FP6-511931

Mind RACES

from Reactive to Anticipatory Cognitive Embodied Systems

DELIVERABLE D2.3

Evaluation Methodology and Metrics

Due date of deliverable:
01 /04/ 2006

Actual submission date:
15/ 05/ 2006

Start date of project:
01 / 10 / 2004

Duration:
36 month

Organization name of lead contractor for this deliverable:
ISTC-CNR

Revision:
REVISED

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	**X**
PP	Restricted to other programmes participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Document identifier:	**DELIVERABLE_WP2_N_3**
Date:	**2006-05-15**
Work package:	**WP2**
Partner(s):	*IDSIA, ISTC-CNR, NBU, NOZE, OFAI, UW-COGSCI*
Lead Partner:	*NBU*
Document status:	*Approved*
Deliverable identifier:	**WP2_D2.3**

Delivery Slip

	Name	Partner	Date	Signature
From	MAURICE GRINBERG	NBU		
Verified	RINO FALCONE	ISTC-CNR		
Approved by	RINO FALCONE	ISTC-CNR		

Files

Software Products	User files
MS Word™	WP2_D2.3.DOC

Project information

Project acronym:	Mind Races
Project full title:	MIND RACES: from Reactive to Anticipatory Cognitive Embodied Systems
Proposal/Contract no.:	IST-511931
Project Manager:	ISTC_CNR
Name:	Rino Falcone
Address:	CNR-ISTC via S. Martino della Battaglia,44 00185 Rome ITALY
Phone:	+39 06 44 595 253
Fax:	Fax: +39 06 44 595 243
E-mail	rino.falcone@istc.cnr.it

TABLE OF CONTENTS

1 PART 1 – Management Overview

1.1 Document Control

This document is a co-production of all the partners mentioned above. All the partners answered a questionnaire and filled tables in order to generate possible evaluation metrics closely related to the scenarios of the project (see D2.1). Then the information was systematized and some general principles were formulated by NBU. The final version was sent for approval and verification on the 02/05/2006.

1.2 Executive Summary

In order to generate evaluation metrics that are appropriate for MindRACES a questionnaire has been created and distributed to the partners. Additional tables have been supplied to help the easier systematization of the metrics proposed. The contributions of all the partners have been processed and a common table with metrics have been generated. On the basis of the presentation of the results of this effort and the discussion among the partners during the regular consortium meeting in Würzburg (April 20-21), a few metrics common to all partners were selected giving the possibility for each partner to add its own evaluation metrics which accounts for the specificity of the architecture used.

1.3 Evaluation in MindRaces

The purpose of evaluation is to assess the ‘added value’ of anticipatory mechanisms in the performance of the robots (real or simulated) on the tasks encountered in the scenarios. The evaluation can be carried on by using different criteria. Such criteria could be:

- **global** (performance in a scenario) and/or **task-oriented**;
- **objective** or **comparative**.

A metric or an expert assessment should be introduced for each objective criterion. Some of the objective criteria like time spent on a task and energy consumed might not be suitable to compare two different cognitive mechanisms because of their different nature, therefore their different time spending and energy consumption. Yet they might be used for comparison of one mechanism in various tasks. In our opinion those objective metrics which estimate cognitive characteristics (anticipation, prediction, learning, etc) and task related characteristics (collected objects number, number of undesired collisions with an opponent count, etc) are suitable for between-mechanism comparison.

The comparative evaluation of each mechanism can be related:

- to the performance to the system without this mechanism (if possible);

- to human behaviour presumably involving similar mechanisms;
- to another alternative anticipatory mechanism;
- to two or more integrated anticipatory mechanisms.

All these mechanisms should be used in the same environment.

1.4 Success questionnaire

1. Select specific tasks or full scenarios which can be used to evaluate the anticipatory mechanisms.
2. What kind of human expert evaluation do you plan to use in your scenarios?
   a. Objective
   b. Based on performance
   c. Related to task achievement
   d. Related to comparison
   e. Other
3. Which objective metrics do you plan to use in your scenarios?
4. What comparison with respect to non-anticipatory mechanisms do you plan?
5. How does integration with another mechanism change the performance in your scenarios?

6. Please compare other models to yours with respect to anticipatory behaviour. What do you base it on?
7. Please compare your mechanism's performance with human behaviour.

1.5 Cognitive aspect metrics template

The table template below should be used for describing metrics and the cognitive aspect to which it is related.. Function field gives the exact algorithm for measuring the performance. It would be better if we had a higher scale of measurement for that purpose (like interval or ratio) but a simple ordinal scale is good enough, because via such a scale we can find out when a mechanism or model performs better than others. Index – short unique descriptor for the metric, which will be used further for its identification. Range - gives output values of the function.

Cognitive aspect	Index	Metrics	Function	Range

1.6 Non-cognitive metrics template

This table is used to give us description of other metrics that were proposed by partners and for which we don't define any cognitive aspect.

Index	Metrics	Function	Range

1.7 Metrics usage template

This table will serve for determining which scenario the proposed metric will be used for.

Scenario	Metrics

PART 2 – Deliverable Content

2 ISTC-CNR

ISTC consider that it is better to select few and easy success metrics for each task (3 or 4 at most) which will be used by all the partners. For example, the metric M1 (similarity between the object found and the target object) in the guards-and-thieves scenario is certainly of interest for an analogical system, but much less relevant for other systems (for example, we do not plan to have "distractors"). Of course, each partner can also use extra metrics for evaluating their own specific cognitive functionalities (for example by comparing it with systems lacking those functionalities). For this reason, ISTC have split between METRICS SHARED BY ALL THE PARTNERS and METRICS RELEVANT FOR ISTC.

ISTC suggest restricting the set of metrics to the most relevant for the task, i.e. related to its goals. For example, it is easier to evaluate the time scale of prediction in "focused" tasks such as find an object or fish-catching than in "broader" ones such as guards-and-thieves which is focused on many other aspects.

2.1 Scenario: GUARDS AND THIEVES

2.1.1 Success questionnaire

1. Select specific tasks or full scenarios which can be used to evaluate the anticipatory mechanisms.

   a. *Guards and Thieves*

2. What kind of human expert evaluation do you plan to use in your scenarios?

   f. Objective (YES)

   g. Based on performance (YES)

   h. Related to task achievement (YES)

   i. Related to comparison (YES)

   j. Other (NO)

3. Which objective metrics do you plan to use in your scenarios?
4. What comparison with respect to non-anticipatory mechanisms do you plan?

   b. In our opinion, the best way to evaluate the cognitive functions is by comparing an anticipatory and a non-anticipatory system in the same task and use the usual metrics. For example, we are comparing two systems (with and without forward models) in the same guard-and-thieves tasks by using the same metrics of success.

5. How does integration with another mechanism change the performance in your scenarios?

   c. We plan to extend our architecture in a vertical (hierarchical) way, by adding planning capabilities and goal-orientedness (which permits e.g. to build and store plans for finding the thieves); and to increase the perceptual capabilities (which permits e.g. to pre-process salient information and to perform epistemic actions).

