Long Term Requirements for Cognitive Robotics

Aaron Sloman and **Jeremy Wyatt** and **Nick Hawes,** *
School of Computer Science, University of Birmingham, UK
http://www.cs.bham.ac.uk/~axs {jlw} {nah}

Jackie Chappell
School of Biosciences, University of Birmingham, UK
http://www.biosciences.bham.ac.uk/staff/staff.htm?ID=90

Geert-Jan M. Kruijff
DFKI GmbH, Saarbrücken, Germany
http://www.dfki.de/~gj

Abstract

This paper discusses some of the long term objectives of cognitive robotics and some of the requirements for meeting those objectives that are still a very long way off. These include requirements for visual perception, for architectures, for kinds of learning, and for innate competences needed to drive learning and development in a variety of different environments. The work arises mainly out of research on requirements for forms of representation and architectures within the PlayMate scenario, which is a scenario concerned with a robot that perceives, interacts with and talks about 3-D objects on a tabletop, one of the scenarios in the EC-funded CoSy Robotics project.

Long term goals

Researchers working in cognitive robotics do not all have the same objectives. Many AI researchers, though not all, are interested in an ill-defined future goal which is roughly characterised as 'human-level' AI. This paper discusses requirements for achieving that goal. Likewise, McCarthy (1996) discussed requirements for achieving 'human-level' intelligence, pointing out that this involves what he called 'the common sense informatic situation', which he contrasted with 'the bounded informatic situation' that characterizes most AI work.

In the bounded informatic situation there is a well defined collection of information that determines a range of correct answers to many questions, such as 'How can you achieve X?', 'What can you do with X?' and many more, whereas in the common sense situation a creative thinker can typically go on and on producing new answers which are easily interpreted as correct answers to the question, even though they may be highly impractical or acting on them may have undesirable side effects – such as measuring the height of a building by jumping off it with a stop watch.

McCarthy lists several characteristics of the common sense informatic situation including:

> *"The known facts are incomplete, and there is no a priori limitation on what facts are relevant. It may not even be decided in advance what phenomena are to be taken into account. The consequences of actions cannot be fully determined. The common sense informatic situation necessitates the use of approximate concepts that cannot be fully defined and the use of approximate theories involving them. It also requires nonmonotonic reasoning in reaching conclusions."*

After further discussion of examples and strategies, he ends with a list of seven problems to be overcome, including giving machines common sense knowledge of the world, designing epistemologically adequate languages (languages rich enough to express what a person or robot can actually learn about the world), elaboration tolerance, nonmonotonic reasoning, contexts as objects, introspective capabilities, and the need for a uniform solution to the problems of formalising action. He later comments on these problems that 'lie between us and human level intelligence', as follows:

> *'Many will find dismayingly large the list of tasks that must be accomplished in order to reach human-level logical intelligence.'*

More problems than meet the eye

Our contention is that the list of problems to be solved is far larger than McCarthy suggests in this paper[1]. One reason why his list is far from comprehensive is that his problems, like many problems considered in AI, involve the ability to think, to answer questions, to give reasons, and to solve problems. All of these are primarily *intellectual* problems.

However, humans can do many things before they can talk, answer questions, and reason about the topics which McCarthy discusses and which many AI systems address. A typical child who has not yet learnt to talk is learning to perceive many things in the environment, to act on them in different ways, and to engage in both solitary and social nonverbal play, experimentation, and exploration, which seems

*This work is supported by the EU-funded CoSy Project, described at http://www.cognitivesystems.org

[1]He discusses additional points in a later unpublished paper (McCarthy 1999)

to be crucially involved in driving development. Many examples are given in Rochat's book (2001).

The same is true of many other animal species, though they do not learn to talk as humans do, and there is little evidence that they think as humans do. For example, none of them have learnt to build rockets. Yet, depending on the species, they can find mates, escape being eaten (sometimes), get food (even food that is attempting to escape), build nests, climb trees, pick berries, jump accurately, protect their young against predators, make a nest by arranging branches in a treetop, and much more. Many of the tasks, including hunting in the dark, and leaping through treetops, are beyond what humans can do. These activities do not require, and competence in their performance is not sufficient for, the development of an external human-like language used for inter-individual communication using a complex recursive syntax, compositional semantics and many varieties of speech acts.

Of course that does not imply that no language of any kind is involved: the animals are clearly acquiring, storing, manipulating, and using information about things in the environment, about processes that can occur and about causal relations between events and processes. This information must be encoded in some format that supports access, transformation, and use. We can call that format a language even though it is not used for external communication between individuals.

