Benchmarking in Developmental Robotics

Frank Guerin (1) f.guerin@abdn.ac.uk
Lauriane Rat-Fischer (2) lratfischer@gmail.com
(1) University of Aberdeen,
King's College, AB24 3UE Aberdeen, Scotland.
(2) Laboratoire Psychologie de la Perception, CNRS UMR 8158,
Université Paris Descartes 45 rue des Saints-Pères,
75270 Paris Cedex 06, France.

December 16, 2013

Abstract

There is at present no standard benchmarking for assessing and comparing the various existing works in developmental robotics. Developmental robotics is more of a "basic science" research endeavour than more mainstream robotics, which is more application focussed. For this reason benchmarking for developmental robotics will need a more scientific basis, rather than a specific application focus. The solution we propose is to benchmark developmental robotics efforts against human infant capabilities at various ages. The proposal here may allow the community to showcase their efforts by demonstration on common tasks, and so to enable the comparison of approaches. It may also provide an agenda of incremental targets for research in the field.

1 Introduction

Work in developmental robotics is primarily inspired by the cognitive development of human infants [19, 2, 6, 14]. There is at present no standard benchmarking approach for assessing and comparing the various existing works in developmental robotics. It may be reasonable to measure the success of bio-inspired solutions with respect to their natural models; this is proposed here.

In order to get inspiration we can look at various other benchmarking efforts in Computer Science, and competitions, and also at the types of testing done on human infants and comparative studies across species. If we compare existing robot/Computer Science benchmarks with the kinds of testing done in studies of animals or human infants we find that animal and human testing is more "scientific", in that it is testing general abilities, based on some hypothesis about abilities [12]. Robot testing and competition, in contrast, tends to be more "applied" in focussing on industrially

relevant and/or quite specific scenarios (for example RoboCup@Work[1] RoboCup Soccer, RoboCup Rescue, RoboCup@Home[2] or PASCAL Visual Object Classes [10]) rather than the underlying competences for general intelligent behaviour. In most cases this is reasonable when the goal is to engineer systems for specific applications, however if the goal is a more general purpose intelligent behaviour (as is the overall goal of developmental robotics [2]), then a more scientific approach to benchmarking is warranted. For this reason our proposal is primarily based on borrowing ideas from testing in infants. However, we also combine this with the robot competition idea, for reasons discussed in the text.

2 Motivating the Benchmarks

We wish to propose tests which are appropriate for the developmental robotics community, meaning that they must capture aspects of *development*. However we must avoid tests that are so difficult that they will be ignored; the type of ongoing development displayed by human infants [27] is likely to be beyond the state-of-the-art for some time to come. In addition we want tests which establish some standard measures, and have some stability from year to year, and for this purpose it seems reasonable to anchor the tests to measures of human infant competences at various ages. Such tests can allow developing artificial systems to compare their systems against a standard benchmark. At the same time we recognise that infant tests may need to be adapted for robots, just as comparative research working with non-human species needs to adapt tests.

If we compare testing robots (benchmarking) and testing of infants or animals, the following points of difference need to be considered:

1. Infants and animals have intrinsic behaviour. E.g. when they see a piece of food or a brightly lit object, it is likely to elicit some behaviour, without any instruction or demonstration being provided; e.g. they may be likely to make an attempt to take possession of it; when they see a string they may well pull it, perhaps just to see what happens. Robots typically have no behaviour unless programmed to do something specific. For some tests, like an immediate test of competence, we can simply instruct the roboticist on what behaviour is needed. For other tests such as developmental ability, robots need to be programmed with intrinsic behaviour (as done by some researchers [23, 30]).
2. The tests applied to animals are often not suitable for programmed robots because they often do not test abilities such as simple transfer and generalisation (there are exceptions of course [4]). Generally people testing infants or animals assume that simple transfer/generalisation comes for free[3], because it typically does in animals and infants; this is not the case in robots. For example infant

[1] http://www.robocupatwork.org/

[2] http://www.robocup.org/

[3] By simple transfer we mean the way that the infant can still be successful if lighting conditions are altered, or the toy to be retrieved is changed slightly, or the table surface texture is changed, etc.

psychologists or ethologists do not try to test if the infant or animal can still succeed if the light source in the room is moved to another side so that the lighting on the objects is changed. Such a seemingly insignificant change can disturb robots because their computer vision algorithms for object recognition can be quite fragile and may be confused by a different shadow. This means that any test on robots must rigorously test the robot's ability to do simple transfer under varying conditions. This is in order to avoid a robot team using a lot of very situation-specific programming to obtain a high test result under very precisely known conditions, but in fact having a system which is vastly inferior to an animal under varied conditions.

