THE NINTH WHITE HOUSE PAPERS
Graduate Research in the Cognitive and Computing Sciences at Sussex

Editors:
Jason Noble & Sara R. Parsowith

CSRP 440

December 1996

ISSN 1350-3162

UNIVERSITY OF
SUSSEX
AT BRIGHTON

Cognitive Science
Research Papers

THE NINTH WHITE HOUSE PAPERS

Graduate Research in the Cognitive and Computing Sciences at Sussex

CSRP 440

Editors:
Jason Noble & Sara R. Parsowith

December 1996

Contents

Dedication

The editors would like to dedicate the Ninth White House Papers to Jo Brook for her many contributions to COGS over the years—particularly LaTeX above and beyond the call of duty. Jo has recently left Sussex to work in Edinburgh and we wish her all the best.

Preface

Each year since 1988, COGS graduate students have been meeting at Sussex University's conference centre, the White House, located at the Isle of Thorns, near Haywards Heath. Students are given the opportunity to give presentations on their work, exchange ideas, and most importantly, socialize. Out of this annual event arises a collection of short papers that have come to be known as the White House Papers. This is the ninth edition.

This summer, all postgraduate students at COGS were invited to submit papers of around 2000 words for inclusion in the Ninth White House Papers. The resulting collection reflects work in diverse areas of research, such as artificial life, cognitive psychology, compiler design, computer supported cooperative work, computer vision, developmental psychology, evolutionary computation, health psychology, linguistics, medical education, philosophy of science, and software design.

This year's workshop was organised by Sara Parsowith who wishes to thank Stephen Eglen and Jo Brook who are both veteran organizers. Their guidance helped ensure that the workshop was a success. The editors would like to thank all the DPhil students who contributed for making the workshop both intellectually stimulating and loads of fun. Thanks also to Theo Arvanitis for being our guest speaker this year. We are grateful to Professor Matthew Hennessy and the COGS Graduate Research Centre for funding the workshop.

Jason Noble
Sara Parsowith
December 1996

In: Noble, J. & Parsowith, S. R. (Eds.). (1996). The Ninth White House Papers: Graduate Research in the Cognitive and Computing Sciences at Sussex. Cognitive Science Research Paper 440, School of Cognitive and Computing Sciences, University of Sussex.

Multiple Intelligences, Instructional Design and Medical Education

Zahra Al-Rawahi
zahraar@cogs.susx.ac.uk

School of Cognitive & Computing Sciences
University of Sussex
Brighton
BN1 9QH

Abstract Gardner claims in his multiple intelligences (MI) theory that every cultural role requires a combination of different abilities regardless of sophistication. He considers a person's mind as a collection of more than one intelligence or ability. How can the combination of these intelligences meet the cognitive needs of different domains and individuals? To answer, it is important to discuss how these different intelligences interact and combine with each other. The aim of this paper is first to discuss the different combinations of intelligences within instructional design for medical education and second to discuss how the theory of multiple intelligences encourages multiple representations of knowledge.

The following issues will be addressed: How MI allows instructional designers to choose appropriate technology to convey their knowledge according to the educational needs of their domain. How MI enables learners to choose the appropriate technology for themselves to process the information according to their own intelligences or learning styles. Furthermore, how MI gives the learners the flexibility to drive their learning. In brief, I shall review how MI theory acts as an essential educational tool for the instructional designer. My research aims to apply Gardner's theory of MI and the cognitive apprenticeship approach to the design of a multimedia Intelligent Tutoring System for teaching medical students.

1 Introduction

In this paper I shall discuss how different intelligences combine and interact with each other. I shall concentrate on the combination of different intelligences within instructional design for medical education. In the second part of the paper I shall show how the MI theory is important for multimedia instructional designers and how it helps them to convey their message in an effective way.

The theory of MI proposed by Gardner in 1983 claims that human beings possess at least seven types of mental functioning or intelligence. He categorised the following intelligences: linguistic, logical-mathematical, spatial, musical, bodily-kinaesthetic, interpersonal and intrapersonal. These work together depending on the problem to be solved, within the limitations imposed by the individual environment. Some activities require higher progress of a single intelligence, while others require development in more than one practice (Gardner, 1993a). According to Gardner "Nearly all cultural roles exploit more than one intelligence; at the same time, no performance can come about simply through the exercise of a single intelligence." (Gardner, 1993b).

2 The combination of different intelligences in medical education

The combination of intelligences is important for any task to be done whatever its sophistication. For a person to perform any task he or she needs different combinations of these intelligences. Mastering these influences the future career of learners.

> It is of the utmost importance that we recognize and nurture all of the varied human intelligences, and all of the combinations of the varied human intelligences. We are all so different largely because we all have different combinations of intelligences. If we recognize this, I think we will have at least a better chance of dealing appropriately with the many problems that we face in the world. (Gardner, 1987)

The next section addresses the combination of intelligences in medicine, and the technology designed to support them. A large number of combinations of intelligences occur during the development of medical skills. Therefore, the following section will study some of these combinations only.

2.1 Linguistic and interpersonal intelligences

Linguistic intelligence refers to an individual's capacity to use either written or spoken language effectively as a vehicle of expression and communication. Interpersonal intelligence refers to the capacity to communicate appropriately and effectively, and to respond to other people and understand their feelings.

Technology has advanced quite effectively in this field: for instance, in electronic mail and groupware. Groupware encourages collaborators at different locations and time zones to communicate and discuss issues through language. There is increasing research on computer supported collaborative writing and on how students can write high quality documents collaboratively.

In medical education the combination of these intelligences is very important. Medical students need to communicate as well as to respond and to understand the feelings of their patients. They also need to communicate effectively and work co-operatively with their colleagues and patients. Interpersonal intelligences can also be combined with musical and visual intelligences: for example, recent developments include communication via audio and video on the computer screen.

2.2 Musical and logical intelligences

Musical intelligence refers to the ability to use and understand music and rhythm with precision. It may be exercised by listening to a variety of sounds and by engaging in rhythmic activities or by composing and conducting music. Bodily intelligence refers to the ability to use body movements in a precise and skilful manner.

These two processes are combined when training medical students in diagnosis by using the audio auscultation technique. Auscultation is the act of listening to internal organs for diagnostic purposes. It is a valuable diagnostic tool that physicians use regularly in their practice, for example by placing a stethoscope on a chest to examine a heart disorder. Heart disorders can cause audible changes in the heart beat. Hence, the physician needs to have very good musical intelligence, to differentiate between different heart sounds and to detect an abnormal heart beat.

Cardiac Auscultation Diagnosis Instruction (CADI) is an Intelligent Tutoring System (ITS) which is under development (Fenstermacher, 1995). This system will be designed to create an environment where students can practice diagnosis using the audio auscultation technique and get feedback on their progress from an expert. The program will give students a brief introduction to the simulation case, then students will be asked to begin working on a scenario. In the scenario students will be provided with different patients' heart sounds, and then they will be asked to draw a graphical representation of that sound.

This program will help students to develop their musical and logical intelligences concurrently. Students will be required to present their understanding logically through the graphical representation of a particular sound. If the student has a problem hearing the sound, then s/he will draw an incorrect chart. The system will recognize the student's problem and will provide suggestions. This example shows how musical and logical intelligences can be combined together and describes a method of developing these two intelligences to medical students.

2.3 Spatial, kinaesthetic and logical intelligences

In medicine this combination is especially important for surgeons. In surgery physicians need to visualize with very high accuracy the anatomic structure of organs. They need to be able to view them from different directions before starting surgery. More generally, medical students need to visualize the appearance and development of a disease prior to and during physical examination of a patient.

Modern radiological techniques like CT, MRI, and PET provide precise spatial representations of patient anatomy. These representations can also be enhanced with computer graphics providing 3-D models that allow students to practice surgery. For instance, the VOXEL-MAN system (Hohne, Pflesser, Pommert, Riemer, Schiemann, Schubert & Tiede, 1995) allows learners to practice surgery in a well-designed environment.

The evolution of virtual reality now adds further possibilities, linking spatial and kinaesthetic ability. For example, the University of North Carolina has developed a virtual reality application to help a physician to position the beams of radiation used in cancer therapy (McLellan, 1994). Another example is the Virtual Reality Assisted Surgeon, which enables surgeons to interactively visualize 3-D renderings of CT and MRI data. It also enables the surgeon to scale, orient and position pre-scanned body imagery on-line in real time from any desired perspective.

Another example of interaction between visual and bodily kinaesthetic intelligences is the MAGNOBRAIN project (Anogianakis, Apostolakis, Harding, Krotopoulou, Nendidis, Psaltikidou, Terpou & Tsakalidis 1995). This is a European research project that is being developed to serve as a clinical tool for bioelectric / biomagnetic source localization and as an aid in clinical diagnosis of brain diseases.

In summary, in practicing surgery the learner needs: a) spatial ability in order to visualize the radiologic scans of the patient in three dimensions, b) logical ability to follow a scientific and logical method of diagnosis, leading to a safe and appropriate treatment, c) bodily abilities to practice surgery with very high accuracy.

2.4 Logical, linguistic and interpersonal intelligences

Scientists, physicians and other medical practitioners use their logical abilities to carry out experiments and analyse data. They rely on their linguistic abilities to interpret and discuss their findings. These two intelligences are dominant in the scientific community. Medical practitioners normally try to deliver their findings logically using language with their colleagues or students. They start by analysing their findings, and then transmit them through language which is either in written (book or papers) or spoken (discussion and seminar) form.

Medical students are required to think logically and discover relationships and patterns between different concepts and diseases. They have to work in a systematic way. Also of use is interpersonal intelligence. It is crucial for them to have an interpersonal ability so they can communicate their understanding and findings with their patients, consultants and other students.

2.5 Visual and linguistic intelligences

Learning to diagnose radiologic images is a difficult task for medical students. Anatomical features in radiologic images are hard to identify hence students can easily misinterpret them and get confused between different pathologies. Therefore, a consistent, appropriate language to describe appearance and abnormalities is crucial for interpretation of X-ray, CT scans, MRI images, etc. Also, it is important for the description and comparison between normal and abnormal cases and for explaining how different organs act in association with each other or individually.

The systematic and precise linguistic interpretation of different pathological findings from visual representations is important because: a) it resolves ambiguity, b) it enhances understanding, c) it allows easy transfer of knowledge.

An ITS which joins these two intelligences is the MR Tutor, developed by the University of Sussex and De Montfort University Institute of Neurology (Sharples, du Boulay, Teather, Teather, Jeffery & du

Boulay 1994). This system is designed to train radiologists in how to interpret MR images of the brain using a structured image description language.

2.6 Technology and combination of all intelligences

The combinations of all seven intelligences can be found on the Internet. The World Wide Web (WWW) gathers together people from different locations with different cultures, backgrounds, abilities and time zones. It joins them in their learning processes and provides multiple representations of knowledge. It encourages individuals to communicate with each other in solving their problems.

An example of medical education that combines all seven intelligences is the interactive patient (Hayes & Lehmann, 1995). It is delivered via the WWW. The interactive patient is an on-line learning-by-doing approach designed by Marshall University School of Medicine. It provides learners with an environment that combines different intelligences and encourages students to use their different intelligences. This program attempts to educate physicians with access to the Internet, and to give them a possibility to obtain continuing medical education without having to leave their town.

The interactive patient provides the learner with a simulated patient. The learner can choose to take the patient history, examine the patient, or request laboratory and radiological tests. Learners can get the requested tests immediately.

An example of a medical specialist who combines all seven intelligences together in his practice is the cardiologist surgeon. The following example shows how cardiologist surgeons use different intelligences in their routine.

A cardiologist needs to have a well developed musical ability. S/he has to be able to investigate different sound disorders of the patient's heart. Also, a cardiologist surgeon needs to develop the skilful, very precise hand movements necessary for practicing surgery. In addition, surgeons use their linguistic and interpersonal abilities to discuss with patients the details of their illnesses. Also, surgeons need to discuss the best way of treating and of practicing surgery with other consultants. Surgeons also use their spatial ability to view patient radiologic images in two and three dimensions, as well as to rotate and view the image from different angles. Moreover, surgeons use logical ability to analyse laboratory tests and patient graphical representations. Therefore, in surgery the medical consultant and physician need to be competent and be able to use all intelligences together.

3 Multiple intelligences and instructional design

The second part of this paper discusses how the theory of multiple intelligences encourages multiple representations of knowledge. The MI theory allows instructional designers to choose appropriate technology to support learning according to the educational needs of their domain. As Gardner proposes, not all individuals are typical in their way of processing information. Therefore, it is important for instructional designers to choose an appropriate learning theory to meet the needs and the abilities of diverse learners in different domains. They may combine two or more intelligences to clarify certain topics and deepen students' comprehension of that task.

Applying MI theory to multimedia can also allow learners to choose for themselves appropriate ways to process their information, by giving the learners the appropriate tools to process information according to their own intelligences or learning styles. Students can choose to get the information as text, graphics and 3-D images, videos, animations, sounds, or simulation. MI theory offers learners the flexibility to drive their learning by giving them an appropriate learning situation and by encouraging them to choose the appropriate way for processing information according to their intelligences. Students also need to be aware of their different intelligences, and to be able to know their areas of strength, and try to develop them further.

Thus, a multimedia instructional designer can develop systems which has several ways of processing information. The choice can be left for the learners to choose their appropriate method. Multimedia systems can be enhanced with artificial intelligence techniques through the design of ITSs. Intelligent multimedia systems can keep track of students' preferred way of learning and can adapt the system according to the learners' learning styles.

In addition, the benefits for using the MI approach in educational system design are: firstly, it is a user-centred approach which is based on providing the user with information in a way that is relevant to their method of processing the information. By using MI theory, learners may choose to process their knowledge spatially via 3-D images, movies, text, sound, etc. Or they can choose to practice doing an experiment through simulation or virtual reality. Second, it is a constructivist approach to learning where the learners have control of the learning.

In designing educational multimedia systems or other learning materials to be used by several people with diverse learning styles, the instructional designers can make use of MI theory as a framework. By using MI theory as a general checklist, the instructional designer can cover different learners with their diverse abilities. This should also ensure that diverse learners will be motivated to use the system, because it meets their learning styles.

4 Conclusion

To conclude, MI theory acts as an essential educational theory for the instructional designer. It allows designers to choose the appropriate technology to convey their knowledge and it helps the learner to choose the appropriate cognitive tools for their abilities. Hence, students will be actively engaged in the process of learning and their self esteem will be raised.

This paper is related to my DPhil research proposal. My research is about developing a networked intelligent multimedia tutoring system for teaching medical students. The system will be based on the learning-by-doing approach. It will take multiple intelligences theory as a general framework, and the cognitive apprenticeship approach as an instructional method. The system will be accessible from the Internet via the WWW navigators.

Acknowledgements

I would like to express my special thanks to Dr. Mike Sharples for his valuable advice and suggestions throughout writing this paper.

References

Anogianakis, G., Apostolakis, M., Harding, G. F. A., Krotopoulou, K., Nendidis, G., Psaltikidou, M. S. P., Terpou, D., & Tsakalidis, A. (1995). MAGNOBRAIN. In Morgan, K., Satava, R. M., Sieburg, H. B., Mattheus, R. & Christensen, J. P. (Eds.), *Proceedings of the International Conference on Interactive Technology and the New Paradigm for Healthcare*, pp. 4-14. IOS Press and Ohmsha.

Gardner, H. (1993a). *Frames of Minds: The Theory of Multiple Intelligences*. New York, Basic Books.

Gardner, H. (1993b). *Multiple Intelligences: The Theory in Practice*, New York, Basic Books.

Gardner, H. (1987). The theory of multiple intelligences. *Annuals of Dyslexia, 37*, 19-35.

Fenstermacher, K. (1995). Cardiac ausculation diagnosis instruction. *http://www.cs.uchicago.edu/ fensterm/CADI/CADI.html*

Hayes, K. & Lehmann, C. U. (1995). The interactive patient. Marshall University School of Medicine. *http://medicus.marshall.edu/medicus.htm*

Hohne, K., Pflesser, B., Pommert, A., Riemer, M., Schiemann, T., Schubert, R. & Tiede, U. (1995). A new representation of knowledge concerning human anatomy and function. *Nature Medicine, 1*(6), 506-511.

McLellan, H. (1994). Virtual reality and multiple intelligences: Potentials for higher education. *Journal of Computing in Higher Education, 5*(2), 33-66.

Morgan, H. (1992). An analysis of Gardner's theory of multiple intelligence. Paper presented at the Annual Meeting of the Eastern Educational Research Association.

Sharples, M., du Boulay, B., Teather, D., Teather, B. A., Jeffery, N. & du Boulay, G. H. (1994). The cognitive basis for an MR image tutor. *Proceedings of East-West Conference on Computer Technologies in Education, Part 1*, pp. 214-219, Crimea.

Appendix: Description of the seven intelligences

Linguistic intelligence Linguistically talented students have the ability to use language effectively. They have the ability to manipulate the structure of language, the meaning of language, and the practical use of language. Linguistic learners are usually mastery readers, they have a great ability to understand and remember what they have read (Gardner 1993a; Gardner 1993b; McLellan 1994).

Logical / mathematical intelligence Logically talented students have the ability to do skilful experiments, work with numbers, ask questions, explore patterns and relationships; they can handle long chains of logical reasoning, and have the ability to solve problems; they learn best by categorising, classifying, working with abstract patterns and relationships (Gardner 1993a; Gardner 199b; McLellan 1994).

Spatial / visual intelligence Spatially talented students have the ability to organize their thinking efficiently in visual and special forms, to orient themselves in three-dimensional space, to grasp the visual world accurately, to perform transformations on their initial realizations and to rebuild viewpoints according to their spatial experience by drawing their ideas graphically (Gardner 1993a; Gardner 1993b; McLellan 1994).

Musical intelligence Musically talented students have an acute sensitivity to music; they can skilfully listen and respond to music; they are good at picking up sounds, remembering songs, noticing pitches and rhythms, and keeping time; they learn best by rhythm, song, and music (Gardner 1993a; Gardner 1993b; McLellan 1994).

Bodily / kinaesthetic intelligence Kinaesthetically talented students have the potential to work skilfully with objects and to control their body movements. They process knowledge through bodily sensations (Gardner 1993a; Gardner 1993b; McLellan 1994).

Interpersonal intelligence Interpersonally talented students have the ability to talk effectively to people and to join groups; They are good at understanding people, leading, organising, communicating. They learn best by sharing, comparing, relating and co-operating (Gardner 1993a; Gardner 1993b; McLellan 1994).

Intrapersonal intelligence The intrapersonally talented students have thorough understanding of their inner selves. They normally work alone and pursue their own interests and goals. They learn best by working on individual projects and having their own space (Gardner 1993a; Gardner 1993b; McLellan 1994).

In: Noble, J. & Parsowith, S. R. (Eds.). (1996). The Ninth White House Papers: Graduate
Research in the Cognitive and Computing Sciences at Sussex. Cognitive Science Research
Paper 440, School of Cognitive and Computing Sciences, University of Sussex.

The Developmental Prerequisites of Self-Presentation

Robin Banerjee
robinb@cogs.susx.ac.uk

School of Cognitive & Computing Sciences
University of Sussex
Brighton
BN1 9QH

Abstract Self-presentation—verbal and non-verbal behaviour intended to control others' impressions of the self—has been the focus of much social psychological research for several decades. Its role in children's social-cognitive development, however, has been largely ignored. This discussion considers the developmental prerequisites of self-presentation. Special attention is paid to the developments in self-awareness, the increasing sophistication of mental-state understanding, and the changes in motivational interests that trigger the self-presentation processes so important for social interaction in childhood and adulthood.

1 What is self-presentation?

We may start with the basic premise that in any social interaction we present ourselves in a certain way, whether we do so consciously or unconsciously. However, the concept of self-presentation becomes vacuous if used to cover all behaviours in a social interaction. Accordingly, most authors have limited their definitions of self-presentation by emphasising the basic motive of *attempting to control others' impressions of the self* (Baumeister, 1982; Goffman, 1959).

Self-presentational motives can manifest themselves in a variety of ways: through speech, through expressive features and gestures, through material displays, and through purposive behaviours (Schneider, 1981). Most of the existing social psychological literature on self-presentation in adults has focused on verbal self-presentations, for example, looking at self-descriptions in mock job interviews (e.g., Jones, Gergen & Davis, 1962). However, it is clear that non-verbal behaviour—facial expression, posture, appearance, clothing, material possessions, altruistic acts, conformity—is an equally important means by which we attempt to manipulate the impressions others have of us (see Schlenker, 1980). All of these self-presentations must obviously be intentional, but they may or may not be conscious. For example, an automatic polite smile triggered by a boss's over-used joke may be motivated by a desire to convey a particular impression of the self to the boss, yet this intention may not have been in conscious awareness at the time of the act (cf. Schank & Abelson's (1977) work on scripts, and Goffman's (1959) work on roles).

Before we move on to the developmental prerequisites of self-presentation, it should be noted that self-presentations need not be deceptive. Self-presentations may or may not match current or "plausible" self-concepts (Rhodewalt, 1986). For example, when we are actively trying to create a favourable impression on a job interviewer, we sometimes may make deceptive claims about ourselves, but will often simply concentrate on selectively projecting what we see as the positive aspects of our self-concepts. However, there is a fine line between being economical with the truth and being deceptive, and it remains to be seen whether this distinction translates into different patterns of development. Further examination of this issue is presented later in this discussion.

2 Can children be self-presenters? Cognitive prerequisites

Despite the vast literature on the role of self-presentation in adult social processes, little attention has been paid to the developmental origins of self-presentation. It seems clear that public face—the projected self-image—is of great importance to preadolescents and adolescents (Fine, 1988), and observational work suggests that even kindergartners use primitive versions of adult face-work strategies to "repair" a damaged public face after being criticized or threatened (Hatch, 1987). Indeed, Sluckin (1981) concluded from his observation of school-children in the playground that children's self-esteem, like adults, depends greatly on their "reputation," or public face. Unfortunately, while a handful of experimental studies have investigated various aspects of self-presentational behaviour in children, none has been informed by a theoretical understanding of the prerequisites of self-presentation. We start with a discussion of the cognitive prerequisites of self-presentation, in the areas of self-awareness and mental-state understanding.

2.1 Self-awareness

At the very least, a self-presenter must be aware of the self as an acting, thinking, and feeling entity distinct from others. This self-awareness is thought by many to show itself first in the form of visual self-recognition. Work by Lewis and Brooks-Gunn (e.g., 1979) using photographs, mirror images, and video images suggests that a recognition of the self through categorical cues (i.e., stable and enduring categorical features of the self) appears in the second year. Also at this time, infants are able to refer to themselves by name, age, and sex, and are starting to reflect not just upon their physical characteristics and actions (e.g., "red hair," "I play"), but also on their current perceptions, cognitions, feelings, and motivations (e.g., "I see a car," "I don't believe it that Brian went to A…", "I'm sad I popped it," "I wanna take nap"—Bretherton & Beeghly, 1982; Dunn & Brown, 1991; Shatz, 1994).

Furthermore, as children grow older, their self-concepts undergo several qualitative shifts. Their self-descriptions will refer not just to enduring physical characteristics and momentary mental states but also to regular activity patterns and stable dispositions (e.g., Yuill, 1993). The capacity to conceive of the self in this way is clearly important, since dispositional characteristics form the subject matter of many self-presentations (e.g., presenting the self as confident, friendly, or generous). Indeed, we may expect that developmental changes in the content of self-descriptions will be directly associated with shifts in the kinds of self-presentational goals generated by children. In a similar vein, we will later see that the nature of the self-concept, and consequently the nature of self-presentations, is also shaped by changes in motivational concerns, as social comparison processes increase in importance (e.g., Butler & Ruzany, 1993).

2.2 Understanding of mental states

In addition to having the capacity to reflect privately on the self, a self-presenter must understand that s/he is a public object that is perceived and evaluated by others. Besides being able to remark on their own beliefs, attitudes, desires, and emotions, 3–4-year-olds seem able to comment on others' mental states as well: Brown (1973) and Wells (1985) report that toddlers make reference to both their own and others' intentions, and research on children's theory of mind and on their understanding of emotion has shown that pre-schoolers regularly refer to others' beliefs and emotions (see Harris, 1989; Perner, 1991a). The use of self-presentation would appear to require the capacity to attribute mental states both to the self and to others, for the child must understand others' beliefs or evaluations about his/her own physical or psychological characteristics. Unfortunately, little systematic research has specifically examined the development of the capacity to attribute evaluations of the self to others. However, we may turn to the flourishing literature on children's understanding of mental states in order to formulate hypotheses about when and how this capacity is likely to emerge. An understanding of mental states is clearly a requirement for self-presentation: to be aware of how one is evaluated by others, one must be able to conceive of the mental states of others.

Research on mental-state understanding has taken many angles. Firstly, observation of children in natural settings indicates that children behave as if they are aware of others' expectations, intentions, and emotions. For example, Reddy (1991) describes how infants in their first year will offer an object and then teasingly withdraw it, thus seeming to deliberately create and play on a false expectation. Similarly, Dunn (1991) writes of the comforting, helping, and joking behaviour of infants in their second year, which is responsive to others' mental states (e.g., distress, goals). Studies of everyday conversations have also demonstrated that children in their second and third year reflect on their own and others' mental states (e.g., Bretherton & Beeghly, 1982), often strategically as part of excuses and justifications for transgressions (e.g., Dunn, 1988). Finally, work on pretence indicates that 2-year-olds are perfectly capable of understanding and behaving in accordance with a framework of pretence set up by others (e.g., Dunn & Dale, 1984). This kind of work provides convincing evidence that young infants have at least some knowledge of others' mental states, but none of these studies have the control over conditions necessary for drawing conclusions about the toddler's capacity to intentionally control others' *mental states* (as opposed to others' *behaviours*). With this in mind, we may now turn to theoretical and empirical frameworks specifically concerned with children's understanding of their own and others' mental states.

2.2.1 Social role-taking

Selman's work on *social role-taking* (e.g., 1980), from the Piagetian tradition, is still a popular theoretical framework; it conceives of the stages the child might pass through in the development of capabilities for understanding and co-ordinating other people's perspectives. According to his theory, only from ages 6 to 8 are children aware that others have social perspectives different from their own. Even then, children at this age are supposedly unable to co-ordinate these viewpoints (focusing instead on just one viewpoint at a time) and are for the most part inaccurate in judging their content. At this stage, then, we would not expect children to demonstrate an understanding of the public self (the self as perceived and evaluated by others). However, Selman indicates that between the ages of 8 and 10 the child acquires the ability to understand and co-ordinate other people's viewpoints, allowing them to reflect upon the self from someone else's viewpoint. This is the age at which, if the theory is correct, we can expect children to become aware of the fact that they are perceived and evaluated by others (and hence be capable of using self-presentation tactics). In support of this, studies by Aloise-Young (1993) and Yoshida, Kojo, and Kaku (1982) showed that 8-year-olds are capable of manipulating self-descriptions for self-presentational goals; 6-year-olds were found to be incapable of such manipulations.

To *understand* self-presentation in others requires the additional ability to reflect independently on various parties' interpretations of each other's viewpoints. This, according to Selman, would require Stage 3 role-taking capabilities: the ability to co-ordinate the perspectives of several interacting parties. According to the role-taking stage theory, this does not develop until around age 11. Again, this approach has found empirical support, this time in the work of Bennett and Yeeles (1990), who found that 11-year-olds but not 8-year-olds provided interpersonal (self-presentational) explanations for self-promotion and ingratiation tactics used by story characters.

2.2.2 Theory of mind

If we were to take the *"theory of mind"* approach, rather than looking at the stages of social role-taking, we would make rather different predictions. In contrast with role-taking theory's concern with rather broadly defined "viewpoints" or "perspectives," the "theory of mind" refers to a conceptual understanding of specific mental states and the processes by which these states interact with each other and with the external world. Most research in theory of mind is concerned with children's understanding of belief; specifically, whether children appreciate the "representational" nature of belief (belief as a representation of reality that may or may not be accurate). The false belief paradigm (Wimmer & Perner, 1983), which requires the child to predict actions on the basis of false beliefs, has been used extensively as an index of theory of mind acquisition. Results reliably suggest a transition at age 4. Furthermore, research (e.g., Sodian, 1991) indicates that children from that age can apply this representational understanding to execute deceptive strategies designed to fulfill certain desires or achieve certain goals (e.g., deliberately misleading a competitor about the location of a prize).

The research on theory of mind so far does not allow us to determine with any confidence whether or not children of this age are aware of others' representations of the self (as opposed to others' representations about the location of an inanimate object). Such representations often involve more than just a simple belief about how reality is; they involve abstract evaluations (e.g., he is kind) that are based on initial perceptions (e.g., he shared the chocolate). There is virtually no evidence on children's understanding of how perceptions of one's actions may be evaluated differently by different audiences. Therefore, while we can be fairly confident that the use of self-presentation cannot occur before an understanding of false belief, it is difficult to know at this stage whether understanding false belief is *sufficient* for understanding evaluative representations about the self.

As mentioned earlier, the *understanding* of self-presentation in others is likely to require a further level of cognitive sophistication. Using the "theory of mind" approach, one may argue that this understanding requires an appreciation of second-order representation, since the child must appreciate one person's intention to manipulate the representations others have of them. There is, however, considerable debate about whether intention is representational in the first place. Perner (1991b) points out that many desires can be understood non-representationally, as hypothetical overt goal-situations. For example, Astington (1991) showed that 3-year-olds conceived of intentions such as wanting to swing in terms of the act of swinging (i.e., the goal-situation), rather than in terms of running to a swing (i.e., the intentional action leading to that goal). However, the self-presentational intention is particularly complex because the goal involves the manipulation of others' *representations* of the self, for example, "as clever". This intention would clearly be difficult to translate into an overt goal-situation. Therefore, we can expect the understanding of self-presentational intention to emerge in parallel with the understanding of second-order false belief, which is also a representation about a representation (Perner & Wimmer, 1985). Evidence suggests that an understanding of second-order false belief may appear as early as age 4 or 5 (Sullivan, Zaitchik & Tager-Flusberg, 1994), and almost certainly within the following two or three years (Perner & Wimmer, 1985).

At this stage, it is difficult to determine whether the development of self-presentation follows the course predicted by the role-taking framework or the course predicted by the theory of mind framework. Closer examination of the studies by Aloise-Young (1993) and Yoshida, Kojo and Kaku (1982) reveals that the tasks in these experiments were not really designed to pinpoint the minimum age for self-presentation. The self-presentational goals the children had to attain in these studies, (presenting the self as modest, and as skilled on some newly-learned ball-throwing game) were advanced, unclear or not explicitly stated, and quite likely not relevant to the youngest children in the samples. Furthermore, the study by Bennett and Yeeles (1990) on children's understanding of self-promotion and ingratiation, is likely to have underestimated children's understanding of self-presentation, since it drew inferences about this understanding from *spontaneous* explanations of story characters' verbal interactions. Children may fare better on tasks where the objective is to predict appropriate behaviour (from a set of choices) for a story character with a given self-presentational goal. Preliminary research of this kind has shown that 5-year-olds are capable of predicting appropriate facial displays for characters with different self-presentational goals. Moreover, this capacity was found to be related to the appreciation of second-order false belief (Banerjee & Yuill, under review). Thus, in line with findings that the social role-taking framework has underestimated children's mental-state understanding, we may find that the theory of mind model will be of greater use in analyses of the development of self-presentation.

Finally, before moving on to the motivational prerequisites of self-presentation, we should re-assess the distinction between plausible and clearly deceptive self-presentations mentioned earlier. From a representational point of view, it seems reasonable to suggest that plausible self-presentations are easier to understand because they do not involve false representations about the self. Rather, they simply entail manipulating others' knowledge states (e.g., make sure that others know something about the self). Leekam (1991) demonstrated that 4-year-olds can understand efforts to manipulate whether others are or are not ignorant about something, and it is likely that even younger children are capable of using such strategies. In fact, Reddy and Simone (1995) observed that 18–22-month-old toddlers will impart knowledge about a new toy selectively to only those people who were not present when the new toy was introduced, suggesting an awareness of who knows what about reality. From this, it may be only a small step to selectively projecting different facets of the self to different people. A full understanding of self-presentation, however, is likely to rely on more sophisticated mental-state understanding, as discussed above.

3 Do children care about self-presentation? Motivational prerequisites

Even if a child is cognitively capable of using or understanding self-presentation tactics, s/he clearly needs to experience or understand the *motivation* to control others' impressions of the self. In other words, the use and understanding of self-presentation presume a concern about social evaluation. Although there are likely to be individual differences in the extent of this concern (cf. Buss's (1980) work on public self-consciousness; Graziano, Leone, Musser & Lautenschlager's (1987) work on self-monitoring), all adults possess and act on self-presentational motivations at least occasionally. When can children be said to have such motivations?

3.1 Shyness and embarrassment

We might approach this question by determining when children first respond to negative social evaluation. Work on shyness and embarrassment in children is limited and inconsistent. Buss, Iscoe and Buss (1979) found that parents reported embarrassment in their children only when they reached age 4 or 5, and took this to indicate that children from this age are concerned about social evaluation by others. While this inference seems to be consistent with the suggestion that the capacity for self-presentation requires an appreciation of representational thought, it clearly cannot be justified purely on the basis of parental reports of embarrassment. Other authors believe that they have found evidence that even young infants are aware at some level of others' evaluations of the self. Lewis, Sullivan, Stanger, and Weiss (1989) argued that infants showed "coy" behaviour ("silly smiles," gaze-aversion) in front of mirrors, when over-complimented, and when asked to dance in front of an experimenter. This may show an awareness that others are viewing the self and that others may overtly react to the self, but whether or not this indicates an awareness of or concern with others' evaluations is an open question.

Perhaps the best test of children's concern about social evaluation comes from the work of Bennett and Gillingham (1991). They found that 8–11-year-olds, but not 5-year-olds, reported embarrassment when presented with hypothetical situations designed to elicit self-consciousness, despite the fact that the audience in these situations was not overtly negative. Since the audience was not actively deriding the child, the embarrassment must have derived from a concern about the audience's *evaluation* of the self. This does seem to suggest that 8-year-olds will be motivated to control others' impressions of the self. However, this study required children to verbalize feelings in hypothetical situations, and it is quite possible that 5-year-olds would have the same concern for social evaluation as 8-year-olds if they actually found themselves in those situations. Indeed, the evidence of Hatch (1987), that kindergartners make efforts to repair a damaged public face (e.g., after being criticized or threatened), strongly suggests that 4-5-year-olds *are* concerned about others' evaluations of the self. The difficulty with this kind of observational work, of course, lies in untangling the concern about others' *behaviours* towards the self from the concern about others' *evaluations* of the self, and further research will therefore be necessary to confirm that children of this age are not only capable of but are also motivated to manipulate social evaluation of the self.

3.2 Social comparison

A further development that is likely to be of great importance in provoking motivational interest in others' evaluations of the self, is the increasing preoccupation with social comparison as children start to interact with peers. Gottman and Parkhurst (1980) demonstrated that children aged 3 to 5 years old regularly compared themselves with others, determining similarities, establishing differences, and exploring others' beliefs, feelings, and desires. Social comparison is likely to be of great importance in the development of self-presentation, particularly since it is implicated in self-concept formation. Harter (1983, 1988) notes that others' reactions to the self form an important evaluative component of self-definitions. This self-evaluation by way of comparison with others seems likely to enable the child to ascertain what kinds of self-presentations are socially desirable.