Moreover, the language may require many of the important characteristics of a human language apart from (a) the linear form of meaningful structures that is a precondition of acoustic transmission, (b) the communication of such structures from one individual to another, and (c) the use of symbols and syntactic forms to indicate the type of speech-act (Austin 1962) e.g. whether it is an order or a request. In particular, animals that learn about objects and processes of varying complexity, and perform purposive actions of varying complexity may use internal forms of representation that involve syntactic structure and compositional semantics. If that is true of non-linguistic animals, perhaps something like it is true of pre-linguistic humans, and will need to be true of pre-linguistic robots.

None of that implies that *only* a language of the general form based on a function-argument structure (referred to as 'Fregean' in (Sloman 1971)) is needed. Different forms suit different subtasks. In particular the requirements of 'online control' of actions, and some forms of prediction, reasoning and planning may favour 'analogical' forms of representation, including some with continuous variation, though often they will need to be used in conjunction with more symbolic forms.

What Mechanisms Are Required?

If there are such information-processing competences in other animals then they may have existed in pre-linguistic evolutionary precursors to humans. Of course for such an internal language to exist, and to be usable in the ways indicated, there must be mechanisms that manipulate the information structures, and they must interact with other mechanisms involved in perception, action and learning.

Our hypothesis is that the ability to see, manipulate, think about, design, and build 3-D structures is far more complex in its computational requirements than has previously been recognised. It may, for example, require a collection of abilities including various kinds of representations of processes involving (a) multiple rigid and flexible objects and parts of objects with changing relationships, (b) irregular shapes with complex surfaces not all of which are visible at the same time (c) different kinds of materials of which objects can be composed with different resulting causal properties, that need to be understood in order to make use of those objects (d) and multiple simultaneous causal interactions.

At present no AI systems come close to having this collection of capabilities. There appears to be something deep, evolutionarily old, and largely unexplored which is shared with a subset of other animals and which lies behind many characteristically human capabilities, including the ability to use language, to reason hypothetically, to think about what others are doing, to design new complex physical structures, to think about complex action sequences and compare alternatives, before executing them.

This old collection of competences includes being able to perceive, understand (at least partially), and in some cases act on 3-D structures and processes of varying kinds and complexity. For example, we can see processes involving flexible objects, or multiple moving objects such as a group of dogs playing in a field, a herd of deer frightened by a predator, fingers on a hand peeling a banana. We do not yet understand what explains such abilities, especially the ability to manipulate complex and flexible objects, but if it is partly a result of evolution then (a) mechanisms that are involved in such human competences may be shared with other animals, (b) mechanisms that originally evolved for a sighted species may still be available for humans born blind, and (c) the internal forms of representation and mechanisms for using them used by non-linguistic species, and prelinguistic children may be prerequisites for the development of language as we know it. At the very least, if those mechanisms provide a basis for perceiving, thinking, and acting on the world then they make it possible for animals to have something to communicate about. Moreover, some of the considerations above suggest that the common view that symbolic intelligence depends on the development of human language, e.g. (Steels 1996), has things the wrong way round.[2]

Deeply Embodied Agents

This may seem to be nothing more than a reiteration of familiar slogans of the last few decades (perhaps most famously (Brooks 1991)) which have been used as a basis for attacking symbolic AI and focusing on neural nets, dynamical systems and physically implemented robots. These include such slogans as that human-like intelligence needs to be embodied (Anderson 2003), semantic competence needs to be 'grounded' in sensory information, and that subsymbolic mechanisms are the basis of intelligence.

[2] As argued on more general grounds in (Sloman 1979).

However, what we are proposing is importantly different from work inspired by such slogans, and also requires advances in vision and the representation of 3-D structures and processes that are beyond the current state of the art and are not usually discussed by proponents of embodiment or opponents of symbolic AI.

In particular much recent work on embodied systems[3] assumes that all the information processing that goes on is based on learnt or innate knowledge about sensorimotor relationships, and that insofar as there is anything that goes beyond the mappings between particular modes of sensory input patterns and particular patterns of motor output signals it is merely multi-modal, using correlations between different sensor or motor modalities. On that view a system learns and uses correlations between the patterns of processing that occur in different sensory subsystems and can therefore usefully combine modalities, for instance using visual information to disambiguate tactile information or vice versa.