3 Dimensions and Branches of Abilities to Test

The central idea of this paper is to compare artificial systems with infants; e.g. a robot may have some abilities which are comparable to an infant of 6 months, 9mts, 12, 15, or 18 months. If one is comparing a robot with an infant at any age, one can compare on a number of dimensions:

Dim.1 The immediate ability on a one-shot chance, i.e. can it do the task right now. (Note that in line with Sec. 2 even this test must involve some variations in situation to thoroughly test a robot.)

Dim.2 The wider generalising ability; e.g. if it can pull one string to get the toy, what if we change it to a wide ribbon instead of a string, or even a towel? Infants tend to be good on this dimension and robots not. This encompasses object variety, task variety, environment conditions, body conditions.

Dim.3 The rapid learning ability of the robot/infant; a 9-month-old infant will not be able to pull a string to get a toy if she has no practice, but at this age she may be able to learn it in a single day, or within two days. This learning ability encompasses the ability to learn from human demonstrations as well as through autonomous exploration to discover ways to overcome problems e.g. by variation of motor control programs, or strategy change (e.g. to transfer from a different source skill to apply to the target problem).

Dim.4 The developing ability (slow learning) of the infant/robot: a 9-month-old infant may fail many tasks, but has the developing ability to get better at these tasks in a timescale of months. Furthermore the infant will get better without any specific training. It has the intrinsic motivation to do the "right" type of object play to create new learning experiences in a suitable order. This developing ability is only exhibited to a very limited degree, if at all, by present day robots. This developing ability encompasses many sub-abilities, such as the discovery of new affordances.

It is necessary to consider these dimensions because if one were to create standard tests (borrowed from the infant literature) with a predefined

set of objects, then it might be possible to make a robot outperform a 9-month-old infant on all of them. But the robot is really only outperforming on the first dimension above; on just about every other dimension the infant is vastly superior. We need tests that test all dimensions, ideally.

At the same time we cannot set unreasonable expectations for robots. So it may (at the present time) be pointless to compare infants and robots in terms of ability to discriminate object by haptic exploration [5, 31] because the hardware currently available to robots is inadequate.

In addition to (and orthogonal to) the dimensions above we can also design tests on various "branches of development" [34, p. 101] corresponding to domains of competence. Borrowing from Uzgiris and Hunt [34][4] may be a good starting point here (we use the term "agent" to apply to robot or infant):

Scale I The Development of Visual Pursuit and the Permanence of Objects. The simpler end of this scale involves sustained following of an object visually; the advanced end involves making accurate inferences about where an object is after it is put in a container and the container is seen undergoing successive displacements.

Scale II The Development of Means for Obtaining Desired Environmental Events. The simpler end of this scale includes repeating an action which just caused an interesting effect (the agent may have done this accidentally, and all that is required is that the agent be capable of repeating this). Visually directed grasping also appears towards the simpler end of this scale. The advanced end includes using a stick to obtain an object that is out of reach on a horizontal surface, without a demonstration needing to be provided.

Scale IIIa The Development of Vocal Imitation. The simpler end of this scale includes repeating sounds that the agent has just produced itself; the advanced end involves direct repetition of new words.

Scale IIIb The Development of Gestural Imitation. The simpler end of this scale involves matching the agents own movements to body movements presented, for familiar body movements; the advanced end involves imitation of a gesture which is known to be unfamiliar to the agent, even for gestures which cannot be seen by the agent on its own body, for example, pulling the ear.

Scale IV The Development of Operational Causality. The simpler end of this scale involves some tests which are identical to those on Scale II, but while Scale II goes on to focus on the development of an ability to plan to achieve goals, and to make adjustments to means actions, Scale IV focuses on understanding that there are sources of causality external to the agent. The advanced end of Scale IV involves the agent handing a toy to a human to request that a particular demonstration be repeated, or attempting to activate a mechanical toy by actively seeking out the correct manipulation of some part of the toy that could activate it.

[4]There are not many examples of works in developmental robotics which compare with infant scales of development, although Kido et al. [17] do compare with the Kyoto Scale of Psychological Development. This scale however appears to be only available in Japanese.