6. Please compare other models to yours with respect to anticipatory behaviour. What do you base it on?

   d. Comparison will mainly be done with respect to the cognitive functions and their relations (for example by exploiting the "Absolute Complexity" and "Absolute

Capabilities" criteria proposed by NOZE –with respect to their proposal, we intend "core objects" as "cognitive functionalities")

7. Please compare your mechanism's performance with human behaviour.
    e. We don't plan to do that

2.1.2 Cognitive aspect metrics (Shared by all the partners)

Cognitive aspect	Index	Metric	Function	Range
Prediction / Anticipation	GTC1[1]	Performance measurement (*even if for some systems it is difficult to exactly separate expected/unexpected*)	A combination of the **number of unexpected opponents/ rival meetings**	(0 1]
Prediction / Anticipation	GTC2	Valuables protection/collection	This metric is different for treasure-hunters and guards. For the guards it is based on the number of objects preserved and for the treasure-hunters on the number of objects violated or *collected and preserved* (see Scenario 2.3) ***Objects preserved count / objects ever owned count***	[0 1]

2.1.3 Cognitive aspect metrics (Relevant for ISTC)

Cognitive aspect	Index	Metric	Function	Range
Prediction / Anticipation	GTC-ISTC1	Hierarchical structure	Comparison between the performance of the architecture with 1, 2 and 3 layers (by using the usual metrics) **Es.: GTC1-2layers/GTC1-1layer**	(0 1]
Prediction / Anticipation	GTC-ISTC2	Replanning	Usefulness of replanning **Number of replans in successful tasks / total number of replans**	(0 1]

[1] GTC = Guards and Thieves Cognitive

Prediction / Anticipation	GTC-ISTC3	Intention Reconsideration	Usefulness of intention reconsideration **Number of intention reconsiderations in successful tasks / total number of intention reconsiderations**	(0 1]
Prediction / Anticipation And Learning	GTC-ISTC4	Epistemic Actions And Agent's Learning	Efficacy index for Epistemic Actions **Valuables found with explicit Epistemic Action execution (ongoing actions) / Global amount of found Valuables**	[0,1]

2.1.4 Non-cognitive metrics (Shared by all the partners)

Index	**Metric**	**Function**	**Range**
GTNC1[2]	Number of sessions which ended with a successful completion of the task (for the guard: capture the thief; for the thief: collect the valuable)	***Number of successful tasks / Number of tasks***	[0 1]
GTNC2	Resources used in the task completion	A weighted combination of the time spent on task, of the trajectory length and of the energy consumed (a distance in this 3D space): ***1 / (1 + distance)***	(0 1)
GTNC3	Number of the target objects collected	***Number of collected objects/ Number of objects to collect***	[0 1]

2.1.5 Non-cognitive metrics (relevant for ISTC)

index	**Metric**	***Function***	**Range**
GTNC-ISTC1	Energy consumption in function of simulated time	*Considering energy as the main agent' s resource, here we exploit* ***Energy = E(t)***	[0 1]
GTNC-ISTC2	Temporal distribution of 'instrumental' resources	*Each agent has limited resources that can be allocated to three operations: Control Rate, Range of*	[0,1]

[2] GTNC = Guards and Thieves Non Cognitive

		Vision, Speed. We can monitor how many resources are used: ***Control Rate + Range of Vision + Speed***	
GTNC-ISTC3	Average cost of the task in terms of used resources	**Total Amount of resources spent / Valuables collected (number of Tasks accomplished)**	(0 1]
GTNC-ISTC4	Average cost of the task in terms of travel distance	**Travel distance / Valuables collected (number of Tasks accomplished)**	(0 1]

2.1.6 Evaluation plan for ISTC-CNR:

As described in the deliverable D2.2, ISTC-CNR is implementing the same Guard and Thief Scenario using two different frameworks.

1. As for the first one, ISTC will build up a 3-layered architecture representing the guard in the guards and thieves scenario. They are currently building the first prototype architecture, only involving the first layer, and comparing it with some simpler systems lacking some or all the anticipatory capabilities.
2. In the second system ISTC defined a set of Anticipatory and Affective behaviours (including surprise and caution) developing 'families' of software agents with different competences. Some of their specific metrics aim at comparing agents with or without anticipatory and affective capabilities.

Both prototypes will be evaluated (by using the metrics GTC1, GTC2, GTNC1, GTNC2, GTNC3) in the first phase (April 2006 – September 2006). The second prototype is also being tested using GTNC-ISTC1.

In the second phase (October 2006 – March 2007), the first prototype will also include the second layer; it will thus also be evaluated by using the metrics GTC-ISTC1 and GTC-ISTC2. This work will be reported in the deliverable D4.2. The second prototype will be evaluated also according to GTC-ISTC4.

In the third phase (April 2007 – September 2007), the first prototypes will also include the third layer; the prototype will thus also be evaluated by using the metric GTC-ISTC3. This work will also be reported in the deliverable D4.3.

This evaluation work proceeds in parallel with the comparison with other partners' systems in the same scenario and with the integration (with additional cognitive capabilities), i.e. WP6.

2.2 Scenario: “finding and looking for” . Tasks: “robotic arm + camera for reaching”

2.2.1 Success questionnaire

1. Select specific tasks or full scenarios which can be used to evaluate the anticipatory mechanisms.
    a. *Scenario: “Find and looking for”. Tasks: identifying and reaching targets with a robotic arm + a camera (targets can be: involving/non-involving memory; single/ sequences; static/dynamic).*
2. What kind of human expert evaluation do you plan to use in your scenarios?
    a. None
3. Which objective metrics do you plan to use in your scenarios?
    a. Learning times
    b. Capacity to achieve the task (yes-no)
    c. Accuracy of performance (average error between targets and reached points)
    d. Speed of accomplishment of tasks (0-T)
4. What comparison with respect to non-anticipatory mechanisms do you plan?
    a. Learning times
    b. Capacity to achieve the task (yes-no)
    c. Accuracy of performance (average error between targets and reached points)
    d. Speed of accomplishment of tasks (0-T)
5. How does integration with another mechanism change the performance in your scenarios? (…the previous three metrics still do!)
    a. Learning times
    b. Capacity to achieve the task (yes-no)
    c. Accuracy of performance (average error between targets and reached points)
    d. Speed of accomplishment of tasks (0-T)
6. Please compare other models to yours with respect to anticipatory behaviour. What do you base it on? (…the previous three metrics still do!)
    a. Learning times
    b. Capacity to achieve the task (yes-no)
    c. Accuracy of performance (average error between targets and reached points)
    d. Speed of accomplishment of tasks (0-T)
7. Please compare your mechanism’s performance with human behaviour.
    a. We will carry out a comparison of performance of the model with that of animals and humans subjects engaged in similar experimental tests. The previous three metrics are still good for this comparision:
    b. Learning time
    c. Capacity to achieve the task (yes-no)
    d. Accuracy of performance (average error between targets and reached points)
    e. Speed of accomplishment of tasks (0-T)