Social comparison should continue to be instrumental in motivating self-presentational behaviour throughout childhood, since peer groups play a significant role in evaluative processes through to preadolescence and adolescence. Fine (1988, p.217) points out that the deliberate manipulation of others' impressions becomes "critical to long-term popularity" during preadolescence. As children grow older, it is likely that the criteria that form the basis for social comparison, (e.g., running fast, being "tough," having particular material possessions) will also determine the content of verbal and non-verbal self-presentations. Future research must trace not only this development, but also the increasing sophistication and complexity of self-presentations, as children learn from experience what personal characteristics are desirable for which target audience.

4 Conclusion

Research on children's social-cognitive development is a rapidly-growing area, but it has so far neglected the implications of this development for social processes such as self-presentation. The existing literature provides a useful foundation for formulating hypotheses about when we can expect children to be capable of using—and motivated to use—self-presentation tactics. It is the task of future research to test such hypotheses.

References

Aloise-Young, P. A. (1993). The development of self-presentation: Self-promotion in 6- to 10-year-old children. *Social Cognition, 11*, 201-222.

Astington, J. W. (1991). Intention in the child's theory of mind. In Frye, D. and Moore, C. (Eds.), *Children's Theories of Mind: Mental States and Social Understanding*, pp. 157-172. Lawrence Erlbaum, Hillsdale, NJ.

Banerjee, R., & Yuill, N. (under review). Children's understanding of self-presentational display rules: Associations with mental-state understanding.

Baumeister, R. F. (1982). A self-presentational view of social phenomena. *Psychological Bulletin, 91*, 3-26.

Bennett, M., & Gillingham, N. (1991). The role of self-focused attention in children's attributions of social emotions to the self. *The Journal of Genetic Psychology, 152*, 303-309.

Bennett, M., & Yeeles, C. (1990). Children's understanding of the self-presentational strategies of ingratiation and self-promotion. *European Journal of Social Psychology, 20*, 455-461.

Bretherton, I., & Beeghly, M. (1982). Talking about internal states: The acquisition of an explicit theory of mind. *Developmental Psychology, 18*, 906-921.

Brown, R. (1973). *A First Language*. Harvard University Press, Cambridge, MA.

Buss, A. H. (1980). *Self-Consciousness and Social Anxiety*. Freeman, San Francisco.

Buss, A. H., Iscoe, I., & Buss, E. H. (1979). The development of embarrassment. *Journal of Psychology, 103*, 227-230.

Butler, R., & Ruzany, N. (1993). Age and socialization effects on the development of social comparison motives and normative ability assessment in kibbutz and urban children. *Child Development, 64*, 532-543.

Dunn, J. (1988). *The Beginnings of Social Understanding*. Blackwell, Oxford.

Dunn, J. (1991). Sibling influences. In Lewis, M. & Feinman, S. (Eds.), *Social Influences and Socialization in Infancy*, pp. 97-109. Plenum Press, New York.

Dunn, J., & Brown, J. (1991). Relationships, talk about feelings, and the development of affect regulation in early childhood. In Garber, J. & Dodge, K. A. (Eds.), *The Development of Emotion Regulation and Dysregulation*, pp. 89-108. Cambridge University Press, Cambridge.

Dunn, J. & Dale, N. (1984). I a Daddy: 2-year-olds' collaboration in joint pretend with sibling and with mother. In Bretherton, I. (Ed.), *Symbolic Play: The Development of Social Understanding*, pp. 131-158. Academic Press, Orlando, FL.

Fine, G. A. (1988). Friends, impression management, and preadolescent behavior. In Handel, G. (Ed.), *Childhood Socialization*, pp. 209-233. Aldine de Gruyter, NY.

Goffman, E. (1959). *The Presentation of Self in Everyday Life*. Doubleday, Anchor Books, New York.

Gottman, J. M., & Parkhurst, J. T. (1994). A developmental theory of friendship and acquaintanceship processes. In Collins, A. (Ed.), *Minnesota Symposium on Child Psychology, Vol. 13, Development of Cognition, Affect, and Social Relations*. Lawrence Erlbaum, Hillsdale, NJ.

Graziano, W. G., Leone, C., Musser, L. M., & Lautenschlager, G. J. (1987). Self-monitoring in children: A differential approach to social development. *Developmental Psychology, 23*, 571-576.

Harris, P. L. (1989). *Children and Emotion: The Development of Psychological Understanding*. Blackwell, Oxford.

Harter, S. (1983). Developmental perspectives on the self-system. In Hetherington, E. M. (Ed.), *Handbook of Child Psychology, Vol. 4, Socialization, Personality, and Social Development*, pp. 275-386. New York, Wiley.

Harter, S. (1988). Development processes in the construction of the self. In Yawkey, T. D. & Johnson, J. E. (Eds.), *Integrative Processes and Socialization: Early to Middle Childhood*, pp. 45-78. Lawrence Erlbaum, Hillsdale, NJ.

Hatch, J. A. (1987). Impression management in kindergarten classrooms: An analysis of children's face-work in peer interactions. *Anthropology and Education Quarterly, 18*, 100-115.

Jones, E. E., Gergen, K. J., & Davis, K. E. (1962). Some determinants of reactions to being approved or disapproved as a person. *Psychological Monographs, 76*,(2). Whole no. 521.

Leekam, S. R. (1991). Jokes and lies: Children's understanding of intentional falsehood. In Whiten, A. (Ed.), *Natural Theories of Mind*, pp. 159-174. Blackwell, Oxford.

Lewis, M., & Brooks-Gunn, J. (1979). *Social Cognition and the Acquisition of Self.* Plenum Press, New York.

Lewis, M., Sullivan, M. W., Stanger, C., & Weiss, M. (1989). Self-development and self-conscious emotions. *Child Development, 60*, 146-156.

Perner, J. (1991a). *Understanding the Representational Mind.* MIT Press, Cambridge, MA.

Perner, J. (1991b). On representing that: The asymmetry between belief and desire in children's theory of mind. In Frye, D. & Moore, C. (Eds.), *Children's Theories of Mind: Mental States and Social Understanding*, pp. 139-156. Lawrence Erlbaum, Hillsdale, NJ.

Perner, J., & Wimmer, H. (1985). "John thinks that Mary thinks that ...": Attribution of second-order beliefs by 5 to 10-year-old children. *Journal of Experimental Child Psychology, 39*, 347-371.

Reddy, V. (1991). Playing with others' expectations: Teasing and mucking about in the first year. In Whiten, A. (Ed.), *Natural Theories of Mind*, pp. 159-174. Blackwell, Oxford.

Reddy, V., & Simone, L. (1995). Acting on attention towards an understanding of knowing in infancy. Paper presented at the BPS Developmental Section Annual Conference, University of Strathclyde, Glasgow, 8-11 September.

Rhodewalt, F. T. (1986). Self-presentation and the phenomenal self: on the stability and malleability of self-conceptions. In Baumeister, R. F. (Ed.), *Public Self and Private Self*, pp. 117-142. Springer-Verlag, New York.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, Plans, Goals, and Understanding: An Enquiry into Human Knowledge Structures*. Lawrence Erlbaum, Hillsdale, NJ.

Schlenker, B. R. (1980). *Impression Management: The Self-Concept, Social Identity, and Interpersonal Relations*. Brooks/Cole, Belmont, CA.

Schneider, D. J. (1981). Tactical self-presentations: Toward a broader conception. In Tedeschi, J. T. (Ed.), *Impression Management Theory and Social Psychological Research*, pp. 23-40. Academic Press, New York.

Selman, R. L. (1980). *The Growth of Interpersonal Understanding.* Academic Press, New York.

Shatz, M. (1994). *A Toddler's Life: Becoming a Person.* Oxford University Press, New York.

Sluckin, A. (1981). *Growing Up in the Playground: The Social Development of Children.* Routledge, London.

Sodian, B. (1991). The development of deception in young children. *British Journal of Developmental Psychology, 9*, 173-188.

Sullivan, K., Zaitchik, D., & Tager-Flusberg, H. (1994). Preschoolers can attribute second-order beliefs. *Developmental Psychology, 30*, 395-402.

Wells, G. (1985). *Language Development in the Pre-School Years*. Cambridge University Press, Cambridge.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representations and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*, 103-128.

Yoshida, T., Kojo, K., & Kaku, H. (1982). A study on the development of self-presentation in children. *Japanese Journal of Educational Psychology, 30*, 120-127.

Yuill, N. (1993). Understanding of personality and dispositions. In M. Bennett (Ed.), *The Child as Psychologist: An Introduction to the Development of Social Cognition*. Harvester Wheatsheaf, Hemel Hempstead.

In: Noble, J. & Parsowith, S. R. (Eds.). (1996). The Ninth White House Papers: Graduate Research in the Cognitive and Computing Sciences at Sussex. Cognitive Science Research Paper 440, School of Cognitive and Computing Sciences, University of Sussex.

Scientific Induction and Transfer of Learning

Hilan Bensusan
hilanb@cogs.susx.ac.uk

School of Cognitive & Computing Sciences
University of Sussex
Brighton
BN1 9QH

Abstract Transfer of learning is a hot topic in machine learning. It is about the ability of learning systems to take advantage of learning more than one task. This work attempts to point at a possible analysis of some features of scientific induction by means of the framework and results emerging from this area.

1 Introduction

One of the most impressive feature of human inductive performance is the ability to generate predictively successful hypotheses from a relatively small amount of instances. This success, however, depends on how familiar the environment is. In other words, it depends on what can be exploited from the knowledge acquired in related tasks. This process of making use of the information gathered from other induction tasks to decrease the amount of data needed has been called "transfer of learning" in the psychological literature (Ellis, 1965) and has recently received the attention of the machine learning community, for instance in Murre (1995), Pratt (1994) and Silver and Mercier (1996). Transfer of learning is the process where the learning of one task improves the performance of the learner in related tasks learned either simultaneously or subsequently.

In this short paper, I intend to point at some phenomena described by methodologists and philosophers of science that can be described as instances of transfer of learning in scientific induction. In the next section I shall attempt to show how transfer is a means to improve the induction bias for a given task. In the remainder of the paper I shall concentrate on scientific transfer among simultaneous and sequential inductive tasks.

2 Transfer of learning: a feature of efficient induction

Transfer of learning might be seen as a case where solving more than one problem is easier. The basic process of transfer can be characterized as a bias adjustment process. The learner's bias—which can be seen as the set of assumptions of the learner or its prior probability distribution—is the tendency of the learner towards some hypotheses. Since the space of possible hypotheses in non-trivial learning problems is usually too large for an exhaustive search, some restrictions should be made. Normally, these restrictions come from the representation limitations of the learner—some hypotheses cannot be represented or are hard to express in terms of the learner's representation scheme and therefore the learner is biased against them. We refer to the set of hypotheses that the learner has the means to represent as being the hypothesis space of the learner—this is sometimes called the hard bias of the learner for it is always confined to the scope of its hypothesis space.[1] Apart from these hard biases, learners have also

[1] The size of the hypothesis space—and therefore the strength of the bias—can be measured by the so-called VC dimension of a learner (Blumer, Ehrenfeucht, Haussler, & Warmuth, 1989; Haussler, 1988) that expresses the greatest number of data points for which there is a consistent hypothesis in the hypothesis space. A small VC dimension means a hypothesis space with rather limited expression power and therefore a strong bias.

what can be called soft biases or straight preferences that can be expressed as a probability distribution over the hypothesis space. Biases, of course, are needed by any sort of learners—human or machine ones—and in both cases they are part of the induction mechanism.

For both humans and machines, the difficult part is to find the right (soft and hard) bias for a given learning problem. In machine learning, this difficulty is made apparent both in choosing the right learner for a given task and in finding the right initial configuration of a learner for a given task (the right network architecture for instance). In both humans and machines, different tasks require different biases but usually we don't know in advance which bias is the most convenient one.[2] The human solution to the difficulty is to rely on related previous experience. Hence, we don't consider the hypothesis that the number of tunnels in the railroad between New York and Pittsburgh is a function of the number of bananas we eat on the way and we don't assume that next year all the emeralds will look blue.[3] Similarly, we understand that if some disciplines are taught earlier in school, they can help the learning of others, later.

This transfer strategy seems to be the key to our inductive efficiency; we don't need to start every induction task from scratch because we can import knowledge from elsewhere. We can conjecture that our inductive success in our environment is therefore partly a consequence of our previous lifetime experiences in this environment, which has adjusted the biases accordingly to make safe generalizations easier, and partly a consequence of evolution—the previous experience of our species. Quine once said (Quine, 1994) that the problem of how we are better than a random guesser in induction tasks can be partially answered by evolution. A general answer would have that the key for success is the presence of adequate biases for our current environment. This process of bias finding can be illuminated by the recent machine learning research on transfer. This research can also help to understand another area of human induction that is seldom or never accounted for in terms of transfer: scientific theory construction. Although scientific induction might be considered in a continuum with everyday inductive tasks, it has a number of characteristic features, due, among other reasons, to its relatively greater complexity. We might therefore want to ask how and to what extent scientific induction makes use of transfer to determine the appropriate biases.

3 Science and synchronic transfer: conscilience

Baxter (1995) has provided a framework for transfer and bias learning where the fundamental result is a proof that the bound on the minimum number of examples required per task for learning multiple tasks is lower than the minimum number of examples required to learn each task. This is so when the bias needed for one task can be usefully applied to the others—the tasks being therefore related. We can say that the relevant internal representations for the task class is learned during the learning (and transfer) process and is useful to complete the learning of the class.

According to Baxter's framework, we can measure the transfer accuracy of a learning system either by considering its generalization accuracy in the tasks it deals with or by measuring its performance in other, related tasks. These two ways of measuring accuracy suggest two different transfer situations: a situation where different tasks are learned simultaneously (often called multitask learning), and another where tasks are presented in sequence. In the first case, internal representations are acquired by the simultaneous contribution of all tasks whereas in the second case the bias learned in some tasks is applied to the others. We can call the first situation synchronic transfer and the second diachronic transfer. These two modalities of transfer can be usefully mapped into two different processes that take place in scientific theory construction, as we shall see in the remainder of this paper.

A typical example of synchronic transfer in the recent machine learning literature can be found in Caruana's multitask leaning network (Caruana, 1996). In this system, various tasks are used to help the

[2] As might be intuitively clear, there is no universal bias. In fact, each bias is bound to perform as poorly in some tasks as it performs well in others. This can be seen as a consequence of the results of Schaffer (1994) and Wolpert (1995).

[3] These examples are references to similar ones used in Hanson (1969) and Goodman (1983) respectively.

learning of a main task. The multitask network is a three layer feed-forward network with one output node for each task to be learned. The inputs for the different tasks are provided to the network and the shared hidden layer aims to represent the structure common for all tasks—the internal representation or the common bias for the task class. The network is trained by backpropagation (Rumelhart, Hinton, & Williams, 1986). Caruana reports results in learning problems such as that of predicting the mortality chances of pneumonia patients given the prior to hospitalization test results as input and after-hospitalization test results as helping tasks (predicted by output nodes). The performance is reported to be better than the performance of the single output node network with the same inputs.

The process of learning internal representations by multitask learning and synchronically transferring knowledge among the tasks can illuminate what goes on when a scientific theory is used to add confirmation to an empirical law. There is a long discussion in philosophy of science about the dispensability of theories (Hempel, 1965; Craig, 1953; Putnam, 1962; Ramsey, 1990) where one part claims that scientific theories can be dispensed with and we can recontruct science by using merely empirical laws. This assumption is challenged by Putnam (1963) by saying that theories are needed to discover and confirm some empirical laws and therefore they are at least heuristically necessary. Putnam's example is the statement P saying that "when two subcritical masses of uranium 235 are slammed together to produce a single supercritical mass there will be an explosion" before the first large nuclear explosion. By that time, the only support for P was to be found in the nuclear theory, by itself supported by some empirical evidence. The nuclear theory is part of the inference that enabled the prediction of P before any large nuclear explosion.

Theories, therefore, are used not only to unify the different pieces of evidence and empirical laws but also as a heuristic guide for discovery and confirmation. In other words, a theory might be seen as an internal representation or a bias common to a task class. Sometimes, as in Putnam's example, the bias is so strong and adequate that no further evidence is needed to learn a different empirical law. To view theories as internal representations or as biases for a class of empirical phenomena might help to explicate conscilience—the scientific reliance upon theories for learning empirical laws (see Hesse (1968) for a discussion of conscilience). Conscilience is what explains the additional support gained by Kepler's second law due to its unification with the law of falling bodies by Newtonian mechanics. Theories, by providing a learning bias, guide the discovery of related empirical phenomena. Theory construction can therefore be compared to bias learning and understood as second order induction. In a specific analogy with Caruana's multitask network, we can view the need for a theory as something that adjusts the bias for the main task aimed at, i.e., the empirical law that is to be found. In any case, the inference from various related empirical laws to the construction a theory can be seen as a clear case of synchronic transfer.

4 Science and diachronic transfer: entrenchment

In contrast to synchronic transfer, diachronic transfer occurs by profiting from already learned related tasks. Therefore, transfer performance should be measured by accuracy in novel tasks. An example of diachronic transfer architecture among the neural network learners can be found in the consolidation system (Silver & Mercier, 1995). The system is composed of two feed-forward networks: a task network and an experience network. The former uses backpropagation to learn a class of tasks, some of them related to each other while the latter is trained to correlate the provisory weights of the former network during its training stage of a task—called signature weights—to the converged weights for each task. The experience network is therefore a $n/n+\sqrt{n}/n$ network where n is the number of connections in the task network. The knowledge acquired in the learning of the first group of tasks is therefore preserved in the experience network that is then used to provide the task network with an adequate converged weight vector for a task, which is in turn used to provide the initial weights for the task network. The initial weight vector is the bias that is learned by the experience network from the weights achieved by the task network in the initial sequence of tasks. Transfer takes place when weight vectors for previously learned tasks are used as initial weights for a new, related task. The authors report being able to reduce the number of examples required for safe learning of a sequence of Boolean problems.

In diachronic transfer, an already learned bias is used to speed up learning. This is precisely what seems to happen when scientists make use of previously accepted theories to guide generalizations. The previously accepted related laws and theories are used to determine the shape and the language for the new laws and theories. Theory construction is guided by previously accepted theories that, given the continuity of science, act as a bias by preventing some conclusions while emulating others. Boyd nearly describes the process as a diachronic transfer when he considers the body previously accepted scientific theories as

> . . . establishing principles of scientific rational inductive reasoning which, sometimes, dictate conclusions which we must accept, given that we accept a particular theory. [. . .] Existing theoretical knowledge often sets a sharp constraint on the methodologically acceptable responses to new data. . . (Boyd, 1985, p. 247–8)

This restriction on the space of acceptable theories is what Goodman (1983) has called the entrenchment of hypotheses. Accepted scientific hypotheses should be relevantly similar to the ones already accepted: they should, for instance, be formulable by using the predicates already used in the existing theories.

Diachronic transfer machine learners—for instance, the consolidation system—can help to clarify entrenchment. What is transferred to the new task is what could be learned from the previous ones; the experience network, in the consolidation system, speeds up learning by improving the initial points for the search in the weight space. The bias learned is used to guide the learner in new tasks and this is why entrenchment should be understood as a tendency towards accepted theories. Also, if we understand this as being a soft bias, we can take scientific revolutions—large ruptures with the accepted body of theoretical knowledge—to be the generation of hypotheses that are not among the high-ranked preferences given by the bias (the previous scientific knowledge) due to the incapacity of the preferred hypotheses to fit the data. In any case, the methodological import of accepted knowledge can be illuminated by seeing it as a bias and the whole process as transfer.

References

Baxter, J. (1995). Learning internal representations. In *Proceedings of the Eighth International Conference on Computational Learning Theory* Santa Cruz, CA, USA. ACM Press.

Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M. (1989). Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, *36*(4), 929–965.

Boyd, R. (1985). The logician's dilemma. *Erkenntnis*, *22*, 197–152.

Caruana, R. (1996). Algorithms and applications for multitask learning. In Saitta, L. (Ed.), *Proceedings of the Thirteenth Conference on Machine Learning*, pp. 87–95 San Mateo, CA, USA. Morgan Kaufmann.

Craig, W. (1953). On axiomatizability within a system. *The Journal of Symbolic Logic*, *18*(1), 30–32.

Ellis, H. (1965). *The Transfer of Learning*. The Macmillan Company, New York, USA.

Goodman, N. (1983). *Facts, Fiction and Forecast*. Harvard University Press, Cambridge, Mass., USA.

Hanson, N. R. (1969). *Perception and Discovery*. Freeman and Cooper Co, USA.

Haussler, D. (1988). Quantifying inductive bias: AI learning algorithms and Valiant's learning framework. *Artificial Intelligence*, *32*(2), 177–222.

Hempel, C. G. (1965). The theoretician's dilemma. In *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, pp. 173–226. Free Press, New York, USA.

Hesse, M. (1968). The conscilience of inductions. In Lakatos, I. (Ed.), *Inductive Logic*, pp. 232–246. North Holland Publishing Company, Amsterdam, The Netherlands.

Murre, J. (1995). Transfer of learning in backpropagation and in related neural network models. In J. Levy, D. Bairaktaris, J. B., & Cairns, P. (Eds.), *Connectionist Models of Memory and Language*. UCL Press, London, UK.

Pratt, L. Y. (1994). Experiments on the transfer of knowledge between neural networks. In S. Hanson, G. Drastal, R. R. (Ed.), *Computational Learning Theory and Natural Learning Systems, Constraints and Prospects*, pp. 523–560. MIT Press.

Putnam, H. (1962). What theories are not. In Nagel, E., Tarski, A., & Suppes, P. (Eds.), *Logic, Methodology and Philosophy of Science*, pp. 240–251. Stanford University Press, Stanford, CA, USA.

Putnam, H. (1963). Degrees of confirmation and inductive logic. In Schilpp, A. (Ed.), *The Philosophy of Rudolf Carnap*. Open Court, Chicago and LaSalle, Ill., USA.

Quine, W. V. O. (1994). Natural kinds. In Kornblith, H. (Ed.), *Naturalizing Epistemology*, pp. 57–76. MIT Press, Cambridge, Mass., USA.

Ramsey, F. P. (1990). Theories. In Mellor, D. (Ed.), *Philosophical Papers*, pp. 112–137. Cambridge University Press, Cambridge, UK.

Rumelhart, D., Hinton, G., & Williams, R. (1986). Learning internal representations by error propagation. In Rumelhart, D., McClelland, J., & the PDP Research Group (Eds.), *Parallel Distributed Processing: Explorations in the Micro-structures of Cognition. Vols I and II.* MIT Press, Cambridge, Mass.

Schaffer, C. (1994). A conservation law for generalization performance. In *Proceedings of the Eleventh Conference on Machine Learning* San Mateo, CA, USA. Morgan Kaufmann.

Silver, D. L., & Mercier, R. E. (1995). Toward a model of consolidation: The retention and transfer of neural network knowledge. In Associates, L. E. (Ed.), *Proceedings of the INNS World Congress on Neural Networks*, Vol. III, pp. 164–169. http://www.csd.uwo.ca/ dsilver/.

Silver, D. L., & Mercier, R. E. (1996). The parallel transfer of task knowledge using dynamical leaning rates based on a measure of relatedness. Tech. rep., Department of Computer Science, The University of Western Ontario, London, Ontario. http://www.csd.uwo.ca/ dsilver/.

Wolpert, D. (1995). Off training set errors and a priori distinctions between learning algorithms. Tech. rep. 95-01-003, Santa Fe Institute, Santa Fe, NM, USA.

In: Noble, J. & Parsowith, S. R. (Eds.). (1996). The Ninth White House Papers: Graduate Research in the Cognitive and Computing Sciences at Sussex. Cognitive Science Research Paper 440, School of Cognitive and Computing Sciences, University of Sussex.

Health Anxieties and the “Worried Well”: Locating and Defining an Elusive Population

Kate Cavanagh
katecav@cogs.susx.ac.uk

School of Cognitive & Computing Sciences
University of Sussex
Brighton
BN1 9QH

Abstract The threats associated with serious illness make it a natural object of fear. Some recent literature has focused upon the “Worried Well”, a term most recently used to refer to those individuals who suffer from varying degrees of concern about HIV and AIDS. There is considerable inconsistency and discrepancy in the literature as to the features of this group, and how this group might be best conceptualized for the purposes of research and ultimately treatment. In order to elucidate the characteristics of this population it is necessary to define the characteristics of those considered to be worried well, and to further locate these characteristics within pre-existing frameworks of health anxiety. Herein it is suggested that studies of previously described populations for whom the ‘fashionable’ illness of the era became the focus of health anxieties, may provide a fruitful basis for the construction of a theoretical and practical understanding of the worried well.

1 The impact of HIV and AIDS on research

The recognition of Acquired Immune Deficiency Syndrome (AIDS) in 1981, and the later discovery of its probable cause Human Immunodeficiency Virus (HIV), have had an unprecedented impact upon medical, psychological and sociological research. In addition to the neuropsychological sequalae of the syndrome, and attempts to understand the best methods by which to prevent further infection, the recognition of fears about HIV and AIDS has prompted considerable discussion.

2 The recognition of the worried well

The “Worried Well” (WW), are an elusive population. This term has been adopted in various contexts in psychological and medical literature. Although sometimes used in the context of more generalized hypochondriacal states, the term WW has more recently been associated with those individuals suffering from various degrees of concern about HIV and AIDS. However, even within this field to whom the term may be applied is unclear. The lack of a clear definition of whom might be considered WW exacerbates the difficulty in establishing the aetiological and presenting features of this group, and additionally causes problems for implementing effective treatment of these potentially distressing and disabling fears. Beyond the individual discomfort posed by these threats of illness, the financial implications of concerns about HIV and AIDS are far from inconsequential. A study of 270 general practitioners in London suggests that a considerable workload in primary care comprised of patients who obsessively fear contracting HIV (King, 1989). This review will attempt to highlight the discrepancies in published literature to date, and will outline some issues which need further consideration in order to further our understanding of the aetiological, maintaining and presenting features of the WW, to facilitate support for those persons vulnerable to such maladaptive reactions to a major public health concern and for effective research in this area to be undertaken. The WW require increased clinical and research attention from the psychiatric community (Faulstich, 1987). In order to elucidate the characteristics of this population it is necessary to clarify who constitutes the WW, if indeed any identifiable population exists.

3 Identifying the worried well

A review of the literature to date has cast doubt upon the clarity of any widely held clinical or theoretical definition of this group. Since the mid-1980's a variety of authors have posited definitions of who might be considered WW, however there is little agreement in this area. With the broadest catchment Dilley (1988) has suggested that the term WW has become applicable to nearly "every informed sexually active person". King (1993) has endorsed this wide application for the term including those with "varying degrees of concern about having contracted HIV". However this broad meaning may have shifted, influenced by authors such as Faulstich (1987) who infers that only those with histories of high risk behaviour, those "individuals who realize their lifestyles have put them at risk for infection" should be considered WW, a definition supported by Fornstein (1984). Harowski (1987) has expanded this definition to include HIV positive individuals, whose concerns are somatopsychic and realistic. However, Miller et al (1988) have suggested that the WW must remain "infection free, as verified by serological testing and clinical assessment", considering that without established seronegativity individuals could not be labelled "well". In contrast to Faulstich (1987) and Fornstein (1984), Goldmier (1987) has excluded individuals in high risk populations expressing the "essential feature is a background of minimal or no risk". Green (1989) has also considered the WW to have had "objectively little or no risk". Davey and Green (1991) have concluded that the WW are a "heterogeneous group". Indeed it is clear that attempts to pinpoint the WW within the current literature are essentially asinine. Adhering to any of these definitions in the light of the paucity of research in this area, and lack of coherent theoretical justification for these assertions would be extremely tenuous. As no further specificity is appropriate, herein the WW are conceptualized as those individuals suffering from extreme concerns about HIV and/or AIDS, in the absence of other dominant psychopathology.

4 Locating the worried well: the utility of existing frameworks of health anxiety

Existing frameworks of health anxiety might provide some foundation for locating and defining the WW. Hypochondriacal concerns and inappropriate health care use have been a focus of interest within psychological and medical literature for some time. The threats associated with serious illness would seem to make it a natural object of fear. However, hypochondriasis is one of the most poorly understood and inadequately researched phenomena in psychology. Indeed the empirical study of hypochondria is a recent phenomenon, and much of what was written pre-1990's was based upon clinical observation and assumptions. It has been suggested that the pejorative connotations associated with the term hypochondria may have restricted in-depth investigation of the phenomena. Health anxieties have been constructed within psychological and medical literature in several ways. The forth edition of the American Psychological Associations Diagnostic and Statistical Manual (1994) provides a "medical model" and classifies hypochondriasis as a somatoform disorder, the essential feature of which is the preoccupation with, fears of having or idea that one has a serious illness in the absence of any identifiable organic pathology. For diagnosis these concerns must interrupt work, academic or social functioning, and have been present for more than six months. The cognitive-behavioural formulation of health anxiety, presented in recent literature by authors such as Salkovskis and Warick (Salkovskis and Warick, 1986; Warick, 1989; Warick and Salkovskis, 1990), focuses upon preoccupations with bodily health including the precipitation of concerns out of proportion to existing justification in association with the (mis)interpretation of bodily sensations and the pursuit of reassurance. This formulation, in accord with the ideas of Ryle (1948), incorporates the notion of a continuum running from mild, transient concerns through to morbid health preoccupation. Whilst having wide ranging utility within clinical practice and providing a succinct theoretical basis for research, these notions of health anxiety provide little insight into the aetiological mechanisms underlying concerns about HIV and AIDS.

5 How might we conceptualize the worried well?

One appropriate theoretical road down which it may be profitable to travel in the pursuit of the aetiology of “worried wellness” is that of previous populations displaying similar characteristics. Within the recent literature relating to the WW there has been some considerable debate as the specific role of HIV and AIDS in worried wellness. Early correspondence in this debate inferred a causative role of AIDS in worried wellness (Riccio and Thompson, 1987; Miller et al, 1985; O’Brien and Hassenyeh, 1985; Jacob, 1987; Windgassen and Soni, 1987). However the weight of the argument has favoured the conceptualization of fears about HIV infection and AIDS as representing contemporary themes which are incorporated into pre-existing vulnerabilities to psychopathology (Jacob et al, 1989). O’Brien (1987) suggests that in the case of worries about AIDS, it is not something specific to AIDS that results in psychological disturbance, but a result of the fact that the devastating impact of AIDS has become a part of public consciousness. This conceptualization of fears about HIV and AIDS has been endorsed by several authors (e.g. Lipkin, 1986; Riccio and Thompson, 1987).

6 The multifaceted aetiology of fears about specific illness

The literature pertaining to specific illness fears has endorsed a picture of multiple factors (Bianche, 1971; Ryle, 1948; Straker, 1951). Four main aetiological features have been repeatedly elucidated in the literature. Firstly, individuals upbringing and developmental factors have been implicated, such as “physical disease” orientated environments (Marks, 1987). It seems plausible that individuals prone to focus upon bodily sensation and to attend to somatic symptoms will be precipitated from environments in which such a focus is the norm (Zborowski, 1952). Lipowski (1969) has highlighted the influence of early conditioning in the interpretation of bodily symptoms as threats. Secondly, guilt has been discussed as a factor that may exacerbate illness phobia (Marks, 1987). Using an adaptation of Scanbergers (1959) Guilt Questionnaire, Bianchi (1971) found disease phobics felt significantly more guilty than matched controls. Thirdly, health history of the self and significant others, for example of 31 “cancerophobics” studied by Ryle (1948), 12 had lost a close friend or relative to the disease, or had intimate acquaintance with such a case. The notion of identification with the illness has also been considered in cases of fears about HIV and AIDS, in which members of high risk groups are all too aware that partners and friends are seropositive or becoming ill and dying from AIDS. Finally, awareness of the disorder through publicity campaigns intended to assist in the dissemination of health promoting behaviour. Illness phobics tend to focus their worries, to some extent upon the current “fashionable” illness within the community at large, or within the subculture to which they adhere (Marks, 1987), or are in synchrony with the medical preoccupations of the age (Bianchi, 1971). This endorses previous authors discussions, for example Laughin (1956) commented upon the impact of cultural influences on the “popularity” of certain phobias. Such phenomena might be illustrated by the increase in presentation of concerns about tuberculosis following publicity campaigns (Pope, 1911), description of nosophobia as the fear of apoplexy in the 1890’s (Tuke, 1892), and in elevated cases of venerophobia in populations with high incidence of gonorrhoea (German and Arya, 1969).

7 Factors implicated in the aetiology of fears about HIV and AIDS

It is extremely difficult to construct a stable picture of how such factors may interact in the construction of fears about HIV and AIDS. Case studies of the WW, and others with varying degrees of concern about HIV and AIDS, have indicated wide ranging cultural, personal and sexual histories in those presenting with concerns (however, whether published case histories represent a representative sample of the WW population, if such an entity exists, is questionable). Thus we might turn to previous populations presenting with concerns about a specific illness in order to ascertain an appropriate course of research in this area.

8 Parallels between the worried well and syphilophobics

The closest analogy to the WW has been drawn with "syphiliphobics" (those presenting with concerns about syphilis)(Knapp and Vandecreek, 1989; Vuorio et al, 1990). Syphilophobia was first reported in medical literature in 1586 (cited in MacAlpine, 1957), and became a common complaint during the 18th and 19th century (Baur, 1988). The similarities in presentation between those labelled syphilophobic and those considered WW are manifold. Parallels between syphilis and HIV/AIDS in terms of modes of transmission and characterization by stages, including a long latency period, and a final stage of physical and mental deterioration and ultimately death can be drawn. Additionally, similarities between the conceptualization of these illnesses within the cultures in which they arose might be noted.

9 The cultural values, illness and the media

Like cholera in the 1830's, syphilis and HIV have been associated with so called "immoral" lifestyles and resultant blaming of the victim (Dworkin and Pincu, 1993; Herek, 1990). Muir (1991) highlights the image of HIV and AIDS associated with immorality, illegality, infidelity and illness. Fears of contagion are heightened by "plague" metaphors used (Sontag, 1990), and like syphilis, AIDS has become the contemporary metaphor for corruption, decay, and malignant destructive consummate evil (Enlow, 1984). This powerful characterization, created by media interpretation and public attention within a framework of the dominant religious and cultural values of the era, might exacerbate concerns about illness. HIV and AIDS have received media coverage of an unparalleled intensity in the history of disease (Davey and Green, 1991). Whilst health promotion techniques have undoubtably had a positive impact in terms of curbed transmission rates for HIV, the impact of these campaigns on those individuals predisposed, or susceptible to, excessive concerns about illness has been neglected in research. Increases in presentation with concerns about HIV and AIDS following high impact campaigns is not surprising. However as findings indicate that there has been little increase in those testing HIV positive (e.g., Beck et al., 1990) might indicate that media campaigns may have played some role in precipitating irrational fears, and that the research area would benefit from some empirical focus into this possibility.

10 Conclusion

Our understanding of fears about illness in general, and HIV and AIDS in particular is far from complete. It would seem imperative that a widely accepted theoretical structure for such concerns is outlined, in order to facilitate powerful empirical research. Considerable research is needed in order to ascertain the likely aetiologic candidates for these potentially distressing and disabling fears.