Scale V The Construction of Object Relations in Space. The simpler end of this scale involves looking alternately at two different objects at different locations; the advanced end involves constructing a tower of objects, or relating in space a tool and an object initially separated, to perform a tool action, or making a detour to reach an object that rolled behind some furniture, or detecting the absence of a person from their habitual location.

Scale VI The Development of Schemes for Relating to Objects. The simpler end of this scale involves visual inspection of an object; intermediate behaviours include applying motor schemes which are appropriate to the particular characteristics of the objects (often called exploiting affordances in contemporary robotics); the advanced end involves understanding the social uses of objects, and verbal naming of objects.

All of these scales seem quite suitable for testing in developmental robotics, with the exception of vocal imitation perhaps, because it may be too easy for artificial systems.

For each scale we can define a battery of suitable tests for various infant ages, where infants of that age performing at the median of a typical sample would pass all the tests. Now to take a concrete example: suppose we wish to benchmark a robot against a 9-month-old "median infant", taking the domain of competence of means for obtaining effects (Scale II), which would include tasks such as pulling a support object to retrieve a supported object. Suppose further that the robot is only being compared on dimension Dim.1, and that the robot has an overall success of 76% on the battery of tests (the median infant would be expected to score 100%), then one could abbreviate and say the robot had 9mt-Dim.1-ScaleII:76%(2013). The same robot might also score 12mt-Dim.1-ScaleIIIa:100%(2013); i.e. it is performing like a 12-month-old infant in Dimension 1 on the scale for vocal imitation abilities. Roboticists could quote these scores for comparison. This also presumes that the robot must also be able to pass all tests in Scale II and IIIb for every younger age. For example a visual recognition system that may perform well on naming objects cannot be scored on Scale VI if it cannot also physically manipulate objects (a skill associated with an earlier age on that scale). If a robot is being tested across each of the six scales then one could average the score, but this is unlikely to be a useful measure because few roboticists would attempt competence in all domains; most will specialise.

4 Implementation and Replicability

There is a tension between the requirement for replicable tests and the need to have unseen tests for robots. It would not make sense to publish details in advance of a test; even if one has hundreds of objects for testing, a robotics team could prepare in advance by training on all of them. This is well known since the early days of AI, because Turing's test inserts a human to ask questions from an unlimited repertoire so that pre-scripted responses will not work [32]. It is also recognised in the RoboCup@Home

project: "As uncertainty is part of the concept, no standard scenario will be provided in the RoboCup@Home League."[5]

Tests therefore need a test centre to do the assessment. Like the various RoboCup competitions. Details can be released after the event and people can try a post-test to see what score they get, but it cannot be compared with an in-test score.

Ideally the benchmark scores of a particular robotics system would be stable from year to year, but in practice this is impossible because the test materials must necessarily change. Some systems might get lucky and get a high score in one year because the particular materials and situations were suitable for them. This is why a quoted score should be appended with the year of the test. However there is an anchor point to keep the tests broadly similar from year to year because it is a measure against infants; infants are a fixed reference point for benchmarking. For example if we take tool use ability in Scale II at 18 months then we could foreseeably continuously devise new tests for the next ten years which would be doable by infants but which would seriously challenge robots. For example, as robots become able to do some of the 2014 tasks more competently then new materials such as plastic bags or play-dough may be introduced. Robots performing well in 2014 might perform poorly in 2024 due to changing materials, even though the tests would be of roughly equivalent difficulty for an infant. To design varied tests for ten years (across the four dimensions) we would need some control longitudinal studies in human infants, because there is insufficient data available at present on infant capabilities. Thus there is a limit of our proposition in the present paper unless we get some concrete data in the near future from psychology. An additional issue for stability of the tests is that we only know so much about infants at present, so the reference point may change slightly with ongoing psychological research.