2.2.2 Cognitive aspect metrics

Cognitive aspect	Index	Metric	Function	Range
Memory	RC1[3]	Capacity to solve tasks requiring memory of past events	How long can be the time spanning events relevant for acting	[0, T]
Attention	RC2	Capacity to gate-out distractors	How many targets in the scene can the system tackle	[1, N]
Dynamic actions	RC3	Capacity to track moving targets	Speed of moving targets that the system can track	[0, S]
Prediction of future target position	RC4	Capacity to anticipate future position of moving targets	Error between anticipated and actual position of targets	[0, E]
Prediction of right time for action	RC5	Capacity to anticipate time of future events	Error between time prediction (action execution) and the actual time of the target event	[0, T]

2.2.3 *Non-cognitive metrics*

Index	Metric	Function	Range
RNC1	Learning times	Time needed to improve and achieve steady performance in solving the task by exploiting any of the following cognitive functionalities	[0, T]
RNC2	Capacity to accomplish the task	Evaluate if system can accomplish the specific type of task	Yes-no
RNC3	Accuracy of performance	Measure error between targets and reached points	[0, R]
RNC4	Reaction times; Time for accomplishment	Measure how long the system takes to accomplish the task	[0, T]

[3] RC = Reaching Cognitive, RNC = Reaching Non Cognitive

3 NBU

3.1 Success questionnaire

1. Select specific tasks or full scenarios which can be used to evaluate the anticipatory mechanisms.

Looking for an object; Thieves and guards

2. What kind of human expert evaluation do you plan to use in your scenarios?
    a. Objective
    b. Based on performance
    c. Related to task achievement
    d. Related to comparison
    e. Related to the complexity of the task achieved

3. Which objective metrics do you plan to use in your scenarios?

Please look at the sections below.

4. What comparison with respect to non-anticipatory mechanisms do you plan?

Comparison with a model performing full space search. Comparison with our model with the anticipatory mechanisms ‘turned off’.

5. How does integration with another mechanism change the performance in your scenarios?

A possible integration regarding selective attention could be done with LUCS. Mechanisms for selective attention should speed up the anticipatory processes and make them more reliable. See d615 for details.

6. Please compare other models to yours with respect to anticipatory behaviour. What do you base it on?

We plan comparison of our model with connectionist ones (like AKIRA of ISTC).

7. Please compare your mechanism’s performance with human behaviour.

We have done some experiments with humans which show psychological plausibility of our model. We plan to do more experiments in future. Moreover, we would like to compare if possible the complexity of the tasks that can be achieved by human subjects and the robots, including comparison of time for completion.

3.2 Cognitive aspect metrics

Cognitive aspect	Index	Metric	Function	Range
Recognition / Mapping	M1	Similarity between the object found and the target object	Similarity using distance in the 3D space (shape x size x colour) ***1 / (1 + distance)***	(0 1]
Recognition / Mapping	M2	Belonging of the object found to the target objects class	Similarity using distance in an area in 3D space (shape x size x colour)	(0 1]

			1 / (1 + distance)	
Prediction / Anticipation	M4	Target object's position prediction/anticipation adequacy	Number of trials before finding the right object ***1 / Trial numbers***	(0 1]
Prediction / Anticipation	M7	Performance measurement	A combination of the number of unexpected opponent / rival meetings	(0 1]
Prediction / Anticipation	M8	Valuables protection	This metric is different for treasure-hunters and guards. For the guards it is based on the number of objects preserved and for the treasure-hunters on the number of objects violated or *collected and preserved* (see Scenario 2.3) ***Objects preserved count / objects ever owned count***	[0 1]

3.3 Non-cognitive metrics

Index	**Metric**	**Function**	**Range**
M3	Number of sessions which ended with a successful completion of the task	***Successive tasks count / all the tasks count***	[0 1]
M5	Resources used in the task completion	A weighted combination of the time spent on task, of the trajectory length and of the energy consumed (a distance in this 3D space): ***1 / (1 + distance)***	(0 1)
M6	Number of the collected target objects	***Number of collected objectst / Number of all the objects to collect***	[0 1]

3.4 Metrics usage

Scenario	**Metrics**
1.1. Finding a specific object	**M1 M3 M4**
1.2. Finding class of objects by class description	**M2 M3 M4**
1.3. Finding an object into a labyrinth	**M1 M3 M4 M5**

	Guard	Treasure-hunter
2.1. Treasure-hunters and guards (simple) 2.2. Treasure-hunters and guards (complex)	**M7 M8**	**M1 M3 M4 M5 M6 M7**
2.3. Wild west (several treasure-hunters and no guard (sheriff))	**M1 M3 M4 M5 M6 M7 M8**	

4 NOZE

Evaluation in autonomous cognitive and multi agent systems is a complex task that was often addressed with different strategies depending on the system itself and on the environment in which it is set up to work. Collected information about different evaluation methodologies [1][2][3] that also take into account industry standards considerations and successfully case of study. In the first section of this document NOZE will first point out what they found interesting and important to add to their evaluation metrics system. In the second section NOZE try to summarize an additional set of metrics to extend the overall partners proposals.

4.1 Different Metrics Strategies

Evaluation of autonomous system can be basically divided into 2 categories:

- **Environment independent:** the metric tries to identify an absolute performance and quality measure of the system that depend only on the system characteristics and not on the operational environment in which the system is acting.

- **Environment dependent:** the metric considers relevant the contextual performance and quality achievement of the system in a specific simulated or real environment.

These kind of metrics can be complementary and can be used together to evaluate any autonomous systems. It is important to point out that *Environment dependent* analysis is often a simpler challenge that lead to very good results when the goal of the annalists is to guarantee the achievement of clearly defined critical tasks (robots in rescue environments, controllers in complex machinery such as cars ...). However, in MindRACES an evaluation of anticipatory mechanisms both as environment independent and environment dependent is needed; an example of the first case is the "predictive power", an example of the second case is "the advantage of having anticipatory control of action, or perception, etc.". In the next section a generalized metric related to the *Absolute Complexity and Capabilities* of the autonomous system that is acting into the specific scenario is proposed. Other kinds of Environment Dependent metrics are also proposed both in cognitive and non-cognitive categories. A new category, computational metrics, is also discussed.

4.2 Proposed Metrics

Jean Piaget said: “To understand is to discover, or reconstruct by rediscovery...”[5], that lead to question ourselves if the capacity of an autonomous system to act in an environment (often uncertain) depend[4] on it capabilities of *self-adapt*[1] and *self-innovate*[2]. How many different kind of possible and plausible solutions to a problem can be proposed by the system having no

[1] Ability to modify its own working parameters to adapt its own behaviour to unknown inputs or to perform better with known inputs.
[2] Ability to add new mechanisms to its own repertoire to deal with unknown and known inputs.

information about the environment in which the problem will be solved? How much information need the system to evaluate a possible[5] solution to a problem? How much information stored into the system can be modified by incoming information? Is the system able to provide new solutions to previously solved problem? These questions, and many others of this kind, are obviously related to Cognitive Systems Capabilities but they don't need any particular environment representation to be discussed. If we consider, for example, a classical A* algorithm developed to solve an Approximate Travel Salesman problem one knows before any particular instance of the problem that the algorithm **has to store information about all the nodes and the relative weights of the graph that should be traversed. That fact depends on the A* itself** and not on a specific environment instance. The algorithm needs *complete information* about the problem space and so we can know a-priori that with *incomplete information* it will not be able to solve correctly any Approximate Travel Salesman problem instance. In that sense NOZE tries to identify a small set of simplified metrics that try to analyze their systems quality. These metrics will also try to take into consideration Psychological and Biological plausibility as in the Biometrical approach[3]. Here NOZE present a set of questions that characterize these measures.

