special node types, namely:

Register nodes are simplified concept nodes with a single GEN gate. Their purpose lies in the execution of node scripts, where they may act as pointers into node spaces.

Sensor nodes transmit values from the environment or from somatic or cognitive states of the agents by their activation.

Actuator nodes trigger external actions or basic functions within the agent, when activated. Actuators may encapsulate arbitrary low level functionality of the agent, such as movement commands or tools to interact with the environment. **Node factories** are special actuator nodes that clone the currently active node structure and can thus create new nodes.

Associators establish or strengthen the links between currently active nodes, **dissociators** weaken or remove them. Nodes that are not longer linked are removed from the node space.

Node spaces can be nested too, and it is possible to conceptualize the complete agent as a single node space with numerous sub-spaces.

Outlook

MicroPsi is still in a prototypic stage of implementation, and many aspects of the original Psi theory and of

[3] In Dörner's original concept, this is handled by a structure resembling neural networks, whereby assemblies of five neurons (such a group is called 'quad') represent features, objects, situations and actions. Nonetheless, these networks should not be confused with biological neural structures and can be described more accurately as influence or belief networks (Good 1961, Shachter 1986). Dörner's implementation makes mostly use of pointer lists, leading to a programming style that is somewhat similar to LISP.

several planned extensions have not been realized yet. The current architecture consists of a server that connects to both a set of agents, a simulated world and a set of user interfaces. A timer component can be used to control the speed of the simulation. The visualization of the simulated world using a 3D engine is planned for a later stage.

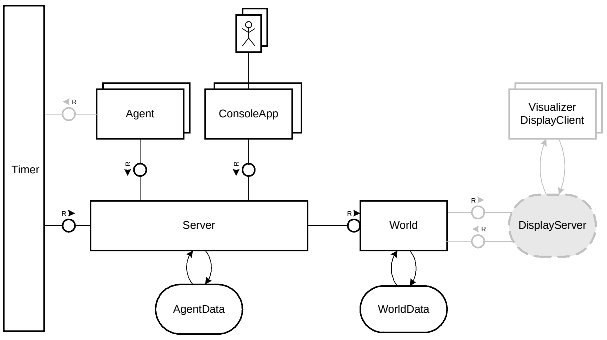


Figure 5: MicroPsi Agent and Server.

Experiments with MicroPsi agents are planned when a stable implementation is reached and will be concerning especially category and hierarchy generation, communication and social behavior.

Acknowledgments

This work would not have been possible if not for many interesting and helpful discussions with Prof. D. Dörner and members of his group. This also applies to the students of our own workgroup, especially R. Vuine, who contributed many ideas and a lot of practical work to the design and implementation of MicroPsi.

References

Anderson, J.R., & Lebière, C. (1998): Atomic Components of Thought. Hillsdale, NJ: Lawrence Erlbaum.

Bach, J. (2002). Enhancing Perception and Planning of Software Agents with Emotion and Acquired Hierarchical Categories. In *Proceedings of MASHO 02, German Conference on Artificial Intelligence KI2002*, (pp. 3-12)

Bratman, M. E. (1987). *Intention, Plans, and Practical Reason*. Harvard University Press, Cambridge, Mass.

d'Iverno, M., Kinny, D., Luck, & M. Wooldridge, M. (1998). *A Formal Specification of dMARS.* Proceedings of the Fourth International Workshop on Agent Theories, Architectures and Language, LNAI 1365, 155-176: Springer

Dörner, D. (1999). *Bauplan für eine Seele*. Reinbeck: Rowohlt

Dörner, D., & Schaub, H. (1998). Das Leben von PSI. Über das Zusammenspiel von Kognition, Emotion und Motivation. http://www.uni-bamberg.de/~ba2dp1/psi.htm

Dörner, D., Bartl, C., Detje, F., Gerdes, J., Halcour, D., Schaub, H., & Starker, U. (2002). *Die Mechanik des Seelenwagens. Eine neuronale Theorie der Handlungsregulation.* Bern, Göttingen, Toronto, Seattle: Verlag Hans Huber.

Franklin, S. (1999). *Action Selection and Language Generation in "Conscious" Software Agents.* Proc. Workshop on Behavior Planning for Life-Like Characters and Avatars, i3 Spring Days '99, Sitges, Spain

Georgeff, M.P., & Ingrand, F.F. (1988): *Research on Procedural Reasoning Systems.* Technical Report, AI Center, SRI International, Menlo Park, CA Good, I. J. (1961). *A Causal Calculus*. British Journal of the Philosophy of Science, 11:305-318

Klix, F. (1984). Über Wissensrepräsentation im Gedächtnis. In F. Klix (Ed.): Gedächtnis, Wissen, Wissensnutzung. Berlin: Deutscher Verlag der Wissenschaften.

Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). *SOAR: An Architecture for General Intelligence.* Artificial Intelligence, 33: 1-64

Lenz, M., Burkhard, H.-D. (1998*). Case retrieval nets: Basic ideas and extensions.* Technical report, Humboldt University, Berlin

Newell, A. (1990). Unified Theories of Cognition. Cambridge, Mass.: Harvard University Press

Olson, I.R., & Jiang, Y. (2002): Is Visual Short-Term Memory Object Based? Rejection of the "Strong Object" Hypothesis. Perception and Psychophysics, 64, 1055-1067

Rao, A.S., & Georgeff, M.P. (1998). Decision procedures for BDI logics. *Journal of Logic and Computation*, 8

Shachter, R. D. (1986). *Evaluating influence diagrams*. Operations Research, 34:871-882

Sloman, A. (1992). *Towards an information processing theory of emotions*. http://www.cs.bham.ac.uk/~axs/cog_affect/Aaron.Sloman_IP.Emotion.Theory.ps.gz

Sloman, A. (1994). *Semantics in an intelligent control system*. Philosophical Transactions of the Royal Society: Physical Sciences and Engineering. Vol 349, 1689, 43-58