Such theories are also often combined with claims about how the physical properties of sensor and motor devices play a crucial role in reducing the computational load, e.g. properties such as mass, momentum, and elasticity, of moving physical components (as in the old example of a 'compliant wrist' (Cutkosky, Jourdain, & Wright 1984)).

Parts of the physical environment often do some of the work, for instance allowing changes in the environment during performance of a task to determine what should be done next so that it is not necessary to have internal mechanisms recording how far the work has proceeded. There is now much research based on these ideas sometimes described as research on systems that don't use representations[4].

We can use the label 'deeply embodied' to describe species whose information about the environment is implemented only in sensorimotor competences closely coupled with aspects of the physical environment. By 'sensorimotor' we refer to information about the contents of patterns of signals from sensors and to motors within the organism. The patterns may be more or less abstract, but are only patterns found in those signals, including statistical associations, referred to as 'sensorimotor contingencies'.[5]

Informationally Disembodied Agents

This contrasts with information about the content of the environment (e.g. about 3-D surfaces in different orientations, with varying curvature), which is capable of being sensed or acted on in many ways, but can be represented 'amodally', i.e. independently of how it is sensed or acted on. An organism or a machine that can acquire, manipulate and use such a-modal information could be described as partly 'informationally disembodied'.

We do not dispute that there are many species that are 'deeply embodied' in the sense we have defined. As far as we know that is true of most invertebrates, and possibly all fish. It may also be true of many mammal and bird species. However, at present the lack of evidence of more abstract processing capabilities, using amodal 'objective' forms of representation may simply be a consequence of researchers failing to look for such evidence. For example, the portia spider may be an example. Portia sometimes pounces on its prey from above. In order to get to a suitable position, it spends some time scanning the environment, and in some cases selects a round-about route which takes it through locations that it could not see during the scanning process. Moreover some of the time while getting to a suitable new position, its prey is out of site. All that suggests, but does not prove, that portia builds and uses some sort of representation of a 3-D environment containing things that are not always sensed by it. More research is needed to find out exactly what is going on. Perhaps portia is partly informationally disembodied.

Humans certainly are (though perhaps not in early infancy), and perhaps some other animals have also evolved information processing mechanisms that are largely disconnected from sensor and motor signals and that this is what has given them greater power and flexibility, which is most extreme in the case of humans. In other words, whereas intelligence was almost entirely embodied during most of evolution it has recently, in a small number of species come to be importantly disembodied. But for this there would be no mathematicians, philosophers, programmers, or historians, and perhaps not even nest-building birds or hunting mammals. (Note: the contrast between sensorimotor based and objective representations is not the same as a contrast between subsymbolic and symbolic representations.)

Species differences

It is informative to reflect on some differences between biological species. Some of them are remarkably competent soon after birth. Examples are chicks that break open their shells and almost immediately go pecking for food; and deer that within hours, or even minutes, after birth can walk to the nipple to suck, and even run with the herd to escape predators. We can describe these as having mostly genetically preconfigured competences. There are other species that seem to be born or hatched almost helpless yet by the time they are adults develop a rich and flexible collection of skills which seem to be far more cognitively demanding than those achieved by the precocial species. In some cases they seem to be much better able to adapt to changes in the environment, including learning to solve problems that none of their ancestors could have encountered, e.g. in human domestic contexts.

The latter are often described as altricial species. However, as suggested in (Sloman & Chappell 2005) it is better to make the contrast we are focusing on in terms of varieties of *competences* rather than varieties of *species*. Even the altricial species typically have some competences that are

[3] See the collection of papers on situatedness and embodiment(Ziemke 2002)

[4] Though that is merely a symptom of excessively narrow definition of 'representation', since anything that acquires, manipulates and uses information is using something that encodes or represents whatever is in that information.

[5] This label is ambiguous, and is sometimes taken to include what we would call *objective* 'condition/consequence contingencies' involving things in the environment rather than signals in the animal or robot, as pointed out in
http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0603

genetically preconfigured (like sucking in humans), and others that may be described as 'meta-configured' meaning that there is an innate ability to acquire certain classes of skills through interaction with the environment, in such a way that the precise competences acquired will depend on the environment and how the individual interacts with it. Thus members of a species may have an indefinitely large variety of meta-configured competences that they have the *potential* to develop though each individual will develop only a small subset of them. (Compare Waddington's epigenetic landscape.) In particular, we conjecture that