4.2.1 Absolute Complexity

1. How many *Core Objects*[3] that determine the cognitive behaviour of your system are present in your formal model? **(1,N_0]**
2. How many different kinds of *Relevant Relations*[4] (Dependency, Use, Containment, Association, Inheritance) exist between *Core Object* in your formal model? **(1,N_1]**
3. How many *Relevant Relations* exist between *Core Objects* in your formal model? **(1,N_2]**
4. How many *Relevant Relations* determine changes in *Core Objects* status attributes in your formal model? **(0,N_3]**
5. How many *Core Objects* are able to exploit their *Relevant Relations* without relying on the operational context/environment type? **(0,N_4]**
6. How many *Core Objects* are biologically inspired or represents mechanisms that are biologically plausible? **(0,N_5]**
7. How many *Core Objects* are psychologically inspired or represents mechanisms that are psychologically plausible? **(0,N_6]**
8. How many *Relevant Relations* are biologically inspired or represents mechanisms that are biologically plausible? **(0,N_7]**
9. How many *Relevant Relations* are psychologically inspired or represents mechanisms that are psychologically plausible? **(0,N_9]**

[3] A Core Object is a fundamental entity (like a Class in Object Oriented Design) identifiable in your formal and computational model (even if not Object Oriented) that execute primary cognitive functionalities

[4] A Relevant Relation is one of the following : Dependency, Use, Containment, Association, Inheritance; it exhists between Core Objects and determine how its behaviour interact within your formal and computational model.

10.How many different source of inputs is your system able to exploit (numerical, propositional, image, sound, ...)? **(0,N_{10}]**

The overall metric is :

AC = 1 / exp(C)

where **C = 1 / Sum$_{[N0,...,N9]}$**

Absolute Capabilities

1. 1. Is your system able to operate with incomplete information about a given environment? **Yes [1] , No [0]**
2. 2. Is your system able to give different solutions to the same problem/task? **Yes [1] , No [0]**
3. 3. Is your system able to learn from one type of input source exploiting this information in front of a different kind of input source? **Yes [1] , No [0]**
4. 4. Is your system able to manage or to take into account its own resources (memory and time) to fulfill tasks? **Yes [1] , No [0]**

5. Is your system able to communicate in any human understandable formalism its current behaviour? **Yes [1] , No [0]**
And the overall metric is:

AB = Sum$_{[1,...,5]}$ / 5

These lead to:

ACC = (AC + AB) / 2

Note that the overall metric measure could be discarded in favour of a tabular representation of each feature with its own weight. Vectors produced so far could then be ordered in a taxonomical way. For example:

Absolute Capabilities:

Q1	**Q2**	**Q3**	**Q4**	**Q5**
1	*1*	*0*	*0*	*1*

4.2.2 Cognitive and Non-cognitive Environment Dependent Metrics

NOZE suggests that proposed measures based on the *Time Metric* (i.e. Learning, Time to complete the task, ...) will be changed in favour of other kind of context (environment and hardware) independent entities. Here are some examples:

Cognitive aspect	**Index**	**Metric**	**Function**	**Range**
Learning	C5**	Learning steps	**1 / (1 + learning steps)**	**(0 1)**

Where **1 step** correspond to **1 call** to the most relevant learning procedure of its own computational model (for example in a simple feed forward neural network the call to a backpropagation procedure; in a propositional based system the call to any procedure that update the content of its own knowledge base).

Index	**Metric**	**Function**	**Range**
G3	Relative time being in non-desired state (arrested, a chaser, etc)	**1 /exp (-number_of_state_transitions_from_desired_to_undesired_state))**	**[0 1]**

4.2.3 Computational Metrics

Evaluation of the computational load and resources usage of its own system is also an important measure of its quality. Here NOZE present a simple metric to take into consideration that aspect:

Cload = 1 / KbOfMemoryAllocated(n)

where *KbOfMemoryAllocated(n)* is the average value of the memory allocation of the program in Kb between *n* different simulations in a specific environment.

Final Considerations

NOZE points out that these kinds of metrics can be very useful from an industrial perspective. This metrics characterize features that are appealing to potential partners interested in the development of applications based on the models of NOZE giving them also a view of the complexity of their computational systems.

Bibliography

[1] Luis O. Arata; **Interactive Measures and Innovation** ; Department of Fine Arts, Language and Philosopy; Quinnipiac University, Hamden

[2] I.C. Ulinwa; **Understanding a Machine through Multiple Perspective Analysis** ; Walden University, Minnesota

[3] A. Meystel, J. Albus, E. Messina, D. Leedom; **Performance Measures in Intelligent Systems** ; PERMIS'03

5 OFAI

5.1 Success questionnaire

1. Select specific tasks or full scenarios which can be used to evaluate the anticipatory mechanisms.

   Game Room scenario and *Guards and Thieves* scenario

2. What kind of human expert evaluation do you plan to use in your scenarios?
   f. Objective
   g. Based on performance
   h. Related to task achievement

3. Which objective metrics do you plan to use in your scenarios?

   For details, see below in the document.

4. What comparison with respect to non-anticipatory mechanisms do you plan?

   Not applicable for our kind of architecture.

5. How does integration with another mechanism change the performance in your scenarios?

   This can only be measured in respect to the used mechanisms and therefore can be determined by the performance measures described by the respective partner.

6. Please compare other models to yours with respect to anticipatory behaviour. What do you base it on?

   We would base our comparison on the behaviour that the agents show externally and which are reviewed by a human observer.

7. Please compare your mechanism's performance with human behaviour.

   We think that our research and whole field of cognitive systems research has not reached a developmental stage that enables us to do that yet.

5.2 Cognitive aspect metrics

Cognitive aspect	Index	Metric	Function	Range

Recognition / Mapping	M1	Similarity between the object found and the target object	Similarity using distance in the 3D space (shape x size x colour) ***1 / (1 + distance)***	(0 1]
Recognition / Mapping	M2	Belonging of the object found to the target objects class	Similarity using distance in an area in 3D space (shape x size x colour) ***1 / (1 + distance)***	(0 1]
Prediction / Anticipation	M4	Target object's position prediction/anticipation adequacy	Number of trials before finding object at the right place ***1 / Trial numbers***	(0 1]
Prediction / Anticipation	GM1[6]	Performance measurement	A combination of the number of unexpected events in the agent – object interaction ***Number unexpected events / Interaction Trial Count***	(0 1]
Prediction / Anticipation	GM2	Successful hunting	***Number of Times the prey has been caught / Trial numbers***	[0 1]

5.3 Non-cognitive metrics

OFAI use the metrics M3 and M5 originally proposed by NBU and split them for the purpose their three developmental stages into more specific metrics (GM3-GM8); for obvious reasons M6 does not apply to their scenario.

