Collaborating with Neural Network-based Artificial Intelligence Agents

Emma Byrne
Computing Science
Middlesex University
The Burroughs, London
NW4 4BT, U.K.
e.byrne@mdx.ac.uk

Dan Diaper
DDD SYSTEMS
26 St. Marks Road,
Bournemouth, Dorset,
BH11 8SZ, U.K.
ddiaper@ntlworld.com

Christian Huyck
Computing Science
Middlesex University
The Burroughs, London
NW4 4BT, U.K.
c.huyck@mdx.ac.uk

ABSTRACT
CABot is a novel project to build neuro-cognitive agents. One aim of the CABot project is to develop agents that collaborate with a human user to carry out tasks in a 3D environment by means of a natural language conversation. As the agent is based on human neural and cognitive models, users may expect human-like collaborative skills from CABot. We believe that: (a) generations of CABot agents may be artificially intelligent but may nevertheless lack the full range of human-like collaborative abilities; and (b) even with human-level collaborative abilities, collaboration between AI and human (and AI and AI) will be no less challenging than collaboration between human and human. Many of these challenges were foreseen with respect to earlier forms of AI. CSCW and HCI are likely to inform the solutions to these challenges as AIs become widespread in collaborative settings.

Author Keywords
Artificially intelligent agents, collaboration, human-agent interaction.

ACM Classification Keywords
H.5.3 Information Interfaces and Presentation: Group and Organization Interfaces—*Collaborative Computing*; I.2.0 Artificial Intelligence: Cognitive Simulations; I.2.9 Artificial Intelligence: Distributed Artificial Intelligence—*Intelligent Agents*

INTRODUCTION
"Expert systems' reliability will, in the end, be strongly determined by the quality of their dialogue with users and the compatibility between what the user believes the expert system knows and what it actually knows." [8]

Substitute "collaborative, artificially intelligent agents" for "expert systems" and Diaper's conclusions of more than twenty years ago are equally applicable today. The potential, and still unsolved, issues concerning how people interact with a truly intelligent, non-human agent remain.

Now is the right time to ask questions of where our likely future technological developments, in this case in Artificial Intelligence (AI), might lead. For example, how they may help and hinder: individual, organizational, social, political, economic, ethical and aesthetic desideratum; and just what might be these, often conflicting, desires?

This paper will discuss several of these concerns, and motivate them with the example of a novel intelligent agent: CABot. The Cell Assembly robot (CABot) is a project that aims to develop intelligent agents that emulate human neuro-cognitive architecture. CABot1 [17] was an initial prototype, a first generation intelligent agent that could collaborate with human users in a virtual 3D environment. It is able to parse natural language instructions and to act on those instructions within the virtual world. All the CABot agents are designed with a commitment to neurophysiological fidelity to the human brain and functional fidelity to the human mind. As a result, the agents not only perform in a cognitively human way, but these neural network systems also perform in a similar time to human performance. For example, the CABot2 parser functions in human-like times when parsing natural language sentences (e.g. [18]).

However, there are still human-like behaviors that completed CABot agents do not exhibit. For example, neither CABot1 or 2 incorporates a theory of mind, nor do they have a concept of negotiation. Whilst CABot1 and 2 can solve several interesting linguistic challenges, they do not have a particularly sophisticated approach to pragmatics or argumentation. Collaboration between humans is challenging, even when those humans possess these abilities. Collaboration with an apparently human intelligence that nevertheless lacks these features is likely to lead to problems ranging from mere frustration to, in safety critical applications, human deaths and serious injuries.

These concerns are not new, and many of them were foreshadowed in early work on expert systems. Expert systems in the 1980s were defined as being systems that could emulate some aspects of human expert thought and behavior (e.g. [23]), i.e. they did not have to work like either the hu-

IUI 2009, February 8-11, 2009, Sanibel Island, FL, USA.

man brain or mind, but just copy some human behaviors (in restricted environments).

In contrast, the overall neural and cognitive fidelity approach of the CABot project aims to create AIs that are human-like in both form and function. Systems are tested against human performance and look to match it. As a result, human users collaborating with these novel AIs might expect human-level collaborative skills. The unresolved issues of people and expert systems communication and understanding are even more germane for CABot-like neural agents than to the old expert systems. This position paper lays out several of the challenges that the authors foresee as human agents begin to collaborate with artificially intelligent agents, or indeed when artificially intelligent agents begin collaborating among themselves.