Replicability is desirable in scientific results, and indeed tests can be replicated once the details of the materials of a particular year are published. However in a fast moving field like robotics tests may rapidly lose their relevance in the longer term. This is abundantly evident in areas such as the DARPA Grand Challenges which have already moved through three distinct challenges to keep at the leading edge of current capabilities (in Grand Challenge 2004 no car even came close to finishing the course, but in Grand Challenge 2005 five teams finished). Also in computer vision: the PASCAL dataset has constantly evolved and the series is now finished[6]. Where benchmarks do not evolve the community working on them can become stuck in a rut. Paul Cohen in his assessment of what makes a good challenge suggests that "The challenge should be administered frequently, every few weeks or months, and the rules should be changed at roughly the same frequency to drive progress toward the long-term goals." [9] Note his highlighting of the "long-term goals"; i.e. if the rules are not changed the efforts to succeed at a challenge may end up in clever engineering for a narrow goal which has lost sight of the original goals that motivated the creation of the challenge. One is also reminded of

[5] http://www.robocupathome.org/rules

[6] http://pascallin.ecs.soton.ac.uk/challenges/VOC/

the well-known adage that "When a measure becomes a target, it ceases to be a good measure."

5 Detail on Benchmarks and Testing

In this section we look at Scale II in some detail and give examples for tests in each dimension.

5.1 Detail of Scale II

Note we are not sticking to Uzgiris and Hunt exclusively here, in particular some ages are adjusted to reflect more recent research. Note that all the tests here are for what we are referring to as a "one-shot chance" (Dimension 1). Even though in practice two attempts may be tried, and the agent may be allowed 20 seconds to prepare the attempt and to do a little groping/readjustment of behaviour during the attempt, keeping roughly inline with standard practice in infant assessment.

Infant Age Months	Test Behaviours
2	Hand-eye coordination; watching the hand
3	Able to repeat motor schemes that brought an interesting result; e.g. attempts to keep a hanging toy in motion by repeated movements
6	100% success on visually directed grasping of small (hand-sized) objects [3, p. 174]
8	Execution of one action (means) preparatory to executing another, and adaptation of the means action: Drops an object held in the hand in other to free the hand for grasping another. Removes an obstacle blocking a desired toy Pulls a supporting object in order to retrieve an out-of-reach toy.
9	Use a common behaviour pattern as means for multiple ends: Use some form of locomotion to retrieve an object needed for another activity.
10	More advanced anticipation of effects of means action: Does not pull a supporting object when the toy does not rest directly on it.
12	Able to exploit alternate means to obtain objects: Retrieves and out-of-reach toy across a horizontal surface by means of a string tied to it.
13	Retrieve a toy not directly in sight by pulling a string vertically, typically requiring bimanual releasing and re-grasping.
18	Use a stick-like tool to retrieve an out-of-reach toy across a horizontal surface.

20	Display foresight by using an appropriate means in a problem situation; e.g. putting a necklace into a tall container, foresee the likely fall of the container and adopt a successful approach from the start.
22	Recognition of hindrance toward an end, implying advanced representation; e.g. does not attempt to stack a solid ring (among other rings) onto a peg.

5.2 Testing Dimension 1

Recall that Dimension 1 is the immediate ability on a one-shot chance, i.e. can the agent do the task right now. To test visually directed grasping for example the rough size of the object can be specified in advance of the test, but the particular objects will not be known in advance, nor would the other conditions such as the surface on which the objects are placed or the other objects in the near vicinity. In the infant case infant-attractive objects are usually used such as a toy with a coloured light inside. This same technique could be used in the robot case to indicate to the robot which toy it should grasp, if there is more than one in view. For advanced behaviours such as using a stick to retrieve a toy, slightly varied sticks and toys and surfaces may be used in order to make it difficult for a very rigid robot behaviour to be unreasonably successful.

The introduction of variety here has a parallel with research on infants in that tasks are also varied in infant studies; with infants, the target object must sometimes be changed to increase the motivation to retrieve it and keep the infants' attention for the task, whereas with robots, the target is changed to ensure that the robot is capable to generalize. Infants have an intrinsic "attraction for novelty" [24] which cannot be disabled, so experimenters must use it to keep attention. Robots can of course also be programmed with an "attraction for novelty", but this is not required for Dimension 1, but rather for Dimension 4 (below).

For the even more advanced tasks such as "Displaying foresight by using an appropriate means in a problem situation" an array of tasks, specified without completely precise detail in advance of the test, can be presented, provided all are within the realm of a typical 22-month-old infant's capabilities. This puts 22mt-Dim.1-ScaleII well beyond the capabilities of any present day robot. However there are present day robots which could obtain a reasonable (perhaps 50%(?)) score on 6mt-Dim.1-ScaleII.