References

American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders,* 4th ed., Washington DC.

Baur, S. (1988). *Hypochondria: Woeful imaginings*. University of California Press, Berkeley.

Beck, E. J., Donegan, C., Cohen, C. S., Moss, V., & Terry, P. (1990). An update on HIV testing in a London STD clinic: Longterm impact of the AIDS media campaigns. *Genito-Urinary Medicine, 66*, 142-147.

Bianchi, G. N. (1971). The origins of disease phobia. *Australian and New Zealand Journal of Psychiatry, 5*, 241- 257.

Davey, T. & Green, J. (1991). The worried well: ten years of a new face for an old problem. *AIDS Care, 3*, 289-293.

Dilley, J. W. (1988). Psychiatric sequalae of HIV. In Paine, L. (Ed.),*AIDS: Psychiatric and psychosocial perspectives*. Crook Helm, London.

Dworkin, S. H. & Pincu, L. (1992). Counselling in the era of AIDS. *Journal of Counselling and Development, 71*, 257-281.

Faulstich, M. E. (1987). Psychiatric aspects of AIDS. *American Journal of Psychiatry, 144*, 551-556.

Fornstein, M. (1984). AIDS anxiety and the worried well. In Nichols, S. E. & Ostrow, D. G. (Eds.), *Acquired Immune Deficiency Syndrome*. American Psychiatric Press, Washington DC.

German, G. A. & Arya, O. P. (1969). Psychiatric morbidity amongst a Uganda student population. *British Journal of Psychiatry, 115*, 1323-1326.

Goldmeir, D. (1987). Psychosocial aspects of AIDS. *British Journal of Hospital Medicine*, 232-240.

Harowski, K. J. (1987). The worried well: Maximising coping in the face of AIDS. *Journal of Homosexuality, 14*, 299-306.

Herek, G. M. (1990). Illness, stigma and AIDS. In Costa, P. T. & Vanden Bos, G. R. (Eds.), *Psychological aspects of serious illness: Chronic conditions, fatal diseases and clinical care, Master Lectures*. APA Press, Washington DC.

Jacob, K. S., John, J. K., Verghese, A. & John, T. J. (1987). AIDS-phobia. *British Journal of Psychiatry, 150*, 412.

Jacob, K. S., John, J. K., Verghese, A. & John, T. J. (1989). The fear of AIDS: Psychiatric symptom or syndrome. *AIDS Care, 1,* 35-38.

King, M. B. (1989). Psychological and social problems in HIV infection: Interviews with general practitioners in London. *British Medical Journal, 299,* 713-717.

King, M. B. (1993). *AIDS, HIV and mental health*. Cambridge University Press, Cambridge.

Knapp, S. & Vandecreek, L. (1989). Fear of AIDS: Its meaning and implications for clinical practice. *Journal of Contemporary Psychotherapy, 19,* 239-247.

Laughlin, H. P. (1956). *The neuroses in clinical practice*. Saunders, London.

Lipkin, B. (1988). Pseudo AIDS, AIDS panic or AIDS phobia. *British Journal of Psychiatry, 152*, 425.

Lipowski, Z. J. (1969). Psychosocial aspects of disease. *Annals of Internal Medicine, 71,* 1197-1202.

MacAlpine, I. (1957). Syphilophobia: A psychiatric study. *British Journal of Venereal Disease, 33*, 92-99.

Marks, I. M. (1987). *Fears, phobias and rituals*. Oxford University Press, New York.

Miller, D., Acton, T. M. J. & Hedge, B. (1988). The worried well: Their identification and management. *Journal of the Royal College of Physicians of London, 22*, 158-165.

Miller, D., Green, J., Farmer, R. & Carrol, G. (1985). A pseudo AIDS syndrome following from fear of AIDS. *British Journal of Psychiatry, 146*, 550-551.

Muir, M. A. (1991). *The environmental contexts of AIDS*. Praeger, New York.

O'Brien, G. & Hassenyeh, F. (1985). AIDS panic: AIDS induced psychogenic states. *British Journal of Psychiatry, 146*, 91.

O'Brien, L. S. (1987). Not a case of pseudo AIDS. *British Journal of Psychiatry, 151,* 127.

Pope, C. (1911). A note on tubercular phobia. *The medical fortnightly, 39*, 205.

Riccio, M. & Thompson, C. (1987). Pseudo AIDS, AIDS panic or AIDS phobia? *British Journal of Psychiatry, 151*, 863.

Ryle, J. A. (1948). The 21st Maudsley lecture: Nosophobia. *Journal of Mental Science, 94*, 1-17.

Salkovskis, P. M. & Warick, H. M. C. (1986). Morbid preoccupations, health anxiety and reassurance: A cognitive behavioural approach to hypochondriasis. *Behavioural Research and Therapy, 24*, 597-602.

Schanberger, W. J. (1959). A factorial investigation of some theoretical distinctions between anxiety and guilt feelings. *Studies in Psychology and Psychiatry, Catholic University of America, 10*, 1-14.

Sontag, S. (1990). *AIDS and its metaphors*. Penguin, London.

Straker, M. (1951). Sickness fears: A manifestation of anxiety. Treatment Service Bulletin, 6, 197-199.

Tuke, D. H. (1892). *A dictionary of psychological medicine*. Churchill, London.

Vuorio, K. A., Aarela, E. & Lehtinen, V. (1990). Eight cases of patients with unfounded fear of AIDS. *International Journal of Psychiatry in Medicine, 20*, 405-411.

Warick, H. M. C. (1989). A cognitive behavioural approach to hypochondriasis and health anxiety.*Journal of Psychosomatic Research, 33*, 705-711.

Warick, H. M. C. & Salkovskis, P. M. (1990). Hypochondriasis.*Behaviour Research and Therapy, 28*, 105-117.

Windgasses, E. & Soni, S. (1987). AIDS panic.*British Journal of Psychiatry, 150*, 126-127.

Zubrowski, M. (1952). Cultural components in response to pain. *Journal of Social Issues, 8*, 16-30.

In: Noble, J. & Parsowith, S. R. (Eds.). (1996). The Ninth White House Papers: Graduate Research in the Cognitive and Computing Sciences at Sussex. Cognitive Science Research Paper 440, School of Cognitive and Computing Sciences, University of Sussex.

Some False Starts in the Construction of a Research Methodology for Artificial Life.

Ezequiel A. Di Paolo
ezequiel@cogs.susx.ac.uk

School of Cognitive & Computing Sciences
University of Sussex
Brighton
BN1 9QH

Abstract This article briefly reviews some guidelines for building a research methodology in Artificial Life given by Miller (1995). A formal argument is presented to point at some problems arising from the systematic application of these guidelines given the current state of affairs in theoretical biology, and some practical arguments are proposed against the downsizing strategy adopted by Miller.

1 Introduction.

Take any recent collection of papers on simulation of adaptive behavior or artificial life (see Brooks & Maes, 1994) and you will find a lot of very imaginative, inspired and often methodologically messy pieces of work. Some of them address legitimate scientific questions, others seem to be more like proofs of concept, others still, just nicely implemented simulations. This problem has already been pointed out by Miller (1995) and some of its causes (mainly from a "sociological" perspective) have been analysed by him.[1]

Miller correctly points out that many of the problems encountered in the way current ALife research is done are attributable to the "computer science influence" in the field. Poor scholarship, poor follow-through of research and poor communication with researchers in other fields working on related issues are some of the problems he mentions. Miller also identifies six methodological heuristics for "doing A-Life as good theoretical biology." All of them, except one, result from a (healthy) exercise in the application of scientific common-sense: identify a valid problem that may benefit from explorations with computer techniques, survey the background material on that problem, contact/collaborate with people already working on similar problems, publish your results in a media recognized and respected by those people, take advantage of the flexibility and power of computer simulations and perform systematic explorations into the dynamics of the problem. While it is good to state these guidelines explicitly, there is not much novelty in them for researchers with a sufficient capacity for self-criticism.

What I consider Miller's core suggestion, from which a research methodology could begin to take form, is the following: "*Develop a well-targeted simulation that extends current biological models and yields directly comparable results.*" This is a very interesting proposition because is not as obvious as the others and, in my view, it deserves some critical comments.

Here I propose to defend the position that this suggestion may indeed be practised with good results but that it can *not* be placed at the core of a methodology for ALife, whether we see ALife as a collection of novel techniques or as a scientific discipline in its own right.

[1] For a detailed methodological analysis on a specific piece of work see Noble and Cliff (1996).

2 ALife as a tool for theoretical biology.

Complex computer simulations do not define a new science by themselves, in any case, they are a technique which may provide new ways of doing an existing science. In the case of ALife, Miller assumes that this science is theoretical biology and I will not argue with that for the moment.

As with any new technique there is always the hope of resolving old controversies over issues that have long remained the battleground of academic disputes (think for example of Zeno's Paradox and the invention of the calculus of infinite series). This may or may not happen in the case of ALife, but the chances of it happening will remain small unless an appropriate methodology is defined for this task. The starting point for this definition must be an understanding of the general character of the sciences in question (which, as we will see, will not necessarily be complete by achieving an understanding of the way these sciences are actually practised in our particular time.)

According to Miller:

> A powerful way of using A-Life simulations is to take an existing formal model from theoretical biology and relax the assumptions (preferably one at a time) that were required to make the mathematics tractable. (Miller, 1995)[2]

There is little doubt that such a method will tend to enrich current models in theoretical biology with new answers too hard (or even impossible) to obtain analytically. However, more often than not, this same method will tend to charge the ALife work with many of the methodological and philosophical assumptions of those same theoretical models. This would not be much of a problem in principle, except for the fact that most analytical models in theoretical biology, particularly those dealing with problems in evolution, are representatives of an extreme (yet mainstream) position in the spectrum of biological research and it is rather difficult to see how many of the philosophical presuppositions of these models can be "relaxed" in practice. In fact, these presuppositions are not viewed by most researchers as assumptions, but they are rather "accepted as such". This is a fact that is not confined to biology; it can be identified in any science, including mathematics. According to Stolzenberg, "Despite a certain superficial similarity, there is all the difference in the world between (1) *accepting* something as being what it appears to be and proceeding (in life, as an experiencing being) on that basis; and (2) merely exploring the consequences of *the assumption* that some thing is what it appears to be" (Stolzenberg, 1984).

It is necessary to stress that the problem does not arise directly from Miller's recommendations, but it originates in the current status of polarized positions in biology. So, it is not the quality of Miller's methodological solution that I am questioning in this section, but rather its appropriateness given the current state of affairs.

My purpose is to elaborate on the question of why a direct association with current formal models in theoretical biology may not be such a good idea (even if only as a "starting point") for building a methodology, and why some of the true potential of computer simulation techniques may be wasted in this way. The latter question will be addressed in the next section. The argument for the first part is rather simple and involves only the recognition of two facts and the acceptance of a conjecture:

Fact 1: There are important controversies in current biological research especially (but not only) in evolutionary theory. Compare Dawkins (1982), Gould (1989), Goodwin (1994), Gould and Lewontin (1979), Kauffman (1993), Lewontin (1983, 1984), to cite a few representative examples. These debates can be broadly characterized as arising from different views on the role of natural selection in evolution.

Fact 2: Current formal models in theoretical biology in their vast majority concentrate around only one position in these controversies (game-theoretic, statistical and economical models, population

[2] Note that this *faces* the issue of building a good simulation but *avoids* the issue of building a good model, a task that is entrusted to whoever developed the formal model in the first place. More on this on the next section.

genetics, etc.: concepts such as fitness and adaptation are never questioned, agent and environment are separate things, the latter being of a much more static nature.)

Conjecture: ALife and computer simulation techniques in general, if properly applied, may have the potential for resolving scientifically at least some of the controversial issues. Computer explorations could go right to the root of the debate, showing, for instance, the (im)plausibility of other mechanisms besides natural selection having an important role in evolution, or describing ontogenetic processes as a result of agent-environment co-determination,[3] thus helping to provide links between developmental and evolutionary biology.

Conclusion: From here it follows that an association with formal models in theoretical biology, and in general with *any* particular position in the controversy, will, to say the least, reduce drastically ALife's chances of fulfilling its conjectured potential. This is mainly because it is not possible to refute or modify the assumptions that have tacitly permeated into the ALife model and have been "accepted as such" due only to methodological reasons.

This conclusion is true even when the method of association itself involves the relaxation of the original model's assumptions (and, in a way, a departure from that model). Some of the (often hidden) assumptions of the model will inevitably permeate any simulation experiment based on that model; at the very least the need to answer the same questions involving the same theoretical constructs the model is supposed to be aimed at will remain. This seems trivial but we realize it is not when we consider that, from a different theoretical perspective, those same questions may have never been asked or those theoretical constructs may not even exist.

It is important to stress that this argument does not rely on any preference as to which position in the controversies we should adopt. For all we know, this same argument could be applied to any scientific discipline in a similar situation. Researchers adhering to different positions should therefore find the argument equally compelling (of course extreme representatives of some positions will not even recognize the existence of a controversy, but that is another matter). On the other hand some researchers may find Miller's guidelines very useful particularly if their lines of research parallel similar existing lines in theoretical biology. As I said, this is not an argument against using this methodological approach in a particular piece of work. This methodology may fit perfectly well in individual cases, but my claim is that we have to be careful not to use it as a milestone for ALife's methodology in general.

It can be said that this argument is contingent on a specific state of affairs and that there is nothing wrong with Miller's proposition in principle. At first glance this is true, but, since we are discussing methodological guidelines for scientific research, it is not acceptable that these guidelines should strongly depend on the current state of affairs. A methodology should be robust enough not to be modified by each single new result or new interpretation of empirical facts. The very fact that the suggested guidelines are weakened by a question of contingency makes, paradoxically, the argument to be one of principle.

The practical reason behind the argument is to point to the fact that given that a putative new research tool may provide new feasible avenues of research within a scientific discipline and therefore change the way that particular science is practised, even permitting researchers to address questions they could not hope to answer (or even ask) before, then the methodology for this new tool should not be strongly constrained *a priori* by the way that science is *actually* practised. This is because the praxis depends fundamentally, although not always in an obvious or recognizable way, on the existing tools and methods of research. It requires a conceptual effort to try to abstract from the praxis and from the constraints imposed by it (e.g., which questions are not even asked because we can't use any of our tools to try to answer them, or what assumptions are adopted only on a pragmatic basis) and try to picture which questions would be addressed if the identifiable constraints had been removed and the new tool were to be incorporated. This picture could be flawed but the Wright brothers would not have built their flying machine if they did not try to think beyond bicycles.

[3] See Oyama (1985), Levins and Lewontin (1985), Varela, Thompson, and Rosch (1991).

3 Other arguments

The preceding argument should suffice to settle the question that Miller's guidelines cannot be used to define a research methodology for ALife. However some people may find this argument too formal, addressing an issue that is too distant and of little importance to them in practice, and argue that following these guidelines actually helps them to do better research. This is a fair observation and I repeat that I agree with people taking advantage of such a simple and clear methodology, as long as they remain aware of its consequences and limitations.

There are, however, much more proximate issues, that suggest a loss of potential for ALife associated with the systematic application of Miller's recommendations. For reasons of space I will only sketch roughly some of them.

Verbal models: When I first approached the question of how to use computer simulations to address scientific problems I realized the enormous potential that arises from the computer modelling of verbal arguments. These kinds of arguments abound especially in biology and the reason they remain verbal is that they usually involve statements about quite complex processes very hard to formalize mathematically. But this does not mean that a computer model cannot be built. Some examples of ALife-style work on these lines include a quite large collection of papers on the interactions between ontogenetic plasticity and phylogeny (Baldwin effect), the illustration of properties like self-repair, self-maintenance and autopoiesis in ensembles of formal-chemical organizations (Fontana, Wagner, & Buss, 1994; Fontana & Buss, 1996), the demonstration with very simple evolutionary models showing self-organized criticality of the viability of punctuation patterns in the time distribution of species extinction (Bak & Paczuski, 1995), and simple models of parapatric speciation (Di Paolo, 1996). In these examples, coming from researchers working in different disciplines, knowledge about the plausibility of a verbal argument is gained with a simple computational model that reproduces a situation similar to the one exposed in the argument.

Direct modelling: Unfortunately, doing computer explorations in areas where no formal model exists implies constructing our own model directly. The charm of Miller's guidelines is that this step is avoided. But if we must face it, then it is necessary to answer many questions. These are precisely the questions that a good research methodology should address. How does this model reflect the actual phenomena? What can be concluded from explorative incursions into the modelled dynamics? Fortunately this is not the first time in history that these or similar questions arise. Analogous problems can be identified in the use of mathematical models in science. Some of these issues are addressed by Jackson (1996).

Examples of classic direct computational modelling in physics include the work of Fermi, Pasta, and Ulam (1965) and Lorenz (1963) among others. Some physicist are keen on the idea that dynamic phenomena in nature can be better described with computational rather than mathematical models (Feynman, 1965; Fredkin, 1990).

It is to be expected that the true potential of computer simulation techniques will lie precisely in those areas where little can be gained from the use of existing formal models, and I must recognize that this is in accordance with the spirit expressed in Miller's article, but I also think that many problems with existing formalizations could benefit from *radical* computer modelling as well, which does not involve the mere relaxation of assumptions in a formal model, but the direct questioning of presuppositions "accepted as such" in these models.

Why theoretical biology?: If ALife-style computer explorations need not rely on existing formal models provided by theoretical biology, what is it that constraints ALife to this particular discipline? As soon as the ALife researcher needs to build her own models into her simulations a whole new spectrum of scientific problems is opened which may undergo the same treatment. From self-organizing historical processes in general to the dynamics of social interactions and economic structures, there is a vast space of open problems that can benefit from ALife's approach. Many researchers are also interested in cognitive science problems (especially as applied to the design of control architectures for robots), which, while strictly a biological matter, are hardly addressed by people working in theoretical biology.

4 Conclusions: looking for a starting point.

Miller addresses the question of ALife's methodology by progressing in an ever more restrictive approach, first specifying the role of ALife only as a tool of research, then restricting its use to problems in theoretical biology and finally to those problems in which an existing formal model can be found. With each step the question of methodology seems to become less fuzzy and easier to solve. I have shown, however, that even if one agrees with all this there are still problems in the universal applications of his guidelines.

My opinion is that the question of methodology must not be resolved by restricting it to a tractable size. While I believe that work in ALife can be used successfully as a tool for extending formal models in theoretical biology. I also believe that it can be used to do research in areas where no formal model exist and also "belonging" to other scientific disciplines.

Taking all this into consideration we must then consider the question of whether ALife is best treated as a tool or as a (potential) scientific discipline in its own right. This is a very difficult question with no straight answer.[4] My opinion is that we will "make" the answer in the following years, but I don't see what benefit we will get by *defining* it in one way or another from a methodological perspective.

So, where should we look for a methodology for doing science with ALife? The first consideration that we must make is that a methodology is not something that is discovered and expressed in golden rules; it is rather something that is constructed through a history of failures and successes in scientific research. If we want to make this methodology explicit, then the first thing we must do is to understand the character of the scientific explanations that are used in current scientific research in biology and cognitive science, to start with, and try to envision what kind of scientific explanations ALife's work would provide and how they would compare to current types of explanations in those sciences. By performing this necessary step we can try to answer whether ALife models, besides being used as new sources of information, can also be used as new sources of *understanding*.

5 Acknowledgements

This work has benefited from Phil Husband's and Inman Harvey's valuable comments. The opinions expressed here remain my own. The author is grateful to the *Consejo de Investigaciones Científicas y Técnicas de la República Argentina* and the Argentine Ministry of Education for their support.

References

Bak, P., & Paczuski, M. (1995). Complexity, contingency and criticality. *Proceedings of the National Academy of Sciences, USA*, *92*, 6689 – 6696.

Brooks, R., & Maes, P. (1994). *Artificial Life IV.* The MIT Press, Cambridge, Mass.

Dawkins, R. (1982). *The Extended Phenotype.* Oxford University Press.

Di Paolo, E. A. (1996). A computational model of speciation in non-uniform environments without physical barriers. Cognitive science research paper 412, School of Cognitive and Computing Sciences, University of Sussex.

Fermi, E., Pasta, J., & Ulam, S. M. (1965). Introduction to studies of non-linear problems. In *Collected Papers of Enrico Fermi, V. 2.* Univ. of Chicago Press.

Feynman, R. P. (1965). *The Character of Physical Law.* MIT Press, Cambridge, Mass.

Fontana, W., & Buss, L. (1996). The barrier of objects: from dynamical systems to bounded organizations. In Casti, J., & Karlqvist, A. (Eds.), *Boundaries and Barriers*, pp. 56 – 116. Addison-Wesley.

[4] Is mathematics a tool?

Fontana, W., Wagner, G., & Buss, L. W. (1994). Beyond digital naturalism. *Artificial Life*, pp. 211 – 227.

Fredkin, E. (1990). Digital mechanics. *Physica D*, *45*, 254 – 270.

Goodwin, B. C. (1994). *How the leopard changed its spots: the evolution of complexity*. Weidenfeld and Nicolson, London.

Gould, S. J. (1989). *Wonderful Life*. Hutchinson Radius.

Gould, S. J., & Lewontin, R. C. (1979). The spandrels of san marco and the panglossian paradigm: a critique of the adaptationist programme. *Proceedings of the Royal Society London*, *B 205*, 581 – 598.

Jackson, E. A. (1996). The second metamorphosis of science: a second view. Working paper 96-06-039, Santa Fe Institute.

Kauffman, S. A. (1993). *The Origins of order*. O. U. P. New York.

Levins, R., & Lewontin, R. C. (1985). *The Dialectical Biologist*. Harvard University Press.

Lewontin, R. C. (1983). The organism as the subject and object of evolution. *Scientia*, *118*, 63 – 82.

Lewontin, R. C. (1984). Adaptation. In Sober, E. (Ed.), *Conceptual issues in evolutionary biology: an anthology*. The MIT Press, Cambridge Mass.

Lorenz, E. N. (1963). Deterministic non-periodic flow. *Journal of Atmospheric Science*, *20*, 130 – 141.

Miller, G. E. (1995). Alife as theoretical biology: how to do science with computer simulation. Cognitive science research paper 378, School of Cognitive and Computing Sciences, University of Sussex.

Noble, J., & Cliff, D. (1996). On simulating the evolution of communication. Cognitive science research paper 420, School of Cognitive and Computing Sciences, University of Sussex.

Oyama, S. (1985). *The Ontogeny of Information*. Combridge University Press.

Stolzenberg, G. (1984). Can an inquiry into the foundations of mathematics tell us anything about mind?. In Watzlawick, P. (Ed.), *The Invented Reality*, pp. 257–308. W.W. Norton.

Varela, F. J., Thompson, E., & Rosch, E. (1991). *The Embodied Mind: Cognitive Science and Human Experience*. The MIT Press.

In: Noble, J. & Parsowith, S. R. (Eds.). (1996). The Ninth White House Papers: Graduate Research in the Cognitive and Computing Sciences at Sussex. Cognitive Science Research Paper 440, School of Cognitive and Computing Sciences, University of Sussex.

An Exemplar-Based Recognition and Recall System Using an Interpretation Process

Robert Ellis
roberte@cogs.susx.ac.uk

School of Cognitive & Computing Sciences
University of Sussex
Brighton
BN1 9QH

Abstract The research focuses on the act of interpretation as being central to memory function. Interpretation of input information is defined as the process of associating prior memory to alter input information, which in turn associates with prior memory. An exemplar-based process model of memory, called TeSME, is being developed which implements this cycle of association in order to create recognition and recall effects. Parallels are drawn with previous exemplar-based models of memory and AI learning systems to show how TeSME and the interpretation process are a logical progression of these systems.

I have been trying to modify and carry over computer-based psychological models of memory into more applicable AI memory-based systems. I have concentrated on what I would call the act of interpretation as being central to memory function, and thus of potential benefit to AI.

I got the idea that the act of recognition (or a scene's interpretation) was key to expert decision-making from decision theory research. I thought that perhaps this interpretation was a basic property of memory, a basic process one step above memory association. No AI machine learning system seemed to me to encapsulate this concept in a cognitively plausible fashion, so I looked to computer models of memory from the psychological literature.

I have a computer program whose design is loosely based on Hintzman's (1986) inter-trace resonance version of the Minerva-2 memory model. It also shares some of the properties of the later versions of the Context Model (GCM-ISW; Aha & Goldstone, 1990; Aha, Kibler & Albert, 1991). These computer programs simulate the processes of recognition and recall in human memory. Both store input exemplars intact, and recall loosely involves finding the exemplar which best matches the cue (association). Inter-trace Minerva-2 and GCM-ISW differ from earlier exemplar models by effectively altering the input cue during the recognition/recall process, altering the result of the association in ways which mirror experimental observations.

Hintzman's (1986) Minerva-2 model is an exemplar (or instance) based system which stores exemplars as separate vectors of features. During recall, a cue vector is presented, in which the response features are empty. Each stored exemplar (containing both cue and response features) contributes to a single response vector, the contribution of each depending on its similarity to the input vector. If they are all fairly similar the result is a jumble of response features which don't resemble anything. Normally, some are more similar than others and the response vector will have its response feature section filled appropriately. However, this will often contain some noise and Hintzman noted that by using the response vector as a second input cue, the second response would be a lot cleaner, because it would resemble the appropriate stored exemplars more closely and the noise generating exemplars would get ruled out. This modification worked reasonably well and was named "inter-trace resonance". Hintzman then discovered that this process would allow Minerva-2 to model other experiments, including particular context effects which the earlier system (and perhaps GCM-ISW) could not.

From my perspective, Minerva-2 takes the initial input vector, 'interprets' it to produce an initial response vector, then uses the interpreted input as a basis for the desired memory recall effects. Unfortunately, Minerva-2's vector representation is rather limiting, with each feature being either present, don't know or not present. In contrast, GCM-ISW (Aha & Goldstone, 1990) has a measure of activation strength or salience of particular features. A newly input exemplar's feature activations can be adjusted through comparison with pre-stored exemplars (e.g., reducing the strength of features which do not match with any exemplar). This in turn affects the closeness of match with particular exemplars. Aha and Goldstone (1990) note that this version of the GCM model can perform categorization better than models which do not assume that each input exemplar has a potentially different feature salience, determinable by memory response. From my perspective, GCM-ISW's memory interprets newly input exemplars by modifying these weights, deriving the required behaviour.

Heit (1992) used the context model to model the effects that chains of association have on memory recall. The initial response from a cue was cycled back as a second cue, deriving a further response which could link dissimilar cue and response exemplars together. Heit then showed that there was evidence for this chaining in recall experiments.

In effect, my TeSME system combines GCM-ISW's ability to adjust the salience of features, with Minerva-2's ability to add new features to the input. When a new exemplar is input to TeSME, the interpretation process iteratively compares it with pre-stored exemplars, which alters the content of the input, which in turn alters the result of each comparison. After a set number of cycles, the input is said to have been interpreted and its final contents can be used to calculate recognition and recall information. The interpreted input can then be placed, as a new exemplar in memory, ready to effect the contents of future inputs.

TeSME is a logical progression of both Minerva-2 and GCM-ISW, but the structure of these earlier systems prevented their extension in this direction. I think the purpose of TeSME should be to highlight the possibility that such an interpretation process is a basic, but multi-faceted, function of human memory.

In building the TeSME system, I have been exploring its properties, some of which were not originally foreseen. In summary:

- Recognition and recall measures can be derived from the product of the interpretation process.
- Learning is the addition of interpreted input exemplars to memory.
- Forgetting is merely the removal of the oldest exemplars in memory—if they were useful, their 'effects' live on in later stored interpreted exemplars.
- The addition of new exemplars to the input may present a hypothetical attention process with something to do, i.e., "is this feature really there?"
- The iterative nature of the interpretation process means that stored exemplars compete against each other to influence the input. For example, adding a new feature means that another competing stored exemplar is no longer similar. Context effects and chains of association between stored exemplars can occur.

References

Aha, D. W., & Goldstone, R. L. (1990). Learning attribute relevance in context in instance-based learning algorithms. *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, pp. 141–148. Lawrence Erlbaum Associates, Hillsdale, NJ.

Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning, 6*, 37-66.

Heit E. (1992). Categorisation using chains of examples. *Cognitive Psychology, 24*, 341–380.

Hintzman, D. L. (1986). “Schema abstraction” in a multiple-trace memory model. *Psychological Review, 93*, 411–428.

In: Noble, J. & Parsowith, S. R. (Eds.). (1996). The Ninth White House Papers: Graduate
Research in the Cognitive and Computing Sciences at Sussex. Cognitive Science Research
Paper 440, School of Cognitive and Computing Sciences, University of Sussex.

Face Recognition Using Radial Basis Function Neural Networks

A. Jonathan Howell and Hilary Buxton
{jonh,hilaryb}@cogs.susx.ac.uk

School of Cognitive & Computing Sciences
University of Sussex
Brighton
BN1 9QH

Abstract This paper presents experiments using an adaptive learning component based on Radial Basis Function (RBF) networks to tackle the unconstrained face recognition problem using low resolution video information. Firstly, we performed preprocessing of face images to mimic the effects of receptive field functions found at various stages of the human vision system. These were then used as input representations to RBF networks that learnt to classify and generalize over different views for a standard face recognition task. Two main types of preprocessing (Difference of Gaussian filtering and Gabor wavelet analysis) are compared. Secondly we provide an alternative, 'face unit' RBF network model that is suitable for large-scale implementations by decomposition of the network, which avoids the unmanagability of neural networks above a certain size. Finally, we show the 2-D shift, scale and y-axis rotation invariance properties of the standard RBF network. Quantitative and qualitative differences in these schemes are described and conclusions drawn about the best approach for real applications to address the face recognition problem using low resolution images.

1 Introduction

The human face poses several severe tests for any visual system: the high degree of similarity between different faces, the extent to which expressions and hair can alter the face, and the large number of angles from which a face can be viewed in common situations. A face recognition system must be robust with respect to this variability and generalize over a wide range of conditions to capture the essential similarities for a given human face. It is only recently that work on biologically-motivated, statistical approaches to face recognition has begun to deliver real solutions. One of the main problems that these approaches tackle is dimensionality reduction to remove much of the redundant information in the original images. There are many possibilities for such representations of the data, including principal component analysis, Gabor filters and various isodensity map or feature extraction schemes. A well known example is the work of Turk & Pentland (Turk & Pentland, 1991), on the 'eigenface' approach, which is widely acknowledged to be useful for practical application. However, the need for representations at a range of scales and orientations causes extra complexity and updating the average eigenface (used for localization) when new faces are added to the dataset are problems for this scheme. These difficulties have been overcome to some extent in later work by various researchers (Pentland, Moghaddam, & Starner, 1994; Petkov, Kruizinga, & Lourens, 1993; Rao & Ballard, 1995). In particular, it seems that appropriate preprocessing of input representations for a face recognition scheme can overcome the problems of lighting variation and multiple scales. Other sources of variation such as face orientation, expression, occlusion etc. still remain.

In our work we use an adaptive learning component based on RBF networks to tackle the unconstrained face recognition problem. We want our face recognition scheme to generalize over a wide range of conditions to capture the essential similarities of a given face. The RBF network has been identified as valuable model by a wide range of researchers (Moody & Darken, 1988; Poggio & Girosi, 1990b;

Girosi, 1992; Musavi, Ahmad, Chan, Faris, & Hummels, 1992; Ahmad & Tresp, 1993; Bishop, 1995). Its main characteristics are first, its computational simplicity (only one layer involved in supervised training which gives fast convergence), and second, its description by a well-developed mathematical theory (resulting in statistical robustness). RBFs are seen as ideal for practical vision applications by (Girosi, 1992) as they are good at handling sparse, high-dimensional data (common in images), and because they use approximation which is better than interpolation for handling noisy, real-life data. RBF networks are claimed to be more accurate than those based on Back-Propagation (BP), and they provide a guaranteed, globally optimal solution via simple, linear optimization. An RBF interpolating classifier (Edelman, Reisfeld, & Yeshurun, 1992), was effective and gave performance error of only 5–9% on generalization under changes of orientation, scale and lighting. This compares favourably with other state of the art systems such as the Turk & Pentland scheme. In contrast to more deterministic methods using warping based on registration of features, eg (Craw, Costen, Kato, Robertson, & Akamatsu, 1995), our approach uses simpler preprocessing, but learns to discriminate using the RBF networks to overcome occlusion arising out of head rotation.

Cognitive studies of the way human faces are perceived (for example (Bruce, 1988)) can contribute to the design of systems that automate this kind of visual processing. There is support for having 'face recognition units' (FRUs) for recognising familiar faces (Bruce & Young, 1986; Bruce, 1988; Bruce, Burton, & Hancock, 1995). This idea is partly captured by the standard RBF techniques described next where the first layer of the network maps the inputs with a hidden unit devoted to each view of the face to be classified. The second layer is then trained to combine the views so that a single output unit corresponds to the individual person. We have taken this idea further and have developed a 'face unit' network model, which allows rapid network training and classification of examples of views of the person to be recognized. These face units give high performance and also alleviate the problem of adding new data to an existing trained network. We use the various views of the person to be recognized together with selected confusable views of other people as the negative evidence for the network. Our face units have just 2 outputs corresponding to 'yes' or 'no' decisions for the individual. This is in contrast with Edelman (Edelman et al., 1992) who did not use such negative evidence in their study. We show that this system organization allows flexible scaling up which could be exploited in real-life applications.

2 The RBF network model

The RBF network is a two-layer, hybrid learning network (Moody & Darken, 1988, 1989), with a supervised layer from the hidden to the output units, and an unsupervised layer, from the input to the hidden units, where individual radial Gaussian functions for each hidden unit simulate the effect of overlapping and locally tuned receptive fields. They use the vector norm distance, $|\mathbf{i} - \mathbf{c}|$, equivalent to $\sum_{x=1}^{N}(i_x - c_x)^2$, between the N-dimensional input vector $\mathbf{i}$ and hidden unit centre $\mathbf{c}$ (N being the number of input units). The output value can be seen to approach a maximum when $\mathbf{i}$ becomes most similar to $\mathbf{c}$. The input vectors are unit-normalized.

Each hidden unit has an associated σ (sigma) 'width' value which defines the nature and scope of the unit's receptive field response[1]. This gives an activation that is related to the relative proximity of the test data to the training data, allowing a direct measure of confidence in the output of the network for a particular pattern. In addition, if the pattern is more than slightly different to those trained, very low (or no) output will occur.

The output o for hidden unit h (for a pattern l) can be expressed as:

$$o_h(l) = \exp[-\frac{|\mathbf{i}(l) - \mathbf{c}_h|^2}{2\sigma_h^2}], \tag{1}$$

the hidden layer output being unit-normalized, as suggested by (Hertz, Krogh, & Palmer, 1991). For

[1] It is equivalent to the standard deviation of the width of the Gaussian response, so larger values allow more points to be included.