Index	**Metric**	**Function**	**Range**
GM3	Developmental stage 1: Number of session in which the robot gained a set of useful behaviours by interacting with objects in its environment, i.e. reaching a previously defined developmental level	***Successive tasks count / all the tasks count***	[0 1]
GM4	Developmental stage 2: Number of session in which the robot successfully built an abstraction layer of network layer one	***Successive tasks count / all the tasks count***	[0 1]
GM5	Developmental stage 3:	***Successive tasks count / all the***	[0 1]

[6] GCM = Game room Metrics

	Number of session in which the prey was successfully caught	***tasks count***	
GM6	Developmental stage 1: Resources used in the task of the robot gaining a set of useful behaviours by interacting with objects in its environment, i.e. reaching a previously defined developmental level	A weighted combination of the time spent on task, of the trajectory length and of the energy consumed (a distance in this 3D space): ***1 / (1 + distance)***	(0 1)
GM7	Developmental stage 2: Resources used in the task of the robot successfully building an abstraction layer of network layer one	A weighted combination of the time spent on task, of the trajectory length and of the energy consumed (a distance in this 3D space): ***1 / (1 + distance)***	(0 1)
GM8	Developmental stage 3: Resources used in the completion of the “catching the prey” task	A weighted combination of the time spent on task, of the trajectory length and of the energy consumed (a distance in this 3D space): ***1 / (1 + distance)***	(0 1)

5.4 Metrics usage

Scenario	**Metrics**
1.1. Finding a specific object	**M1 M3 M4**
1.2. Finding class of objects by class description	**M2 M3 M4**
1.3. Capturing basic knowledge and learning to interact	**M1 M3 M4 M5 GM1 GM3 GM6**
2.1. Successful building of an abstraction layer of network 1	**M4 GM1 GM4 GM7**
3.1 Capturing the prey	**M1 M2 GM5 GM8**

6 UW-COGSCI

6.1 Success questionnaire

1. Select specific tasks or full scenarios which can be used to evaluate the anticipatory mechanisms.

Finding a specific object or members of a class of objects in the game room or the house:
The goal is to develop the capability of finding a specific object. Thus, the system is supposed to interactively evolve an object representation and then, due to a preferably internal drives, interact with the object in question. Clearly, to be successful, the system first needs to have this object in reach to accidentally interact with it and find this desirable. Later, then, the system "wants" to interact with that object action (arising motivation/emotion) and thus starts searching and/or approaching and/or touching and/or interacting with the desired object or object type.

Important additional criteria besides the already mentioned (outlined in the original NBU document, such as learning time, accuracy, speed, etc.) are the ***autonomy of the system***, the ***amount of pre-programming necessary/pre-programmed capabilities provided***, the ***learning flexibility*** of the system and the ***plasticity*** of the system.

Autonomy of system: The question is how much the system relies on external teachers and/or supervised feedback signals. Questions such as if the system is able to realize on its own how it can approach (touch, interact with) an object need to be considered. In the perspective of anticipatory behavioral control (Hoffmann, 1993, 2003) and the ideomotor principle, it would be most desirable to develop a system that autonomously starts to interact with an outside world by initial random movement patterns and then, due to these patterns, learns how to initialize and control goal-directed movement patterns.

Amount of pre-programming: Proper comparisons between systems are only valid, if the amount of pre-programming is considered as well. A system in which object recognition is hard coded is certainly interesting but much less interesting (and less "cognitive") than a system that develops an object recognition system solely based on observation of and interaction with the outside environment. Thus, the goal is to develop – albeit probably initially mediocre – systems that learn representations and movement patterns solely based on autonomous interactions with the environment. For this, learning algorithms are necessary, that are biased towards the development of object recognitions. Of course, anticipatory mechanisms seem to be the key to successfully do so.

Moreover, once object recognition is available, it needs to be considered how much of the "desire" to interact with an object needs to be pre-programmed. Most desirable, again, would be the integration of a motivational module that triggers goal representations that then trigger goal-directed actions. Thus, a hierarchical modular architecture needs to be designed (Poggio, Bizzi, 2004) that allows for the propagation of goal representations and consequent action initiation and control. The amount of pre-programming thus lies in the detailed pre-programming of the modules and their interaction with each other.

The most well-known system of that kind is Deb Roy's Ripley architecture (Roy, Hsiao, & Mavridis, 2004; Roy, 2005) in which the system maps words to an internal representation and finally to gripping actions. The system, albeit very successful and impressive, is completely pre-programmed and in this sense not cognitive and only hard-codedly anticipatory (and thus gets a very low score on the pre-programmed criterion, that is, high amount of pre-prog.).
Learning flexibility: In relation to the pre-programming and most likely an advantage of less pre-programmed systems can be expected to be their *learning flexibility*, that is, the flexibility to develop different internal representations according to the observed outside environment. For example, a system might be trainable to search for RED objects, however, a system might also be trainable to search for an object with a color property, for which a motivational module then decides, which one is desirable. Another example would be a system that first searches for red objects and encounters only either triangular or squared objects that are red. Next, the system might be confronted with a malfunctioning color sensor so that it should search either only for triangular or squared objects, according to the previously encountered correlation. Such emergent behavior is highly desirable and confirms system flexibility.

Plasticity: Besides the flexibility to learn what is encountered in the environment, it is also very desirable to have a life-long learning system that is able to adjust its behavior and internal representation to persistent changes in the environment. Locations of food sources may change, food item properties may change, behavior of objects may change, sensory accuracies and biases may change. An important criterion for a highly cognitive learning system is its capability to adjust to such changes.

2. What kind of human expert evaluation do you plan to use in your scenarios?
   i. Objective
   j. Based on performance
   k. Related to task achievement
   l. Related to comparison
   m. Other

Clearly, all the traditional forms of evaluation will be used including predictive accuracy (in terms of object location, object identity), learning speed (number of interactions with the environment), computational effort (CPU times, space requirements), noise robustness (noise in sensors and actuators), and scalability (more degrees of freedom, larger (sensory) input space). These criteria should then be put into relation with the aforementioned amount of autonomy and amount of pre-programming in the system as well as the degree of system flexibility and plasticity.

3. Which objective metrics do you plan to use in your scenarios?

The metrics certainly then will have to be concretized according to the specific scenario implementations.

4. What comparison with respect to non-anticipatory mechanisms do you plan?

Eventually, comparisons with hard-coded approaches may yield further data. However, currently our task is to create such an architecture that develops object representations autonomously.

5. How does integration with another mechanism change the performance in your scenarios?

Several mechanisms will be integrated in the architecture including notions of predictions, confidence in these predictions, resulting surprise, attention, etc. The addition of each of these mechanisms will be evaluated. Each of the additions should yield advantages according to the aforementioned evaluation criteria.

6. Please compare other models to yours with respect to anticipatory behavior. What do you base it on?

We base our models on the principles of anticipatory behavioral control and the ideomotor principle, that is, that behavior is continuously goal-directed and triggered by goal-representations.

7. Please compare your mechanism’s performance with human behavior.

The aspects above are all based on observations made in cognitive psychology and related fields. Concrete comparisons with human behavior will hopefully be possible once the initial architecture is available. Direct comparisons, however, are currently not planned.