The scale elaborated by Uzgiris & Hunt proposes a precise coding based on two to six main categories of behaviours for each task. The behaviours that are considered critical for achievement of a step in the scale are clearly marked. For example, the execution of visually directed grasping of small objects is coded into four main behavioural categories: reaching but not grasping the object, grasping the object when both object and hand are in view simultaneously, grasping the object when only the object is in view, and grasping the object after anticipating the contact between the hand and the object. The three last categories are considered as critical for showing an agent's success, but are not succeeded all at

the same age in infancy. Another example, the use of stick as means for retrieving an object is coded into five behavioural categories: playing with the stick, trying to reach for the toy directly, playing with stick and toy without retrieving, obtaining the toy by means of the stick after demonstration, and without demonstration. Either one of the last two behaviours is an adequate criterion for showing an agent's success; i.e. success after having encouraged infants to do the task, and/or having showed them how to perform the task, is coded as success. This is also appropriate for our robot tests, but it must be an independent tester who does the demonstration, and not a member of the robot's team.

As already mentioned above, the scale proposed by Uzgiris & Hunt is just a starting point for several reasons. First, the coding may have to be re-adapted for robot testing, in the same way that a coding is sometimes adapted in comparative research when working with non-human species. Second, an update with more recent literature in developmental psychology and comparative research may be necessary for some of the tasks proposed here. Going back to the execution of visually directed grasping, an important aspect in which developmental psychologists have been particularly interested in the last decade is the agent's ability to visually and motorically anticipate the grasping action. In contrast, Uzgiris and Hunt have removed the category containing the "anticipatory" component in their scaling analysis, because they observed it too infrequently. Since their work, developmental psychologists have shown that planning an efficient grasping of an object develop later in infancy, e.g. [20, 21, 8]. Measures such as eye-gaze, velocity of the reaching hand, and pre-orientation of the hand before grasping have been used to evaluate infants' planning abilities of successful grasping.

Concerning the use of stick as means for retrieving, recent work in the literature has provided a detailed analysis of comparable tool use behaviours (with a rake-like tool instead of a stick (e.g. [22]). The coding was based on a 5-categories coding slightly different from the one proposed by Uzgiris & Hunt, with a detailed description of 26 typical behaviours embedded within the categories. In this analysis, success after observation was coded apart from spontaneous success without observation of the target action, which partly explains that the authors found successful tool use at a later age than specified by Uzgiris & Hunt. (22 months = first spontaneous success, 18 months = first success by observation [28].)

5.3 Testing Dimension 2

Recall that Dimension 2 is the wider generalising ability; e.g. if the task is pulling a string to get a toy, we may change to a wide ribbon instead of a string, or even a towel. If the task is retrieval with a stick (18 months) a variety of non-stick objects can be presented to see if the robot chooses from among them with the same skill as an infant of 18 months (for example the too-short, or too-flexible objects used in Brown's study [4]). There exists substantial comparison data with infants for a rake tool[7]. The task may also require the tool canonically used for task A to

[7]Work in progress by one of the authors in Univ. Paris-Descartes

be used for task B, such that robots should not be limited by functional fixedness any more than infants are. Variations encompass object variety (both target object and tool object), task variety, environment conditions (e.g. lighting, table surface, clutter), body conditions (a weight can be attached to the robot's arm). The specific variations will not be published in advance of the test.

Like Dimension 1, Dimension 2 is also a one-shot chance. For the simplest tests (i.e. corresponding to youngest infants), like grasping, Dim.1 is no different to Dim.2. For intermediate tasks, like pulling a string, Dim.2 has added variability, but no new unseen tasks. For the most advanced tasks, like "Displaying foresight by using an appropriate means in a problem situation" an array of tasks, unseen in advance of the test, can be presented, provided all are within the realm of a typical 22-month-old infant's capabilities.

5.4 Testing Dimension 3

Recall that Dimension 3 is the rapid learning ability of the robot/infant. This learning ability encompasses the ability to discover ways to overcome problems e.g. by variation or strategy change. Dimension 3 tasks are the same as Dimension 2, but we propose that the agent is allowed to practice and try repeatedly for six hours on four successive sessions. In between sessions there is no further programming of the robot allowed, but it may have maintenance, and may carry out some further processing itself.