- there is a kind of intelligence that pre-linguistic humans share with many other species, though not with precocial species, which involves developing a rich variety of competences through interaction with a complex and varied 3-D environment
- this intelligence, at least in young humans, and to varying degrees in several other species, involves
  - perceiving complex 3-D structures and processes in which 3-D structures and relationships change in an enormous variety of ways,
  - understanding how those processes depend on different aspects of a complex situation, including:
    - the materials of which things are made (kinds of stuff),
    - their shapes and sizes,
    - the relationships between their parts, including relationships between parts of surfaces of different objects, e.g. fingers and a cup,
    - how objects can be manipulated and the consequences of possible actions
  - learning about and using information about causal relationships between events and processes, including constraints on what can happen, where two kinds of causation are involved (Sloman 2005):
    - Humean causation: correlation based, concerned with conditional probabilities (as in Bayesian nets)
    - Kantian causation: structure based and deterministic – concerned with relationships between spatial structures for instance.
  - planning actions in advance and predicting their consequences,
  - wondering what will happen if something is done
  - understanding commonalities between actions done in different ways, and by different individuals
- The development of competence based on this kind of intelligence uses innate 'meta-level' capabilities of various kinds including:
  - The ability to perceive complex structures and processes composed of previously perceived smaller parts, and to select descriptions of some of them to be stored for reuse (so that recursion allows increasingly complex structures to be represented)
  - The ability to abstract away from sensory and motor details so that the stored specifications have some generality and objectivity, e.g. representing a block being placed on another block in terms of changing relationships between surfaces, edges and corners of the blocks rather than changing patterns of sensory input and motor signals
  - The ability to use these stored representations to formulate new goals, new questions and new hypotheses that can drive processes of exploration and testing. For instance, any representation of a physical state of affairs can become a specification for the goal of making the state of affairs exist. Any such representation can generate a variety of questions to be investigated including the question whether what is represented exists, and more complex questions regarding what would make an incomplete representation represent something true.[6]
  - To notice unexpected side-effects of such processes and use them as the basis for various kinds of learning (including modifications or extensions of the ontology used)

Viewer Independent Affordances

The abilities listed above subsume much of what is meant by saying that animals perceive positive and negative affordances in the environment, i.e. possibilities for action and constraints on action. We need to generalise that idea to include an example of informational disembodiment: perception of 'vicarious affordances', i.e. affordances for other individuals (e.g. the young of the species whose affordances may need to be enhanced, or prey whose affordances must be obstructed).

In children these abilities are manifested in playing with many different kinds of manufactured toys as well as natural substances such as water, sand, mud, parts of their own bodies, hair or fur on animals, and many kinds of food – solid, liquid, more or less viscous, more or less sticky, more or less tough requiring cutting, tearing or chewing.

Although each episode of interaction with the environment involves sensory and motor arrays of signals, humans are able, in many cases, to form modality-neutral representations of what is going on in a subset of the environment, and to learn generalisations involving those abstracted subsets. Moreover, they can also re-assemble different kinds of information in new combinations in order to make sense of novel percepts using old knowledge and also in order to plan new actions or design new physical structures using old components in novel configurations (e.g. you can imagine the consequences of building a wall out of packs of butter).

This phenomenon can be summarised by saying that through creative, controlled exploration and play, children discover 'orthogonal competences', that are recombinable in new ways.

Some other animals also seem able to do this, though to a lesser degree. A spectacular example was Betty the New Caledonian crow making hooks out of wire in the Behavioural Ecology Research Group, University of Oxford (Weir, Chappell, & Kacelnik 2002).[7] Another example was the observation of tool use in lemurs reported in (Santos, Mahajan, & Barnes 2005), who state

> *'Our overall pattern of results, which suggests that lemurs solve the cane-pulling task like other tool-using primates, poses a puzzle for the view that differences in*

[6] As discussed in
http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0507

[7] As demonstrated in videos available at this site:
http://users.ox.ac.uk/~kgroup/tools/tools_main.html

primates' natural tool use reflect differences at the level of conceptual ability. Our results, instead, provide support to an alternative account – namely, that many primates share an ability to reason about the functional properties of different objects, irrespective of whether they exhibit tool use in their natural behavioral repertoire.'

A 'Disembodied' Grasping Concept

Many animals grasp things with their mouths or beaks. Since eyes are fairly rigidly related to mouths the process of grasping is closely correlated with specific optical flow patterns. When some species evolved paws or hands that could be moved independently of the mouth in order to manipulate objects, the variety of visual patterns corresponding to functionally the same sort of action, such as we might call pushing, prodding, pulling, or grasping, exploded. Thus there was a distinct advantage to be gained by developing a form of representation of the process that merely involved changing 3-D relationships between 3-D surfaces of objects (e.g. two movable body parts such as palm and thumb, upper and lower jaw, left and right hand, and the thing grasped between the two body parts), independently of how they were sensed, or how the motion was caused.