THE CELL ASSEMBLY ROBOT

CABot is a project with the long-term goal of building intelligent agents by following human neural and psychological models. This project has developed several prototype agents (the major versions are CABot1 and 2), another is currently under development (CABot3), and others are being planned (e.g. CABot4). These early systems are built from simulated, artificial fatiguing Leaking Integrate and Fire (fLIF) neurons that are a model of natural neurons [16].

Cell Assemblies (CAs) [14] emerge from these fLIF neurons via unsupervised learning, and when ignited, these CAs represent the content of the agents' short-term knowledge (e.g. declarative semantic knowledge) and how it thinks over time (e.g. learning, rule following, decision making). As in the human brain, fLIF neurons can be members of more than one CA and the existing CABots support both specialized function, corresponding to current models of human neural functional allocation, and less specialized cognitive computational capabilities.

The basic CABot architecture is based on current models from psychophysics and experimental cognitive psychology, and from linguistics and psycholinguistics. Because of current hardware limitations, some of the recent CABot system demonstrations have been programmed, rather than learnt, to conform to such models [18]. However, the fLIF neuron-based, cognitive architecture is able to learn what is programmed in the CABot agent. This has been tested in other systems. For example, the CABot2 parser is programmed to store the semantics of lexical items as overlapping CAs following the Wordnet hierarchy [22]. Other simulations demonstrate that these overlapping CAs can be learned to automatically derive the hierarchies [16].

The CABot1 and 2 systems use a freely available virtual reality environment [1]. The agents are "robots" within the virtual world, tasked to assist the human user in the environment. In the following section, the system is described, starting with the agent's inputs from its virtual environment and from the user. The agent processes such inputs and i) produces natural language output to the user and ii) operates within the virtual environment. This is what Anderson [3] calls end-to-end behaviour.

The CABot agents receive, via JPEG files, visual input from the virtual world that they inhabit. The user receives the same view on their screen, from the user's own perspective. To interpret this visual input, the agent has a simulated retina that codes the stimulus as spatial frequencies corresponding to different receptive field sizes and so provides an output similar to that sent from the retina of the human eye, via the optic nerve, to the visual cortex. The 3D visual virtual world is admittedly very simple but CABot1 and CABot2 can navigate around their world using visual input. Both agents can recognize objects, and some versions of CABot1 can learn new shapes. Labels for these new shapes are learnt via interaction with the user. This begins to address the symbol grounding problem: how do arbitrary mental symbols come to represent real-world objects [13].

To interact with the agent, the user enters natural language keyboard inputs (we have not yet attempted to implement speech) and the CABot2 agent uses semantic, lexical and parse systems, based on current psychological and linguistic models, to understand the user inputs. There is some evidence that not only are sentences parsed in a way similar to how people do it, but that the CABot2 parser performs in times similar to those that people take when parsing [18].

Current visual information is combined with the processed language. Together, visual and linguistic inputs are used to learn new declarative knowledge, or to learn or generate appropriate behavior. The agent can create plans, which while simple, are not merely a linear means-end analysis. Furthermore, it can generalize its rules to create new behaviors where necessary.

It is not uncommon for AI researchers to think of building AI systems as assembling building blocks. The CABot approach is different in that what it builds are, potentially real-time, simulations of how the brain is structured and operates and then shows how, as an agent, it performs cognitive and linguistic functions in ways similar to how people perform. Whilst the agent's environment and tasks are simple at the moment, two things encourage the belief that this research will scale-up.

Firstly, because they use similar neural hardware (digitally simulated fLIF neurons) and similar organization (CAs) to the human mind, the "building blocks" of CABot are cognitive functions based on models of human cognition. It is not necessary to build sets of new "building blocks" to enable new abilities as CABot is autonomous and self organizing. In addition to the wide range of different types of cognitive activity CABot has simulated, further cognitive abilities are possible using the same architecture.

Secondly, CABot is capable of independent, self directed learning. Present it with a novel visual object, for example, and the agent can learn to discriminate it from other objects it has previously learned, and to recognize it later. It can generalize its learning and categorize new stimuli [16]. Using

both its visual and its language systems, CABot can ground its linguistic representation in its visual world by varying synaptic strength between neurons following a few simple organizational principles at the CA level of description.