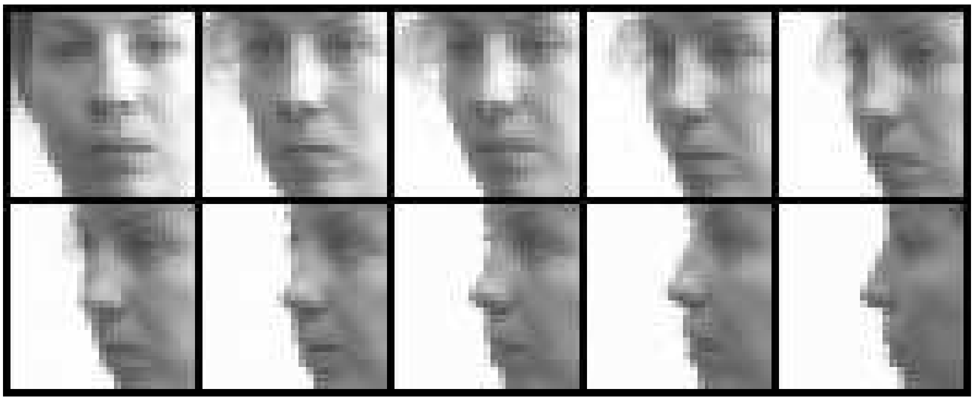


Figure 1: Entire 10-image range (rotating around the y-axis) for one person before preprocessing

output unit i, the output is:

$$o_i(l) = \sum_h w_{ih} o_h(l). \tag{2}$$

Whilst the weights w_{ih} can be adjusted using the Widrow-Hoff (Widrow & Hoff, 1960) delta learning rule, the single layer of linear output units permits a matrix pseudo-inverse method (Poggio & Girosi, 1990a) for their exact calculation. The latter approach allows almost instantaneous 'training' of the network, regardless of size[2]. The RBF network's success in approximating non-linear multidimensional functions is dependent on sufficient hidden units being used and the suitability of the centres' distribution over the input vector space (Chen, Cowan, & Grant, 1991).

2.1 'Face unit' RBF model

For the following tests, two types of network were used: a 'standard' RBF model and a 'face unit' RBF model. The standard network is trained with all possible classes from the data with a 'winner-takes-all' output strategy, whilst the 'face unit' network produces a positive signal only for the particular person it is trained to recognize. For each individual, a 'face unit' RBF network can be trained to discriminate between that person and others selected from the data set, using 'pro' and 'anti' evidence for and against the individual. Details can be found in (Howell & Buxton, 1995c). Although this second approach increases complexity, the splitting of the training for individual classes into separate networks gives a modular structure that can potentially support large numbers of classes, since network size and training times for the 'standard' model quickly become impractical as the number of classes increases.

3 Form of test data

Lighting and location for the training and test face images in these initial studies has been kept fairly constant to simplify the problem. For each individual to be classified, ten images of the head and shoulders were taken in ten different positions in 10° steps from face-on to profile of the left side (see Figure 1), 90° in all. This gave a data set of 100 8-bit grey-scale 384×287 images from ten individuals.

A 100×100-pixel 'window' was located manually in each image centred on the tip of the person's nose, so that visible features on profiles, for instance, should be in roughly similar locations to face-on. This 'window' region was sub-sampled to a variety of resolutions for testing. Full details are given in (Howell & Buxton, 1995a). The resolution of the images is represented as '$n \times n$', a resolution of 25×25 being used for the work reported here. The ratio of training and test images used is represented as 'train/test', eg '20/80', where 100 images were in the data set and 20 were used for training and 80 for test. The 'face unit' network size is denoted by '$p+a$', where p is the number of 'pro' hidden units, and

[2] A network of 250 hidden units and 10 outputs, *ie.*2500 parameters, which required several hours of Sparc 20 processing time for gradient descent can be computed in a small fraction of a second.

a is the number of ‘anti’ hidden units. Tests were made on a range of network sizes from 1+1 to 6+12 (which are effectively 2/98 and 18/82 networks).

3.1 Pre-processing methods

Although the RBF network was able to learn the dataset without preprocessing, *ie.*on pure grey-level values (Howell & Buxton, 1995b), the authors see preprocessing of the images as a valid and important intermediate step, highlighting relevant parts of the information, and adding an essential invariance to illumination (Marr & Hildreth, 1980).

Two main techniques are used for the preprocessing of the images: Difference of Gaussian (DoG) filtering and Gabor wavelet analysis at a range of scales. One way of thinking about these input representations and mapping them onto our RBF networks is to use the analogy with visual neurons. The receptive field of such a neuron is the area of the visual field (image) where the stimulus can influence its response. For the different classes of these neurons, a receptive field function $f(x,y)$ can be defined. For example, retinal ganglion cells and lateral geniculate cells early in the visual processing have receptive fields which can be implemented as Difference of Gaussian filters (Marr & Hildreth, 1980). Later, the receptive fields of the simple cells in the primary visual cortex are oriented and have characteristic spatial frequencies. Daugman (Daugman, 1988) proposed that these could be modelled as complex 2-D Gabor filters. Petkov et al (Petkov et al., 1993) successfully implemented a face recognition scheme based on Gabor wavelet input representations to imitate the human vision system. Our earlier studies (see (Howell & Buxton, 1995b)) showed that these later stages of processing make information more explicit for our face recognition task than the earlier DoG filters.

The experiments presented here concentrate on two specific applications of these techniques:

- DoG convolution with a scale factor of 0.4, with a reduced range of grey-levels. The sampled values were thresholded to give zero-crossings information. A 25×25 image gave 21×21 convolved values, *ie.*441 samples per image.
- Gabor ‘A3’ sampling (for details, see (Howell & Buxton, 1995b)), with a full range of grey-levels. Data was sampled at four non-overlapping scales from 8×8 to 1×1 and three orientations (0°, 120°, 240°) with sine and cosine components. A 25×25 image gave 510 coefficients per image.

4 Generalization over views (y-axis Rotation) by the RBF network

Fixed selections of images used for training to keep the experiments as constrained as possible. Table 1 shows both the standard and face unit RBF network models able to generalize very well over the different views with either the DoG or Gabor preprocessing method.

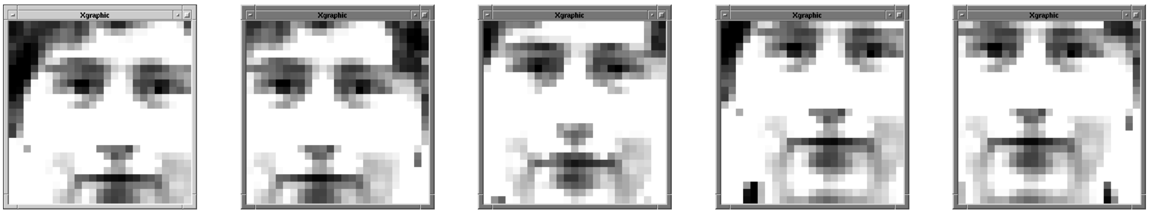
Network	Pre-processing	Initial %	% Discarded	% After Discard
Standard 50/50	DoG	88	28	100
	Gabor	94	30	100
6+12 Face Unit	DoG	92	35	95
	Gabor	95	25	100

Table 1: Effect of pre-processing methods on **original** dataset

5 Shift and scale invariance properties of the RBF network

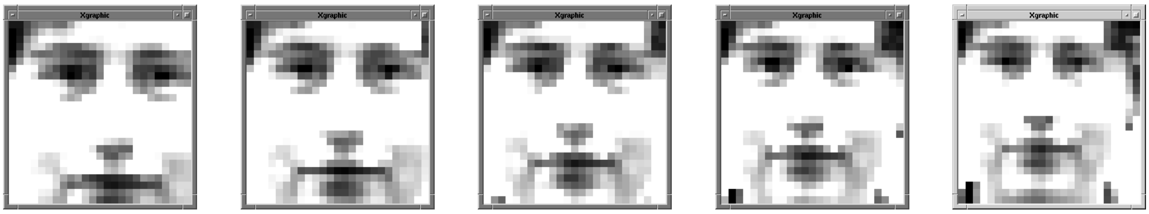
Two further data sets were created to test the RBF network’s generalization abilities:

- A shift-varying data set with five copies of each image: one at the standard sampling ‘window’ position, and four others at the corners of a box where all *x*,*y* positions were ±10 pixels from the centre (see Figure 2).



(a) (b) (c) (d) (e)

Figure 2: **Shift-varying** data for the ‘face on’ view of one individual: (a) top left (b) top right (c) normal view (d) bottom left (e) bottom right



(a) (b) (c) (d) (e)

Figure 3: **Scale-varying** data for the ‘face on’ view of one individual: (a) +25% (uses 111×111 window) (b) +12.5% (107×107) (c) normal view (100×100) (d) –12.5% (94×94) (e) –25% (87×87)

- A scale-varying data set with five copies of each image: one at the standard sampling ‘window’ size, and four re-scaled at ±12.5% and ±25% of its surface area, ranging from 87×87 to 111×111 (see Figure 3).

5.1 Inherent invariance - training with original images only

These experiments used only the original from each group of five for training, using all the varied ones (and the remainder of the original ones not used for training) for testing. This gives a measure of the intrinsic invariance of the network to shift and scale, *ie.*the invariance not developed during training by exposure to examples of how the data varies.

Network	Pre-processing	Initial %	% Discarded	% After Discard
Standard 100/400	DoG	14	84	21
	Gabor	35	82	47
10+20 Face Unit	DoG	51	30	51
	Gabor	57	38	52

Table 2: Effect of pre-processing methods on **shift-varying** dataset (the original from each group of five used for training)

Network	Pre-processing	Initial %	% Discarded	% After Discard
Standard 100/400	DoG	58	63	78
	Gabor	77	46	95
10+20 Face Unit	DoG	69	40	69
	Gabor	83	36	88

Table 3: Effect of pre-processing methods on **scale-varying** dataset (the original from each group of five used for training)

5.2 Learnt invariance - training with shift and scale varying images

These experiments again used a fixed selection of positions for training examples, using all five versions of each original image. This gives the network information about the shift and scale variance during training to help in learning this kind of invariance.

Network	Pre-processing	Initial %	% Discarded	% After Discard
Standard 250/250	DoG	72	46	94
	Gabor	85	35	98
30+60 Face Unit	DoG	84	32	93
	Gabor	90	24	97

Table 4: Effect of pre-processing methods on **shift-varying** dataset (full groups of five used for training)

Network	Pre-processing	Initial %	% Discarded	% After Discard
Standard 250/250	DoG	83	34	98
	Gabor	90	26	97
30+60 Face Unit	DoG	91	24	97
	Gabor	93	20	98

Table 5: Effect of pre-processing methods on **scale-varying** dataset (full groups of five used for training)

6 Observations

Several points can seen from the results:

- The RBF network is shown to be able to generalize well in a non-trivial task classifying *y*-axis rotated faces (3-D complex shapes).
- Gabor preprocessing is shown to give a more generally useful input representation than the DoG preprocessing.
- Not suprisingly, the multi-scale Gabor preprocessing is shown to give greater scale invariance than the DoG preprocessing.
- The Gabor preprocessing is also shown not to fail catastrophically on the tougher shift invariance tests, unlike the DoG preprocessing.
- The RBF network is shown to have an inherent scale invariance on these tasks that does not need to be explicitly learnt from examples.
- In contrast, RBF networks do not have an inherent shift invariance, but this can be learnt from appropriate training data.
- The 'face unit' RBF network is shown to be superior to the standard network in terms of lower discard proportions for a particular level of generalization performance.

Although only ten individuals are being classified here, this type of network has been shown to work well with greater numbers of classes. For instance, the Olivetti Research Laboratory database of faces[3] with 400 images of 40 people can be distinguished with a high level of performance - with Gabor preprocessing, 95% can be correctly recognized after discard (see (Howell & Buxton, 1996)).

[3] available via *ftp*, for further information: `http://www.cam-orl.co.uk/facedatabase.html`

7 Conclusion/future work

In summary, the locally-tuned linear Radial Basis Function (RBF) networks showed themselves to perform well in the face recognition task. This is a promising result for the RBF techniques considering the high degree of variability introduced by the varying views (y-axis rotation) of a person's face in these data sets. By centering our sampled faces on the nose of the profile views, we can regard the partial occlusion as simply missing features from the other side of the face. This is in accord with known results from Ahmad & Tresp (Ahmad & Tresp, 1993) who trained a variety of nets to recognize stationary hand gestures from computer-generated 2-D views (polar co-ordinates) of fingertips. They obtained good generalization for 3-D orientation and showed that RBF nets were able to cope well even when much of the data was missing. Although their standard test data was handled well by a BP net, it performed badly with missing features and suffered a serious falling off in performance as more elements were lost. They showed, however, that a Gaussian RBF net (of the kind we used in our studies) could cope well, having a success rate of over 90% even with 50% of the features missing. This behaviour is very useful for coping with occlusion and other factors which lead to incomplete visual data.

We are now testing to see if the degree of view, scale and shift invariance that can be learnt by the RBF nets is sufficient to cope with data isolated from real-time video by a general purpose motion tracker. We are also studying invariance to facial expression and refining an automated 'face-finder' routine. This is necessary for the next stage of development in which people are to be identified in natural image sequences with the usual variations in illumination as well as position, scale, view and facial expression. The statistical nature of the information successfully captured by RBF nets to do the classification task may also be effective for the face localization task. It is clear from the work of Turk & Pentland (Turk & Pentland, 1991) and Bishop (Bishop, 1995) and others using statistically based techniques that this is the key to good performance and the RBF techniques are mathematically well-founded, which gives a clear advantage in engineering a solution to our application problems. Current work (Howell & Buxton, 1996) is tackling a much more unconstrained recognition task using faces tracked in real-time and gathering enough information to classify them accurately with good generalization to other image sequences containing familiar people.

References

Ahmad, S., & Tresp, V. (1993). Some solutions to the missing feature problem in vision. In Hanson, S. J., Cowan, J. D., & Giles, C. L. (Eds.), *Advances in Neural Information Processing Systems*, Vol. 5, pp. 393–400. Morgan Kaufmann.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition.* Oxford University Press.

Bruce, V. (1988). *Recognising Faces.* Lawrence Erlbaum Associates.

Bruce, V., Burton, A. M., & Hancock, P. J. (1995). Missing dimensions of facial distinctiveness. In Valentine, T. (Ed.), *Cognitive and Computational Aspects of Face Recognition: Explorations in face space*, pp. 138–158. Routledge.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, *77*, 305–327.

Chen, S., Cowan, C. F. N., & Grant, P. M. (1991). Orthogonal least squares learning algorithm for radial basis function networks. *IEEE Transactions on Neural Networks*, *2*, 302–309.

Craw, I., Costen, N., Kato, T., Robertson, G., & Akamatsu, S. (1995). Automatic face recognition: combining configuration and texture. In *Proc. Int. Workshop on Face and Gesture Recognition*, pp. 53–58 Zurich, Switzerland.

Daugman, J. G. (1988). Complete discrete 2-D gabor transforms by neural networks for image analysis and compression. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *36*(7), 1169–1179.

Edelman, S., Reisfeld, D., & Yeshurun, Y. (1992). Learning to recognize faces from examples. In *2nd European Conference on Computer Vision*, pp. 787–791 Genoa, Italy.

Girosi, F. (1992). Some extensions of radial basis functions and their applications in artifical intelligence. *Computers Math. Applic.*, *24*(12), 61–80.

Hertz, J. A., Krogh, A., & Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation*. Addison-Wesley.

Howell, A. J., & Buxton, H. (1995a). Invariance in radial basis function neural networks in human face classification. *Neural Processing Letters*, *2*(3), 26–30.

Howell, A. J., & Buxton, H. (1995b). Receptive field functions for face recognition. In *Proc. 2nd Int. Workshop on Parallel Modelling of Neural Operators for Pattern Recognition (PAMONOP)* Faro, Portugal.

Howell, A. J., & Buxton, H. (1995c). A scaleable approach to face identification. In *Proc. Int. Conference on Artificial Neural Networks (ICANN'95)*, Vol. 2, pp. 257–262 Paris. EC2 & Cie.

Howell, A. J., & Buxton, H. (1996). Towards unconstrained face recognition from image sequences. In *Proc. 2nd Int. Conf. Automatic Face and Gesture Recognition* Killington, Vermont.

Marr, D., & Hildreth, E. (1980). Theory of edge detection. *Proc. R. Soc. London*, *B207*, 187–217.

Moody, J., & Darken, C. (1988). Learning with localized receptive fields. In Touretzky, D., Hinton, G., & Sejnowski, T. (Eds.), *Proceedings of the 1988 Connectionist Models Summer School*, pp. 133–143. Morgan Kaufmann.

Moody, J., & Darken, C. (1989). Fast learning in networks of locally-tuned processing units. *Neural Computation*, *1*, 281–294.

Musavi, M. T., Ahmad, W., Chan, K. H., Faris, K. B., & Hummels, D. M. (1992). On the training of radial basis function classifiers. *Neural Networks*, *5*, 595–603.

Pentland, A., Moghaddam, B., & Starner, T. (1994). View-based and modular eigenspaces for face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 84–91.

Petkov, N., Kruizinga, P., & Lourens, T. (1993). Biologically motivated approach to face recognition. In *Proceeding of International Workshop on Artificial Neural Networks*, pp. 68–77.

Poggio, T., & Girosi, F. (1990a). Networks for approximation and learning. In *Proceedings of the IEEE*, Vol. 78, pp. 1481–1497.

Poggio, T., & Girosi, F. (1990b). Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, *247*, 978–982.

Rao, R. P. N., & Ballard, D. H. (1995). Natural basis functions and topographic memory for face recognition. In *Proceeding of International Joint Conference on Articial Intelligence (IJCAI'95)*, pp. 10–17 Montréal, Canada.

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, *3*(1), 71–86.

Widrow, B., & Hoff, M. (1960). Adaptive switching circuits. In *1960 IRE WESCON Convention Record*, Vol. 4, pp. 96–104. IRE, New York.

In: Noble, J. & Parsowith, S. R. (Eds.). (1996). The Ninth White House Papers: Graduate Research in the Cognitive and Computing Sciences at Sussex. Cognitive Science Research Paper 440, School of Cognitive and Computing Sciences, University of Sussex.

You'll Never Walk Alone in Vygotsky's Zone

Rosemary Luckin
rosel@cogs.susx.ac.uk

School of Cognitive & Computing Sciences
University of Sussex
Brighton
BN1 9QH

Abstract The title of this paper emphasizes the need for collaboration between the more able and the less able partner in a learning relationship which aims to create and maintain a "Zone of Proximal Development" (ZPD). If correctly constructed during instructional interaction, even after that interaction has ceased, the collaborator should be "invisibly present" as a helper in the learner's performance. One way to use the notion of a ZPD in software design is to investigate how the computer can perform the role of the more able partner in a collaborative relationship. This paper looks at the implications of using the ZPD as a basis for software design with particular reference to the problems involved in quantifying and adjusting the assistance offered to the user in the process of providing collaborative support.

1 What Vygotsky wrote about the ZPD

There are two presentations of the ZPD available in English translation. Each takes a slightly different approach and emphasis. The first source is "Thought and Language"(1986). The introduction of the concept of the ZPD here is set in the context of an investigation into the search for a working hypothesis to explain the development of scientific concepts in childhood. In particular, Vygotsky is concerned with the clarification of the relationship which exists between instruction and mental development. An area which Vygotsky feels has particularly lacked attention is the measurement of a child's mental development. Previously this had been done in terms of the child's ability to solve standardized problems unassisted. However, Vygotsky suggests that this method only measures the completed part of the child's development and that this is not the whole story. The ZPD is presented as a dynamic assessment metric, designed to assess the child's potential through their collaborative performance capability as opposed to their individual performance ability. The second place where Vygotsky discusses the ZPD is in "Mind in Society" (1978). Here the concept of the ZPD is introduced as a response to questions about the nature of the relationship between learning and development when a child is at school. In order to understand the relationship between learning and development it is insufficient to determine a single developmental level representing the development that has already taken place. Success depends upon the determination of at least two developmental levels, in addition to the actual developmental level, the level which a child can attain with assistance must be identified. The ZPD defines the mental functions that have not yet matured. In order to understand the child' s mental development it is essential to identify the two levels: "the actual developmental level and the zone of proximal development." (Vygotsky 1978). The creation of the ZPD is the essential feature of learning, it 'awakens' the internal developmental processes which can only operate when the child is interacting. A fundamentally important aspect of the ZPD from both versions is the necessity for collaboration or assistance from another, more able partner. However there is no detailed account of the form that this assistance should take.

> Different experiments might employ different modes of demonstration in different cases: some might run through an entire demonstration and ask the children to repeat it, others

> might initiate the solution and ask the child to finish it, or offer leading questions.
> (Vygotsky, 1978)
>
> . . . the first step in a solution, a leading question, or some other form of help.
> (Vygotsky, 1986)

The ZPD leaves open the nature of the collaboration required and the specification of what is learned during the instructional interaction (Wood & Wood, 1996).

2 Pinning down the nature of collaborative assistance

The suggestion that the ZPD is poorly specified and somewhat opaque is not new. Wertsch (1985), for example, tried to clarify and extend the notion of the ZPD. Wood (1976), (see Wood, Bruner & Ross 1976) coined the term "scaffolding" to describe tutorial assistance. Wood et al. (1976) present effective scaffolding as something more than the provision of hints and graded help. It involves simplification of the learner's role. The path the child takes through the ZPD is based upon their appropriation of what the teacher makes available to them. The more able partner in the collaborative relationship needs to organize the child's activity in a way which prevents them from being overwhelmed by uncertainty. The model of effective teaching proposed by Wood et al. (1976) involves the use of the contingent teaching strategy which requires that the more able partner adopt the following approach:

> When the child makes an error, then immediately take over more control . . . However, if the child is successful in following instruction relinquish some control.

There have been several other attempts to pin down the nature of the assistance that teachers can provide for children as they learn. Gallimore and Tharp (1988) for example use the term "Means of Assisting" and portray teaching as offering assistance "at points in the ZPD at which performance requires assistance." This leaves the specification of these points as a vital component of the process. Whilst there is variety in these different approaches, the goal is the same: tuning the assistance offered by the more able partner to the needs of the less able.

In addition to the problems of clarification, there is a growing trend towards the recognition that the established schools of thought such as the Apprenticeship and Situated Cognition paradigms (Lave 1988, Lave and Wenger 1991, Brown, Collins Duguid 1989) may have overlooked something vital. For example Bliss, Askew and Macrae (1996) examined scaffolding in classrooms to address the question of the applicability of this teaching technique to school contexts. They conclude that scaffolding in school is more difficult than was initially imagined. The scientific thought which is required at school is not the same as the everyday, spontaneous thought which has adapted well to the scaffolding approach. They suggest a new focus on the learner, whose level of development must be diagnosed through dialogue, and the domain knowledge of the curriculum. Before scaffolds can be constructed, or even planned, a careful analysis of the domain which indicates potential links to the child's existing, intuitive knowledge is essential. A similar point is made by Hedegaard (1996). She criticizes the situated cognition approach for its over reliance upon social practices in everyday life. Within a Vygotskian approach the difference between everyday learning and school learning are recognized. Children's school learning must be anchored in everyday life situations, subject-matter areas and the learner's development. The Situated and Apprenticeship models pay due attention to the first of these anchors but insufficient to the second and third (Hedegaard 1996). When discussing the merits of the Learner-Centred approach to education, Norman and Spohrer (1996) acknowledge the benefits of focusing on the learner and authentic problem solving activities BUT a structured analysis of the curriculum content is still necessary.

3 A framework for software design

One way to use the zpd in software design is to investigate ways in which the computer can offer assistance to the learner, the aim being that it can then play the role of the more able partner in a learning relationship. This assistance needs to pay attention to the nature of the activity which is offered to the child and the help which is made available to her as she attempts to complete this activity. Any adjustments to either the activity, the learner's role in the activity, or the help she is offered needs to be in line with the system's beliefs about the child's ZPD. The activity must encourage strenuous mental effort on the part of the child, but it must be possible for her to achieve success with some help from the system.

Two further constructs are of use in clarifying what needs to be specified, these are :

- The Zone of Available Assistance (ZAA)
- The Zone of Proximal Adjustment (ZPA)

The ZAA describes the variety of qualities and quantities of assistance which the system can make available to the child; it is a measure of the system's versatility in general for any child. However, the assistance which is selected and actually offered to the child needs to be matched to that particular child's ZPD. This is where the Zone of Proximal Adjustment (Murphey 1996) comes into play. The ZPA is the subset of the ZAA which is specific to the particular child with whom the system is interacting. The aim for the software designer becomes that of maximising the system's versatility: its ZAA, and providing a means of targeting the ZPA so that it as close as possible to the child's ZPD. In this way the ZPA is the system's view of the assistance side of the ZPD partnership. This entails the system assessing the child's ZPD and then adjusting to take account of this assessment.

The original descriptions of the ZPD, the interpretations of subsequent research and the analysis of those observing classroom interactions can all inform software design. Designing software to address the implications of the ZAA and ZPA requires that attention be paid to the following:

- A careful analysis of the subject matter of the domain is necessary to ensure that the child is introduced to a fabric of systematically organized scientific concepts with potential links between the concepts in the curriculum to be taught and the concepts already experienced by the child.
- The introduction of the scientific concepts in a systematic manner. Whilst there needs to be flexibility in terms of the child's role and the system's assistance, there need to be some overall guiding goals within the intended sequence of instruction.
- The identification of 2 levels: the child's independent ability and her collaborative capability. In other words, a means of representing the system's beliefs about the child's ZPD. The only evidence available as the basis for assessment is the child's performance on the activities she completes with the system. Measuring the collaborative support which she needed in order to succeed provides a means of measurement.
- The provision of collaborative assistance both in terms of the adjustment of activities and the provision of help. This assistance needs to be capable of being tuned to the system's beliefs about the child's ZPD.
- Recognition of the role the computer plays within the current classroom culture.
- An interface which allows for and promotes interaction.

Neither partner in the learning relationship should "walk alone". As well as a medium for the child and system to create a joint situation definition and communicate, it must act as the channel for the system to formulate beliefs about how the child is doing and for the system to show the child how to succeed.

4 References

Bliss J., Askew M. and Macrae S. (1996). Effective Teaching and Learning: Scaffolding Revisited. *Oxford review of Education, Vol. 22 (No1)*, pp. 37-61.

Brown J. S., Collins A. and Duguid P. (1989). Situated Cognition and the Culture of Learning. *Educational Researcher*. Jan-Feb 1989.

Tharp R. G. & Gallimore R. (1988). *Rousing Minds to Life: teaching, learning and schooling in social context.* C.U.P.

Hedegaard M. (1996). *Situated Learning and Cognition - Theoretical Learning and Cognition*, presented at the 2nd Conference for Socio-Cultural Research, Geneva.

Lave J. & Wenger E. (1991). *Situated Learning: Legitimate Peripheral Participation.* Cambridge University Press, New York.

Lave J. (1988). *Cognition in Practice: Mind, mathematics, and culture in everyday life.* Cambridge University Press, New York.

Murphey T. (1996). *Proactive Adjusting to the Zone of Proximal Development: Learner and Teacher Strategies.* Presented at the 2nd Conference for Socio-Cultural Research, Geneva.

Norman D. A. & Sophrer J. C.(1996). Learner Centred Education *Communications of the ACM, Vol. 39*, No. 4.

Vygotsky L. S. (1978).*Mind in Society: The Development of Higher Psychological Processes.* Harvard University Press, Cambridge, Mass.

Vygotsky L.S. (1986).*Thought and Language.* The M.I.T. press, Cambridge, Mass.

Wertsch J. (1984). *Culture, Communication and Cognition: Vygotskian Perspectives.* Cambridge University Press, Cambridge.

Wood D., Bruner J. S., and Ross G. (1976). The Role of Tutoring in Problem Solving. *Journal of Child Psychology and Psychiatry, Vol. 17*, pp. 89-100.

Wood D., Shadbolt N., Reichgelt H., Wood H. and Paskiewitz T. (1993). *EXPLAIN: Experiments in Planning and Instruction.* Dept of Psychology, University of Nottingham.

Wood D. and Wood H. (1996). Vygotsky, Tutoring and Learning *Oxford review of Education, Vol. 22 (No 1)*, pp. 5-16.

In: Noble, J. & Parsowith, S. R. (Eds.). (1996). The Ninth White House Papers: Graduate Research in the Cognitive and Computing Sciences at Sussex. Cognitive Science Research Paper 440, School of Cognitive and Computing Sciences, University of Sussex.

Automatic Acquisition of the Argument Structure and Semantic Preferences of Verbs.

Diana McCarthy
dianam@cogs.susx.ac.uk

School of Cognitive & Computing Sciences
University of Sussex
Brighton
BN1 9QH

Abstract An important aspect of a verbal lexical entry concerns the structural and semantic relationships between a verb and its arguments. This includes the surface syntactic expression of arguments, alternations between these expressions and semantic preferences between 'predicate' (verb) and arguments. Natural Language Processing (NLP) systems rely on this information in order to produce parses and handle ambiguity. Establishing and maintaining this knowledge manually for any non-trivial NLP application is extremely labour intensive. I shall present approaches which can acquire this information automatically from corpora of naturally produced text.

1 Introduction

Lexicon development is frequently cited as the bottle-neck for developing Natural Language Processing applications (Zernik, 1991). Developing and maintaining lexicons for a specific application is a costly process. Developing a more general lexicon suitable for any task has, on the whole, been abandoned for the forseeable future on account of the substantial labour that would be required and the unmanageability of the resultant system that would arise because of ambiguity. Automatic acquisition of lexical knowledge is seen as the solution, however methods to obtain lexical information still need to be identified. The work described here will contribute towards the lexical acquisition tools being developed for the EU funded SPARKLE project (SPARKLE, 1996).

The structural and semantic relationships between a verb and its arguments are vital components of a verbal lexical entry because natural language processing (NLP) systems require this information to avoid ungrammatical parses and to handle ambiguity. This information includes subcategorization frames, selectional restriction and diathesis alternations. These aspects of verbal argument behaviour all lie at the bridge between syntax and semantics.

Subcategorization frames are the range of syntactic argument structures that a predicate appears with. For verbal predication this is the surface realization of the underlying roles of arguments and will therefore be related to the verb's meaning. Thus in example 1A the verb 'gave' requires a Noun Phrase (NP) in subject position which expresses the 'agent' of the 'give' action, a NP in object position, the 'theme' of what is given, and a prepositional phrase with the preposition 'to' expressing the 'recipient' of the action. In contrast in 1B the verb 'poisoned' requires only the subject NP and an object NP and in 1C the verb 'collapsed' takes only a subject NP.

1A (John) gave (a bone) (to the dog).

1B (John) poisoned (the bone).

1C (The dog) collapsed.

This information is required by natural language processing systems in order to avoid spurious ungrammatical parses.

Diathesis alternations are regular variations in these surface expressions of arguments. For example the verb ‘gave’ can alternate between the two subcategorization frames exemplified in 2A and 2B. This is known as the dative alternation and occurs for verbs sharing the appropriate semantic components such as verbs of ‘giving’ e.g. ‘give’, ‘loan’, and ‘serve’ and those expressing instantaneous cause of ballistic motion e.g. ‘bash’, ‘hurl’ and ‘throw’ (Levin, 1993).

2A John gave a bone to the dog.

2B John gave the dog a bone.

These alternations provide useful organizational information for an NLP lexicon in that the alternating patterns need only be stored once and participating verbs can be stored with the base forms only and an indication of the alternations that apply. Storing generalizations is helpful in terms of efficiency and ease of maintenance but also may help establish relationships between word senses and the structures they participate in.

Selectional restrictions are the semantic constraints that apply to the combination of arguments with a predicate. For example in 3 both A and B are grammatically correct but only A seems semantically plausible because tables do not ‘give’ things and sausages do not usually receive things and certainly do not receive poems. Selectional restrictions are useful for word sense disambiguation, amongst other things. For example in 4 it is extremely unlikely that the reader will interpret ‘bank’ as having the sense RIVERBANK.

3A John gave the dog a bone.

3B The table gave the sausage a poem.

4 He robbed a bank.

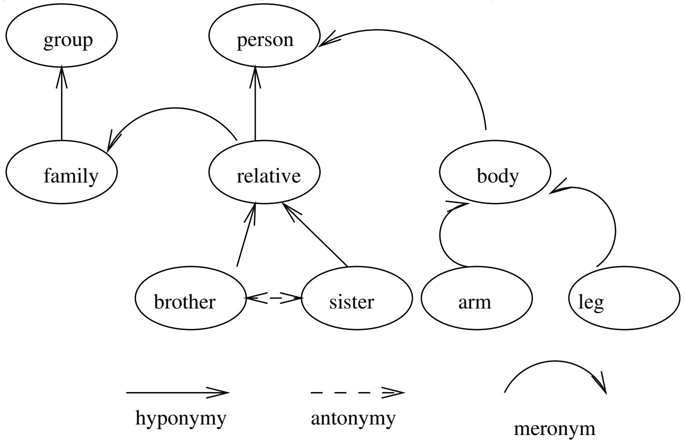
This paper describes approaches I hope to employ in automatically acquiring both the selectional restrictions and diathesis alternations of verbs. I shall use machinery for the automatic acquisition of subcategorization developed by other contributors (Briscoe & Carroll, 1995) on the SPARKLE project. These phenomena lie at the border between syntax and semantics and therefore as well as the subcategorization frame information I shall require a representation and method for semantic tagging. The next section covers the options I have for semantic tagging and section 3 and section 4 outline how I will use this information to acquire the selectional restrictions and diathesis alternations respectively and the final section provides a summary.

2 Semantic representation

The options I have for semantic representation are distinguished primarily by whether the knowledge is provided by humans or automatically. Machine Readable Dictionaries (MRD) and Thesauri (MRT) typically provide a set of semantic codes which have been used by some researchers (Yarowsky, 1992), (Cowie, Guthrie, & Guthrie, 1992), (Basili, Pazienza, & Velardi, 1993), (Resnik, 1993b), (Resnik, 1993a) however the disadvantage is that the codes are provided by humans for very general purposes and therefore the sense distinctions are likely to be inappropriate given any specific application and data set (Kilgarriff, 1996).

One frequently used resource is WordNet which unlike MRDs is organized in terms of word senses rather than words. A WordNet class represents a lexical concept and contains a group of words that are synonymous for the particular sense governed by the class. The classes are then linked together hierarchically where the links indicate semantic relationships such as hyponymy in nouns. See figure 1

Figure 1: A Network Representation of a Segment of WordNet.



for an example of what a small section of the hierarchy might look like as a network representation (Felbaum, Gross, & Miller, 1991), (Miller, Beckwith, Felbaum, Gross, & Miller, 1993).

The alternative is to use automatic clustering of words based on distributional information as this means that the classification is tailored to the data at hand. Pereira et al (Pereira, Tishby, & Lee, 1993) do this using co-occurrence data from specific syntactic relationships classifying nouns acting as direct-objects by the distribution of verbs that they occur with. Pereira et al then perform hierarchical clustering using this distributional information with relative entropy as the similarity metric between classes. A similar approach uses proximity to the target word instead of syntactic relationships for the co-occurrence data. Scütze does this using the cosine between co-occurrence vectors as the similarity metric (Schütze, 1992). The problem with automatic clustering is that the words in the classes produced are not always semantically similar. For example from Pereira et al's work although coherent clusters such as one including the words 'conductor','vice-president','chairman',and 'director' are obtained, clusters are also formed with less obviously related words such as 'state', 'modern' and 'farmer'.