Finding a modality independent, objective representation of the process would allow generalisations learnt in one context e.g. about the effect of twisting or pulling on something that is being grasped, to be transferred to many other contexts. In particular, it could be transferred to contexts where possible actions of another agent are being considered: the perception of what we called 'vicarious affordances' above.

This ability to abstract from the sensorimotor specifics to the representation of objective relationships and processes in the environment may have required major changes in brain mechanisms so as to allow a representation of a relatively unchanging environment to be maintained persisting across a host of different sensory and motor signal patterns. Thereafter many actions could be represented in terms of the 'objective' processes occurring in that environment and many generalisations about such processes could be far more economically expressed than the corresponding collections of sensory motor associations.

Whether this required the innate mechanisms driving exploration and learning to have a built in pre-disposition to use representations of 3-D structures and motions, or whether the need for that emerged through some very general learning capability (e.g. an information-compression capability) remains an open question. However it is unlikely that all those millions of years of evolution in a 3-D visually perceived spatial environment (with moving, rigid and non-rigid objects) did not deeply influence the processes of learning even in the visually impaired. For instance there may be innate explicit or implicit information about the environment having a 3-D structure in which changes occur.[8] Further discussion of these issues can be found at www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0601

[8]Whether animals that use other sensory modalities, e.g. bats, share the same amodal representations is an open question.

Vision problems

One aspect of the kinds of competence and learning of the competences that we are discussing is the role of vision. Vision research over the last few decades has been very disappointing as regards the requirements discussed here. In the 1978 paper of Barrow and Tenenbaum and other work at that time, some important ideas about the perception of spatial structure and motion were beginning to be explored, that were later abandoned mostly in favour of work on recognition, tracking, and localisation, with little or no work on understanding of 3-D spatial structure. There has been a great deal of progress on specialised image processing tasks, driven by collections of benchmarks which have little or nothing to do with the ability to act in the environment, for instance benchmarks concerned with recognition of faces, or types of animals or vehicles, but without perceiving any spatial structure.

In cases where there is acquisition of information about surface structure, e.g. using laser range finders, there is little or no *understanding* of surface structure. Rather the information is stored in a form that can be demonstrated by projection of images from new viewpoints, not by its use in *action* (e.g. through affordances). In contrast consider what a human can see regarding possible actions in this image (which has low resolution, poor lighting and much noise).

Figure 1: Most people clearly see a rich collection of affordances, despite the noise and low resolution of the image.

Most people looking at that picture can select different parts of the objects that could be grasped using finger and thumb or a tool bringing two surfaces together. Moreover, they can see roughly the orientation required for the grasping surfaces for different grasping locations. For instance the orientation required varies smoothly all around the rim of the cup, along the edge of the saucer, on the handle, etc. If the cup is grasped without the grasping surfaces first being aligned with the relevant portions of the surface of the cup the result will be enforced rotation of the cup if the fingers are firmly compressed.

Those are among the affordances for action visible in the shape of the surface of the cup. There are many more affordances concerned with lifting the cup pouring something into or out of it, sipping from it, throwing it, etc.

However, as far as we can tell after extensive enquiry there are no AI vision systems that can perceive surface structure in such a way as to produce an understanding of the impli-

cations for actions.

Moreover, since the variety of types of surfaces and 3-D orientations of graspable, touchable, pushable, pullable surface fragments is astronomical any attempt to learn about such affordances by storing sensorimotor correlations will founder on a combinatorial explosion.

So, for real progress in this field to occur, substantially new kinds of AI vision systems will be needed, involving novel forms of representation for information about surface structure, properties of materials and affordances.

Perceiving processes

The situation is even more complex than has been so far described. All spatial objects in addition to having spatial relations to other objects also have many parts that have spatial relations to each other and to other objects. For example a cube has faces, edges and vertices that are related to one another and if there is another cube in the vicinity the faces, edges and vertices of the two cubes will all be related. Moreover, not only is this a geometrical fact about the cubes it is also the case that such relations can be perceived, even if not all of them will be perceived in any given situation.

We can express this by saying that perception can involve multi-strand relationships. Moreover, when processes occur, many such relationships can change in parallel. Thus perception of changing scenes can involve representation of several concurrent processes.