It is proposed that CABot4 will be a conversational agent. In addition to the behaviors of CABot2 and 3, CABot4 will also be able to produce natural language text for the user to read. This should enable a much more sophisticated form of interaction. CABot is not the first attempt at creating a cognitive architecture. Other cognitive architectures, such as ACT-R [2] and Soar [21] could be considered symbolic AI systems. However, these systems have no fidelity at the neuron level. In contrast, all computation in CABot is carried out entirely via fLIF neurons of a reasonable biological fidelity. As a result, CABot is distinct from ACT-R and Soar in that these fLIF neurons can organize themselves into CAs that learn a wide range of cognitive functions.

INTERACTING INTELLIGENT AGENTS

Diaper [6], argues that HCI, and therefore CSCW, are primarily engineering disciplines, ultimately being focused on solving problems concerning people and computers. Both HCI and CSCW, like cognitive science, KBS and other forms of AI, are highly interdisciplinary, drawing from, inter alia, psychology, sociology, philosophy and linguistics, as well as from computer science. CSCW's profound contribution has been to shift the focus of HCI from a computer system's single direct end user to how groups of people can work together using computer systems (groupware).

The potential of groupware with AI capabilities was recognized even in CSCW's early days. Connolly and Edmonds [5] suggest the concept of an agent "has emerged as central" to combining CSCW and AI. Skipping forward a decade, Hollnagel [15], in the context of cognitive task analysis, allows some AI agents a potential status similar to human ones, being elevated to autonomy from being merely tools. Such AI agents, like people, have goals and means of achieving them.

Reviewing, and while supporting, Hollnagel's proposals, Diaper [10] suggests a more extreme position, based on his task analysis and general systems analysis approaches (e.g. [9, 11]), where a "work system" is an agent composed of a collection of components in a general system which function together to perform work (defined as changing the application domain). In the context of task performance, such agents are often ephemeral, indeed, they are really a systems analyst's model of how work is achieved. Critically, it is the agent, rather than its components, that possesses goals, even when such components are intelligent entities, human or machine.

It may be a useful level of analysis to treat a person, or an AI device, as an agent. Nevertheless, the concept of an agent need not be co-extensive with human or machine entities treated in such a singular manner. Indeed, in analyzing how groups of people perform work, it is usually necessary to understand not what individual people do, but what they do collectively.

Well documented in the CSCW literature are the problems and difficulties with group working, and with successfully implementing groupware of an adequate sophistication to support the requirements that CSCW researchers have identified. Diaper (1986a) identified the even greater problems when some of the things in a system were intelligent but not human. If people have difficulty understanding what they do alone at a task, and even more so when working with other people, then how much harder will it be when alien AIs are added to the mix?

Several aspects of intelligence are particularly useful in achieving productive collaboration: a theory of mind that allows individuals to usefully speculate about the beliefs, desires and intentions of their collaborators; an ability to negotiate or otherwise accept joint decision-making; and some pragmatic competence that allows the individual to parse implicit as well as explicit meaning in language.

If CABot is to become a useful collaborative agent, these types of intelligence may be required. Some future generation of CABot will eventually learn these abilities in the same way that it learns other cognitive functions. In the interim, is it preferable to give CABot "prosthetic" versions of these abilities, using external systems that support argumentation for example? Or can CSCW systems bridge the gap between an agent that lacks these abilities and human users that expect them?

CSCW AND AI

CSCW issues rather than the technological problems will provide the main future research challenges for CSCW. To illustrate the sorts of issues that are known from CSCW research to be problematic, four examples are selected and described below: conflict; democratic decision making; argumentation; and role definition. In each case, it is argued, that adding intelligent AI systems leads to further complications, and it is concluded that recognizing these sooner, rather than later, is desirable and wise.

First, a great deal of collaborative work involves conflict [12]. Conflict, per se, is not a bad thing, and in many instances it is desirable and needs to be exploited if collaborations are to be successful. As well as differing in many other ways, such as in knowledge, skills, personality and ethics, different people have different agendas. This is reasonable when they have different responsibilities. Collaboration is most challenging when these conflicts fail to be resolved.