I propose to use an automatic method for deriving a semantic classification which will then be used for semantic tagging. The semantic representation will be produced using syntactic relations as this appears to be more informative for common words than proximity based methods (Grefenstette, 1993). Unlike Pereira et al I hope to use evidence from more than one syntactic relationship on the assumption that this will provide me with more semantically cohesive clusters.

The process of tagging necessitates a decision on word sense disambiguation. Since these clustering methods amalgamate evidence from different word senses it is possible to simply use the cluster a word type belongs to, or has the highest probability of belonging to in cases of probabilistic membership (Pereira et al., 1993). However this will be misleading for many words with more than one meaning. Instead it is possible to look at each instance of a word in context and determine the most appropriate cluster from that context rather than just looking up the word form in the semantic representation. In the work proposed a similar approach will be taken to that of Schütze (Schütze, 1992) by finding the closest cluster given evidence from the context words occurring with the instance to be disambiguated. These context words will occur in a specific syntactic relationship with the target word rather than within a fixed distance as Schütze does. The correct cluster will be determined by the contexts that these context words themselves typically appear in.

3 Selectional restrictions

I take the view of selectional restrictions as graded tendencies that exist between predicates and arguments rather than hard and fast rules which amount to violations when broken. This view has been put forward by Wilks (Wilks, 1986) who termed them 'preferences' rather than 'restrictions'. In this way, whilst semantic interpretations that have strong associations between predicate and argument are preferred, it is still possible to allow interpretations that have weaker associations. Thus in example 4 though the association of the verb 'rob' with the 'FINANCIAL INSTITUTION' sense of 'bank' is strongest it is not inconceivable to think of a viable situation such as where a thief was stealing rare plant species from the side of a river. In this way rather than trying to find the exact semantic features required by arguments of verbs I shall instead use a statistical measure of association to estimate the preference of verbs for particular arguments.

It would be an unmanageable task to try and store associations between verbs and all the individual words occurring in argument positions in a lexicon. Some degree of generalization will be needed and for this I shall use the word classes provided by the semantic representation described above. My work resembles that of both Resnik and Ribas (Resnik, 1993a) (Ribas, 1995) in that I shall probably use an adaptation of their basic measure of association given in equation 1.

$$A(v,c) = P(c|v) log \frac{P(c|v)}{P(c)} \tag{1}$$

Resnik and Ribas both use the hyponymy relation between WordNet classes as their hierarchical semantic classification. In their work the frequency of a class is calculated using the occurrence of any word which is identified as belonging to that class or the occurrence of any words belonging to its descendant classes, i.e. hyponyms. Notice that the count of all possible senses of a word are incremented with an occurrence of that word.

$$log \frac{P(c|v)}{P(c)} \tag{2}$$

Equation 2 gives the mutual information between a class and a verb. This measure is frequently used in statistical approaches and provides evidence for the amount of information two entities provide about each other. In this case the entities are a verb (v) and a particular WordNet class (c). If the two co-occur significantly then the posterior probability will be larger than the prior probability. The mutual information will therefore be positive and its magnitude will be determined by the strength of the relationship. If there is no correlation then knowing 'v' has occurred tells us nothing about the occurrence of 'c', the posterior probability will not be different from the prior and the mutual information will be 0. It is also possible to have a negative mutual information where the occurrence of the context 'v' means that the chance of 'c' is less than the prior probability. Since mutual information by itself would be biased towards small classes as these are easier to fit to particular contexts, Resnik weights this mutual information by the posterior probability as that tends to favour larger classes. Thus we arrive at a class that is 'general enough' to characterize the nouns occurring as direct-objects to the verb.

Unlike Resnik and Ribas, I shall use an automatically derived semantic classification based on the data at hand rather than an MRD. I shall also apply some word sense disambiguation to the data before applying the association score, the lack of sense distinction gave rise to a major source of error in their work.

In my approach I shall use the shallow parsing provided by other contributors on the SPARKLE project to identify argument heads on which I can apply the semantic tagging and obtain occurrence data for semantic class at particular argument positions, irrespective of verb, contrasted with co-occurrence with the specific verbs for which selectional preference is being acquired. The semantic representation will be a hierarchical one and like Resnik and Ribas an occurrence of a class at a particular argument position will add weight not only to that class but also to all of its parent classes. Unlike them, I hope to use the semantic tagging to obtain class frequencies relevant to the word sense at hand and not to all potential senses of the word. The class with the highest association score for a particular verb will be concluded to have the strongest degree of selectional preference but other classes can also be stored with degree of association as an indication of the continuum of preference.

4 Diathesis alternations

The results of both subcategorization frame and selectional restriction acquisition are planned to provide the basis for deriving the set of diathesis alternations available to verbs. The work that has been performed on extracting diathesis alternations has been on the whole performed manually (Nicholls, 1994) or semi-automatically using MRDs (Sanfilippo, 1994). To my knowledge the only other researcher who has attempted to get diathesis alternations automatically from naturally occurring text is Resnik (Resnik, 1993a). He looks at the set of implicit-object alternations where the direct-object can be dropped in English. for example:

5A Joe ate the sandwich.

5B Joe ate.

His approach rests on the assumption that objects are dropped where they are more easily inferable from the verb. For example the object of the verb 'eat' is more easily inferable than that of 'need'. He presumes that the strength of the selectional constraint of a verb for its direct-object will indicate participation in this 'object-drop' alternation. To measure this strength he uses the association measure described in equation 1 on page 51 between verbs and their direct objects which are obtained using the Penn Treebank parses (Santorini, 1991a), (Santorini, 1991b). The results are promising but there are a few evident problems. One source of error is that of erroneous senses of direct-objects contributing to the wrong WordNet class. Additionally it seems that in some cases factors other than inferability are involved.

Although I believe that Resnik's approach is very useful for hypothesising such implicit-object alternations, for other alternations it may be even harder to establish the exact links between features of verb meaning and syntactic behaviour. I wish to develop a method for identifying verbs participating in a much broader range of alternations. Like Resnik, I plan to acquire alternations automatically using naturally occurring data but I intend to examine surface phenomena rather than relying solely on a theory of what aspects of meaning cause the alternations. I hope to observe the semantic type that specifies selectional restrictions shifting between different argument positions for different subcategorization frames. I will use this as evidence that the two argument positions in the respective frames occupy the same underlying role and that the surface expression of that role alternates between the two frames. Thus if I extract from my corpus subcategorization patterns 6A from examples like 'John broke the door' and 6B from examples like 'the window broke', with selectional constraints as shown attached to the constituent phrases, I can use this to infer an alternation as in 6C.

6A NP-person V-break NP-physical object.

6B NP-physical-object V-break.

6C NP1 break NP2 $\leftrightarrow$ NP2 break.

The two frames can be compared to establish which alternation they exhibit with reference to a set of alternation patterns specified a priori.

5 Summary

The main theme of my proposal is to use an automatic classification of nouns derived from distributional evidence to provide generalizations for hypothesising selectional constraints between verbs and their arguments. The selectional constraints are to be specific to the type of subcategorization frame as different frames can have different underlying roles occupying the slots. By observing shifts in the preferences of argument slots it is hoped that a wide range of diathesis alternations can be associated with individual verbs. Automatic acquisition of this type of information would be a significant contribution to lexicon development.

References

Basili, R., Pazienza, M. T., & Velardi, P. (1993). Hierarchical clustering of verbs. In Boguraev, B., & Pustejovsky, J. (Eds.), *The Acquisition of Lexical Knowledge from Text. SIGLEX ACL Workshop*, pp. 70–81 Columbus Ohio.

Briscoe, T., & Carroll, J. (1995). Automatic extraction of subcategorization from corpora. Not yet published.

Cowie, J., Guthrie, J. A., & Guthrie, L. (1992). Lexical disambiguation using simulated annealing. In *Proceedings of the 14th International Conference of Computational Linguistics. COLING-92*, Vol. I, pp. 359–365.

Felbaum, C., Gross, D., & Miller, G. (1991). WordNet: A lexical database organised on psycholinguistic principles. In Zernik, U. (Ed.), *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*. Lawrence Erlbaum Associates.

Grefenstette, G. (1993). Evaluation techniques for automatic semantic extraction: comparing syntactic and window based approaches.. In Boguraev, B., & Pustejovsky, J. (Eds.), *The Acquisition of Lexical Knowledge from Text. SIGLEX ACL Workshop.*, pp. 143–153 Columbus Ohio.

Kilgarriff, A. (1996). I don't believe in word senses. To be published in Computational Linguistics.

Levin, B. (1993). *English Verb Classes and Alternations: a preliminary investigation.* University of Chicago Press, Chicago and London.

Miller, G., Beckwith, R., Felbaum, C., Gross, D., & Miller, K. (1993). *Introduction to WordNet: An On-Line Lexical Database.*

Nicholls, D. (1994). French and English diathesis alternations and the LKB. Tech. rep., ACQUILLEX-II.

Pereira, F., Tishby, N., & Lee, L. (1993). Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguists.*, pp. 183–190.

Resnik, P. (1993a). *Selection and Information: A Class-Based Approach to Lexical Relationships.* Ph.D. thesis, University of Pennsylvania.

Resnik, P. (1993b). Semantic classes and syntactic ambiguity. In *Proceedings of the ARPA Workshop on Human Language Technology.*, pp. 278–283. Morgan Kaufman.

Ribas, F. (1995). On learning more appropriate selectional restrictions.. In *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics.*

Sanfilippo, A. (1994). Word knowledge acquisition, lexicon construction and dictionary compilation.. In *Proceedings of the 15th International Conference of Computational Linguistics. COLING-94*, Vol. I, pp. 273–277.

Santorini, B. (1991a). *Bracketing Guidelines for the Penn Treebank Project.* University of Pennsylvania, treebank@unagi.cis.upenn.edu.

Santorini, B. (1991b). *Part-of-Speech Tagging Guidelines for the Penn Treebank Project.* University of Pennsylvania, treebank@unagi.cis.upenn.edu.

Schütze, H. (1992). Dimensions of meaning. In *Supercomputing.*

SPARKLE (1996). SPARKLE - shallow parsing and knowledge extraction for language engineering. Web Page http:/www.ilc.pi.cnr.it/sparkle.html.

Wilks, Y. (1986). An intelligent analyzer and understander of english. In Grosz, B., Spark Jones, K., & Webber, B. (Eds.), *Readings in Natural Language Processing*, pp. 264–274. Morgan Kaufmann. Originally appeared in CACM 18(5).

Yarowsky, D. (1992). Word sense disambiguation using statistical models of Roget's categories trained on large corpora.. In *Proceedings of the 14th International Conference of Computational Linguistics. COLING-92*, Vol. II, pp. 456–460.

Zernik, U. (1991). Introduction.. In Zernik, U. (Ed.), *Lexical Acquisition : Exploiting On-Line Resources to Build a Lexicon.* Lawrence Erlbaum Associates., Hillsdale NJ.

In: Noble, J. & Parsowith, S. R. (Eds.). (1996). The Ninth White House Papers: Graduate Research in the Cognitive and Computing Sciences at Sussex. Cognitive Science Research Paper 440, School of Cognitive and Computing Sciences, University of Sussex.

Can Artificial Life Explain the Evolution of Communication?

Jason Noble
jasonn@cogs.susx.ac.uk

School of Cognitive & Computing Sciences
University of Sussex
Brighton
BN1 9QH

Abstract The evolution of communication is one of many research issues that might be investigated using the new techniques of artificial life. This paper reviews the conceptual issues that must form the theoretical groundwork for any such research, including a definition of the term "communication", the possible impact of simulation work on the question of the origins of human language, the relevance of Dennett's (1987) "intentional stance", and a discussion of the varieties of scientific explanation. There follows a brief summary of work in theoretical biology on the selective pressures associated with communication, and an argument from Miller (1995) that ALife work should begin with models from theoretical biology.

Consider such phenomena as aggregational signals, alarm signals, food signals, territorial and aggressive signals, appeasement signals, courtship and mating signals, and signalling between parents and offspring (Lewis & Gower, 1980). Clearly, some of the animals in the world communicate with each other. Within the evolutionary paradigm, we assume that each of these animals has ultimately descended from a non-communicating ancestor—so how and why did communication behaviour get started? And what sort of evolutionary pressures have led, in some cases, to increasingly complex communication systems? I propose using the methodology of artificial life ("ALife," Langton, 1987) to come up with some answers to these questions.

But what do we mean by "communication" anyway? And if we use ALife methods to study it, what *sort* of answers can we expect—can ALife be real science? What philosophical perspective should we take on the subject? How useful is existing work in biology, game theory or linguistics going to be? Will such research shed any light on the development of human language?

Consideration of my central thesis, that an ALife simulation in which communication evolves will aid our understanding of the real thing, leads immediately to these further questions. Some of them need to be answered as a matter of logical necessity if the thesis is to stand, while others are more matters for clarification. This paper outlines an approach to these preliminary issues.

1 Defining communication

Definitions can be boring stuff, and at first glance "communication" appears so straightforward as to be hardly worth defining. The naïve definition of communication is easy: to communicate is to transmit information; to *tell someone something*. The speaker might be lying, or the listener might have already known what the speaker tells them, but these look like marginal cases.[1] Communication is all about getting a message across.

[1] The use of the terms "speaker" and "listener" here should not be taken as an assertion that all communication is linguistic and verbal. I will discuss in section 2 our tendency to take, as I have done here, an anthropomorphic view of communication. In section 3 I will argue that such anthropomorphism may not be as unscientific as it at first appears.

It is noteworthy that the naïve definition has not been satisfactory for most evolutionary biologists. Burghardt (1970) defined "communication behaviour" as a behaviour on the part of a signaller that is likely to influence the receiver in a way that benefits, in a probabilistic manner, the signaller or some group of which it is a member. Krebs and Dawkins (1984) defined a "signal" as an action or structure which increases the fitness of an individual by altering the behaviour of other organisms. Notice that both of these definitions focus on observable behaviours, and implicitly invoke the idea of natural selection from the outset. The focus on observable behaviour is characteristic of empirical science, and demonstrates a reluctance to engage with the mentalism implicit in the naïve view. The invocation of natural selection obviously limits these definitions to evolved systems only.

Are these definitions good enough? Do they capture all the phenomena we think of as communication, and exclude the phenomena that we don't? A typical animal signal, e.g., territorial scent marking, presumably influences the behaviour of at least some other organisms (discourages them from entering the territory) and thus increases the fitness of the signalling animal (reduced competition for resources). In this sense the definitions may be adequate. But they cast a very broad net. A gazelle that runs fast may influence the behaviour of the cheetah that is chasing it: the cheetah may give up the chase sooner, having judged that the gazelle will probably get away. So, the gazelle's behaviour has influenced the cheetah's behaviour and consequently improved its own fitness. However, it seems strange to describe the gazelle as *communicating* with the cheetah in this instance.

Hasson (1994) pointed out this problem and tried to improve the definition offered by Dawkins and Krebs with the additional proviso that the signal must reduce fitness in contexts other than interactions with other organisms. This would exclude "running very fast" from being a signal, because it's an ability that is useful in a variety of contexts. Maynard Smith and Harper (1995), in turn, point out a problem with Hasson's correction:

> For example, merely being large may alter the behaviour of opponents in contests, and may well be costly in other contexts, but we would not wish to classify large size as a signal. . . There seems no alternative, therefore, to including in the definition the notion that a signal has features specifically adapted to alter the behaviour of others.

What Maynard Smith and Harper are saying is that signals are behaviours or structures that not only influence other animals, but whose *purpose* is to do so. This point will be taken up in section 3, but for now it is enough to note that once the concept of purpose has been raised then we are not so far form the naïve view, whereby signals are *for* passing messages from one cognizer to another.

2 An aside: the origins of human language

It is safe to say that our own linguistic capacity lurks behind all of our intuitions about communication, and therefore those intuitions may lead us astray when we consider signalling systems much more primitive than our own. Bearing that danger in mind, to what extent can we regard human language as continuous with, and having evolved from, animal signalling systems? If we take evolution seriously, human language had to develop from *something*, but the interesting question is whether there was some sort of qualitative leap or whether our linguistic behaviour is just a more complex version of what other animals do.

Unfortunately, there is a long history of undisciplined speculation on exactly this topic. Bickerton (1994) relates the story that as long ago as 1866 the Linguistic Society of Paris became so tired of wildly speculative papers on the origin of language that they imposed a ban, which apparently stands to this day.[2] Barring time travel, it is hard to imagine how one could collect enough data to decisively settle the question of how human language really got started.

However, I think there is enough data to suggest that human language is *not* just another, albeit complex, animal signalling system. Our very familiarity with our own language abilities, combined with a

[2] My thanks to Larry Trask for this anecdote and other material in this section.

desire not to appear species-chauvinist, and recent research interest in areas such as dolphin and whale communication and the teaching of American Sign Language to chimps, has sometimes led people to underestimate the differences between us and the rest of the animal kingdom. Chomsky (1968, 1975), among others, has demonstrated that there really is an enormous gulf between animal signalling systems and human language: we can talk about past and future events, we can talk about conditional relationships, we can predicate arbitrarily many verb-phrases of any noun phrase, we have a rich grammar, and we can produce and understand sentences that we have never heard before. There is no evidence for anything like this kind of sophistication in any other animal.

Bennett (1976) gives an excellent philosophical account of what it might take to move from a signalling system to a language. Drawing on the work of Grice, he suggests that second-order intentionality is a necessary condition for language. That is, to be a language user:

- you have to have intentions,
- you have to be aware that your communication partners have intentions, and
- you have to be aware that your communication partners are aware that you have intentions.

I find Bennett's ideas very promising. However, Bennett is of course referring to organisms whose communication system is right on the edge of being a true language—I am not so optimistic as to believe that the simulated organisms of near-future ALife work will be in such a position.

3 The intentional stance

In section 1, I noted the tendency of some evolutionary biologists to define communication in such a way as to avoid intentional, mentalistic language. However, I think that non-intentional accounts of communication are incoherent—communication just *is* an intentional-level concept. As soon as you are talking about a sender, a receiver and a message or signal, you cannot ignore questions about the meaning of the signal and the purposes of the sender. Even the most rigorous, mathematical treatments (notably Shannon & Weaver, 1949) are suggestive of intentionality in both senses of the word: the sender *intends* to communicate something to the receiver, via a signal that *means* something.

Dennett (1987) outlines what he calls the "intentional stance"; the notion that it can be a useful explanatory tool to view certain complex systems (e.g., evolved organisms) as if they were rational, intentional agents. I am in broad agreement with Dennett's thesis, and I feel that the phenomenon of communication practically cries out for this approach.

I should make clear at the outset that I am not making the bold claim that *animals have intentions*. I am saying that it can be very useful to view them as if they did. Take the phenomenon of predator alarm calls in vervet monkeys (Seyfarth, Cheney, & Marler, 1980). Currently we have no way of detecting intentions "inside" the monkeys, but nor do we have a way of establishing their absence. The point being, our only criteria for judging the various stories we might tell about monkey phenomenology is their explanatory usefulness. I might claim that vervet monkeys are mainly influenced by astrological concerns, but if my story has no demonstrable connection with what the monkeys actually do, then it should not be taken seriously.

In principle, we think we can explain what monkeys do in terms of neurology, and even in terms of chemistry and ultimately physics. However, none of those stories are very practical at the moment, and there is no guarantee that they ever will be. Nevertheless, it turns out that we *can* talk quite usefully about what a monkey intended to achieve by calling, and what a particular call means. And the intentional explanations are quite compelling. For instance, the monkeys do not give alarm calls when they are alone and see a predator. They have also been observed to use an alarm call to lie: one vervet called the leopard alarm when a fight with a rival group was going badly, which sent all the combatants scurrying into the trees, thus buying time for his side. This sort of sophisticated behaviour makes a simple, "objective" stimulus-response account of vervet signalling look naïve.

Dennett's main point is that we should not resist this temptation to adopt the intentional stance towards complex systems like vervet monkeys. It is a practical basis for further inquiry and experiment to treat them, at least provisionally, as rational, intentional agents. Dennett suggests that this is an example of a more general strategy in science of "changing levels of explanation and description in order to gain access to greater predictive power or generality" (Dennett, 1987, p. 239), and compares it to the abstraction of the concept of *food* in biology. The underlying details of physics, chemistry or whatever are legitimately ignored because they do not contribute to the usefulness of explanatory accounts at the higher level.

So where do we draw the line? Can we ascribe intentions and meaning to just anything? Dennett suggests that there will be no clear division between cases where the intentional stance is appropriate and cases where it is not. As we move to less and less complex systems, mechanistic accounts of their function will become progressively more plausible. Relatively simple organisms (and, arguably, all simulated organisms) are an interesting case in this respect. When dealing with complex, higher animals like vervet monkeys, it does not strain our intuitions too far to see them as having intentions (i.e., plans or goals). On the other hand, if we consider a caterpillar of species *A* that mimics another, poisonous species *B* (mimicry being a signal in the eyes of most evolutionary biologists), the intentional description feels a little less comfortable. Where does the *intention* to look like the other species reside? Surely not in the mind or brain of the caterpillar?

No—Dennett argues that evolved systems are often so much like designed systems that it is fruitful, as an explanatory device, to view them as having been designed to function well in their particular environment (notwithstanding Gould & Lewontin, 1979). Therefore the "intention to look like species *B*" is a useful theoretical abstraction, but there is no agent to whom we can ascribe that intention. As a bit of sloppy explanatory shorthand, we might want to say that the intention lies with mother nature or some other imaginary but rational designer. The approach is not as strange as it sounds, although clearly we should take care not to confuse it with the claim that there really *is* a designer. In the case of the caterpillar, if we deny ourselves the intentional perspective, then we are left wondering "why those particular skin pigments?", "why that body length?", and so on, with no answers to these questions in sight. On the other hand, if we adopt the intentional stance, then we can hypothesize that the purpose of the morphology is to look like a member of species *B*, and that the purpose of looking like a member of species *B* is to not get eaten.

In terms of ALife work on communication, Dennett's arguments should serve as a reminder that the behaviour of the simulated organisms may be *usefully interpreted* as intentional without that being a claim about their phenomenology, and that, a priori, we should remain open-minded about whether the best explanation of a particular instance of communication is mechanistic or intentional in character.

4 Varieties of explanation

If we set up a simulation in which some sort of communication evolves, what sort of explanation do we think that might give us for the real thing? Or is that just going to be another communication phenomenon, itself in need of explanation? The discussion of the intentional stance, above (section 3), touches upon these questions but does not answer them. I regard the problem of explanation as extremely important for ALife, particularly given that the field is a new one and that there is not yet a consensus about the appropriate conclusions to be drawn from simulations of evolving systems.

In a classic paper, Nagel (1961) outlined four different modes of explanation:

Deductive: Explicandum follows logically from the premises.

Probabilistic: Explicandum *likely* given the premises.

Functional (teleological): Something is explained by saying what it is *for*.

Genetic (historical): Something is explained by listing the contingent events that led to the current state of affairs.

The first category is characteristic of mathematics and physics, and recalls the deductive-nomological model of explanation put forward by Hempel and Oppenheim (1948). At least one observation statement and at least one natural law combine to form an explanation for the observed event. For example, the stone was released at the top of the tower; unsupported objects will accelerate under the influence of a gravitational field; thus, the stone fell to the ground. Probabilistic explanations are of essentially the same form, with the admission that the world can be fuzzy sometimes. I suspect that ALife has the potential to supply probabilistic explanations at a particular level of generality, e.g., the artificial organisms were confined to an area with finite resources; population levels usually go through a boom and bust cycle when resources are finite; thus, population cycling was (probably) observed. Given that Hempel and Oppenheim's original formulation has been the subject of much criticism, the extent to which this sort of explanation would be satisfactory is not yet clear.

It will be apparent from the time I have devoted to Dennett's intentional stance that I find functional and teleological explanation a promising avenue. Whereas the history of science shows a movement from teleological to physicalist explanations, I think there are sound arguments (Sober, 1993; Wright, 1994; Cummins, 1994; Dennett, 1987) that where complex, evolved systems are concerned, teleological explanations are sometimes the best available. Furthermore, the decision to adopt an intentional stance, like the decision to make a working assumption that adaptationism is true, or the decision to subscribe to evolutionary theory in general, is just that: a *decision*. It is not a hypothesis. As Maynard Smith (1983) says of the theoretical biology literature, "in using optimization, we are not trying to confirm (or refute) the hypothesis that animals always optimize; we are trying to understand the selective forces that shaped their behaviour." Similarly, teleological explanation is a stance we adopt in order to formulate theories at a particular level, and the program will stand or fall based on the utility of its results in the long term.

Hendriks-Jansen (1996), believing that ALife can contribute to the naturalistic understanding of animal behaviour, is most interested in Nagel's fourth category of explanation. Genetic[3] explanations are characteristic of evolutionary biology, and it has often been observed that evolutionary theory does not give much power to make predictions. Hendriks-Jansen accepts this, arguing that the explanation of evolved mechanisms is always going to be a long list of particular facts. When evolution is occurring within a computer program, all conceivable variables, across the entire span of the evolution of a trait, are potentially available for analysis. Thus ALife and historical explanation certainly seems a good marriage. However, there is something unsatisfying (for me, at least) about explaining every evolved phenomenon with a different list of arbitrary events. If it is at all possible, ALife should seek to extract general, or even law-like, explanatory statements from the results of simulations. Historical explanations may function as a supplement to this enterprise.

Nagel's typology is not necessarily complete, and it is also not the only way of classifying the different explanatory styles we actually use. Clark (1996), for example, gives a more fine-grained analysis of explanatory styles used in cognitive science. Clark, along with other authors such as van Gelder (1991), describes an apparently novel class of explanatory story relevant to the ALife paradigm: dynamical or emergent explanation. The idea is that dynamical systems theory (Abraham & Shaw, 1992)—in practice this usually means a system of differential equations—is the most effective framework for capturing and explaining the complex set of relationships between the internal mechanisms of a cognizer and between the cognizer and its environment. In a situation (real or simulated) where communication is evolving, the environment becomes doubly complex in that it now includes the evolving signals of other organisms; explanations in terms of dynamical systems theory may be the only way of grappling with this sort of complexity.

I have described some forms of explanation as more or less satisfying than others; Braithewaite (1953, p. 316) noted that "any proper answer to a 'Why?' question may be said to be an explanation of

[3] Genetic as in genesis, not as in genes.

a sort"; Bromberger (1962, 1966) and van Fraassen (1980) have emphasized pragmatically that theories do not explain, but that human speakers *use* theories to explain things to each other. These observations should remind us, as should Wittgenstein (1953), that in the attempt to analyse explanation as it occurs in the practical human business of science, absolute clarity and mathematical formalism are likely to prove elusive. I do not intend these comments as a sympathetic nod in the direction of relativism or anti-realism; however, I think that a comment by Feyerabend (1991, p.141) captures something of the essence of scientific explanation: "All you can do, if you really want to be truthful, is *to tell a story*, a story that contains... vague analogies to other stories in the field or in distant fields."

5 Communication: current arguments in theoretical biology

Empirical work in biology on the subject of animal communication tends to describe a particular type of signalling within one species or between two species (for reviews, see Harper, 1991; Krebs & Dawkins, 1984; Lewis & Gower, 1980). In some cases this work represents years of careful observation and experiment, and it certainly provides a good observational base for the construction and validation of ALife or other theoretical models. I am more immediately interested, however, in the theoretical biology literature (in section 6 I will try to explain why). Theoretical biology of recent decades has been heavily influenced by game theory (von Neumann & Morgenstern, 1953) and is typified by work such as Maynard Smith's (1982) *Evolution and the Theory of Games*. The approach is to posit a high-level, functional model of some animal behaviour (e.g., a breeding strategy, or, in our case, a signalling system) and then, using mathematical techniques, investigate the parameter values for which that system would be evolutionarily stable. In Maynard Smith's terms, we model the game the animals are "playing", and then try to identify evolutionarily stable strategies (ESSs) for that game. The argument is that we should only expect to find relatively stable systems in nature (although Maynard Smith has admitted that this approach necessarily misses out on the dynamics of evolution; see section 6).

An example of the approach applied to communication is the work of Zahavi (1975, 1977), whose verbal arguments were later validated mathematically by Grafen (1990). They argue that signalling associated with sexual selection will tend to be honest, and the signals costly. This has become known as the "handicap principle". To illustrate: assume that males signal their reproductive fitness to females through some phenotypic trait. Whereas we might expect males to exaggerate their quality, females will be selected on the basis of their ability to discriminate between high and low quality males. The ESS will be one where a signal is expensive for the male to produce (e.g., the peacock's tail). Thus deceptive communication becomes impossible: if the male can afford to display that signal, he really is of high quality. Through logical and mathematical argument, Zahavi's counter-intuitive idea—that sexual selection can be selection for a handicap—is established.

However, there is a conflicting tradition in theoretical biology that emphasizes the extent to which animal signals are likely to be dishonest. Krebs and Dawkins (1984), in a classic chapter entitled "Animal signals: Mind-reading and manipulation" (see also Wiley, 1983), suggest that communication arises because it is advantageous for animals to be good at predicting the behaviour of other animals in their environment. Thus we might expect that at some point the ancestors of *Canis familiaris* learned that bared teeth in another animal was a good predictor of an imminent attack. Once this "mind-reading" ability was established, it could then be used against other animals, so to speak, to manipulate their behaviour: a dog could bare its teeth, with no intention of carrying out a costly attack, and cause other dogs to retreat or submit. Krebs and Dawkins use this sort of story to account for the origin of threat displays. Another way of putting it is that deceptive communication will evolve when it is beneficial to have one's behavioural intentions falsely predicted.

A recent paper by Maynard Smith and Harper (1995) helps to resolve the apparent contradiction between the two views of communication, and suggests some useful terminological distinctions. Maynard Smith and Harper first distinguish between the potential subjects of an animal signal:

- Self-reporting signal

- Other-reporting signal

The vast majority of animal signals are egocentric, i.e. report something about the internal state of the signaller. Aggressive or territorial signals and sexual displays are clearly giving information (or misinformation) about the signalling animal. The much-studied bee dance, on the other hand, is giving information about the environment.

Maynard Smith and Harper next discuss the process by which the reliability of a signal is maintained. Game theory suggests that signals must be of some benefit to the signaller or they will not get started. If a signal manipulates the behaviour of others, to the advantage of the signaller, then the *signaller's* behaviour will be selected for, but what about the behaviour of the receivers? Why do they pay any attention to a signal that serves to take advantage of them? What stops a signalling system from degenerating into constant cheating? Maynard Smith and Harper suggest the following mechanisms:

- Minimal signal
- Cost-added signal
- Index

Minimal signals are signals whose cost is no greater than that required to transmit the signal efficiently. These will only be reliable if "signaller and receiver place the outcomes of the interaction in the same rank order". In other words, minimal signals only work if the two animals are co-operating. This will typically apply between members of social species, like ants or primates, but may even occur in predator-prey relationships. Consider a sparrow looking directly at a domestic cat as if to say, "I've seen you; don't bother trying." The sparrow "tells" the cat that the game is up, because once the cat has been seen, it is no longer in the interests of either animal for the stalking to go on. The bird will fly away before the cat can reach it, but it is a better result for both if the bird can go on feeding undisturbed and the cat can save its energy.

Cost-added signals are signals whose cost is *greater* than what is required to simply transmit the information. The intensity of the signal is correlated with some quality of interest to the receiver. Simplistically, these signals are reliable because, as signallers and receivers co-evolve, cheating becomes prohibitively expensive (recall Zahavi, 1975). A typical example is the peacock's tail.

Indices, like cost-added signals, also have an intensity that is correlated with a quality of interest to the receiver, but the correlation is a result of physical necessity rather than co-evolution. The example given by Maynard Smith and Harper is that of *Panthera tigris*: tigers mark their territory by scratching the bark of trees as high up the trunk as possible. Another tiger entering the area can judge the size of the resident tiger from the height of the claw marks. Clearly, the system will not collapse into cheating as it is physically not possible for the tigers to exaggerate.

Maynard Smith and Harper also neatly sum up the relationship between the intuitive idea that communication is about transmitting information, and the definitions from evolutionary biology based on altering the behaviour of other animals:

> Essentially, there is a connection because it is not evolutionarily stable for the receiver to alter its behaviour unless, on average, the signal carries information of value to it.

The combination of this formulation with the idea that there can be different processes that maintain the reliability of signals represents the main contribution of Maynard Smith and Harper. The apparent conflict between Zahavi (1975) and Krebs and Dawkins (1984) is resolved. Firstly, there are going to be occasions when it is in the interests of two animals to co-operate; that is, to signal truthfully. In those cases we can expect minimal signals, and in the real world that probably means quite subtle signals that we may have trouble detecting.

When the animals are potentially in conflict, then there is a pressure to cheat (signal dishonestly), but cheating will only work in the long run if the receivers keep believing you, and the receivers will only continue to trust the signal if, in the long run, it pays for them to do so because there are lots of honest versions of that signal around. As Maynard Smith and Harper put it, “There are a lot more worms than angler fish lures.” Thus, dishonesty will be an ESS under the right circumstances.

Finally, there are going to be situations such as the peacock’s tail, where the males are under pressure to cheat and exaggerate their quality, and have presumably tried various ploys and short-cuts over evolutionary time, but they are of course co-evolving with the females, who are under even greater pressure not to get fooled. Honesty, too, can be an ESS.

6 A biologically informed methodology for artificial life

I hope that it is obvious from section 5 that the field of theoretical biology contains much of value to an ALife study of the evolution of communication. Miller (1995) has argued strongly that work in theoretical biology and related fields is the best starting point for those who wish to model communication and other biological phenomena *in silico*, and furthermore that there has in the past been a naïve Orientalism among ALife researchers who believed themselves to be looking at issues that had never been considered before. Certainly a disappointing proportion of ALife papers turn out to be little more than “proof of concept” displays. Miller suggests that one way to get beyond this is to begin with specific, already-explored models from theoretical biology, and to convert them from their existing form into that of an ALife simulation.

To make the mathematics tractable, models in theoretical biology contain many unrealistic, limiting assumptions. For instance, it is typical to assume random mating and infinite populations, and to make no attempt to capture the effects of physical space or geometry. Miller’s insight is that ALife makes it possible to start with such a model, the dynamics of which are already reasonably well-understood, and then relax the assumptions one by one, thus building up an increasingly realistic understanding of the phenomenon. There are some examples of this approach in the recent ALife literature: de Bourcier and Wheeler (1994) look at aggressive signalling and territoriality, stating that their methodology “is pitched at an intermediate level between, on the one hand, abstract theories based on mathematical models and, on the other hand, empirical observations in complex environments” (p. 464). Hurd (1995) is also interesting in this respect; he re-examines Grafen’s work on the handicap principle using both a simplified game-theoretic model *and* a computer simulation.

Theoretical biologists like Maynard Smith (1982) are often well aware that game theory’s emphasis on stable structures has meant that the more interesting dynamics of evolution are not addressable. While the problem of just how one can be said to explain or understand a complex, dynamical system is still an open question (see section 4), ALife at least provides some tools and a foothold.

References

Abraham, R., & Shaw, C. (1992). *Dynamics—The Geometry of Behavior.* Addison-Wesley, Redwood, CA.

Bennett, J. (1976). *Linguistic Behaviour.* Cambridge University Press, Cambridge.

Bickerton, D. (1994). Origin and evolution of language. In Asher, R. E. (Ed.), *The Encyclopedia of Language and Linguistics*, pp. 2881–2883. Pergamon Press, Oxford.