It is even more complex than that, since there may be different levels of process, at different levels of abstraction that are seen simultaneously. (Grush (2004) comes close to saying this.) For instance if one cube resting on another is slightly unstable and is being pushed into a more stable position then at one level of abstraction there is merely a relative movement in a certain direction, and at another level of abstraction a decrease in instability, and if this is part of some larger configuration other changes such as production of symmetry in the large configuration. Some of the changes will be metrical (changing distances, orientations, shapes) others topological (changing between touching and being separated, between being inside and being outside, etc.). Moreover, at the same time as these relatively abstract changes are being perceived there may be very detailed perception of changing metrical relations between surfaces that are moving close to one another but need to be accurately aligned for the task to succeed.

It seems to follow from all of this that a human-like robot will need a visual system that is capable of simultaneously representing or simulating multiple concurrent processes of different sorts and different kinds of abstraction in partial registration with retinal images, or to be more precise with optic arrays, since retinal images keep changing with saccades.[9]

Understanding causation

As indicated briefly in the list of requirements, above, humans, though probably not new-born infants, can not only perceive processes of the sorts described above in which many things change concurrently – they can also visualise or imagine them when they are not happening. This is an important part of the ability to make plans or to solve problems involving spatial configurations (as discussed in (Sloman 1971)).[10] An example would be visualising the consequences of attempting to lift the cup in Figure 1 by holding a pencil horizontally under the handle and lifting it, or pushing a pencil horizontally through the hole in the handle and lifting it. Visualising the consequences involves not only the ability to transform a representation of the 3-D scene but also the ability to apply constraints corresponding to rigidity of the material, features of the shape, the influence of gravity, the weight of the material, the behaviour of liquids (if there is some liquid in the bottom of the cup) and possibly inertia, depending on how fast the pencil is lifted.

Insofar as all this includes the ability to propagate constraints in changing representations and to understand why certain changes necessarily produce others (given the constraints) it provides the basis for a kind of causal understanding that is structure-based and deterministic (Kantian), rather than purely correlational (Humean/Bayesian) type of causation.

This also seems to be the basis for much human mathematical competence especially in learning about geometry and topology. The full implications of this will need to be developed on another occasion, though some preliminary discussion can be found in (Sloman 2005). There is much to be done that will go a long way beyond what current AI systems do, as regards using vision, and the ability to visualise for a wide variety of tasks.

Methodology for long range research

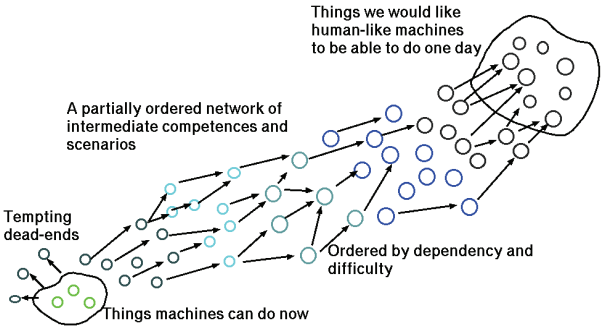

Figure 2. A roadmap/graph of scenarios, from(Sloman 2006)

The discussion above, though still requiring a great deal of work, is an example of a kind of methodology which we commend to the research community. This involves setting up a network containing a large number of robot scenarios of varying degrees of difficulty and depth, starting from detailed descriptions of machines doing things that AI systems are nowhere near doing at present, and working back through scenarios that are less and less complex and demanding.

[9] Vision as multiple concurrent simulations is discussed in: http://www.cs.bham.ac.uk/research/projects/cosy/papers/#pr0505

[10] A more detailed specification of 'fully-deliberative' competence is under development at http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0604

This can produce a partially ordered collection of requirements for future systems, ordered by dependency relations and difficulty, where early items are not very far from what we can do now and more remote items correspond to goals that can only be achieved in the distant future. In order to ensure some generality the long range scenarios should include not just one type of performance but a considerable variety, so that the robot's capabilities can be tested in many contexts.

It is important that the scenarios are not chosen by any one research group, but involve contributions from a wide research community proposing tasks that everyone can agree are hard, everyone can agree would be worth getting robots to do, whether for scientific or engineering reasons, and for which everyone can agree human capabilities (or in some cases other animals) provide proofs of possibility. People should be able to agree on the relevance of such long term goals even if they disagree as to which mechanisms, formalisms, architectures, research methodologies are appropriate for achieving the goals, and even if they disagree on the relative importance of different goals. Since the specification of requirements in terms of scenarios is independent of the explanatory mechanism that may enable the requirements to be met, people do not need to agree on mechanisms in order to develop test scenarios jointly.