6.2 Cognitive aspect metrics

Please describe the cognitive aspect metrics. Here is an example.

Besides the proposed metrics in the original NBU document (that, of course, need to be then adjusted to the actual task at hand), UW propose to also evaluate the following (C0a, C0b, C6, C7,C8):

Cognitive aspect	**Index**	**Metric**	**Function**	**Range**
System Autonomy	*C0a*	*Amount of supervision necessary*	***1/(1+degree of supervision) (sorry, I guess the degree of supervision is not so easy to quantify)***	***(0,1]***
Amount of pre-programming	*C0b*	*Amount of hard-coded task-relevant representations and programs*	***Also hard to quantify – the more learning due to interaction and observation the better.***	
Prediction / Anticipation	C1	Predicting/anticipating rival position (avoiding collision or catching)	**1 / (1 + collisions count)**	**(0 1]**
Prediction / Anticipation	C2	Rival / object position prediction preciseness	**1 / (1 + least mean square error)**	**(0 1]**
Recognition /	C3	Similarity between the	**1 / (1 + distance)**	**(0 1]**

Mapping		object found and the target object in the object description n-dimensional space		
Recognition / Mapping	C4	Signal recognition	**1 / (1 + signal inadequate acts count)**	**[0 1)**
Learning	C5	Learning time	**1 / (1 + time to learn)**	**(0 1)**
Flexibility	*C6*	*Flexibility to learn in different (but similar) environments, different but similarly successful representations and behavioral patterns*	***Amount of flexibility: Success of survival in different environments.***	
Plasticity	*C7*	*Capability of continuous adaptation*	***Change in the environment and speed to adaptation relative to the change at hand: speed/change***	
Noise Robustness	*C8*	*Capability of filtering / handling noise*	***Influence of noise on other performance criteria***	

6.3 Non-cognitive metrics

UW has no additional metrics besides the ones proposed by NBU:
Please describe your global metrics. Here is an example.

Index	**Metric**	**Function**	**Range**
G1	Time to complete the task	**1 / time**	**(0 1]**
G2	Time being in non-desired state (arrested, a chaser, etc)	**1 / (1 + time in non-desired state)**	**(0 1]**
G3	Relative time being in non-desired state (arrested, a chaser, etc)	**1 – (time in non-desired state) / (total time)**	**[0 1]**
G4	Objects/treasures collected count	**(objects collected count) / (all the objects number)**	**[0 1]**

6.4 Metrics usage

Generally all the metrics above apply to the game room and house scenario and the task to look for and retrieve objects (or types of objects) in these scenarios.

References:

Hoffmann, J. (1993). *Vorhersage und Erkenntnis: Die Funktion von Antizipationen in der menschlichen Verhaltenssteuerung und Wahrnehmung*. [Anticipation and cognition: The role of anticipations in human behavioral control and perception]. Hogrefe, Göttingen, Germany.

Hoffmann, J. (2003). Anticipatory Behavioral Control in Butz, M. V., Sigaud, O., & Gérard, P. (Eds.) *Anticipatory Behavior in Adaptive Learning Systems: Foundations, Theories, and Systems*. Springer-Verlag, Berlin-Heidelberg, 44-65.

Poggio, T., & Bizzi, E. (2004). Generalization in vision and motor control. *Nature*, 431, 768-774.

Deb Roy. (2005). Grounding Words in Perception and Action: Insights from Computational Models. *Trends in Cognitive Science*, 9(8):389-96.

Deb Roy, Kai-Yuh Hsiao, and Nikolaos Mavridis. (2004). Mental Imagery for a Conversational Robot. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 34(3):1374-1383.

7 IST

7.1 Cognitive aspect metrics

Cognitive aspect	Index	Metric	Function	Range
Emotion	MC1	Quality of emotions elicited, expression success.	Questionnaire to persons observing the scenario	[0 1]
Recognition	MC2	Distinction between object and other "similar"/distracting objects	When it's in front of object perform several trials with the different objects ***Right Recognitions / Number of Trials***	[0 1]
Prediction / Anticipation	MC3	Ignoring non-pretended objects without explicit evaluation of the object	Number of attention shifting stimulus versus Number of focus on object for evaluation ***1 – (Evaluations / AttentionShifts)***	[0 1]
Prediction / Anticipation	MC4	Response to "interesting" stimulus	Number of stimulus to the robot vs. number of attention shifts ***AttentionShifts / Number of Stimulus***	[0 1]

7.2 Non-cognitive metrics

Index	Metric	Function	Range
MN1	Number of Times ball was successfully found	***Successive tasks count / all the tasks count***	[0 1]
MN2	Time to find the ball	When ball is found ***1 / (1 + time from stimulus to recognition)***	(0 1]
MN3	Credibility of the agent	Questionnaire to persons observing the scenario. Evaluation on emotion, anticipation and	[0 1]

		credibility observed.	
MN4	Time to build the word	Time to complete the word	time
MN5	Games won	Percentage of games won in the games played	[0, 1]

7.3 Metrics usage

Scenario	Metrics
Finding and Looking for: Fetch That Object	**MC1 MC2 MC3 MC4 MN1 MN2 MN3**

8 LUCS

LUCS propose a two fundamental performance metrics that can be used in all the different scenarios.

1. **Average tracking error**. This is relevant for both the Guards and Thieves and Predicting in a Dynamic World. In the first case, it concerns tracking of thieves or guards and in the second the tracking of the dynamical object such as a ball. The environment and mechanisms can be changed in different ways to evaluate performance under different conditions. The coordinate system for the error measurements depends on the task, for example, image coordinates or world coordinates. Also different prediction times can be used to track future locations.
2. **Average time to complete task**. This measure is relevant in all scenarios although the tasks are different. Assuming other factors such as robot hardware and computational power is equal in all tests, the remaining influence on the task depends on the mechanisms used. The environment and mechanisms can be changed in different ways to evaluate performance under different conditions. This measure is also relevant for learning situations where convergence of the learning system is assumed to be the end of the task. For simulations, an arbitrary time-base can be used to still measure time rather. This assumes that the physical or simulated operations of the robot is the limiting factor and not computer speed.

These performance measures are definitively non-cognitive, but they have the great advantage that they are easily recorded and can be used for sound statistical analysis. It is LUCS' conviction that non-cognitive measures will also favour the more cognitive solutions if the environments are sufficiently realistic/complex. The factors that are not reflected in the measure can be address by using different environments and different tasks.

Direct cognitive evaluation is also interesting, but they are often not objective. It is hard come up with more cognitive metrics that does not automatically favour one model over another. They can still be useful for illustrative purposes however.

8.1 Success questionnaire (Average tracking error)

We first consider the case of average tracking error:

1. *Select specific tasks or full scenarios which can be used to evaluate the anticipatory mechanisms.*

   f. Tracking of thief or guard position/prediction of future location.

   g. Tracking the dynamic object/predicting future location

2. *What kind of human expert evaluation do you plan to use in your scenarios?*
   - k. Objective YES
   - l. Based on performance YES
   - m. Related to task achievement YES
   - n. Related to comparison YES
   - o. Other

3. *Which objective metrics do you plan to use in your scenarios?*
   - h. average tracking error
   - i. average prediction error for different prediction times

4. *What comparison with respect to non-anticipatory mechanisms do you plan?*
   - j. comparison with identical systems without prediction; e. g. feedback control

5. *How does integration with another mechanism change the performance in your scenarios?*
   - k. This remains to be seen! We will test different combinations of mechanisms.