Note that the ages presented in Sec. 5.1 need to be changed for Dimension 3. For example use of a string to retrieve moves from 12 months to 9 months (there is no large-group study result for this, it is based on a small sample). Unfortunately we have very few results from infants on how well they can learn over the timespan of one, two or three days. There is significant anecdotal evidence that they do have this capability [24, 25, 26]. In one exceptional study Chen and Siegler [7] did apply their "microgenetic" method to assess infants' (one group was 18-26 months) ability to learn over relatively short timescales. To highlight the fact that such studies are rare in infancy the authors stressed that in this study they were "applying to toddlers a type of process analysis that has proved fruitful in studies of older children."

5.5 Testing Dimension 4

Recall that Dimension 4 is the developing ability (slow learning) of the infant/robot. In contrast to previously described dimensions, the easiest way to test a robot on this dimension may require that the codes of the agent are made available and explained to the testers, and that it can be verified that the robot does not have a pre-scripted unfolding of more advanced abilities, which are themselves handcoded. Rather it is required that the robot possess some general development abilities, so that through interaction with the environment it extends its competences. To score 100% on 9mt-Dim.4-ScaleII the robot must pass the code inspection test, and additionally must be intrinsically motivated to play with objects in an environment, such that after several hundred hours (corresponding to

the months elapsed for infants) it can go on to pass Dimension 2 tests for 12 months, 18 months, etc. up to 22 months. A robot that does less than this will be scored in proportion to how far it goes, for example if it only achieves 100% on month 12 tasks but not beyond, then it scores $3/(22-9) = 23\%$, i.e. 9mt-Dim.4-ScaleII:23%.

5.6 Example of a Test Competition for 2014

In this section we suppose that a test is to be defined for 2014, in order to sketch out what would be plausible to test at the present time. A test in a particular year must focus on a level of ability close to what is achievable.

Scale I (Visual Pursuit and the Permanence of Objects) may not be interesting to test at the present time as we are not aware of any roboticists working on systems which can track the location of hidden objects undergoing displacements; also the lower ends of this scale may be too easy for robots because they can be programmed with basic knowledge of objects. Scale II (Means for Obtaining Desired Environmental Events) is plausible to test up to 20 months, although poor performance can be expected on the more advanced tasks. Scale IIIb (Gestural Imitation) is plausible to test at all ages and is likely to reveal that robots score highly [16]. Scale IV (Operational Causality) may be skipped due to a lack of effort on this in current developmental robotics. Scale V (Object Relations in Space) is interesting to test mainly because of the difficult demands that would be placed on vision and motor control components; "understanding space" in a basic functional way is relatively straightforward given that a robot can be programmed with the knowledge of three-dimensional Euclidean space. However it remains challenging to perform such tasks as visually recognising spatial relationships between objects from vision [29, 11], or for example understanding when one object may be suitable as a container for another, or constructing a tower of objects by placing in equilibrium, or appropriately spatially positioning a tool relative to an object. Scale VI (Schemes for Relating to Objects) is interesting to test at all ages; it is a suitable scale to test work on affordance learning [33] for example, and also reference to objects in shared interaction [13].

Tests for Dimensions 1 and 2 are relatively straightforward to formulate, and should be of interest to the community. Tests for Dimension 3 can be formulated mostly by relying on anecdotal evidence on infant abilities due to the lack of formal studies available. Dimension 3 tests are of great interest to the community due to their developmental flavour, and also because there is considerable research effort devoted to rapidly learning from demonstrations or from exploration (see for example the EU FP7 Xperience project[8] [1]). Dimension 4 tests are the most challenging for robotics, but there are sufficient examples existing to make their inclusion worthwhile already. For example affordance learning [33] can produce a qualitative change in the way that a robot would interact with an object, and so represents a step through ages in Scale VI. Also there are examples of development from simple sensorimotor coordinations through to more advanced object manipulations [18, 15].

[8] http://www.xperience.org/

6 Discussion: Is it a good test?

In this section we will evaluate our test with reference to Cohen's assessment of what makes a good challenge [9]. Firstly a test or challenge needs a clear easily understood long term goal, which this has: matching the competence of a 22 month old human on all six scales, in all dimensions. Secondly it needs an organisation to support it and revise rules to keep inline with the long-term goal; this it does not have yet[9].

With regard to scoring Cohen stresses the need for giving entrants feedback to help them to understand what worked and what did not, and why. This seems reasonably straightforward for most tests in Dimensions 1-3, but more tricky in Dimension 4; we are not clear on how best to tackle this at present; it must be borne in mind when a detailed scoring system is formulated for Dimension 4 tests.