Our cup and saucer configuration might be in various domestic scenarios with different features, including stacking the relics of a meal into a dishwasher, feeding a baby some milk from a cup, pouring tea, or painting the cup.

When such scenarios, far beyond our current implementation capabilities, have been described, it should be possible to work backward through a branching set of less demanding and more manageable scenarios, which everyone could agree would be stepping stones to the more difficult ones, until we reach scenarios that are within or only a little beyond current capabilities.

This 'backward chaining' research methodology contrasts with the more common 'forward chaining' where people ask what improvements can be added to their current techniques and systems. Improvements are often narrowly focused on performance on fixed benchmark tests rather than extending the variety of things our systems can do.

The problem is that there is no guarantee that the forward chaining improvements will take us nearer to achieving our long term scientific goals, even if they usefully help to solve immediate engineering problems, such as recognising faulty parts on a production line.[11]

Scenario Development Tools are Needed

Unfortunately experience shows that most people find it very difficult to develop the distant scenarios in sufficient detail for them to provide a basis for the backward chaining. So we have recently developed a methodology for generating scenarios by using a 3-D grid of requirements. One dimension of the grid is concerned with types of entities (concrete, abstract, inert, mobile, totally dumb, intelligent, etc.) and another dimension with things that can be done to entities, e.g. perceiving, physically acting on them, referring to them, thinking about them, etc. Those two dimensions produce a grid of types of competence/entity pairs (some of which are empty). The third dimension is complexity, difficulty, and remoteness.[12]

This grid can help us focus on problems involving particular types of competence applied to particular types of entity. Then scenarios of varying complexity can be devised by combining different subsets of the grid. This approach can be used to create long term roadmaps defining scientific challenges that everyone will agree are hard. Progress can then be identified in relation to which portions of the graph of scenarios have been achieved. New classes of benchmarks requiring integration of different combinations of competences can be defined.

The rectangular grid is an oversimplification: some boxes need complex subdivisions, and other boxes will be empty.

It is hoped that joint work on developing a large collection of scenarios that determine long term goals and intermediate milestones can be shared between research communities that normally do not talk to one another because they assume that different mechanisms should be used.

Conclusion

There are many open questions about the kind of long range research described here including questions about the forms of representation needed, the particular sorts of innate mechanisms required to drive the process of exploration and discovery, the kind of architecture in which all these competences can be combined (including many mechanisms that have not been discussed here), and whether we are right in claiming that even congenitally blind or limbless humans need brain mechanisms that evolved in species concerned with perception and manipulation of complex 3-D objects.

A particularly interesting question relates to whether the representational mechanisms that we have postulated (albeit still with great vagueness) in pre-linguistic children and non-linguistic animals anticipated requirements for the linguistic forms of representations that develop later in children and evolved later in our history. The conjecture suggested by this work is that the processes that include development of orthogonal recombinable competences required the evolution of an *internal* language used within an animal that had many of the features (e.g. syntax and compositional semantics) that later developed to support linguistic communication between individuals using external languages.

Meta-semantic competence

Other questions that we have not discussed here include questions about the requirements for an animal or robot not only to be able to represent physical and geometrical structures and processes in the environment but also to represent some of the entities in the environment as themselves having

[11] An early presentation of this methodology, with some other examples, arising out of a DARPA cognitive systems consultation is here http://www.cs.bham.ac.uk/research/cogaff/gc/targets.html

[12] The grid will be presented during the members' poster session at AAAI'06

representational capabilities, such as are required for having goals, beliefs, percepts, plans, intentions and mental states such as anger, fear, liking, enjoyment, etc. A special case of this is adopting what Dennett calls 'the intentional stance' (Dennett 1987). It is a special case because it involves attributing rationality to other individuals. However we can, and do, think of people, other animals, and robots as perceiving, learning, deciding, noticing, being mistaken, etc. without assuming that they are rational.

A pre-requisite for attributing information processing capabilities to something is meta-semantic competence, i.e. being able to use a form of representation that allows reference to other things that use representations, or at least process information. Such meta-semantic competence is not a trivial extension of ordinary semantic competence, because it involves coping with referentially opaque contexts (e.g. where the assumption that X=Y does not justify the substitution of 'Y' for 'X', as Frege noted in connection with 'the evening star' and 'the morning star', in contexts like 'Fred believes the evening star is Mars') (Frege 1960).