6. *Please compare your mechanism's performance with human behaviour.*
   - l. We aim for human-like performance in the different tracking tasks although this is probably too hard in some cases.
   - m. Comparisons with the human cognitive system/brain systems are interesting to make

8.2 Success questionnaire (Average time to complete task)

Second, we consider the case of average time to complete a trask:

1. *Select specific tasks or full scenarios which can be used to evaluate the anticipatory mechanisms.*
   - n. Time to find an object.
   - o. Time to catch/reach an object
   - p. Time to learn/reach a particular performance level

2. *What kind of human expert evaluation do you plan to use in your scenarios?*
   - p. Objective YES
   - q. Based on performance YES
   - r. Related to task achievement YES
   - s. Related to comparison YES
   - t. Other

3. *Which objective metrics do you plan to use in your scenarios?*
   - q. average time to complete task

NOTE: This is a more useful metric than number of successful trials. It allows clearer statistics: mean/standard deviation/t-test/ANOVA etc. The only weakness is that it is necessary that all (or nearly all) trials eventually succeed.

4. *What comparison with respect to non-anticipatory mechanisms do you plan?*
    r. comparison with identical systems without prediction
    s. comparison with random strategies as a base case

5. *How does integration with another mechanism change the performance in your scenarios?*
    t. This remains to be seen! We will test different combinations of mechanisms.

6. *Please compare your mechanism's performance with human behaviour.*
    u. For most tasks, human level performance is not realistic
    v. It us useful and interesting to compare the systems to human performance based on:
        i. structural similarity of the model/mechanisms
        ii. type of errors made
        iii. overall behaviour

9 IDSIA

IDSIA has proposed the METRIC described below. They want to modify the metric a little bit, in the way that they want to use an "average error" measurement over a number (100) of runs. That needs to be made clear, probably.

In the ROOM environment:
- minimize average time (over 100 runs) to find the target object (NOTE: the object can be located in a different spot every run).
- do this with high probability of success (maximize the number of successful finds over 100 runs; maximum time per run is set to say 3 minutes; if it takes more, it is a FAILURE) - minimize transfer problems: maximize the ratio (successful runs on the real robot / successful runs in simulation) This last metric of course only makes sense when success in simulation is already high.

in

In a HOUSE/office environment: same evaluation, only more time per run (5 minutes). NOTE: in the house environment, the agent might itself be located a different starting position every run.
Two additional metrics might be - the amount of real-world experience should be minimized - the amount of CPU time should be minimized (this point is not very important, though)

This is very simple. IDSIA hopes it is enough, and thinks it just is not more complex. Prediction on where objects tend to be, anticipation on expected sensory inputs and memory where the agent has looked/been before all help to improve the score.

10 Summary

In the table below, most of the proposed metrics are summarized. The metrics are sorted with respect to the scenario and the cognitive aspect for which they were proposed. They were not grouped together and the diversity was kept in order to have different choices which will be specified at the moment of their use. Duplicated metrics were excluded and the similarity is mapped to proximity in the tables (consecutive rows). From each group of metrics only one or two will be chosen after concrete consideration of the scenarios and the tasks.

10.1 Guards and thieves scenario

10.1.1 Recognition / Mapping

Index	Metrics	Function	Range	Partner	subjective
M1	Similarity between the object found and the target object	Similarity using distance in the 3D space (shape x size x colour) *1 / (1 + distance)*	(0 1]	NBU,UW	
M2	Belonging of the object found to the target objects class	Similarity using distance in an area in 3D space (shape x size x colour) *1 / (1 + distance)*	(0 1]	NBU	

10.1.2 Prediction / Anticipation

Index	Metrics	Function	Range	Partner	Subjective
M4	Target object's position prediction/anticipation adequacy	Number of trials before finding the right object *1 / Trial numbers*	(0 1]	NBU	
GM2	Successful hunting	*Number of Times the prey has been caught / Trial numbers*	[0 1]	OFAI	
C1	Predicting/anticipating rival position (avoiding collision or catching)	1 / (1 + collisions number)	(0 1]	UW	
M7	Performance measurement(*even if for some systems it is difficult to exactly separate expected/unexpected*)	A combination of the number of unexpected opponents / rival meetings	(0 1]	NBU, ISTC	
GM1	Performance measurement	A combination of the number of unexpected events in the agent – object interaction *Number unexpected events / Interaction Trial Number*	(0 1]	OFAI	

Index	Metrics	Function	Range	Partner	Subjective
M8	Valuables protection	This metric is different for treasure-hunters and guards. For the guards it is based on the number of objects preserved and for the treasure-hunters on the number of objects violated or *collected and preserved* *Number of remaining objects / Number of protected objects by this guard*	[0 1]	NBU	
GTC2	Valuables protection/collection	This metric is different for treasure-hunters and guards. For the guards it is based on the number of objects preserved and for the treasure-hunters on the number of stolen objects or *collected and preserved* *Objects preserved number / objects ever owned number*	[0 1]	ISTC,UW	
L1	Tracking of thief or guard position/prediction of future location; Average tracking error/average prediction error	LMS Error	Real	LUCS	

10.1.3 *Learning*

Index	Metrics	Function	Range	Partner	subjective
C5	Learning steps	1 / (1 + learning steps)	(0 1)	NOZE	YES

10.1.4 *Other*

Index	Metrics	Function	Range	Partner	subjective
M3	Number of sessions which ended with a successful completion of the task (for the guard: capture the thief; for the thief: collection of the valuables)	Successive tasks number / all the tasks number	[0 1]	NBU, ISTC, OFAI	
M6	Number of the target objects collected	Collected objects' number / all objects' number	[0 1]	NBU, ISTC	
M5	Resources used in the task completion	A weighted combination of the time spent on a task and of the energy consumed (a distance in this 2D space): 1 / (1 + distance)	(0 1)	NBU, ISTC, OFAI, LUCS	
G1	Time (steps) to complete a the task	1 / time (steps)	(0 1]	UW, IDSIA, LUCS	
G3	Relative time being in non-desired state (arrested, stuck against a wall, etc)	1 /exp (-number_of_state_transitions_from_desired_to_undesired_state))	[0 1]	NOZE	

Index	Metrics	Function	Range	Partner	subjective
G2	Time being in non-desired state (arrested, a chaser, etc)	1 / (1 + time in non-desired state)	(0 1]	UW	
G3	Relative time being in non-desired state (arrested, a chaser, etc)	1 – (time in non-desired state) / (total time)	[0 1]	UW	
C8 Noise Robustness	Capability of filtering / handling noise	Influence of noise on other performance criteria		UW	
MN3	Credibility of the agent	Questionnaire to persons observing the scenario. Evaluation on emotion, anticipation and credibility observed.	[0 1]	IST	YES
MG1	Complexity of the task that can be dealt with	Several tasks of the same type a ranked by human experts from less to more complex	Number of the most complex task successfully completed	IST	YES