Our proposal already fits well with Cohen's suggestion that the challenge offers a graduated series of challenges and allows incremental building on past efforts. However Cohen notes that it "follows from these principles that the challenge itself should be easily modified, by changing rules, initial conditions, requirements for success...". This point poses some difficulties for our challenge because creating varieties of tests which are doable by infants requires testing with infants, which requires that developmental psychologists are brought on board.

It is desirable that the cost of entry should be low, enabling students to participate. The easiest way to facilitate this would be through simulated challenges in addition to the main challenge (just like RoboCup soccer has a number of different leagues), for example using the iCub simulator.

"We should accept poor performance but insist on universal coverage." [9] Poor performance is accommodated by the facts that the abilities of younger infants can be targets, and also a percentage score is produced; infant level performance is not required. Universal coverage is partly addressed by the varieties of materials and situations which we have proposed in order to test generality and robustness of the robot capabilities. Related to this Cohen advocates "developmental" rather than "divide and conquer" approaches. There is however an element of divide and conquer in our proposal because of the branching into six different scales. This seems to be required because the community seems to want to split up in that way. Researchers interested in gesture, or shared reference for example, are not always interested in other aspects such as means-end behaviour; however a convergence does seem desirable in the long-term.

To sum up: We have proposed some tests for the developmental robotics community which would run in a similar way to the existing RoboCup challenges. We have preferred challenge competitions which change rather than fixed benchmarks, because fixed benchmarks may ultimately reward clever engineering which loses sight of the original goal of developmental robotics. This proposal is a first step which we hope might lead to a series of competitions providing a focus for the developmental robotics community, and a way to measure progress of alternative approaches on common

[9]The Autonomous Mental Development Technical Committee (AMDTC) of the Computational Intelligence Society of the IEEE would be an obvious parent organisation.

tasks.

Acknowledgements

Thanks to Norbert Krüger for comments on a draft.

References

[1] E. E. Aksoy, M. Tamosiunaite, R. Vuga, A. Ude, C. Geib, M. Steedman, and F. Wrgtter. Structural bootstrapping at the sensorimotor level for the fast acquisition of action knowledge for cognitive robots. In *IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, pages 0–0, 2013.

[2] M. Asada, K. Hosoda, Y. Kuniyoshi, H. Ishiguro, T. Inui, Y. Yoshikawa, M. Ogino, and C. Yoshida. Cognitive developmental robotics: A survey. *Autonomous Mental Development, IEEE Transactions on*, 1(1):12–34, 2009.

[3] T. G. R. Bower. *Development in Infancy*. San Francisco : W.H. Freeman, 1982.

[4] A. L. Brown. Domain-specific principles affect learning and transfer in children. *Cognitive Science*, 14(1):107–133, 1990.

[5] E. W. Bushnell and J. P. Boudreau. Motor development and the mind: The potential role of motor abilities as a determinant of aspects of perceptual development. *Child Development*, 64(4):1005–1021, 1993.

[6] A. Cangelosi, G. Metta, G. Sagerer, S. Nolfi, C. Nehaniv, K. Fischer, J. Tani, T. Belpaeme, G. Sandini, F. Nori, L. Fadiga, B. Wrede, K. Rohlfing, E. Tuci, K. Dautenhahn, J. Saunders, and A. Zeschel. Integration of action and language knowledge: A roadmap for developmental robotics. *Autonomous Mental Development, IEEE Transactions on*, 2(3):167–195, 2010.

[7] Z. Chen, R. S. Siegler, and M. W. Daehler. Across the great divide: Bridging the gap between understanding of toddlers' and older children's thinking. *Monographs of the Society for Research in Child Development*, 65(2):i–105, 2000.

[8] L. Claxton, R. Keen, and M. McCarty. Evidence of motor planning in infant reaching behavior. *Psychological Science*, 14(4):354–356, 2003.

[9] P. Cohen. If not Turings test, then what? *AI Magazine*, 26(4), 2006.

[10] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[11] S. Fichtl, J. Alexander, D. Kraft, J. A. Jorgensen, N. Krüger, and F. Guerin. Learning object relationships which determine the outcome of actions. *Paladyn. Journal of Behavioral Robotics. Special Issue on Advances in Developmental Robotics*, 3(4):188–199, 2012.