It is possible that before meta-semantic competence developed, organisms first evolved architectural features and forms of representation that enabled them to monitor and represent some of their own mental states, which could be useful for many purposes including detecting and remedying flaws in thinking, learning or problem-solving processes. Identifying erroneous beliefs or percepts required the invention of meta-semantic competence.

This contrasts with the view that language and social interaction preceded evolution of self monitoring capabilities.

If we ever build systems combining all of the capabilities discussed we shall be able to formulate new precise empirical questions about evolution and about how individual development occurs in children, in addition to being able to make congenial domestic robots to help us in old age.

Acknowledgements

This work was done with the help of colleagues working on the PlayMate scenario in the CoSy robot project funded by the EU Cognitive Systems initiative: see http://www.cs.bham.ac.uk/research/projects/cosy/PlayMate-start.html

Some of the ideas here are related to work by Minsky on his forthcoming book *The Emotion Machine*. We are grateful for comments from the Workshop reviewer.

References

Anderson, M. L. 2003. Embodied cognition: A field guide. *Artificial Intelligence* 149(1):91–130.

Austin, J. 1962. *How To Do Things With Words*. Oxford: Clarendon Press.

Barrow, H., and Tenenbaum, J. 1978. Recovering intrinsic scene characteristics from images. In Hanson, A., and Riseman, E., eds., *Computer Vision Systems*. New York: Academic Press.

Brooks, R. A. 1991. Intelligence without representation. *Artificial Intelligence* 47:139–159.

Cutkosky, M. R.; Jourdain, J. M.; and Wright, P. K. 1984. Testing and control of a compliant wrist. Technical Report CMU-RI-TR-84-04, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA.

Dennett, D. 1987. *The Intentional Stance*. Cambridge, MA: MIT Press.

Frege, G. 1960. On sense and reference. In Geach, P., and Black, M., eds., *Translations from the Philosophical Writings*. Oxford: Blackwell.

Grush, R. 2004. The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences* 27:377–442.

McCarthy, J. 1996. From here to human-level AI. Modified version of invited talk at KR'96, online at http://www-formal.stanford.edu/jmc/human.html.

McCarthy, J. 1999. The Well Designed Child. (Available at www-formal.stanford.edu/jmc).

Rochat, P. 2001. *The Infant's World*. Cambridge, MA: Harvard University Press.

Santos, L. R.; Mahajan, N.; and Barnes, J. L. 2005. How Prosimian Primates Represent Tools: Experiments With Two Lemur Species (Eulemur fulvus and Lemur catta). *Journal of Comparative Psychology* 119:394–403. 4.

Sloman, A., and Chappell, J. 2005. The Altricial-Precocial Spectrum for Robots. In *Proceedings IJCAI'05*, 1187–1192. Available at http://www.cs.bham.ac.uk/research/cogaff/05.html#200502.

Sloman, A. 1971. Interactions between philosophy and AI: The role of intuition and non-logical reasoning in intelligence. In *Proc 2nd IJCAI*. Reprinted in *Artificial Intelligence*, vol 2, 3-4, pp 209-225, 1971, and in J.M. Nicholas, ed. *Images, Perception, and Knowledge*. Dordrecht-Holland: Reidel. 1977.

Sloman, A. 1979. The primacy of non-communicative language. In MacCafferty, M., and Gray, K., eds., *The analysis of Meaning: Informatics 5 Proceedings ASLIB/BCS Conference, Oxford, March 1979*, 1–15. London: Aslib. Online at http://www.cs.bham.ac.uk/research/cogaff/.

Sloman, A. 2005. Two views of child as scientist: Humean and Kantian. Technical Report COSY-PR-0506:, School of Computer Science, University of Birmingham, UK.

Sloman, A. 2006. Introduction to Symposium GC5: Architecture of Brain and Mind – Integrating high level cognitive processes with brain mechanisms and functions in a working robot. Technical Report COSY-TR-0602:, School of Computer Science, University of Birmingham, UK.

Steels, L. 1996. The origins of intelligence.

Weir, A. A. S.; Chappell, J.; and Kacelnik, A. 2002. Shaping of hooks in New Caledonian crows. *Science* 297(9 August 2002):981.

Ziemke, T. 2002. Situated and Embodied Cognition. *Cognitive Systems Research* 3(3). (Editor's introduction to special issue.).