How should conflict between a human and an AI be resolved? One could certainly imagine cases where an AI's preferences and agenda might be the desirable option. As an example, we can entertain a role for an AI to act as a mediator supporting conflict resolution between people - such an AI would always be in conflict with some of the collaborators, since mediation tends to require operations involving things such as bargaining and compromise. Would human collaborators be happy to compromise with an AI? What factors might influence their willingness to accept the AIs input, rather than overruling it?

Conflict may also be resolved through voting. There are arguments that voting is not actually a very good mechanism for small group decision making, with problems such as: empowering the largest minority and not a divided majority; polarization of opinions; choosing the least bad option; giving undue power to floating voters; and strategic voting and bargaining so that decisions are made based on unrelated and irrelevant, and potentially even corrupt, external factors. There are many types of voting systems and polling algorithms that lead to different group behaviors. Particular decisions may be reached because of the details of the democratic system used. Will users accept an enfranchised AI system with a vote of its own? How might those votes be weighted? How should the voting system be designed in order to support combining human and AI votes?

AI systems that understand language and can act on information received from other intelligences may also be open to persuasion. How will people adapt their arguments to a non-human intelligence, on what basis will they be able to influence it? How people will work with future non-human AIs will depend to a great degree on people's beliefs about them.

Furthermore, how might AIs learn to influence their human collaborators? Successful argumentation relies on an understanding of the audience [24]. If an agent's intelligence does not encompass a theory of mind, how can it understand its audience? Some form of prosthetic argumentation support could be considered, such as the systems described in [20], which can segment arguments and identify warrants and conclusions for example. But the question remains: to what degree will human collaborators accept persuasion from an artificial agent?

These three examples (conflict resolution, voting and argumentation) all rest on the answer to a fourth issue: what are appropriate roles for AIs? Roles can be official or unofficial, explicit or implicit, and fairly stable over time or rotating, evolving and ephemeral. It is not going to be easy to design roles for AIs in collaborative work systems, particularly when, in some cases, it is actually desirable for collaborators to take on poorly defined roles temporarily.

Learning is a fundamental property of artificial neural networks. There is no reason why future AI systems of this sort could not learn and evolve their own roles, as people frequently do. Task analyses, ethnographic studies or other data sources demonstrate that people have tacit understanding of their roles, and that they are not good at describing what those roles are. Why would we think an AI would be better than people at describing a role that it had evolved? Furthermore, is it necessary for an AI to be able to describe its role, if it is performing it usefully?

As roles change, co-adaptation is likely to lead to difficulties: in modern approaches to complex systems design in HCI (e.g. [4]), it is recognized that people adapt their behavior to the tools that are available. Collaborative AIs may also adapt to the user. Diaper [7] pointed out a number of problems with auto-adaptive computer systems, a major one is how to explain what has changed in the system to users. Such problems will remain and there seems no reason to suppose that AIs will be any better than people at explaining what they are learning.

There may be many other questions that the CABot agent poses with regard to collaborative working. These questions should be addressed sooner rather than later, as the answers will affect the type of intelligence we need future generations of CABot to include.

DISCUSSION

While our current, simulated neural networks are far too small to display a widely applicable range of behaviors, they do demonstrate the basic sensory, cognitive and response capabilities that a truly intelligent, non-human agent will need to work with people. Furthermore, we expect that chips with a billion neurons will be available within two years [19]. This new hardware will allow simulations of networks four orders of magnitude larger than at present. Developments might be fairly rapid, even within half a dozen years once that hardware is available, because the CABot approach allows its systems to learn and be self organizing and hence they will be genuinely autonomous.

While waiting for such hardware developments, the CABot simulations have demonstrated in miniature that many of the essential human mental capabilities are possessed by the systems that have already been developed, and which are being extended. CABot is a significant step towards intelligent agents that may collaborate with human users, or with other AIs. It is significant both in its architecture, which aims at biological fidelity, and in its behavior, which aims at cognitive fidelity. As such, human users may act towards CABot in the same way that they would act towards other people.

CABot is not yet ready to offer collaborative abilities on a par with its human counterparts. Even if it could, experience shows that collaboration would still be challenging. The results of CSCW and HCI research are therefore essential to the successful adoption of artificially intelligent collaborative agents. The intention of this paper is threefold. First, to sound an alert that the technical capability to finally produce AI agents capable of human level ratiocination may no longer be distant science fiction.