Braithewaite, R. B. (1953). *Scientific Explanation.* Cambridge University Press, Cambridge.

Bromberger, S. (1962). An approach to explanation. In Butler, R. S. (Ed.), *Analytical Philosophy—Second Series*, pp. 72–105. Blackwell.

Bromberger, S. (1966). Why-questions. In Colodny, R. G. (Ed.), *Mind and Cosmos*, pp. 86–111. University of Pittsburgh Press, Pittsburgh, PA.

Burghardt, G. M. (1970). Defining ‘communication’. In Johnston, Jr., J. W., Moulton, D. G., & Turk, A. (Eds.), *Communication by Chemical Signals*. Appleton-Century-Crofts, New York.

Chomsky, N. (1968). *Language and Mind.* Harcourt, Brace and World, New York.

Chomsky, N. (1975). *Reflections on Language*. Pantheon Books, New York.

Clark, A. (1996). Happy couplings: Emergence and explanatory interlock. In Boden, M. A. (Ed.), *The Philosophy of Artificial Life*, pp. 262–281. Oxford University Press, Oxford.

Cummins, R. (1994). Functional analysis. In Sober, E. (Ed.), *Conceptual Issues in Evolutionary Biology*, pp. 49–70. MIT Press / Bradford Books, Cambridge, MA.

de Bourcier, P., & Wheeler, M. (1994). Signalling and territorial aggression: An investigation by means of synthetic behavioural ecology. In Cliff, D., Husbands, P., Meyer, J.-A., & Wilson, S. W. (Eds.), *From Animals to Animats 3: Proceedings of the Third International Conference on Simulation of Adaptive Behavior* Cambridge, MA. MIT Press / Bradford Books.

Dennett, D. C. (1987). *The Intentional Stance.* MIT Press / Bradford Books, Cambridge, MA.

Feyerabend, P. K. (1991). *Three Dialogues on Knowledge.* Blackwell, Oxford.

Gould, S. J., & Lewontin, R. C. (1979). The spandrels of San Marco and the panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society*, *B 205*, 581–598.

Grafen, A. (1990). Biological signals as handicaps. *Journal of Theoretical Biology*, *144*, 517–546.

Harper, D. G. C. (1991). Communication. In Krebs, J. R., & Davies, N. B. (Eds.), *Behavioural Ecology: An Evolutionary Approach* (Third edition)., chap. 12, pp. 374–397. Blackwell, Oxford.

Hasson, O. (1994). Cheating signals. *Journal of Theoretical Biology*, *167*, 223–238.

Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, *15*, 135–175.

Hendriks-Jansen, H. (1996). In praise of interactive emergence, or why explanations don’t have to wait for implementations. In Boden, M. A. (Ed.), *The Philosophy of Artificial Life*, pp. 282–299. Oxford University Press, Oxford.

Hurd, P. L. (1995). Communication in discrete action-response games. *Journal of Theoretical Biology*, *174*, 217–222.

Krebs, J. R., & Dawkins, R. (1984). Animal signals: Mind reading and manipulation. In Krebs, J. R., & Davies, N. B. (Eds.), *Behavioural Ecology: An Evolutionary Approach* (Second edition)., chap. 15, pp. 380–402. Blackwell, Oxford.

Langton, C. G. (1987). Artificial life. In Langton, C. G. (Ed.), *Proceedings of the Interdisciplinary Workshop on the Synthesis and Simulation of Living Systems (ALIFE ’87)*, Vol. VI of *SFI Studies in the Sciences of Complexity*, pp. 1–48. Addison-Wesley.

Lewis, D. B., & Gower, D. M. (1980). *Biology of Communication.* Blackie, Glasgow.

Maynard Smith, J. (1982). *Evolution and the Theory of Games*. Cambridge University Press, Cambridge.

Maynard Smith, J. (1983). Adaptationism and satisficing. *Behavioral and Brain Sciences*, *6*, 370–371.

Maynard Smith, J., & Harper, D. G. C. (1995). Animal signals: Models and terminology. *Journal of Theoretical Biology*, *177*, 305–311.

Miller, G. F. (1995). Artificial life as theoretical biology: How to do real science with computer simulation. Cognitive science research paper 378, School of Cognitive and Computing Sciences, University of Sussex.

Nagel, E. (1961). *The Structure of Science: Problems in the Logic of Scientific Explanation.* Routledge and Kegan Paul, London.

Seyfarth, R., Cheney, D. L., & Marler, P. (1980). Monkey responses to three different alarm calls: Evidence of predator classification and semantic communication. *Science*, *210*, 801–803.

Shannon, C. E., & Weaver, W. (1949). *The Mathematical Theory of Communication.* University of Illinois Press, Urbana.

Sober, E. (1993). *Philosophy of Biology.* Dimensions of Philosophy Series. Oxford University Press, Oxford.

van Fraassen, B. C. (1980). *The Scientific Image.* Oxford University Press, Oxford.

van Gelder, T. (1991). Connectionism and dynamical explanation. In *Proceedings of the 13th Annual Conference of the Cognitive Science Society*, pp. 499–503. Chicago.

von Neumann, J., & Morgenstern, O. (1953). *Theory of Games and Economic Behavior.* Princeton University Press, Princeton.

Wiley, R. H. (1983). The evolution of communication: Information and manipulation. In Halliday, T. R., & Slater, P. J. B. (Eds.), *Communication*, Vol. 2 of *Animal Behaviour*, pp. 156–189. Blackwell, Oxford.

Wittgenstein, L. (1953). *Philosophical Investigations*. Macmillan, New York.

Wright, L. (1994). Functions. In Sober, E. (Ed.), *Conceptual Issues in Evolutionary Biology*, pp. 27–48. MIT Press / Bradford Books, Cambridge, MA.

Zahavi, A. (1975). Mate selection—a selection for a handicap. *Journal of Theoretical Biology*, *53*, 205–214.

Zahavi, A. (1977). The cost of honesty (further remarks on the handicap principle). *Journal of Theoretical Biology*, *67*, 603–605.

In: Noble, J. & Parsowith, S. R. (Eds.). (1996). The Ninth White House Papers: Graduate Research in the Cognitive and Computing Sciences at Sussex. Cognitive Science Research Paper 440, School of Cognitive and Computing Sciences, University of Sussex.

The WWW as an Enabling Technology for Synchronous Collaborative Work

Sara R. Parsowith
sarap@cogs.susx.ac.uk

School of Cognitive & Computing Sciences
University of Sussex
Brighton
BN1 9QH

Abstract There are several products in existence that are designed to support both synchronous and asynchronous collaborative work. However, multi-user systems available on the WWW are predominantly devised for asynchronous co-working. This is largely a function of the infrastructure of the WWW, designed for information sharing among distributed individuals. It can be argued that the asynchronous exchange of information on the WWW is a relatively passive method of working. It is therefore important to examine the most appropriate extensions that enable active, synchronous collaboration on the WWW. Features from existing systems will influence the design of a prototype that supports awareness and hence encourages the development of a shared understanding between co-workers with a specific work goal. The theoretical basis for the design of such a system is rooted in the analysis of group interactions. This paper therefore firstly discusses the processes of shared understandings, negotiation and conflict before outlining the design of the proposed system.

1 Introduction

Wilson (1991) asserts that Computer Supported Co-operative Work (CSCW) combines an understanding of how people work together with the enabling technologies that are designed to support group-work. Research concerned with CSCW system design should therefore be influenced by an analysis of the dynamics of group interactions and the process of collaboration. McConnell (1994) argues that working in a group encourages members to think about their goals and what they are trying to achieve by collaborating with each other. Indeed Roschelle and Teasley (1995) consider collaboration to be the result of a "continued attempt to construct and maintain a shared conception of a problem." Co-operative working thereby helps group members to clarify their own thoughts and ideas through discussion with others. Such discussion fosters the sharing of different perspectives between group members and encourages the development of unified goals. The notion that collaborators need to consciously strive towards effective collaboration is highlighted by Beck (1994). It is argued that although group members need to acknowledge the existence of a given collaboration, this alone is not sufficient; it is possible to be aware of another person's work without actively collaborating with them. It can therefore be proposed that a necessary prerequisite for effective collaboration is a conscious effort to work towards a common goal. Successful collaboration can therefore be characterized by the creation of a collated viewpoint and not just as an exchange of information between individuals.

Ellis, Gibbs, and Rein (1993) argue that groupware products function to "support groups of people engaged in a common task (or goal)". Kaye (1992) notes that this does not require the physical presence or co-location of participants; CSCW enables distributed users to work concurrently on a shared task. Computer systems that are designed to support co-operative work should therefore enhance the process of developing a shared understanding. It is thereby necessary to examine existing tools that are intended to support collaborative co-working in order to evaluate how efficiently they assist collaboration. Firstly, it is important to discuss the fundamentals of collaboration in order to facilitate CSCW system design.

2 Intra-group collaboration

McCarthy, Miles, and Monk (1991) highlight the fact that co-operative working is both advantageous and difficult; co-working is advantageous as the processing capabilities of a group are greater than that of the individual since there are an increased number of viewpoints and skills. However, the presence of multiple viewpoints means that it is necessary to co-ordinate these varying expectations and views. Within the context of co-operative working it is therefore productive to consider how to make maximum use of the increased opinions and abilities available to group members. At the same time the aim should also involve minimising the problems of co-ordinating joint activity and facilitate successful collaboration.

2.1 Common ground

As previously mentioned, communicating individuals come to a group with a set of diverse notions and ideas that need to evolve into a developed shared perspective for successful collaboration. By obtaining "common ground" group members are able to anticipate the beliefs and actions of their co-workers which in turn functions to guide co-operative work. Clark and Schaefer (1989) consider "common ground" to be the mutual knowledge, beliefs and assumptions held by group members. This notion is concerned with the way two or more people relate their common background and experiences to form a composite understanding of the beliefs held by others. Common ground is apparent when individuals understand the differing views held by others and hence are in a position to re-define their own notions in the light of newly presented perspectives. Individuals are still able to hold divergent views. However, there will be an ongoing, developing core knowledge of common understandings between the group as a whole.

2.2 Breakdowns

Having asserted that individuals have obtained a shared understanding, it is necessary to gain insight into the "claims we are making about the mental representations and cognitive states of those two people" (Easterbrook, 1994). Thus a mismatch between one participant's expectations and the actions of another can be due to an error of communication or of perception by either party. Such breakdowns can also occur due to a difference of understanding of the situation. For example, Schrage (1990) points out that confusion often arises because "the same words meant different things to different people." Breakdowns therefore force individuals to "consider explicitly what had previously been assumed: that they share an understanding of the situation" (Easterbrook, 1994), when in fact a misunderstanding has arisen. Thus a breakdown in communication can lead to intra-group conflict.

2.3 Conflicts

Conflict is often identified with fighting, with implications of winning or losing. In a group situation where conflict occurs, participants need to "move from seeing each other as opponents to working together as collaborators" (Tillett, 1991). Resolution can therefore be viewed as a necessary prerequisite to effective collaboration; participants who are in conflict do not have a good basis for negotiation. Conflict itself can yield several positive outcomes. For example, Tillett (1991) argues that these include stimulation of interest and curiosity, prevention of stagnation, and providing encouragement and motivation to examine and solve problems as they arise. Bisno (1988) further argues that intra-group conflict encourages a strong group identity since there is a "development of a sense of solidarity among members of groups engaged in conflict". This coupled with a "heightened enthusiasm and purpose among the conflicting participants" makes it possible for a group consensus to emerge from a conflicting situation. It is consequently logical to employ negotiation tactics when group conflicts arise.

2.4 Negotiation

The process of negotiation involves two or more parties with conflicting interests attempting to come to an agreement on how they can co-operate. Dillenbourg, Baker, Blaye, and O'Malley (1996) define negotiation as the process by which collaborators “attempt (more or less overtly or consciously) to attain *agreement* on aspects of the task domain”. Collaborators can therefore resolve their conflicts by gaining an understanding of what each party comprehends about a given situation. By discussing the breakdowns, group members are actively working towards resolution. This highlights the fact that verbal skills are essential to conflict resolution, which requires both the ability to clearly communicate and to clearly understand communication. CSCW systems should therefore incorporate features that encourage the exchange of dialogue to facilitate opportunities for negotiation.

Having argued that co-ordination breakdowns are likely to occur in collaborating groups and that such conflicts can be productive, it is logical to posit that systems should be designed whereby conflict and negotiation are both encouraged and controlled.

3 Shared understandings and CSCW system design

To obtain a shared understanding, it is necessary for collaborating individuals to have an awareness of the other group members. Dourish and Belotti (1992) define awareness simply as an “understanding of the activities of others”. Awareness therefore enhances the likelihood that individual contributions will be relevant to the group’s activity as a whole. A shared space is therefore a useful tool for the generation of shared understandings; shared spaces “create the aura of copresence: they make collaborators feel like they’re together, even if they’re not” (Schrage, 1990). Schrage (1990) further iterates that such a space is “an excellent tool to list the areas of agreement, disagreement, and solutions for several parties.” CSCW systems should consequently be designed to help maximize the developing awareness that participants have of each other. One method of achieving this is to encourage brainstorming, whereby a group of individuals collectively attempt to generate ideas. This is particularly valuable if participants are jointly pursuing a common goal such as the completion of a co-written document. Roschelle and Teasley (1995) argue that brainstorming encourages students to think laterally and hence expand on incomplete or partial ideas. In addition, Schrage (1990) maintains that computer-augmented brainstorming sessions “provide opportunities to link ideas in unusual ways and let people know what has already been said” which provides further validation of the benefits of shared workspaces.

Ross, Ramage, and Rogers (1995) propose that common ground is achieved when speakers formulate their contributions in relation to their awareness of what the other group members do and do not comprehend. It is therefore necessary to understand the mechanisms that speakers use to establish and maintain common ground during conversations. Plowman (1995) argues that speech encourages the process of collaboration; intra-group talk provides members with explanations about each person’s reasoning and functions to enhance group understandings. It is consequently essential to keep track of conversations since speech results in a quicker, more efficient development of common ground. Spoken communication is thus an essential component of co-operative tasks, even if largely text-based. Computer systems need to be equipped with features that enable ‘speech’ such as the addition of an in-built ‘chat’ facility. This is where a shared space for “chatting” with co-workers is provided, achieved by typing in text and having a ‘conversation’. The additional function of a chat space can enhance the sharing of information since ideas can be ‘discussed’ with others. Private chat facilities further improve functionality by enabling members of a group to speak privately to another individual as and when they require.

4 Non WWW-based systems

A number of CSCW systems have been designed that allow for the development of shared understandings, negotiation and conflicts between collaborators, primarily through communal workspaces.

4.1 The Coordinator

Grantham and Nichols (1993) carried out studies with The Coordinator[1], a computer-based system to facilitate the exchange, clarification and negotiation of commitment between co-workers within organizations. The system design is based on the notion that language is an activity and not merely a transmission of information.[2] The design is also based on the assumptions that conflict arises from misunderstandings, is inevitable and "possibly the belief that conflict itself is productive" (Easterbrook, Beck, Goodlet, Plowman, Sharples, & Wood, 1993). However, The Coordinator was found to be poorly integrated into the work environment. This is largely due to the fact that the system largely ignored whether a person was attempting negotiation or merely avoiding confrontations. The system is poorly designed for encouraging the development of a shared understanding as it neglects negotiation, a crucial component of effective collaboration. The Coordinator thus enables you to communicate effectively with another individual but does not ensure that a successful collaboration will ensue.

4.2 Cognotor Colab

Cognotor[3] was designed to help organize ideas for papers, presentations, talks and reports. The interface consists of WYSIWIS[4] mini-whiteboards. The system aimed to incorporate four stages of idea formulation: brainstorming, the organization of ideas, the evaluation of ideas and generating an outline. Participants are able to choose an empty space on the shared screen and type in ideas that they have; all other participants can see what has been typed. People can therefore annotate items and elaborate on a given idea as they wish.

4.3 Argnotor

Argnotor is an argumentation spreadsheet designed on the premise that many misunderstandings and disputes derive from personal positions, understated assumptions and unstated criteria. Argnotor is consequently designed to make these all explicit and represented in the shared space. Schrage (1990) argues that the software encourages participants to create a shared understanding with their colleagues, both visually and orally.

The afore mentioned products highlight the need for shared workspaces for collaborative work and as such provide implementation guidelines.

5 Synchronous collaboration and the WWW

5.1 The structure of the WWW

The WWW provides access to worldwide distributed information on the internet, enabling people to access globally distributed information by means of hypertext links. WWW interfaces and networks are global and hence are ideal for information sharing among distributed workers. This idea is in accordance with Chiu and Griffin (1995) who consider the WWW to be "an attractive platform for the development of groupware" since the core technology is based on a set of universally accepted standards.

The WWW is designed not for synchronous collaboration but for information *sharing* through static pages. Although asynchronous communication is a good method for enabling co-workers to exchange ideas and drafts, it does not promote critical thinking or discussion. The WWW is typically used in this manner since it is designed to be a worldwide store of information as opposed to an interactive medium. Busbach, Kerr, and Sikkel (1996) argue that the asynchronous exchange of information is a relatively passive method of working since there is no direct interaction involved; thus although the WWW allows people to work co-operatively, following links does not promote an awareness of what other people

[1] Designed by Winograd (1988).

[2] See Winograd and Flores (1986) for an overview of Speech Act Theory.

[3] Designed by Stefik and Seely Brown (1989).

[4] What You See Is What I See.

are doing. The WWW can therefore be perceived as a sprawl of information that requires additional awareness features to support synchronous collaborative work.

There are a large number of systems designed for asynchronous collaborative work already existing on the WWW. A good example is the Basic Support for Cooperative Work (BSCW) shared information system (see Bentley, Busbach, & Sikkel, 1996). The BSCW system is integrated into the existing structure of the WWW; a workspace can be accessed directly with common WWW browsers. However, as previously discussed, asynchronous collaboration is not conducive to promoting awareness of co-workers. 'Real-Time' working enables users to see exactly what other users are doing and enables the concurrent collating of ideas and notes, giving them the perception that they are working "in the same room" (Grenier & Metes, 1992).

Dix (1996) points out that the WWW is already a successful application and that it is important not to lose this success. Hence rather than design a completely new system, it is sensible to merely develop extensions to the WWW in order to accommodate the demands of synchronous collaboration. WWW browsers are not platform specific and thus form a good basis for distributed inter-organizational working. Also there is a copious amount of information available on the WWW making the sharing of distributed information relatively simple. The WWW already has a critical mass of users (Dix, 1996) hence it is important to create 'add-ons' to the web to support group working rather than redesign an already successful application. This is supported by Grudin (1994) who suggests that there is a need to incorporate existing features of single-user applications into CSCW applications in order to take advantage of user familiarity with these particular aspects.

5.2 Systems on the WWW for synchronous collaboration

There are several systems already functioning on the WWW that aim to support synchronous collaboration. Here is a brief overview of some of these:

Frivold, Lang, and Fong (1994) have combined the asynchronous access to information offered by the WWW with a synchronous conferencing tool called COMET (Collaborative Multimedia Environment Technology). Their rationale is that such a combination enables users not only to "browse through a wealth of static information, but also to contact the authors and discuss this information with them as a natural extension of the browsing process." The result is the creation of a shared workspace that permits users to 'talk' to each other as well as see and interact with each other's applications. The system is an example of how to successfully 'bridge the gap' between synchronous and asynchronous methods of working.

Roscheisen, Mogensen, and Winograd (1996) have introduced the concept of SOAPs (Seals of Approval). This is a peer review of ideas and a critique of any shared information, described as being a source of "meta-information" since it involves creating a document containing a rating that describes another document. The shared comments consequently become available on the WWW; icons provide a link to an HTML page with the annotation text. This system is thus a good example of how awareness features can be incorporated into a CSCW system; it provides access to an ongoing store of collaboratively developed information.

Alliance (Decouchant & Romero Salcedo, 1996), permits several users located on different WWW sites to co-operate and produce documents in a structured way. It assigns users with different 'roles' such as the reader role (permits read-only access of a document) and writer (allows modification of a fragment). The same user can have different roles on different fragments. The system highlights the fact that it should be possible to permit or deny access to a given document as necessary. This ensures that collaborating groups can have a degree of privacy when working.

Ingvarsson (1995) discusses how Java[5] can play an active role in extending the WWW for synchronous collaboration. Java eliminates the need to send information from the client to the server for interaction to ensue. The Java language is also well suited to provide interactive content via the WWW,

[5] See http://java.sun.com for an overview of the language.

due to its platform independent nature.

Systems on the WWW designed for synchronous collaboration are starting to emerge:

GroCo is an Electronic Meeting System (EMS) developed by Walther (1995). It consists of shared interactive WWW pages which are displayed in a browser for each participant. The system is implemented in Java and conference applets enable a controlled textual chat between members as well as use of a shared whiteboard. The system is thereby designed to support synchronous work and awareness of other participants.

Another development is TeamRooms (Roseman & Greenberg, 1996). The system uses the metaphor of room spaces. Each 'room' contains a chat tool and a backdrop acting as a shared whiteboard. Applets have been designed to support group work activities such as brainstorming and drawing. When team members are in the room at the same time they see each other through changes in the room's artifacts and by means of multiple telepointers. Awareness is supported by windows that show who is around in a given room.

The Como project is dedicated to Java-based interactive internet communication. It is at present in the prototype phase[6] The product consists of a whiteboard, a chat facility and an appointment scheduler.

6 The proposed system

6.1 Features to implement

Features of the afore mentioned systems have been utilized for proposing the following required features:

Chat facility Conversation appears to be an essential tool for the development of a shared understanding between collaborators. In computer-based tasks the addition of a chat space or a telephone link will ensure that this facility is not lost to individuals that are co-authoring documents. The common ground that develops between people will be continually updated as new information is accrued. Chat applets need to be programmed in order to support 'real-time' conversations and hence provide a mechanism for the exchange of information.

Whiteboard At least one shared whiteboard to encourage shared annotations needs to be implemented in Java and integrated into the system to allow for synchronous editing and brainstorming of ideas.

Security Authentication for users is essential and so the system should provide security features whereby members need to register to enter the system. Also the assignment of 'roles' should be possible so that people can be denied access to particular meetings/documents as required.

Shared viewing Collaborators also need to view a particular WWW site at the same time as other group members. Telepointers will help users reference particular parts of the document.

Awareness Features will be incorporated to enable users to see who is logged on and what changes have been made to documents within a shared workspace. A long term goal is to implement video facilities to maximize awareness of others.

A preliminary design plan is outlined in the Appendix.

7 Conclusion

It has been shown that conflicts between co-operating workers are inevitable since people can hold such varied viewpoints. However, conflicts function to encourage groups to openly discuss their opinions and goals. Breakdowns in communication will occur when members realize that they do not have a shared understanding with the other people in the group. Conflicts force collaborators to re-think their understandings and thereby act as a mechanism to promote shared understandings. This in turn functions to

[6] see www4.informatik.uni-erlangen.de/IMMD-IV/Projects/como/ for details.

enable the group to carry out their designated work goal in the light of newly developed understandings. The theoretical principles that underlie intra-group collaboration can be used to facilitate CSCW system design. Existing systems can show how best to support negotiation, awareness and brainstorming activities. The proposed system is intended to support synchronous collaboration as this is considered to be more effective than asynchronous work for promoting group cohesion.

The WWW is not designed for synchronous collaboration but for information sharing through static WWW pages. To enable synchronous collaboration it is necessary to make use of extensions that are capable of supporting active collaboration between users. The proposed system will be developed in Java to support the synchronous co-editing of documents on the WWW. The aim is to incorporate features that encourage the sharing of ideas and promote negotiation between users. The system will be designed for groups to work towards a common work goal and hence need to gain a shared understanding of what is required. The system should therefore help users reach common ground with respect to their understandings of both the task in hand and the ideas held by the other group members.

References

Beck, E. (1994). Practices of collaboration in writing and their support. Cognitive science research paper 340, School of Cognitive and Computing Sciences, University of Sussex.

Bentley, R., Busbach, U., & Sikkel, K. (1996). The architecture of the BSCW shared workspace system. In Busbach, U., Kerr, D., & Sikkel, K. (Eds.), *CSCW and the Web—Proceedings of the 5th ERCIM/W4G Workshop*, pp. 31–42. GMD.

Bisno, H. (1988). *Managing Conflict.* Sage Publications, London.

Busbach, U., Kerr, D., & Sikkel, K. (1996). Foreword. In Busbach, U., Kerr, D., & Sikkel, K. (Eds.), *CSCW and the Web—Proceedings of the 5th ERCIM/W4G Workshop*, pp. 31–42. GMD.

Chiu, D., & Griffin, D. (1995). Workgroup forum: Tools and applications for WWW-based group collaboration. http://www.w3.org/hypertext/WWW/Collaboration/Workshop/Proceedings/P3.html.

Clark, H., & Schaefer, E. (1989). Contributing to discourse. *Cognitive Science*, *13*, 259–294.

Decouchant, D., & Romero Salcedo, M. (1996). Alliance: A structured cooperative editor on the web. In Busbach, U., Kerr, D., & Sikkel, K. (Eds.), *Proceedings of the 5th ERCIM workshop on CSCW and the Web*, pp. 7–12. GMD.

Dillenbourg, P., Baker, M., Blaye, A., & O'Malley, C. (1996). The evolution of research on collaborative learning. In Reimann, P., & Spada, H. (Eds.), *Learning in Humans And Machines*, pp. 189–211. Pergamon, Elsevier Science, Oxford.

Dix, A. (1996). Challenges and perspectives for cooperative work on the web. In Busbach, U., Kerr, D., & Sikkel, K. (Eds.), *CSCW and the Web—Proceedings of the 5th ERCIM/W4G Workshop*, pp. 143–157. GMD.

Dourish, P., & Belotti, V. (1992). Awareness and coordination in shared workspaces. In *CSCW'92 Proceedings*, pp. 107–114, New York. ACM.

Easterbrook, S. (1994). Coordination breakdowns: Why groupware is so difficult to design. Cognitive science research paper 343, School of Cognitive and Computing Sciences, University of Sussex.

Easterbrook, S., Beck, E., Goodlet, J., Plowman, L., Sharples, M., & Wood, C. (1993). A survey of empirical studies of conflict. In Easterbrook, S. (Ed.), *CSCW: Cooperation or Conflict?*, pp. 1–68. Springer-Verlag, London.

Ellis, C. A., Gibbs, S. J., & Rein, G. L. (1993). Groupware: Some issues and experiences. In Baecker, R. M. (Ed.), *Readings in Groupware and Computer-Supported Cooperative Work*, pp. 39–58. Morgan Kaufmann, San Mateo.

Frivold, T. J., Lang, R. E., & Fong, M. W. (1994). Extending WWW for synchronous collaboration. http://www.ncsa.uiuc.edu/SDG/IT94/Proceedings/CSCW/frivold/frivold.html.

Grantham, C., & Nichols, L. (1993). *Communication: The Matrix of Interaction*, pp. 19–49. Van Nostrand Reinhold, New York.

Grenier, R., & Metes, G. (1992). *Enterprise Networking: working together apart.* Digital Equipment Corporation, US.

Grudin, J. (1994). Eight challenges for developers. *Communications of the ACM*, *37*, 93–105.

Ingvarsson, M. (1995). Extending the WWW to support synchronous collaboration. http://www.sisu.se/ magnusi/thesis/multiuser.html.

Kaye, A. (1992). Learning together apart. In Kaye, A. (Ed.), *Collaborative Learning Through Computer Conferencing: The Najaden Papers*. Springer-Verlag, Berlin Heidelberg.

McCarthy, J., Miles, V., & Monk, A. (1991). An experimental study of common ground in text-based communication. In Robertson, S., Olson, G., & Olson, J. (Eds.), *Proceedings of ACM CHI '91 Conference on Human Factors in Computing Systems*.

McConnell, D. (1994). *Implementing Computer Supported Cooperative Learning*. Kogan Page Ltd., London.

Plowman, L. (1995). The interfunctionality of talk and text. *Journal of Computer Supported Cooperative Work (CSCW)*, *3*, 1–18.

Roscheisen, M., Mogensen, C., & Winograd, T. (1996). Shared web annotations as a platform for third-party value-added information providers: Architecture, protocols, and usage examples. http://www-diglib.stanford.edu/rmr/TR/TR.html.

Roschelle, J., & Teasley, S. (1995). Construction of shared knowledge in collaborative problem solvingolving. In O'Malley, C. (Ed.), *Computer-supported collaborative learning*. Springer-Verlag.

Roseman, M., & Greenberg, S. (1996). Teamrooms: Groupware for shared electronic spaces. In *Human Factors in Computing System: CHI 96 Conference Companion*, pp. 275–276. Addison Wesley.

Ross, S., Ramage, M., & Rogers, Y. (1995). PETRA: Participatory evaluation through redesign and analysis. Cognitive science research paper 375, School of Cognitive and Computing Sciences, University of Sussex.

Schrage, M. (1990). *Shared minds : the new technologies of collaboration.* Random House, New York.

Stefik, M., & Seely Brown, J. (1989). Toward portable ideas. In Olson, M. (Ed.), *Technological support for work group collaboration.* Lawrence Erlbaum Associates Inc, Hillsdale, NJ.

Tillett, G, D. (1991). *Resolving Conflict: a practical approach.* Sydney University Press.

Walther, M. (1995). Groco WWW page. www.dstc.bond.edu.au/ walther/groco.

Wilson, P. (1991). *CSCW: An Introduction.* Intellect, Oxford.

Winograd, T. (1988). A language/action perspective on the design of cooperative work. *Human Computer Interaction*, *3*(1), 3–30.

Winograd, T., & Flores, F. (1986). *Understanding Computers and Cognition.* Ablex, Norwood.

Appendix: Preliminary Design Plans

Aim

To utilize WWW extensions to support the processes underlying synchronous collaborative writing, particularly brainstorming activities and to promote awareness and shared understandings amongst users.

Features

The following features will be implemented:

- The system will be WWW-based
- The programming will be in Java
- Support for brainstorming will be provided
- The system will have built-in awareness facilities
- Shared editing of documents will be supported
- Whiteboard facilities will be implemented
- Real time conversations will be supported
- Concurrent viewing of HTML documents will be possible
- Video communication (eg: MBone) will be incorporated
- Security add-ons will provide authentication of users and restricted access to documents

Requirements

- co-authoring
- brainstorming
- awareness
- chatting
- concurrent document viewing
- videoconferencing

Users

- Distributed groups who are co-authoring written documents both in academia and in industry.

Strategy

A hybrid approach between user-based and software engineering methods of design will be taken. The initial prototype will be evaluated and re-designed as necessary.

In: Noble, J. & Parsowith, S. R. (Eds.). (1996). The Ninth White House Papers: Graduate Research in the Cognitive and Computing Sciences at Sussex. Cognitive Science Research Paper 440, School of Cognitive and Computing Sciences, University of Sussex.

Investigating a Dynamic Mutation Rate Genetic Algorithm

Oliver Sharpe
olivers@cogs.susx.ac.uk

School of Cognitive & Computing Sciences
University of Sussex
Brighton
BN1 9QH

Abstract The genetic algorithm (GA) has proved a useful search tool in many different fields, however there has been much written about the limitations of GAs in the context of optimization. One of the major concerns is that a mismatch between the fitness landscape of the function being optimized, the selection method used, and the mutation rate chosen can lead to far less than optimal performance of the GA. The mutation rate and selection method interact in a manner dependent on the particular fitness landscape. It is often difficult to determine a priori the most appropriate mutation rate for a given selection method. Although the mutation rate can be fine tuned on the basis of empirical studies, generating the necessary data can be an expensive process. In this paper the arguments for using a self-setting and correcting dynamic mutation rate genetic algorithm (DMRGA) are put forward. A specific attempt at just such an algorithm is outlined and some early but quite positive results presented.

1 Introduction

This paper is concerned with mutation rates in genetic algorithms. A fair degree of familiarity with genetic algorithms (GAs) is assumed (for an introduction to GAs see Goldberg (1989)). The paper will relate to GAs only in the context of optimizing a given function, where mutation is the only genetic operator. I will continue to refer to this algorithm as a GA rather than an Evolutionary Algorithm as in later work I intend to reintroduce a recombination operator to the algorithm. At this point in time recombination has been removed to simplify the algorithm and the resultant population dynamics so that they can be more fully understood. The majority of conventional GAs have a fixed mutation rate which is held constant both across the population and over the generations. Various limitations on the effectiveness of such GAs have been explored (Kauffman, 1993) (Muehlenbein, 1992) and are generally related to a mismatch between the mutation rate, the selection method and the fitness landscape being used by the GA to optimize the given function. The major concerns are the mutation/selection catastrophe, the complexity catastrophe and the time it takes the GA to find a solution. The first two involve the mutation rate being too high for a given fitness landscape. Selection provides a converging pressure on the population of the GA, whereas mutation provides a diverging pressure. If the mutation rate is too high then the divergent pressure will be weaker than the convergent pressure and the population will disperse away from good solutions. This is the concern of the mutation/selection catastrophe. The complexity catastrophe is concerned with useful correlations within the landscape structure. Any fitness landscape will have a maximum correlation distance, although this can be infinite. The fitness of any two points with greater distance between them is totally uncorrelated. Hence if the mutation rate is so high that offspring are beyond this distance from their parent, then there will be no correlation between the fitness of the two and the GA is effectively performing random search. This is obviously undesirable. To totally avoid these two catastrophes the mutation rate could be set to a very small value. Then with a large enough population and given enough generations the mutation rate will not be to blame if the global optimum

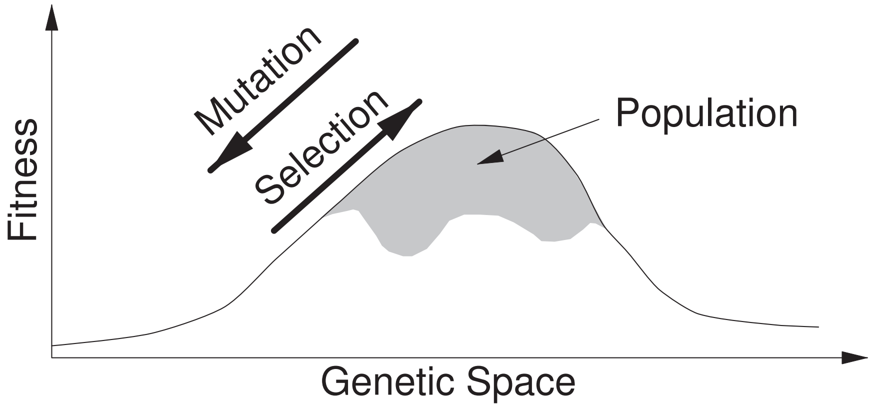


Figure 1: A caricature of the pressures that mutation and selection put on the population.

is not reached. The problem here is that the GA will take too long to run. This could be called the 'life is too short' catastrophe. To speed up the rate of evolution, a higher mutation rate is needed. Hence a balance needs to be found, and the most effective mutation rate will crucially depend on the nature of the fitness landscape (Kauffman, 1993) and the selection method (Harvey, 1994).