10.2 Look for an object scenario

10.2.1 Recognition / Mapping

Index	Metrics	Function	Range	partner	subjective
C4	Signal recognition	1 / (1 + signal inadequate acts number)	[0 1)	UW	
M1	Similarity between the object found and the target object	Similarity using distance in the 3D space (shape x size x colour) 1 / (1 + distance)	(0 1]	NBU, UW	
M2	Level of belonging of the object found to the target object class	Similarity using distance in the class feature space: 1 / (1 + distance)	(0 1]	NBU	

10.2.2 Prediction / Anticipation

Index	Metrics	Function	Range	partner	subjective
C2	Rival / object position prediction precision	1 / (1 + least mean square error)	(0 1]	UW	
RC5	Capacity to anticipate the time of future events	Error between time prediction (action execution) and the actual time of the target event	[0, T]	ISTC	
M4	Target object's position prediction/anticipation adequacy	Number of trials before finding the right object 1 / Trial numbers	(0 1]	NBU	

10.2.3 Learning

Index	Metrics	Function	Range	partner	subjective
C5	Learning time	1 / (1 + time to learn)	(0 1)	NOZE	

10.2.4 Other

Index	Metrics	Function	Range	partner	subjective
C8	Noise Robustness Capability of filtering / handling noise	Influence of noise on other performance criteria	Depending on the metric	UW	
G1	Time (average) to complete the task	1 / average time	(0 1]	UW, IDSIA, LUCS	
G2	Time being in non-desired state (arrested, a chaser, etc)	1 / (1 + time in non-desired state)	(0 1]	UW	
RNC4	Reaction times; Time for accomplishment	Measure how long the system takes to accomplish the task	[0, T]	ISTC	
G3	Relative time being in non-desired state (arrested, a chaser, etc)	1 – (time in non-desired state) / (total time)	[0 1]	UW	
G4	Objects/treasures collected number	(objects collected number) / (all the objects number)	[0 1]	UW	
RNC1	Learning times	Time needed to improve and achieve steady performance in solving the task by exploiting any of the following cognitive functionalities	[0, T]	ISTC	
RNC2	Capacity to accomplish the task	Evaluate if system can accomplish the specific type of task	Yes-no	ISTC	
RNC3	Accuracy of performance	Measure error between targets and reached points	[0, R]	ISTC	
CR	Evaluation of the computational load and resources usage of its own system	Amount of CPU time and memory needed for the completion of the task.		NOZE, LUCS	

10.3 Fetch/catch that object scenario

10.3.1 Recognition / Mapping

Index	Metrics	Function	Range	Partner	subjective
MC2	Distinction between object and other “similar”/distracting objects	When it’s in front of object perform several trials with the different objects: Right Recognitions / Number of Trials	[0 1]	IST	

10.3.2 Prediction / anticipation

Index	Metrics	Function	Range	Partner	subjective

Index	Metrics	Function	Range	Partner	subjective
RC4	Capacity to anticipate future position of moving targets	Error between anticipated and actual position of targets	[0, E]	ISTC	
MC3	Ignoring non-pretended objects without explicit evaluation of the object	Number of attention shifting stimulus versus Number of focus on object for evaluation *1 – (Evaluations / AttentionShifts)*	[0 1]	IST	
MC4	Response to “interesting” stimulus	Number of stimuli to the robot vs. number of attention shifts *Attention Shifts / Number of Stimuli*	[0 1]	IST	

10.3.3 Emotion

Index	Metrics	Function	Range	Partner	subjective
MC1	Quality of emotions elicited, expression success.	Questionnaire to persons observing the scenario	[0 1]	IST	YES

10.3.4 Other

Index	Metrics	Function	Range	Partner	subjective
RC3	Dynamic actions Capacity to track moving targets	Speed of moving targets that the system can track	[0, S]	ISTC	
MN1	Number of Times ball was successfully found	Successive tasks number / all the tasks number	[0 1]	IST	
MN2	Time to find the ball	When ball is found *1 / (1 + time from stimulus to recognition)*	(0 1]	IST	
MN3	Credibility of the agent	Questionnaire to persons observing the scenario. Evaluation on emotion, anticipation and credibility observed.	[0 1]	IST	YES

10.4 General metrics, related to a cognitive model or architecture

Index	Metrics	Function	Range	partner	subjective
N17	Absolute capabilities Overall metrics based on N12-N16	AB = mean(N12, ..., N16)	Real number	NOZE	
N12	Is your system able to operate with incomplete information about a given environment?		Yes [1] No [0]	NOZE	
N13	Is your system able to give different solutions to the same problem/task?		Yes [1] No [0]	NOZE	

Index	Metrics	Function	Range	partner	subjective
N14	Is your system able to learn from one type of input source exploiting this information in front of a different kind of input source		Yes [1] No [0]	NOZE	
N15	Is your system able to manage or to take into account its own resources (memory and time) to fulfill tasks?		Yes [1] No [0]	NOZE	
N16	Is your system able to communicate in any human understandable formalism its current behaviour?		Yes [1] No [0]	NOZE	
C0a	Amount of supervision necessary	1/(1+degree of supervision)	(0,1]	UW-COGS	
C0b	Amount of hard-coded task-relevant representations and programs	Amount of hard-coded task-relevant representations and programs		UW-COGS	

10.5 Selected set of metrics

Using the summary above, we selected the following set of metrics. In the last column the partners who proposed these or similar metric are placed. Integration between them should be considered in the specified scenarios.

10.5.1 Objective

Index	Metrics	Function	Range	Scenario	Partner
A1	Number of successful tasks	Successive tasks number / all the tasks number	[0 1]	ALL	IST,ISTC,OFAI , NBU
A2 *	Time (steps) to complete a the task	1 / time (steps)	(0 1]	ALL	UW, IDSIA, LUCS, ISTC, NBU, OFAI
A3	Anticipating events	A combination of the number of unexpected events in the agent – object interaction *Number unexpected events / Interaction Trial Number*	(0 1]	ALL	OFAI,NBU,UW, ISTC
A4	Similarity between the object found and the target object	Similarity using distance in the 3D space (shape x size x colour) 1 / (1 + distance)	(0 1]	Guard and thieves & Look for an object	NBU, UW
A6	Number of the target objects collected/preserved	Collected/preserved objects' number / all objects' number	[0 1]	Guard and thieves	ISTC,UW,NBU
A7*	Dynamic actions Capacity to track moving targets	Speed of moving targets that the system can track	[0, S]	Fetch the object	ISTC

* Of course the importance of hardware used here is essential. Hence these metrics should be used for models running on the same computers.

10.5.2 Subjective

Index	Metrics	Function	Range	Scenario	Partner
B1	Credibility of the agent	Questionnaire to persons observing the scenario. Evaluation on emotion, anticipation and credibility observed.	[0 1]	ALL	IST
B2	Complexity of the task that can be dealt with	Several tasks of the same type a ranked by human experts from less to more complex	Number of the most complex task successfully completed	ALL	IST
B3	Quality of emotions elicited, expression success.	Questionnaire to persons observing the scenario	[0 1]	ALL	IST