Second, to focus and initiate discussion of what sort of AIs we need, how we should use them and what problems must be overcome if people are to work collaboratively with artificial intelligences.

Third, we believe that incorporating an AI into human ways of working (rather than humans adapting to work with AIs) may be much harder to achieve than it first appears. Identifying these issues now might direct research in HCI and CSCW, we hope, in preparation for future technological developments in AI.

The challenges we foresee, such as conflict and role man-

agement, require new approaches to collaborative working. These approaches must take into account the expectations that human agents will have when interacting with an artificial agent with some human-like abilities. These approaches must also take into account the limitations that the agent has, which will make it difficult to fulfill such expectations.

We believe that the efforts of the CSCW community in supporting collaboration between human agents is ripe for extension into human-AI and AI-AI collaborative systems. This work should begin now, so that the technical and behavioral challenges can be met concurrently.

REFERENCES

1. Crystal space. Website, 2008.
2. J. Anderson and C. Lebiere. *The Atomic Components of Thought*. Lawrence Erlbaum, 1998.
3. J. Anderson and C. Lebiere. *How Can the Human Mind Occur in the Physical Universe*. Oxford University Press, 2007.
4. G. Cockton. Working spheres or engagements: Implications for designing. *Interacting with Computers*, 20(1):279–286, 2008.
5. J. Connolly and E. Edmonds. *CSCW and Artificial Intelligence*. Springer-Verlag, 1994.
6. D. Diaper. The discipline of human-computer interaction. *Interacting with Computers*, 1(1):3–5.
7. D. Diaper. Identifying the knowledge requirements of an expert system's natural language processing interface. In M. Harrison and A. Monk, editors, *People and Computers: Designing for Usability*, pages 263–280. Cambridge University Press, 1986.
8. D. Diaper. Will expert systems be safe? In *Second International Expert Systems Conference*, pages 561–572. Learned Information, 1986.
9. D. Diaper. Understanding task analysis for human-computer interaction. In D. Diaper and N. Stanton, editors, *The Handbook of Task Analysis for Human-Computer Interaction*. Lawrence Erlbaum Associates, 2004.
10. D. Diaper. Joint cognitive task design. *American Journal of Psychology*, 119(2):338–347, 2006.
11. D. Diaper and C. Sanger. Tasks for and task in human-computer interaction. *Interacting with Computers*, 18(1):117–138, 2006.
12. S. Easterbrook. *CSCW: Cooperation or Conflict*. Springer-Verlag, 1993.
13. S. Harnad. The symbol grounding problem. *Physica D*, 42:335–346, 1990.
14. D. Hebb. *The Organization of Behavior*. John Wiley and Sons, 1949.
15. E. Hollnagel. Prolegomenon to cognitive task design. In E. Hollnagel, editor, *Handbook of Cognitive Task Design*, pages 3–15. Lawrence Erlbaum Associates, 2003.
16. C. Huyck. Creating hierarchical categories using cell assemblies. *Connection Science*, 19:1:1–24, 2007.
17. C. Huyck. CABot1: a videogame agent implemented in flif neurons. In *IEEE Systems, Man and Cybernetics Society*, pages 115–120, 2008.
18. C. Huyck. A psycholinguistic model of natural language parsing implemented in simulated neurons. Submitted to Cognitive Neurodynamics, 2008.
19. M. Khan, D. Lester, L. Plana, A. Rast, J. Painkras, and S. Furber. SpiNNaker: Mapping neural networks ont a massively-parallel chip multiprocessor. In *Proceeding 2008 International Joint Conference on Neural Networks*, pages 2850–2857, 2008.
20. P. Kirschner, S. Buckinghan Shum, and C. Carr. *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making*. Springer, 2003.
21. J. Laird, A. Newell, and P. Rosenbloom. Soar: An architecture for general cognition. *Artificial Intelligence*, 33(1), 1987.
22. G. Miller. Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3(4), 1990.
23. P. Sell. *Expert Systems - A Practical Introduction*. Computer Science Series. Macmillan, 1985.
24. S. Stumpf and J. McDonnell. Is there an argument for this audience? In *5th Conference of the International Society for the study of Argumentation (ISSA)*, 2002.