2 How selection and mutation work in a GA

In a GA with only mutation and selection, it is the interplay between the generative nature of the mutation method and the evaluative nature of the selection method that drives the population of the GA across the fitness landscape towards better solutions of the problem expressed by the fitness function (see Figure 1). Hence choice of the mutation and selection methods is crucial in determining, on average, how successful the GA will be. Success of an optimizing GA is usually measured either as the average time needed to find the global optimum, or the average fitness value obtained after a given time. Both measurements have to be average performance values as GAs are a stochastic search method. Also care must be used to ensure that two such values can be meaningfully compared. More will be said on this later.

3 Why bother looking at dynamic mutation rates?

The various catastrophes outlined by Kauffman (1993) and the delicate relation between mutation rates and GA performance (Muehlenbein, 1992) puts a serious question mark over the power of GAs to find novel solutions to complex problems. All these problems are based on a mismatch between the mutation rate and the shape of the local landscape in the context of the given selection method, so a dynamic mutation rate GA could potentially avoid them. However, dynamic mutation rates can have their own problems of a similar nature, such as over-fitting to the landscape, and introducing unexpected and undesirable pressures on the evolutionary system. One of the strengths of a constant mutation rate is precisely that it is independent of the landscape on which it is operating. Hopefully, with careful design and consideration of the dynamic mutation rate mechanism and its interaction with the selective method in use, the problems of over fitting and unwanted evolutionary pressures can be largely avoided.

4 How could a dynamic mutation rate work?

The problem highlighted above is that we know what the effect of the mutation rate is on the success of the GA, but there is no direct way to calculate what the actual mutation rate for a given run should be. A dynamic mutation rate GA could get around this by monitoring the success of the current mutation rate and altering it accordingly. For example, if it could detect the effects of either the mutation/selection catastrophe or the complexity catastrophe then it should lower the mutation rate. Equally if the pace of progress towards an optimal solution is too slow then it should increase the mutation rate. This, as always, is easier said than done, and it is not the only possible approach. Fogarty's successful early work in this area (Fogarty, 1989) varied the mutation rate according to a time dependant schedule rather than attempting to measure the landscape. Another less successful approach, used by Tuson (1995), is to try to measure how successful the various genetic operators are at a given time during a run and then alter the algorithm's parameters accordingly. Whatever the approach there are two main issues that have to be addressed to make a dynamic mutation rate feasible: where should the change be affected and how should the change be calculated?

There are two possible candidates for where to affect the change. Either each individual has its own mutation rate or there is one changing mutation rate for the whole population. Since one of the reasons for having a dynamic mutation rate is to account for variations in the nature of different regions of the fitness landscape, which policy to choose will partly depend on how the population is distributed across the landscape. The algorithm in this paper is likely to set the mutation rate quite high at times, causing the population to disperse, hence different members of the population will be in different regions of the landscape possibly needing different mutation rates for optimal performance. Hence each individual has its own independent mutation rate.

The other problem is how to calculate the changes to the mutation rate. When a change is to occur, the new mutation rate can either be set to a new absolute value, calculated independently of the current mutation rate, or the mutation rate can be increased or decreased in a relative manner. Whatever the choice, care must be taken to ensure that the calculation used does not presuppose any knowledge of the function being optimized. Fogarty (1989) used a generational time dependant schedule to alter the mutation rate, and hence does not measure the landscape in any way. If measurements are to be taken then to avoid over fitting to the landscape, the only items of information that can be safely used are values that are slightly independent of the landscape's structure, such as the current mutation rate, the number of bits in the genome (N) and how long ago the last improvement occurred due to mutation. Tuson (1995) used 'operator productivity', in other words how successful a given genetic operator is at generating fitter offspring, as the gauge by which the probabilities of using a given operator were altered over the time of a run. Unfortunately, Tuson reported that in almost all cases, and under almost all adaptive methods using this gauge, there was a detrimental effect on the performance of the GA.

A final concern is to ensure that the algorithm for calculating the new mutation rate does not get so complicated that it takes a long time to run. This could potentially spoil any benefits of using a dynamic mutation rate GA. With all of these issues in mind, the following is an example of a potentially very useful dynamical mutation rate genetic algorithm (DMRGA).

5 A comparative experiment

For a DMRGA to be worthy of further investigation it must be able to optimize any fitness function at least as efficiently as a conventional GA set with the optimal constant mutation rate for the given fitness landscape. To perform the necessary comparisons of performance a range of fitness functions with varying fitness landscapes is needed. The NK Landscapes, developed by Kauffman (1993), are appropriate for this task as they are capable of testing GAs on fitness landscapes of tunable ruggedness. Although the NK model does not fully capture the complexity of the neutral networks that arise from the fitness functions used in most ALife work, there are currently no simple theoretical models that do. Hence, the NK model is the ideal testbed for the initial trials of this algorithm.

In the model, *N* is the number of gene loci in a genome and K is the number of epistatic interactions between each gene and the N-1 other genes. An NK landscape where K=0 is equivalent to a unimodal fitness function and, at the other extreme, an NK landscape with K=N-1 is equivalent to a random fitness function. Kauffman used these landscapes to explore GAs which use a constant mutation rate. In this paper the same NK landscape model is used to compare the effectiveness of a DMRGA against a range of constant rate GAs each with a different mutation rate. Effectiveness is measured here as the average fitness achieved by the GA after a given time. It has to be average because GAs are a stochastic search method. The GAs will be tested on various landscapes of differing ruggedness. Since two NK landscapes with the same values of N and K can have different distributions of hills and valleys across the search space harbouring optima of differing absolute fitness values, hence a comparison between the success of two GAs over two different landscapes is meaningless. Therefore all the test runs that are to be compared must be performed on exactly the same fitness landscape.

5.1 The control GA

The constant mutation rate algorithm is being used as the control in the experiment against which the effectiveness of the DMRGA can be measured. So the only difference between the control and the experiment GA can be the method of setting the mutation rate. For the sake of context it is necessary, therefore, to outline the selection method and the mutation method (independent of the mutation rate) that is being used by both GAs. As it is the effect of dynamically altering the mutation rate that is being investigated, hence a selection method which has a relatively straightforward effect on the population is most appropriate. Any of the ranking selection techniques, argued above to have no grasp of absolute vertical scale, will suffice. Steady state tournament selection is the method chosen in this paper because it is has the added advantages of being easy to implement and it is often favoured in the ALife community (Sims, 1994) (Harvey, Husbands, & Cliff, 1994) (Reynolds, 1994). The tournaments are of size two and the winner replaces the loser with an offspring that is generated using the same mutation method in both GAs, although the mutation rate may be different. This gives a selective pressure of two choices (one choice the winner and one choice the loser) and hence the expected optimal mutation rate to overcome the selection/mutation catastrophe is an average of 2 bits per individual per generation (Harvey, 1995).

For reasons discussed below, which are related to the workings of the dynamical mutation rate, the mutation method chosen works by having an equal probability of mutating each bit of the genome. This probability is derived from the mutation rate, m, which is expressed as the average number of bits to be mutated per genome per generation. Thus the probability of mutating a given bit is calculated as the mutation rate, m, divided by the length of the genome, N. Hence for a genome with $N = 100$, each bit of the genome has an $m/100$ chance of being mutated when generating a new offspring. This mutation method results in their being a distribution of Hamming distances between parents and their offspring rather than a single value, a property that can be useful in combating Muller's ratchet whilst still exploring away from a local optimum (Harvey, 1994). For each landscape being tested, the control GA is run 30 times for each of 8 different constant mutation rates ranging from $m = 0.1$ to $m = 10.0$. This should ensure that the results obtained contain both the average value attained when the constant mutation rate is set optimally and a demonstration of the effects of setting the constant mutation rate either too high or too low.

5.2 The dynamic mutation rate genetic algorithm (DMRGA)

During the development of this DMRGA the algorithm has gone through many changes as new aspects of the pros and cons of dynamically altering the mutation rate have been realized. The eventual algorithm is a more dynamic approach than that of Fogarty (1989) as the fitness landscape is 'measured' and has an effect on the mutation rate whereas Fogarty's algorithm made changes to the mutation rate according to a fixed schedule. This DMRGA uses a combination of techniques to achieve improved performance, to avoid over-fitting to the landscape and, unlike Tuson's attempt (Tuson, 1995), to ensure that no unwanted pressures affect the evolutionary process.

Each individual goes through three phases in its 'life time' (that is if it survives long enough to experience all three !), 'explore', 'search' and 'leave' (Figure 2). The three phases have different effects on the mutation rate over time, which in turn has effects on the local subsets of the population. The first provides a pressure on exploring subsets of the population to converge onto local optima. The second

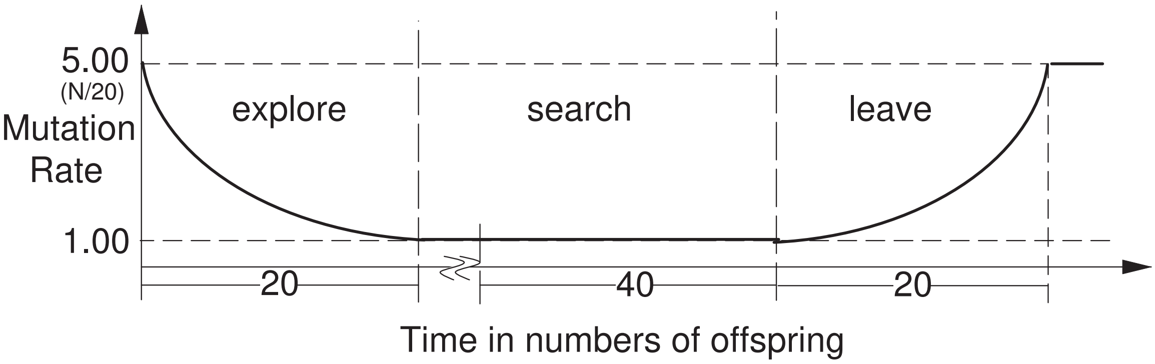


Figure 2: The three phases of an individual's lifetime in the Dynamic Mutation Rate GA

provides a pressure to search the local optima, and the third provides a dispersal or divergent pressure. The third phase counters the 'life is too short' catastrophe and is triggered by a recognition that the current local optimum has been adequately searched. The second phase avoids the mutation/selection and complexity catastrophes by providing a suitably small mutation rate for a long enough period of time to ensure that the current local optimum is adequately searched. This is where the nature of the mutation method (not rate) is most important. The first phase integrates the behaviours of the other two parts as all offspring start in this phase (the mutation rate is inherited from their parent at the time of 'birth' except for the initial population who start with a mutation rate of N/20). Different individuals of the population can be in different phases at the same time, each with its own mutation rate.

5.3 Mechanics of the algorithm

On top of the usual information held for each individual of the population, the DMRGA needs only two extra variables. A **(float)mutation_rate** and a **(int)last_success** variable. Together these two keep track of which phase the individual is in and, of course, the individual's mutation rate.

5.3.1 The explore phase

The explore phase is characterized by **mutation_rate** $>$ 1.0 and **last_success** = 0. The offspring of an individual in the explore phase inherits the identical values of **mutation_rate** and **last_success** as its parent. After every time that an individual in this phase reproduces, its own **mutation_rate** is multiplied by the constant **(float)down** $<$ 1.0. This slowly decreases its mutation rate over time. As soon as the individuals mutation rate is less than or equal to 1.0 it is out of the explore phase and into the search phase.

5.3.2 The search phase

The search phase is characterized by the **mutation_rate** = 1.0 and **last_success** $>$ -1. In this phase the mutation rate is held fixed at 1.0. Each time an individual reproduces the offspring has its mutation rate set at 1.0 but its **last_success** variable is set at double the value of the **last_success** variable of its parent. Both parent and offspring are allowed to know whether or not the offspring is of higher fitness than the parent. If the offspring is fitter, then **last_success** in both individuals is set to 0. Otherwise **last_success** is incremented in the parent. When **last_success** $>$ constant **T**, the individual moves from the search phase into the leave phase. How the value of **T** is set is discussed below, but it's basic purpose is to avoid the 'life is too short' catastrophe by recognizing when the individual's location (probably a local optimum) has been adequately searched and hence it is time to move into the 'leave' phase. To achieve this the algorithm sets **last_success** = -1.

5.3.3 The leave phase

The leave phase is characterized by **last_success** = -1. In this phase the offspring of an individual inherits only the **mutation_rate** of its parent and the offspring's **last_success** is set to 0. After every time that an individual reproduces, its own **mutation_rate** is multiplied by constant the **(float)up** $>$ 1.0. This slowly increases the mutation rate. If the mutation rate is above N/20 then it is set to N/20. This, after all is a large mutation rate, and all of its offspring will also have this mutation rate and hence this will soon generate new individuals that are far from the original parent. This is similar in some ways to seeding the population with randomly generated new individuals. In future work the choice of N/20 as the maximum mutation rate needs to be empirically or analytically confirmed as robust or altered accordingly. There are no criterion for leaving the leave phase except death.

5.4 Setting T

It is the value of T that determines how long an individual will spend searching for the top of the local optimum. Once the individual has had T offspring at the search phase mutation rate, which is an average of 1 bit per genome per generation, it will then move into the leave phase. To set T appropriately it is necessary to work out how long each individual should stay in the search mode to make it highly likely that the top of the local optimum is found by the population before the population disperses from the local optimum.

For the purposes of the experiments in this paper a fairly crude calculation of an appropriate value for T was performed based on the following argument. Assume that a search for the top of a local optimum starts with a single individual that is sitting one bit-flip away from the top of the local optimum. With a mutation rate of an average of 1 bit per genome per generation, roughly 1/3 of offspring will be identical to the parent and roughly another 1/3 will have had exactly one bit flipped. With a genome length of 100 bits, it crudely follows that after having 300 offspring it is likely that one of them will have reached the top of the local optimum. Conversely, if after 300 offspring none have achieved a higher fitness (the sign of reaching the local optimum) then it is likely that the original assumption was wrong and so the original individual was not one bit-flip (or more) away from the local optimum but was actually at the local optimum. Hence the local optimum has already been reached and it is time to allow the population to move away from this local optimum to explore for others.

Of course it is not essential that the original individual produces all of the 300 offspring, as 1/3 of its offspring are identical to itself. Hence it is only necessary to ensure that at least 300 offspring are produced by individuals who are genetically identical to the original individual. Taking into account the doubling of the **last_success** counter it can be shown that with the value of T set to 40 at least 300 such offspring should be produced.

This calculation is a very crude approximation resulting in a usable value for T which will suffice for the task at hand. However, in future work it is necessary to empirically or analytically determine the parameters for optimal performance on all landscapes. If this further work does not show that the parameters for a DMRGA are less sensitive (or are more easily calculable) than setting the mutation rate directly then this whole endeavour is interesting but pointless !

5.5 Setting down and up

The values of the two constants **(float)down** and **(float)up** determine how fast the mutation rate of an individual will decrease in the explore phase to the 1.00 value of the mutation rate during the search phase and then how fast the mutation rate will increase away from 1.00 in the leave phase. In these experiments **down** is set so that if an individual starts with a mutation rate of 5.00 (=N/20 when N=100) then after having 20 offspring the individual's mutation rate will be 1.00 and it will enter the search phase. Similarly **up** is set to increase the mutation rate from 1.00 to 5.00 in 20 generations. Hence **down** $=$ 0.9227 and **up** $=$ 1.0838. Again, further work to formalize the procedure for setting **down** and **up** is essential before this algorithm can become a useful tool.

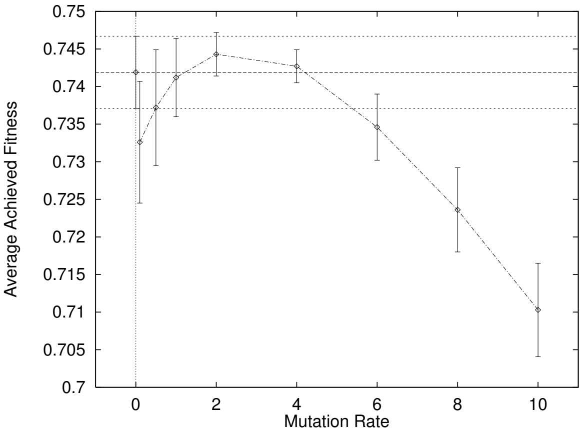


Figure 3: Experiment set up : N=100 K=1 Population=50. The results from the dynamic mutation rate genetic algorithm are plotted at a mutation rate of 0.00 and then extended across the length of the graph.

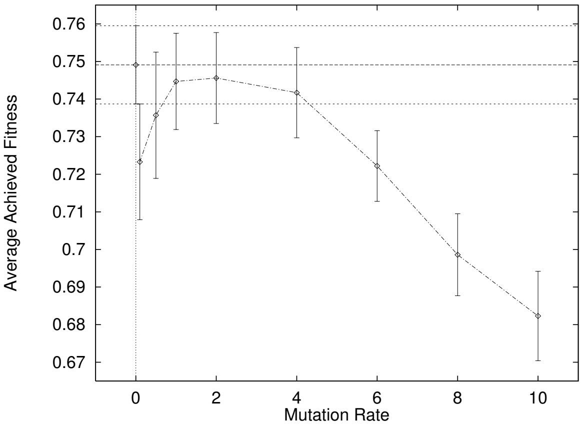


Figure 4: Experiment set up : N=100 K=5 Population=50

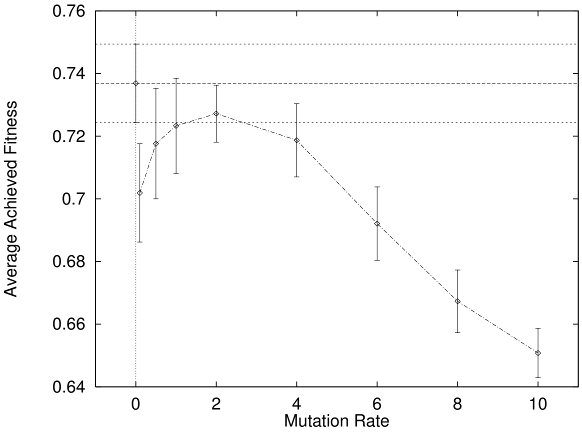


Figure 5: Experiment set up : N=100 K=10 Population=50

6 Results and discussion

This comparative experiment was only run on three different landscapes. All had N=100, but their ruggedness varied from relatively smooth landscape, K=1, to a more rugged landscape of K=5 and the most rugged landscape with K=10. The results from each of these experiments are displayed here as graphs of the average achieved fitness over 30 trials. Each trial consisted of 1000 generations worth of tournaments. The average fitnesses are plotted against the value of the constant mutation rate used by the conventional algorithm. The results from the DMRGA are plotted on the same graph at a mutation rate of 0.00 and then extended across the graph for comparison purposes. From these graphs it is clear that even on the relatively smooth landscape the DMRGA performs as well as the conventional GA (Figure 3). As the ruggedness of the landscape increases so the DMRGA continues to perform as well as, if not better than, the best of the constant mutation rate GAs (Figures 4 and 5). These results show that the DMRGA performs at least as well as the optimal performance of the conventional GA, in this context, with the added advantage that the designer does not have to set the mutation rate. The results also show the potential problems of setting the mutation rate too high or too low since in all sets of results the conventional GA performs significantly less than optimal if the mutation rate is set badly. However it is worth noting that the peak performance of the conventional GA always occurred when using the same mutation rate at the expected rate (an average of 2 bits per genome per generation) for the selection method.

7 A further experiment

The first experiment only used one landscape for each pair of N and K values. However, two NK landscapes with the same value of N and K can have different characteristics, hence the next experiment to perform was to run the same comparison on many different landscapes with the same values of N and K. Comparisons were made on landscapes generated using the same three pairs of N and K values as before. For each pair 15 different random landscapes were generated and tested. Since, from the previous experiment, it is clear that the mutation rate of an average of 2 bits per genome per generation usually performed optimally, in this experiment the constant mutation rate control GA was only run with mutation rates of 1.0, 2.0 and 4.0. Also since it would be interesting to see how each mutation rate technique effects the progress of the population over time during a given run, each test run lasted 1000 generations and the best individual of a given run was recorded every 250 generations for each of the GAs used. Every set up was repeated 30 times to get the average performance.

8 Further results and discussion

The reason for using different landscapes was because they'll have different characteristics from each other. This extends to having different optimal fitness values and a different spread of fitness values across the landscape. Hence it is totally meaningless to compare actual fitness values achieved on different landscapes. So to combine the results from the 15 different landscapes, the data points from each were reduced to the percentage improvement achieved by using the DMRGA. This was calculated as the average fitness value achieved by the DMRGA divided by the average fitness value achieved by the constant mutation rate algorithm with the best performing mutation rate (out of 1.00, 2.00 and 4.00). Hence the resulting figure is the percentage increase that the DMRGA affords on average over the average performance of the best set up conventional GA. These percentage values are plotted over generational time for all three landscape settings.

The results for the N=100,K=1 landscapes (the lower line on the graph of Figure 6) show that the DMRGA performs identically well to the conventional constant mutation rate GA on relatively smooth landscapes. This is a promising result because the mutation rate of a conventional GA is often set to maximize performance on just such smooth landscapes. Over generational time the only significant change is a decrease in the standard deviation, the mean relative performance remains the same.

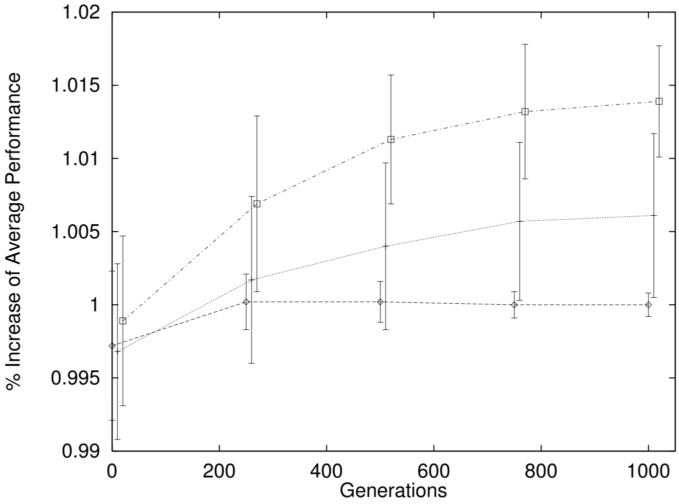


Figure 6: Percentage performance improvement achieved by using the dynamic mutation rate GA over a conventional GA. The top line is calculated from the results for the N=100,K=10 landscapes, the middle line is from the N=100,K=5 landscapes and the bottom line is from the N=100,K=1 landscapes

The N=100,K=10 landscapes give even more positive results (the top line of Figure 6). Here the DMRGA displays a marked improvement over that of the conventional GA. Over generational time this improvement in relative performance increases to a peak improvement of 1.4% over the performance of the best tuned conventional GA. This result may seam quite a small improvement, but apart from the fact that 1.4% of a large fitness value can be a large absolute value in itself, this improvement is in relation to the performance of the best set up conventional GA, hence any comparison to another setting of the constant mutation rate of a conventional GA will display even greater improvement.

The results of the N=100,K=5 landscapes (middle line of Figure 6) are also very promising and give further evidence to suggest that the superior performance of the DMRGA improves in proportion to greater ruggedness, although much more work needs to be done to get a clearer picture of this relation.

All in all the results of this last experiment show that the DMRGA meets the desired goal of a dynamically self-setting mutation rate GA which performs at least as well as a conventional GA under all circumstances, and better in some cases. However there are many parameters whose relation to the success of the DMRGA in these experiments cannot be determined without further investigation.

9 Conclusion

This paper started with an outline why a dynamic mutation rate could have advantages over the conventional constant mutation rate in an optimizing GA. Then a specific implementation was outlined in which each individual of the DMRGA population has its own varying mutation rate, and some comparative results were presented. These results show clearly that the use of a dynamic mutation rate can be beneficial to the effectiveness of the GA to find good solutions. Through the wide range of mutation rates used by the constant mutation rate GA there was also a demonstration of the performance catastrophes of the conventional GA.

The second experiment used the percentage improvement in performance afforded by the DMRGA over the optimal performance of a conventional GA as a measure with which to meaningfully compare the results obtained from different NK landscapes with the same values of N and K. These results show that the DMRGA performs as well as a conventional GA on relatively smooth landscapes, but outperforms the conventional GA on more rugged landscapes. However, what these results do not show is how the performance of the DMRGA will vary with genomes of different length (different values of N) or with populations of different size.

It is only speculative to suggest that the success of the DMRGA at optimizing on an NK landscape means that it will perform well as a search technique in ALife applications. NK landscapes do not capture the full complexity of the fitness landscape of an ALife application such as evolving a neural network controller for a mobile robot. However, a negative result here would have suggested that this technique was not worth considering for any application. Fortunately this is not the case. Although there is much work that is needed to formalize the dynamic mutation rate technique into a usefully portable tool, the early indications are that, by removing the task of setting the mutation rate, it should make a powerful and easy to use addition to the conventional genetic algorithm.

References

Fogarty, T. C. (1989). Varying the probability of mutation in the genetic algorithm. In Schaffer, J. D. (Ed.), *Proceedings of the Third International Conference on Genetic Algorithms and their Applications*, pp. 104–109.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley.

Harvey, I. (1994). Evolutionary robotics and SAGA: the case for hill crawling and tournament selection. In Langton, C. (Ed.), *Artificial Life III*, Vol. XVI, pp. 299–326. Addison Wesley.

Harvey, I. (1995). *The Artificial Evolution of Adaptive Behaviour*. Ph.D. thesis, School of Cognitive and Computing Sciences, Sussex University.
http://www.cogs.susx.ac.uk/users/inman/inman_thesis.html (valid 18/7/96).

Harvey, I., Husbands, P., & Cliff, D. (1994). Seeing the light: Artificial evolution, real vision. In *From Animals to Animats 3: Proceedings of the Third International Conference on Simulation of Adaptive Behaviour*.

Kauffman, S. (1993). *Origins of Order*. Oxford University Press, New York.

Muehlenbein, H. (1992). How genetic algorithms really work: Mutation and hill-climbing. *Parallel Problem Solving from Nature*, *II*, 15–26. http://borneo.gmd.de/AS/ga/publi/gmd_as_ga-92_02.html.

Reynolds, C. W. (1994). Evolution of corridor following behaviour in a noisy world. In *From Animals to Animats 3: Proceedings of the Third International Conference on Simulation of Adaptive Behaviour*.

Sims, K. (1994). Evolving 3d morphology and behaviour by competition. In Brooks, R., & Maes, P. (Eds.), *Artificial Life IV Proceedings*, No. IV, pp. 28–39. MIT Press.

Tuson, A. L. (1995). Adapting operator probabilities in genetic algorithms. Master's thesis, Evolutionary Computation Group, Dept. of Artificial Intelligence, Edinburgh University. http://www.dai.ed.ac.uk/groups/evalg/Projects/MSc/1994_95/andrewt/ (valid 19/7/96).

In: Noble, J. & Parsowith, S. R. (Eds.). (1996). The Ninth White House Papers: Graduate Research in the Cognitive and Computing Sciences at Sussex. Cognitive Science Research Paper 440, School of Cognitive and Computing Sciences, University of Sussex.

Constructing a Dynamic World: A Critique of Gibson's Affordances

Carol Shergold
carols@cogs.susx.ac.uk

School of Cognitive & Computing Sciences
University of Sussex
Brighton
BN1 9QH

Abstract Gibson (1979) put forward a theory of perception which suggests that perception is direct and non-inferential. This realist claim is examined critically, and the behaviour of a room-centring robot (Cliff, Harvey, & Husbands, 1992; Husbands, Harvey, & Cliff, 1995) is described in order to explore problems with Gibson's realist perspective. It is argued that Gibsonian affordances do not provide a useful approach to a changing world. A new conceptual framework is required that focuses on how agents construct relationships with dynamic environments.

1 Introduction

This paper assumes that we want to be able to explain behaviour, the complex dance of animals and people in their environments. The approach to this explanation that I want to take is one which might be described broadly as a 'situated' one which emphasizes the importance of the close interdependence between the agent and its environment. Such an approach is described by Varela, Thompson, and Rosch (1991) and receives a very thorough treatment from a Heideggerian perspective in Wheeler (1996). The relationship between animal and environment is often explored in terms of perception and action, and in this short paper I want to focus critically on the work of J.J. Gibson in *The Ecological Approach to Visual Perception* (1979).

I start by summarising what I think is important in Gibson's approach to perception, concentrating particularly on his notion of "affordances". I then go on to critically examine his realist perspective, attempting to assess how useful it is by exploring the world of a robot evolving in a simple environment. Finally, I look at the problems inherent in applying Gibson's affordances to a dynamic environment and argue that a new conceptual framework is required.

2 An ecological approach to perception

Gibson's approach to visual perception is radical; he doesn't just provide a novel explanation or description of vision, but re-defines what the visual system actually *is*. He includes in the visual system:

> the retina and its neurons, the eye with its muscles and adjustments, the dual eyes that move in the head, the head that turns on the shoulders, and the body that moves around the habitat. The nerves, tracts and centres of the brain that are necessary for vision are not thought of as the "seat" of vision. (Gibson, 1979, p. 309)

This leads to an approach to vision that tries to explain how it is that animals engaged in an active relationship with their environment are able to perceive the environment, in contrast to classic approaches to vision which feature immobilized subjects viewing fixed scenes.

2.1 From spaces to surfaces—invariants

Classic perception theory is based on the idea of space perception; Gibson focuses on the importance of *surfaces*. He argues that vision takes place in the context of an environment composed of many different surfaces, and that light is structured by these surfaces according to invariant laws. This lawfully structured light is then available to the viewer, in what Gibson terms the 'ambient optic array'. For example when an animal is moving relative to its environment, closer objects will seem to move faster than distant ones. This difference in visual flow will allow a moving animal to disambiguate the distances of various objects.

2.2 From properties to ecological meaning—affordances

Gibsonian perception is not of the properties of objects but of their meaning to the animal, and Gibson introduces the notion of the *affordance* to capture this:

> the affordance of anything is a specific combination of the properties of its substance and its surface taken with reference to an animal (Gibson, 1977, p. 67)

Gibson argues that perception is not an inferential process, in which raw sensory information is combined with pre-existing knowledge and memory to allow the animal to infer what the salient properties of the environment are (see for example Marr (1982)). Instead, he sees perception as being already grounded in the possibilities, limitations and demands of the animal's own body:

> When an observer perceives edibility he perceives it in relation to his mouth and teeth and digestive system; when he perceives manipulability he perceives it in relation to his hands ...perception of the environment is inseparable from proprioception of one's own body. (Gibson, 1977, p. 79)

2.3 From representation to realism

This kind of non-representational, non-inferential 'direct' perception seems to offer situated robotics something very appealing; an account of perception which stresses the mutual role of the agent in its environment, and offers an account of meaning based on ecological significance.

The notion of affordances provides a theoretical account of the kind of relationship between agent and world that Brooks described in his papers 'Intelligence without Reason' and 'Intelligence without Representation' (Brooks, 1991a, 1991b). However, whilst one strand of Gibson's work characterizes perception as an active process by which animals come to understand their worlds in terms of their own embodiment, a second strand characterizes perception as a process of picking up a pre-existing ecological reality, characterized and criticized by Rutkowska (1995). Gibson's direct realism has two components. The first is his view that ecological reality is "out there". Problems inherent in this realist position are discussed below (see section 3). The second component is a claim that information about the world can be perceived directly, through a process he called "pick up". He is neutral, even vague, about the *mechanisms* through which this pick up might be achieved, sketching notions such as "resonating". [1]

3 Constructing a world

In order to explore problems inherent in Gibson's realism, I want to focus on a particular example of perceptually guided action in an artificial agent, developed and described by Cliff et al. (1992), Husbands et al. (1995). In these simulations, the visual morphology and recurrent neural net controller of a cylindrical robot are evolved using a genetic algorithm. The environment is a circular arena with white

[1] McDowell draws a useful distinction between the level of internal mechanism, and the level of the animal itself. He argues that for the animal 'its sensory systems are modes of openness to features of its environment' (McDowell, 1994). Gibson is clearly interested in a higher level characterization of perception than the neurophysiological.

floor and ceiling and a black perimeter wall of constant height. Fitness is defined in terms of the proportion of time that the robot spends in the centre of the arena. The robot has two 'eyes', each a single photoreceptor giving a value based on the average brightness of the portion of the environment within its angle of acceptance. The angle of eccentricity (the angle from the centre that the symmetrically placed eyes are located on the cylindrical casing) and the angle of acceptance of the eyes are under evolutionary control.

The robot evolves so that it is able to locate the centre of the arena, and then remain there. For a human subject, relevant invariants would include the fact that distant objects appear smaller, that surface texture becomes denser in relation to the distance of the surface, the relative heights of the wall at different positions in the arena, and so on (Gibson, 1979). However, none of these invariants would be exploitable by a creature with eyes which are basically single value brightness metres. The first challenge to Gibson's realism is to note that 'invariants' only exist in terms of a particular morphology of sensory organ. What kinds of invariants might be available to this robot?

The arena is circular, and so has rotational symmetry. This means that, for a given visual morphology, the visual input for the robot is determined by distance from the centre, r, and the angle that the robot makes with the radius of the cylinder, ϕ (Husbands et al., 1995). There are many possible relationships between this environment, the many possible sensor morphologies and the many possible behaviour patterns of the robot.

Under conditions of fixed illumination, the simplest sort of invariant would be that of a pair of absolute values for the left and right eye that indicated the centre of the arena.

Since the wall is black and the floor and ceiling are white, the robot's visual input will be greatest when the robot has a value of ϕ such that its photoreceptor is oriented towards the wall and r, its distance from the wall, is maximum. In other words, the visual input to a photoreceptor will be greatest when the robot is very close to the wall, but pointing the sensor towards the centre (Cliff et al., 1992). At this position, the dark wall plays a relatively small role in the visual input and the effect of the white floor and ceiling predominates.

A third invariant relationship emerges from the robot's movement. When the robot is close to the wall, any rotation it makes will result in large changes in the amount of light that falls on its photoreceptors. However, the closer the robot is to the centre, the less difference will result. At the exact centre of the arena (where r=0), changing the value of ϕ by rotating will not alter the value of the light at all.

A fourth invariant relationship arises from the fact that the robot has a pair of eyes and carries out the behaviour pattern of rotation. At the centre of the arena, the amount of light falling onto each eye will be equal and will remain equal as the robot rotates. At any other position, this will not be the case.

Varela et al. (1991) are uncomfortable with the realist position in Gibson's claim that invariants exist prior to the animals that exploit them. My contention in this section is that it only makes sense to talk about invariants in the context of a particular sensor morphology and pattern of behaviour. Gibson's examples of invariants are relevant for humans, but would not be relevant for the single pixel eyes of this robot. Equally, there are all kinds of invariants that we are unable to use, such as ultra-violet and infrared light, polarization, magnetic fields, and sonar. However, the most crucial point is that the sequence of sensory input that the robot receives is constructed out of its own behaviour in a given environment. So, in the example above, if the robot was unable to perform a rotation, then the relationships described above would not be useful. The world that the agent perceives is constructed out of its own behaviour and its particular sensor morphologies in a given environment, and Gibson's insistence on the passive "pick up" of "invariants" is misleading because it focuses our attention away from this work of construction, which is a crucial part of an attempt to understand behaviour.

4 Affordances and the dynamics of change

As we saw in the previous section, there is a contradiction inherent in Gibson's work. On the one hand, he argues convincingly that perception arises out of the activities of an embodied agent in its environment.

On the other hand, and even more explicitly, he argues for a position of direct realism.