[12] M. S. Funk. Problem solving skills in young yellow-crowned parakeets (cyanoramphus auriceps). *Animal Cognition*, 5:167–176, 2002.

[13] S. Guadarrama, L. Riano, D. Golland, D. Göhring, Y. Jia, D. Klein, P. Abbeel, and T. Darrell. Grounding spatial relations for human-robot interaction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2013.

[14] F. Guerin, D. Kraft, and N. Krüger. A survey of the ontogeny of tool use: from sensorimotor experience to planning. *IEEE Transactions on Autonomous Mental Development*, 5(1):18–45, 2013.

[15] S. Hart and R. Grupen. Learning generalizable control programs. *IEEE Trans. Autonomous Mental Development*, 3(3):216–231, sept. 2011.

[16] I. Itauma, H. Kivrak, and H. Kose. Gesture imitation using machine learning techniques. In *Signal Processing and Communications Applications Conference (SIU), 2012 20th*, pages 1–4, 2012.

[17] M. Kido, H. Itoh, H. Fukumoto, H. Wakuya, and T. Furukawa. Developing a robot that performs tasks of developmental scales: On gaze control by eye-head coordination. In *SICE Annual Conference (SICE), 2011 Proceedings of*, pages 2488–2491, 2011.

[18] J. Law, P. Shaw, K. Earland, M. Sheldon, and M. H. Lee. A psychology based approach for longitudinal development in cognitive robotics. In *Submitted to Frontiers in Neurorobotics, under review*, 2014.

[19] M. Lungarella, G. Metta, R. Pfeifer, and G. Sandini. Developmental robotics: a survey. *Connection Sci.*, 15(4):151–190, 2003.

[20] M. E. McCarty, R. K. Clifton, D. H. Ashmead, P. Lee, and N. Goubet. How infants use vision for grasping objects. *Child Development*, 72(4):973–987, 2001.

[21] M. E. McCarty, R. K. Clifton, and R. R. Collard. Problem solving in infancy: The emergence of an action plan. *Developmental Psychology*, 35(4):1091–1101, 1999.

[22] J. O'Regan, L. Rat-Fischer, and J. Fagard. Mechanisms leading to tool use: A longitudinal study in human infants. In *Front. Comput. Neurosci. Conference Abstract: IEEE ICDL-EPIROB*, 2011.

[23] P.-Y. Oudeyer, F. Kaplan, and V. Hafner. Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation*, 11(6):265–286, 2007.

[24] J. Piaget. *The Origins of Intelligence in Children*. London: Routledge & Kegan Paul, 1936. (French version 1936, translation 1952).

[25] J. Piaget. *The Construction of Reality in the Child*. London: Routledge & Kegan Paul, 1937. (French version 1937, translation 1955).

[26] J. Piaget. *Play, Dreams and Imitation in Childhood*. London: Heinemann, 1945.

[27] C. Prince, N. Helder, and G. Hollich. Ongoing emergence: A core concept in epigenetic robotics. In L. Berthouze, F. Kaplan, H. Kozima, H. Yano, J. Konczak, G. Metta, J. Nadel, G. Sandini, G. Stojanov, and C. Balkenius, editors, *Proceedings of EpiRob'05 - International Conference on Epigenetic Robotics*, pages 63–70. Lund University Cognitive Studies, 2005.

[28] L. Rat-Fischer, J. O'Regan, and J. Fagard. The emergence of tool use during the second year of life. *Exp Child Psychol.*, 113(3):440–446, 2012.

[29] B. Rosman and S. Ramamoorthy. Learning spatial relationships between objects. *The International Journal of Robotics Research*, 30(11):1328–1342, 2011.

[30] J. Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *the International Conference on Simulation of Adaptive Behavior: From Animals to Animats*, pages 222–227, 1991.

[31] A. Streri and J. Féron. The development of haptic abilities in very young infants: From perception to cognition. *Infant Behavior and Development*, 28(3):290–304, Sep 2005.

[32] A. M. Turing. Computing machinery and intelligence. *Mind*, 59:433–460, 1950.

[33] E. Ugur, E. Oztop, and E. Sahin. Goal emulation and planning in perceptual space using learned affordances, 2011.

[34] I. C. Uzgiris and J. M. Hunt. *Assessment in infancy : ordinal scales of psychological development.* University of Illinois Press, 1975.