> Note that these benefits and injuries . . . these positive and negative affordances, are properties of things *taken with reference to an observer* but not properties of the *experiences of the observer.* (Gibson, 1979, p. 137)

Gibson makes it clear that he includes the 'value' of a given object or fact about the agent's environment in his notion of an affordance. He argues that objects derive their significance through their relationship with the agent, and that perception is of this ecological significance rather than a set of neutral properties. We might thus expect that this bundle of significances would change as the needs of the observer change. However, Gibson also states that:

> The affordance of something does *not change* as the need of the observer changes. The observer may or may not perceive or attend to the affordance, according to his needs, but the affordance, being invariant, is always there to be perceived. (Gibson, 1979, pp. 138-139)

A further contradiction thus becomes apparent when we try to understand situations in which the ecological significance of an affordance for an animal is subject to change. Such changes could arise because of developmental changes; because the animal has learned a new behaviour or behavioural response; and because of rapid time-scale changes in factors such as hunger or the depletion of a patch currently being foraged.

4.1 Behaviours involving learning

As described in section 2.3, the claim that perception is *direct* is at the heart of ecological psychology. We saw that this need not be a claim about the mechanisms involved, but a claim that ruled out perception of qualities and a separate inferential process. If a rat sniffs a food object and perceives it as having a positive affordance, eats the food, is poisoned, and subsequently rejects food of this type, then we would say that the rat has learnt something about the food (see, for example, Davey (1989) for an account of such strategies in the context of an animal's environment). Gibson has two choices here. He can say that the affordance of the food remains the same for the rat, but that the rat then combines the affordance of the food with the learnt quality of the food, that of making it sick. If he does this, he has reverted to a view of perception based on the idea that perception is of qualities, which are then subject to some kind of additional cognitive processing by the animal. The other option is to argue that the *affordance* has changed, that the bundle of ecological significance that is the food item for the rat now includes the affordance of making-me-sick.

4.2 Rapid time scale changes

In the ethology literature there are innumerable examples of rapid time scale changes in the ecological value of given affordances (see Krebs and Kacelnik (1991), McFarland (1989), Stephens and Krebs (1986) for relevant overviews of animal behaviour in terms of optimization). If we take the example of an animal foraging in a natural setting, the ecological significance of a particular food patch will vary according to diverse factors such as:

- the animal's own foraging activity and those of other foragers in depleting the patch
- the threat of predators at the patch
- the need to balance competing requirements such as finding water and keeping warm with foraging at the patch
- the balance between the benefit of collecting food resources with the increased metabolic cost of locomotion when heavier

On considering these kinds of issues, it becomes apparent that the value of an affordance does not remain absolute through the shifting and competing requirements on an agent.

5 Conclusion

The concept of affordances is flawed in several crucial ways by its realism. By focusing on the agent "picking up" a pre-existent reality, it removes attention from the way that an agent's behaviour *creates* the possibility for "invariant" relations to emerge. By its emphasis on the unchanging nature of the relationship between affordance and agent it ignores the way that 'ecological reality' is in constant flux. The next step is to develop a conceptual framework which is able to account for the aspects of behaviour that are most interesting: how agents negotiate environments of constantly shifting possibilities and threats.

References

Brooks, R. A. (1991a). Intelligence without reason. In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence*.

Brooks, R. A. (1991b). Intelligence without representation. *Artificial Intelligence*, *47*, 139–159.

Cliff, D., Harvey, I., & Husbands, P. (1992). Analysis of evolved sensory motor controllers. Tech. rep. CSRP 264, School of Cognitive and Computing Sciences, University of Sussex.

Davey, G. (1989). *Ecological Learning Theory*. Routledge, London.

Gibson, J. J. (1977). The theory of affordances. In Shaw, R., & Bransford, J. (Eds.), *Perceving, Acting and Knowing: Toward an Ecological Psychology*. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston MA.

Husbands, P., Harvey, I., & Cliff, D. (1995). Circle in the round: State space attractors for evolved sighted robots. *Robotics and Autonomous Systems*, *15*, 83–106.

Krebs, J. R., & Kacelnik, A. (1991). Decision-making. In Krebs, J., & Davies, N. (Eds.), *Behavioural Ecology: an Evolutionary Approach (3rd edition)*. Blackwell Scientific Publications, Oxford, UK.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman and Company, New York.

McDowell, J. (1994). The content of perceptual experience. *Philosophical Quarterly*, *44*(175), 190–205.

McFarland, D. (1989). *Problems of Animal Behaviour*. Longman Scientific and Technical, Harlow.

Rutkowska, J. C. (1995). Can development be designed? what we may learn from the Cog project. In Moran, F., Moreno, A., Merelo, J., & Chacon, P. (Eds.), *Advances in Artificial Life: Proceedings of the Third European Conference on Artificial Life*, pp. 383–395 Berlin. Springer.

Stephens, D. W., & Krebs, J. R. (1986). *Foraging Theory*. Princeton University Press, Princeton, NJ.

Varela, F. J., Thompson, E., & Rosch, E. (1991). *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press, Cambridge, MA.

Wheeler, M. (1996). *The Philosophy of Situated Action*. Ph.D. thesis, School of Cognitive and Computing Sciences, University of Sussex, UK.

In: Noble, J. & Parsowith, S. R. (Eds.). (1996). The Ninth White House Papers: Graduate
Research in the Cognitive and Computing Sciences at Sussex. Cognitive Science Research
Paper 440, School of Cognitive and Computing Sciences, University of Sussex.

Optimizing Compilation of Object-Oriented Programming Languages

Samantha Type
samr@cogs.susx.ac.uk

School of Cognitive & Computing Sciences
University of Sussex
Brighton
BN1 9QH

Abstract I will discuss the problems associated with optimizing code written in object-oriented programming languages (in particular C++). Previous research on optimizing compilers has concentrated on optimizing for speed but I will be focusing on optimizing for size. This is because I am looking into the feasibility of programming microchips for embedded applications using C++. For these applications, the size of the executable is more important than the speed (although speed does still matter). The principal problem in optimizing C++ programs is the compilation of virtual functions. Their type is dynamically determined, so the precise target function will not be known until runtime, limiting the scope for possible optimizations as well as requiring extra space for virtual function look up tables. I will discuss some techniques for dealing with this problem.

1 Introduction

I have been looking at optimizing compilation of object-oriented code (particularly C++) for use in the programming of embedded systems—this means that the size as well as the speed of the executable code is important.

An object-oriented language is characterized by support for the principles of inheritance†, abstraction† and encapsulation.[1] Classes† are used in C++ to support both inheritance and encapsulation. For inheritance, one class receives the data-types, variables and functions of a simpler base class. This new derived class can specialize the functions and add new variables and functions of its own. This class in turn can be inherited by other more complex classes. Encapsulation is facilitated by the way that types, data and functions are grouped together to form a class. Abstraction is a process of refining out the unimportant details of an object† so only the essential characteristics describing it remain: these go together to make the class.

I have found that the principle characteristic of the C++ language that causes problems for optimization is virtual functions. The other features of the language, such as inheritance and classes, can be analysed and optimized at compile time. Typically, the type of a virtual function cannot be determined until runtime, making optimization difficult unless some kind of dynamic code generation is introduced. Virtual functions are introduced into the language to enable polymorphism.†

It has been suggested that C++ is no more inefficient than C, the language that it was based on (Spuler, 1992). This is true up to a point: an optimizing compiler could take the source from both languages and produce comparatively efficient code. Unfortunately, problems arise when virtual-functions are introduced. Because of their complex nature and dynamic type, they introduce numerous difficulties for the compiler writer. If we look solely at the optimization of virtual functions, then we should be able to obtain equivalent behaviour and keep all the benefits brought by using the object-oriented programming

[1] Terms marked † are defined in the appendix.

approach. For this reason, I have concentrated my work on the analysis of optimizing techniques for virtual functions.

Optimization techniques currently concentrate on optimizing for the speed of code; this is due to the relatively small cost of space in comparison to 6–8 years ago and the large cost of processor power. Processor power is more expensive so the compiler writers have had to take advantage of what is available and produce code that will run as fast as possible, without being overly concerned about the size of the executable code. Optimizing code for embedded systems has the opposite priorities. The size of the executable code is usually of paramount importance due to the cost of the real-estate on an integrated chip (Liao, Devadas, Keutzer, Tjiang, Wang, Araujo, Sudarsanam, Malik, Zivojnovic & Meyr, 1996). Reducing the size of the code produced by a compiler can therefore keep down the size of ROM and reduce the cost of the chip by several pence. In a market with high turnover and small profit margins, this can make a considerable difference to the profitability of the product.

Some methods concentrating on virtual-function optimization have already been introduced and I will discuss these along with ways in which they can be altered in order to improve the size of the produced code as well as the running time.

In the next section, I will introduce virtual-functions in more detail and show why they are such an important area for optimization. Section 3 will concern the techniques that have already been used. In Section 4, I will outline the future work which will make up my thesis. Finally, in Section 5, I will outline my conclusions. I have included an appendix which gives the definitions of many of the terms used in the paper.

2 Virtual functions

2.1 Introduction to virtual functions

Virtual functions are very powerful and enable polymorphism. They are functions that can be applied to objects of many different types. It is the task of the compiler to find the appropriate version for each call.

Virtual functions are defined in a base class as virtual, and can be redefined in a derived class. The type of the function is declared in the base class and the derived version cannot redefine it. If the derived function differs in the arguments that it takes then the virtual mechanism will not be invoked.

This is a useful mechanism for programmers because it means that code can be written for different classes using the same procedure names (an example might be a print function which when called on objects of the circle class will display a circle and when called for squares will display a square). This makes programming class interfaces easier as well as making the interfaces look the same, and it keeps the implementation of the procedures hidden and therefore more abstract. An example of the code to implement a small hierarchy of classes with virtual functions is shown in figure 1. The virtual function here is the peel function; there could be code for a generic fruit which would involve using your fingers to remove the skin, and this would be specialized in Apple to indicate using a fruit knife to remove the skin.

Virtual functions make the methods† in a derived class preferred over those of the base class, but still allow the base class ones to be used if derived ones have not been defined (van der Linden, 1994). Without the virtual keyword, the output "peeling a base class fruit" would be occur for both cases because p is initially defined as a pointer to Fruit. Without virtual, C++ would not know that the derived class's (Apple's) peel was to be called.

2.2 Implementation of virtual functions

Since the type of a virtual function is dynamic, we need to have some method of determining which function to call. Single inheritance is usually implemented by having each object contain a pointer to a vector (Virtual table or vtable in G++) of function pointers. There is one table for each class, and there is one entry in the table for each method in the class. In this way, the implementation code is shared by all objects of a given class. The table is laid out so that a given function pointer lies at the same offset in

```
class Fruit {
     public:virtual void peel() {

               printf("peeling a base class fruit\n");
          }
          slice();
          juice();
     private:
          int weight, calories_per_oz;
};


class Apple : public Fruit {
     public:
          void peel() {
               printf("peeling an apple\n");
          }
};

...

main() {
     Apple apple;
     Fruit orange;
     Fruit *p;

     p = &apple;
     p->peel();

     p = &orange;
     p->peel();
}

% a.out
peeling an apple
peeling a base class fruit
```

Figure 1: Example of a class hierarchy with virtual function definition (from van der Linden, 1994).

the virtual tables for all subclasses of a class. Each method call can be mapped to a virtual table offset at compile time. At runtime, the call is made indirectly through the pointer at the appropriate offset. Multiple inheritance requires a slightly more complicated scheme with another layer of indirection (van der Linden, 1994).

In an optimizing compiler, the only time a call to a virtual function need generate different code from an ordinary function call is when the function is called through a pointer or reference. Any calls involving objects can be statically bound to the correct function. If the code can be analysed at compile time and objects can be bound to the appropriate functions, then we may be able to do away with the need for some of the virtual function tables and therefore the indirect function calls.

Calls via pointers and references must generate the slower dynamic binding call sequence. This is because calls via pointer or reference could be to a class with others derived from it, which may result in a different function definition.

Virtual tables are generated for each class only if it contains a virtual function. If we can optimize the code so that a class contains no virtual functions then we have successfully saved the space for a virtual table.

2.3 Space requirements

Extra space required by the use of virtual functions is of 2 types:

1. Hidden pointer data member in each object.

2. One table of pointers to functions per class.

No matter how many member functions are virtual, the amount of extra space in an object will be only a single pointer field. Each class has a table of pointers to functions of a size equal to the number of virtual member functions. Although a large number of virtual functions do not increase object size, it does increase executable size because of some extra tables of pointers to functions.

3 Techniques for the optimization of virtual functions

Solving the problem of virtual functions has commonly been tackled with two different approaches. The first is static analysis, called concrete type inference or class hierarchy analysis (Agesen & Hölzle, 1995; Aigner & Hölzle, 1995). The second approach is dynamic analysis (Hölzle & Ungar, 1994; Agesen & Hölzle, 1995; Aigner & Hölzle, 1995).

3.1 Static analysis

This approach works by examining the source code to determine, as far as possible, which versions of virtual functions are likely to be called at each point. Determining the type of the virtual function prior to runtime when its arguments are not pointers to objects (when the compiler cannot determine where the pointer will be pointing) is straightforward. In this situation it will be obvious which version of the function is required because the type of the objects are clear and the virtual function that corresponds will be required. The virtual function call can be replaced with a direct call or pointer to the desired function or with an inline version if it is short.

When the types of the arguments to a virtual function are pointers, the situation is more complicated. The surrounding code has to be examined and the class hierarchies analysed. This can give important type information to the compiler which will use the information to implement pointers directly to functions, or inline in some cases. It is sometimes the case that a probable target can be determined and if this is so then additional code, in the form of a conditional or switch statement, will be included to catch the less likely cases that could occur.

3.2 Dynamic analysis

This works by using information gleaned on a special profile gathering run to determine the likely targets of virtual function calls. This approach will end up generating less code than static analysis because it will only generate code for the paths that do occur rather than for all of those that may occur. It does still rely heavily on the use of switch statements to catch the cases when the unexpected happens. Dynamic analysis has been shown to be more effective than the static approach for hybrid object-oriented languages (such as C++) in the majority of cases, but has the disadvantage of needing extra code for the profile gathering interpreter or compiler.

A further solution, introduced by Calder and Grunwald (1994), suggests using a dynamic code generation approach which involves using information gathered on previous runs to optimize the compiler. This technique, although successful, has not been widely used. I suspect this is due to the complicated nature of the approach. It would not be feasible for the area of code generation for embedded microprocessors because of the extra code that would be needed to collect information on each run.

4 Future work

4.1 Modifying the optimization techniques for embedded systems

The results of methods outlined above have been promising and show that virtual functions can successfully be used in writing C++ code, without having to worry about the cost in terms of the time that it takes to run. However, little attention has been paid to the size of the produced code. When mentioned, a size increase is indicated often of up to 50%. Although this is acceptable in many systems, it is not what we want here. The increase in code size can be attributed to the fact that the principle technique for dealing with virtual functions and improving the code speed is to inline the code. While this makes execution faster, it obviously increases the space required for the code.

Another problem is that, since the microchips for embedded applications are usually being designed for the prototype, a profile gathering run will be impossible unless a simulator is available. This rules out the use of the dynamic technique of optimizing the code. So, it would appear that we are destined to use the static optimization approach. This is fine for discovering possible areas for optimization, but we will need to modify the way that the approach has so far dealt with solving the virtual functions that it has found. This means that a technique based on the static approach but improved to decrease code size rather than increase it is required.

Hölzle and Ungar (1994) show that type inference takes the conservative approach because it examines all of the possible paths through the program—obviously, many of these will not be used. Generating the code for all possible paths also means that more space is required. However, Calder and Grunwald (1994) find that there is often a unique target of a virtual function call, so a direct function call could be inserted without the need for the extensive switch statements. By combining this knowledge, it would seem a feasible solution to statically analyse the C++ code (using class hierarchy analysis), but only change the code for the cases when a unique target can be determined. So where possible, a direct function call will replace a virtual function call (thus saving space and time due to the lack of need for a virtual function table lookup). If virtual calls to a specific function can be eliminated completely then it may even be the case that the lookup table for a given object can be discarded completely. When a unique site cannot be discovered then the code would be left as a virtual function call: this would be slower but a switch statement will not be required so it will not take up any more space. This would be an interesting area to start the research and its usefulness would depend on the number of unique targets that actually do occur in C++ programs.

In order to compare the results of my work with others in the field (for example, to check that code speed is not decreasing), I will implement the virtual function optimizations in conjunction with the traditional techniques that are already commonly used.

4.2 Overall view

My research will be concerned with writing an optimizing compiler for C++. This will concentrate on optimizing for size because I hope to be doing some work in conjunction with a major microprocessor manufacturer, and the code produced would need to go on embedded chips where code size (ROM) is of importance, due to the cost of real estate on the integrated controller chips (Liao et al., 1996). The introduction of the object-oriented programming approach should be of benefit to this field because it encourages modular design and code re-use. It will enable more portable code to be written because the machine dependent parts of the code can be factored out and kept separately from the machine independent parts (Maclean, 1995). This is currently difficult to achieve because of the reliance on C (where the machine dependent and independent parts are inter-mingled) and assembly code programming.

My short-term aim is to do a feasibility study, using the research that I have already looked at to determine whether or not C++ is a feasible programming language for 8-, 16-, and 32-bit chips where the size of the executable is important. I will need to examine the techniques for optimization of code size and even examine the possibility of excluding the use of some features of the language (such as multiple inheritance) that would help to keep the code small while not losing too much of the functionality of the language. Another example would be to allow only one level of indirection.

Plevyak and Chien (1995) have shown that object-oriented programs need both inter-procedural optimization and intra-procedural optimizations (i.e., whole program optimization). This is more important than for procedural code where examining basic blocks can sometimes be enough to satisfy requirements. Obviously any optimizations specifically for object-oriented languages would show best results if used in conjunction with traditional optimization techniques, such as common sub-expression elimination (improving speed and size). The code for these optimizations is readily available and the techniques are well documented so implementing these also should not present any problems.

5 Conclusions

We can make the following conclusions about the use of virtual functions in C++ code:

- Virtual functions are a powerful mechanism for enabling polymorphism. We do not want to have to worry about the cost of using such a useful tool or limit its use in order to keep the size of code to a minimum.
- If we can eliminate the cost of using virtual functions then we will be able use the full functionality of the language and not lose a powerful mechanism for polymorphism.
- In many cases the virtual can be translated into a normal function call.
- Ideally, we would like to be able to reduce the cost of using virtual functions (and C++ as a whole) so that it is of equal efficiency to C.

I have shown the importance of the optimization of virtual functions in the C++ language. I will work on this problem with the aim of reducing the size of the code as a priority but since the chips will be mostly for real-time systems I will aim to maintain a good level of speed too.

References

Agesen, O., & Hölzle, U. (1995). Type feedback vs. concrete type inference: A comparison of optimization techniques for object-oriented languages. *OOPSLA 95.*

Aigner, G., & Hölzle, H. (1995). Eliminating virtual function calls in C++ programs. Technical Report TRCS 95-22, Department of Computer Science, UCSB.

Calder, B., & Grunwald, D. (1994). Reducing indirect function call overhead in C++ programs. Technical Report, Department of Computer Science, University of Colorado.

Hölzle, U., & Ungar, D. (1994). Optimizing dynamically-dispatched calls with run-time type feedback. *SIGPLAN '94*. ACM.

Liao, S., Devadas, S., Keutzer, K., Tjiang, S., Wang, A., Araujo, G., Sudarsanam, A., Malik, S., Zivojnovic, V., & Meyr, H. (1996). Code generation and optimization techniques for embedded digital signal processors. *The First SUIF Compiler Workshop*. Stanford University.

Maclean, S. (1995). Object-oriented programming for embedded systems. Technical Report DSSE-TR-95-1, Department of Electronics and Computer Science, Southampton University.

Plevyak, J., & Chien, A. A. (1995). Compilation of object-oriented programming languages. Technical Report, Department of Computer Science, University of Illinois at Champaign.

Spuler, D. (1992). *C++ and C Efficiency: How to Improve Program Speed and Memory Usage*. Prentice Hall.

van der Linden, P. (1994). *Expert C Programming: Deep C Secrets*. Prentice Hall.

Appendix: Definitions

Here are the definitions of some of the terms used in the paper that may be unfamiliar to readers not active in the area of compiler writing or object-oriented programming. The definitions are based on those in van der Linden (1994).

Abstraction The process of refining away the unimportant details of an object, so that only the essential characteristics that describe it remain. Abstraction is a design activity, the other concepts are the object-oriented programming features that provide it.

Class A user defined type with all the operations on it. A class is often implemented as a struct of data, grouped together with functions that operate on that data. The compiler imposes strong typing - ensuring that these functions are only invoked from objects of the class, and that no other functions are invoked for the objects. Anything in a class is known as a member of the class.

Inheritance Deriving one class from another such that all of the other's characteristics are automatically available. Being able to declare types which share some or all of the characteristics of previously declared types. Being able to share some characteristics from more than one parent type.

Method Member functions of a class (the operations) are also known as methods.

Object A specific variable of a class type, just as j may be a specific variable of type `int`. An object is also known as an instance of a class.

Polymorphism Refers to the ability to have one name for a function or an operator, and use it for several different derived class types. Each object will carry out a different variant of the operation in a manner appropriate to itself. It starts with overloading a name, reusing the same name to represent the same concept with different objects. It is useful because it means that you can give similar things similar names. The polymorphism comes in when the runtime system selects which of these identically named functions is the right one.

Virtual Functions These are more complicated and are the subject of the paper. More details and a definition are given in the text of the paper.

In: Noble, J. & Parsowith, S. R. (Eds.). (1996). The Ninth White House Papers: Graduate Research in the Cognitive and Computing Sciences at Sussex. Cognitive Science Research Paper 440, School of Cognitive and Computing Sciences, University of Sussex.

How Do I Check My Software Designs?

Joseph A. Wood[†]
joew@cogs.susx.ac.uk

School of Cognitive & Computing Sciences
University of Sussex
Brighton
BN1 9QH

Abstract Reviewing software designs is both a *hard, error* problem, and worth automating. This problem is often tackled by calculating various metrics relating to modular structure, in particular cohesion and coupling. We present a novel approach based on statistical cluster analysis. This is illustrated by looking at a software design for a set of traffic lights at a cross-roads.

1 Introduction

Modern software systems are very large and complex: sizes of hundreds of person years of effort are not uncommon. We need to manage and control the interactions which occur in such systems—this can be achieved via modular construction.

Modules enable information hiding, and hence reduce unwanted interactions between components. Moreover, such an approach simplifies the problem by breaking the problem into smaller sub-problems. Modules also ease the problems of managing the production process by identifying required components. This also helps by allowing easy identification of what has and has not been completed.

We also know from empirical studies that the cost of correcting software problems tends to rise by at least an order of magnitude as we progress along the production process. Therefore, we are particularly interested in the early stages, such as requirement capture, specification and design.

We are interested in the later part of the complete design, when the complete system is available for consideration.

2 The problem

The traditional method of checking software designs, still the mainstay in industry, is a series of design reviews. Design reviews have several disadvantages:

- Hard work;
- Requires skilled labour;
- Error prone;
- Time consuming;
- Very expensive;
- Frequently shallow.

[†] Supported by a CASE award from the Engineering and Physical Sciences Research Council in association with British Telecom Laboratories (ML 464531).

Not surprisingly, a number of researchers are keen to find ways to use computers and automated techniques to relieve human involvement.

An obvious first question involves how a design is presented—since it is clear that a design is only a blueprint of something which is going to be built.

Looking at current best practice provides only limited guidance. There is a wide spectrum of notations, ranging from the highly mathematical to natural language with varying degrees of graphical support. Mathematics can be difficult to understand, and relatively expensive to use, such that it is best kept for specific parts of a project. Natural language is easy to produce but frequently suffers from a lack of precision. Automatic processing of natural language is still an area of research, and is not yet a safe foundation for automatic validation of software designs.

Graphics offer a good communication channel between individuals taking advantage of humans' well developed visual processing skills and can be easily sketched for discussion between individual engineers or group presentations. It should be clear, that each graphical notation employs it own special syntax and semantics, however these tend to be easily explained to other people.

There are however disadvantages to purely graphical notations involving both production and machine processing. Although most drawing packages can be used to produce graphics, such general packages are not tailored to specific notations, so they can be time consuming to use (especially for changes) and cannot enforce a notation's syntax rules. Processing such general drawings presents as much as a challenge as natural language processing. Consider for example, a notation which expects a rectangle, but finds instead four individual lines which (perhaps) do not quite meet at the corners. Moreover, there is no standard file format for representing graphics. These difficulties have led to the creation of a number of proprietary drawing packages, each tailored to specific graphical notations,[1] each with their own different file format.

Graphical notations owe much of their power to their (apparent) simplicity, which is achieved at the expense of not representing much detail.

For these reasons, and having consideration for typical languages used in large software systems, we decided to use HOOD[2] (Hierarchical Object-Oriented Design) as a design representation. HOOD has both a graphical notation and a purely textual (superset of the graphical) representation. The HOOD method encourages good software engineering practices, such as decomposition and information hiding. HOOD's primary purpose is to identify the structure of a solution rather than the precise details of the implementation,[3] for which, of course, programming languages are available. Further, the HOOD textual notation has a well-defined syntax and a number of rules for consistency checking which have been automated.

3 Validating designs

There are several aspects to validating a design, which include:

- Is the notation used correctly?
- Have all the required components been identified?
- Have the components been designed in sufficient detail to allow implementation?
- Is the design consistent, both textually and functionally?
- Is the design easy to understand?
- Is the design easy to update?

[1] There are a few meta-drawing packages which can be customized to specific notations as required.

[2] HOOD is a registered trademark of the HOOD User Group.

[3] Hence encouraging programming by contract.

- Will it *work*?

Some of these questions can be answered by 'compiler-type' tools, but some are very hard problems even for humans.

One approach adopted by several researchers, ourselves included, is to develop measures of the design's structure, and *hope* that these measures capture intangible properties of the design such as complexity, understandability and ease of modification etc.

Major objections to this approach follow from the obvious impossibility of using similar measures to capture different properties, and further, why should one metric be a good predictor of several different properties? Additionally, the intangible nature of these properties makes them impossible to define, let alone measure. Such objections are, of course, valid and cause for concern. However, ceteris paribus, the more complex a design becomes the less attractive it becomes. This may be due to being harder to understand, change and debug, etc.

The two most common properties looked for are cohesion and coupling. *Cohesion* measures how well an object[4] has a singleness of purpose, i.e., it has one single, well defined purpose, to which every part of the object contributes. *Coupling* measures how inter-dependent two objects are. Not surprisingly, we would like a system to have strong cohesion and loose (weak) coupling. It is clear that, in some sense, these two properties are closely related, but it is far from obvious exactly what this relationship is. Consider, for example, a single object (identified at some level of decomposition). As a single object, it should have high cohesion, i.e., all its parts contribute to but a single purpose. Now decompose the object into a set of component objects, these must have loose coupling, and yet still contribute to a single purpose.

4 Cluster analysis

Cluster analysis is a statistical technique for examining the similarities and differences between objects. Given a set of n objects $(x_1,\ldots,x_n)$, each of which is characterized by a vector $\langle p_1,\ldots,p_m\rangle$ in a m-dimensional space, we can calculate a distance between any two objects (x_i,x_j, say) as a function of their position vectors. Initially, we can place each object in a distinct cluster. We then seek a way to merge clusters such that, in some sense, the size of each cluster is minimized and the distance between clusters is maximized. Finally, after $n-1$ iterations of this process, we finish with a single cluster. We must emphasize that different distance functions and agglomeration methods can give rise to different cluster hierarchies; nor is the cluster for a given pair of distance function and agglomeration method unique.

Our interest in cluster analysis is to not seek any kind of 'best' design solution, but rather to look for significant clashes between the allocation of objects to modules (in the design) verses their clustering closeness. Our aim is not to tell the designer what should be done, but rather to highlight potential inconsistencies.

4.1 Example

In this section we shall look briefly at a proposed design (Robinson, 1992) for a small system to control a set of traffic lights at a cross-roads. Note, the module `Road typedef`, does not appear in Robinson's design, as the data type `road` is defined in the `Traffic Light System` module. However, such usage violates one of HOOD's rules.

Figure 1 shows a road junction, which has two roads (AC and BD) crossing at right angles. Each road, has a set of traffic lights, and sensors planted in the road to detect the presence of waiting traffic. Once traffic has been allowed to pass a newly changed traffic light, it must continue to flow for a given time, after which, if there is traffic waiting on the alternative road it must be given a turn. As usual, the traffic lights must be co-ordinated so that they change in the correct sequence and do not permit

[4] We are here not just referring to object-oriented systems, but to objects in the general world.

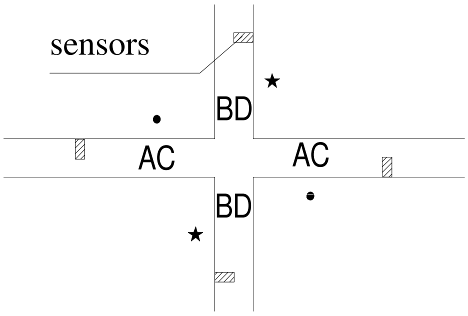


Figure 1: Road layout at traffic junction

traffic to impact whilst crossing the junction. The figure is based on left-hand drive cars, but this has no significance in the design.

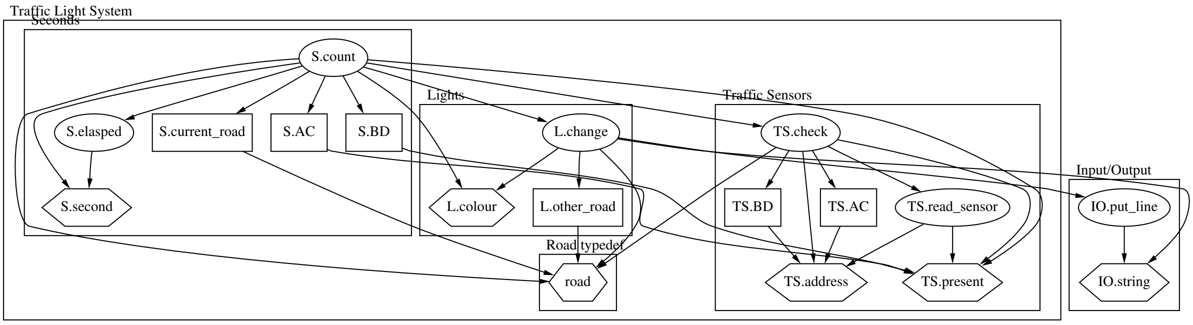


Figure 2: Inter-dependencies between design components

Figure 2 shows a graphical representation of the links between components in a proposed solution to the problem outlined above. The oval shapes represent procedures, the small boxes are (module) variables, the hexagons are types, and the large enclosing boxes represent the module hierarchy. The lines with arrows represent the inter-dependencies between components. An obvious first reaction to Figure 2 is to find it confusing. We believe that in a high quality software design, this disorganized nature would not be apparent. It suggests an inappropriate decomposition of the problem. Intuitively, we feel that there is quite high coupling between `seconds` and other objects in the `traffic light` module, in particular `S.count` seems to have a large number of dependencies.

Figure 3 shows one possible hierarchy clustering. This suggests that there is a difference between the operation `S.count` and the other components which bears out our initial impression.

5 Summary and conclusion

Reviewing large software designs is a hard problem, which is both intellectually and economically worth supporting with automated techniques.

Our approach to this problem is to question the designers' proposed modular structure. We use cluster analysis to find the similarity between components and look for significant difference between closeness in the design and the hierarchical clustering results.

This approach avoids the difficulty of comparing coupling and cohesion inherent in some other approaches.

Although we have only conducted small initial experiments, the outcome looks promising. We have used the HOOD design notation as a basis, but its contribution is in focusing attention on the relationship

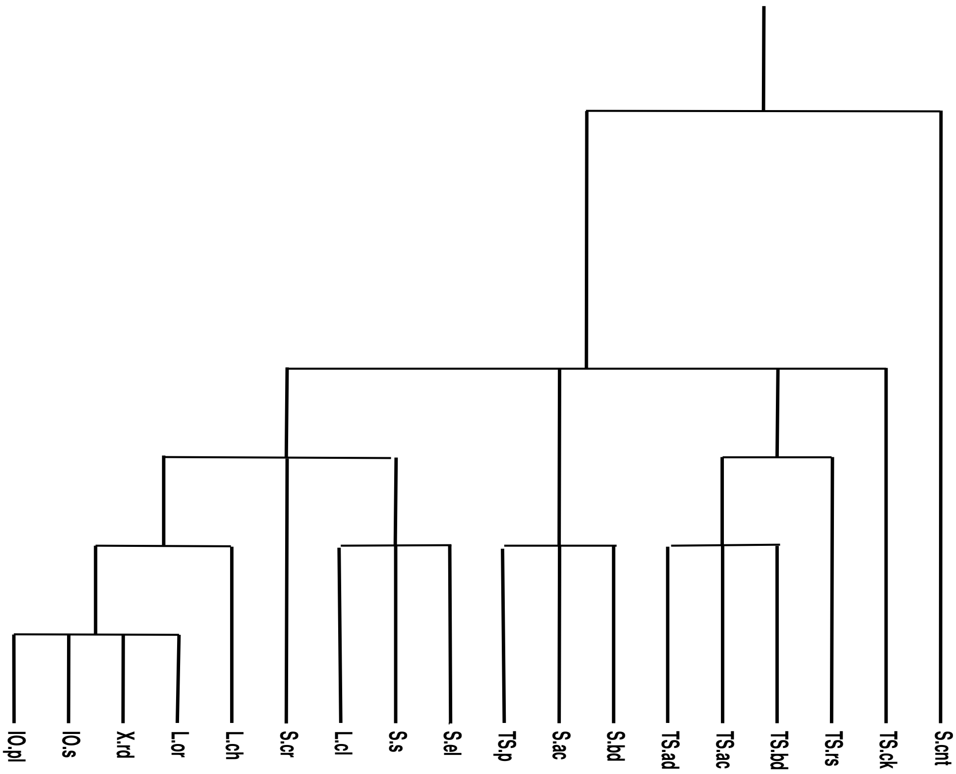


Figure 3: Cluster analysis of traffic junction design

between design components, rather than being necessary a prior.

The use of cluster analysis for examining software designs is unusual, see for example Neil and Bache (1993) and Hutchens and Basili (1993). Neil and Bache (1993) are interested in the organization of programs, as we are, however Neil and Bache perform their analysis at the source code level not the design level.

References

Hutchens, D. H., & Basili, V. R. (1993). System structure analysis: Clustering with data bindings. In Shepperd, M. J. (Ed.), *Software Engineering Metrics, Volume 1: Measures and Validations*, McGraw-Hill international series in software engineering, chap. 5, pp. 83–98. McGraw-Hill Book Company, Maidenhead, England. Reprinted from IEEE Transactions on Software Engineering, 11(8); 749–757, 1985.

Neil, M., & Bache, R. (1993). Data linkage maps. *Software Maintenance: Research and Practice*, *5*, 155–164.

Robinson, P. J. (1992). *HOOD: Hierarchical Object-Oriented Design*. Prentice-Hall object-oriented series. Prentice-Hall, Hemel Hempstead, England.