AISB Quarterly

No.133

The Newsletter of the Society for the Study of Artificial Intelligence and Simulation of Behaviour

Artificial Intelligence and Cognitive Systems

I became involved in AI during the 1970s, when I was in graduate school, because I wanted to understand the nature of the mind. This seemed as though it were one of the core questions of science, on an equal footing with questions about the nature of the universe and the nature of life. Artificial intelligence, with its computational metaphor, offered the only clear course for tackling this challenging problem, and the progress made in the field's first 20 years, since its founding at the 1956 Dartmouth meeting, seemed impressive enough to promise rapid strides toward a broad computational theory of mental phenomena.

When I arrived at Carnegie Mellon University in 1975, and for the next 15 years, AI research drew upon a number of assumptions about the field's goals and about the approaches that might achieve them. In this essay I review these assumptions, the reasons they made sense, and the additional reasons, many of them sociological, they have fallen into disfavor among many AI researchers. After this, I consider whether they have a role to play in the future of the field and, if so, how we can encourage their increased use. I will refer collectively to these assumptions as the paradigm of *cognitive systems*, a term championed by Brachman and Lemnios [1].

High-Level Cognition

One key idea in this paradigm was that AI revolves around the study of *high-level cognition*. When we say that humans exhibit intelligence, we are not referring to their ability to recognize concepts, perceive objects, or execute complex motor skills, which they share with other animals like dogs and cats. Rather, we mean that they have the capacity to engage in multi-step reasoning, to understand the meaning of natural language, to design innovative artifacts, to generate novel plans that achieve goals, and even to reason about their own reasoning. During AI's first 35 years, much of the discipline's research dealt with these issues, and the progress during that period arguably increased our understanding of the mind.

This idea is still active in some AI subfields, such as planning and automated reasoning, although each has developed its own specialized methods, but, unfortunately, other subareas have effectively abandoned their initial concern with high-level cognition. For instance, machine learning, despite its early interest in complex tasks, now focuses almost exclusively on classification and reactive control, whereas natural language processing has largely replaced its original emphasis on understanding with text classification and information retrieval. These shifts have produced short-term gains with many applications and clear performance improvements on their narrowly defined tasks, but I question whether advances on these fronts tell us much about the nature of intelligence. A few researchers who take the cognitive systems perspective (e.g., Friedman, Forbus, and Sherin [2]; Scally, Cassimatis, and Uchida [3]) continue to address high-level behavior, but we need far more work in this important area.

Structured Knowledge

Another important assumption in early AI was that *structured knowledge* plays a central role in cognition, which in turn relies on the ability to represent and organize that knowledge. These claims depend on the fundamental insight—arguably the foundation of the 1956 AI revolution—that computers are not simply numeric calculators but rather general symbol manipulators. As Newell and Simon [4] state clearly in their physical symbol system hypothesis, intelligent behavior appears to require the ability to interpret and manipulate symbolic structures. The most impressive successes in AI's 55 year history, including the many examples of fielded expert systems, have relied on this fundamental capability.

Nevertheless, over the last 20 years, many branches of AI have retreated from this position. The increased popularity of statistical and probabilistic methods has reduced the fragility of traditional symbolic techniques, but only at great losses in representational power. Some subfields have almost entirely abandoned the use of interpretable symbolic representations, caring only about performance, however achieved. This trend is reminiscent of the behaviorist movement in psychology, which rejected the postulation of internal cognitive structures. Other subfields, like knowledge representation and constraint satisfaction, have retained a focus on symbols but limit the formalisms they consider for reasons of efficiency or analytical tractability.

Such developments constitute a step backward from the physical symbol system hypothesis, and they distract from efforts to fathom the complex nature of intelligence. Some scientists, such Fahlman [5] and Schubert, Gordon, Stratos, and Rubinoff [6], continue to assume less constrained formalisms, but we need more than a small cadre working in this arena.

Nowhere is this attitude more prevalent than in machine learning. Early work here dealt with the acquisition of symbolic cognitive structures, and there was a widespread assumption that mechanisms should produce easily interpreted declarative knowledge for use in reasoning, problem solving, or understanding. Machine learning initially aimed to support acquisition of the full range of structures used in knowledge-based systems, as contrasted with pattern recognition, which emphasized more constrained tasks like classification or categorization. As I have described elsewhere

In this issue:

Features and Reviews

Society

Father Hacker

AI and Cognitive Systems (cont.)

The AISB Quarterly
is published by the Society
for the Study of Artificial
Intelligence and Simulation
of Behaviour.

ISSN 0268-4179

Editor
Dr David Peebles
Department of Behavioural
and Social Sciences,
University of Huddersfield
aisbq@aisb.org.uk
http://www.aisb.org.uk/
aisbq/index.shtml

Advertising and Administration
Dr Katerina Koutsantoni
aisbq@aisb.org.uk
SSAISB Executive Office,
Institute of Psychiatry,
King's College London,
Addictions Sciences
Building, B3.06,
4 Windsor Walk, Denmark
Hill
London SE5 8AF
Tel: +44 (0)20 7848 0191
Fax: +44 (0)20 7848 0126

[7], in the late 1980s a number of factors converged to change this situation, including the influx of pattern-recognition techniques, the call for evaluation using metrics like classification accuracy, and the UCI repository's emphasis on attribute-value notations, which was well suited to statistical approaches. Early applications of machine learning [8] also focused on supervised learning with attribute-value notations, and the arrival of the data-mining movement and the World Wide Web in the mid-1990s demonstrated that many commercial problems fit this limited framework. Both emphasize induction of statistical predictors from large data sets, forgetting the original charter of machine learning was to acquire structured knowledge from limited experience.

System-Level Research

A third theme that characterized much early AI work was an emphasis on *system-level* accounts of intelligence. Because researchers envisioned comprehensive theories of the mind, they naturally recognized the need for their programs to comprise a number of interacting components. Many AI systems were given distinctive names that served as shorthand for a constellation of mutually supportive mechanisms. A related trend was the development of high-level programming languages, such as Prolog [9], each with a distinctive syntax that reflected its theoretical assumptions about intelligence. These two ideas merged in Newell's [10] notion of a *cognitive architecture*, which provided an infrastructure for building intelligent agents.

Despite these promising beginnings, by the 1990s many researchers had come to focus their energies on component algorithms rather than integrated systems. This resulted partly from AI finding its primary home in computer science departments, which gave higher status to the study of algorithms. Another influence was the emphasis on conference publications, which provided sufficient space to describe algorithms but not enough for system-level accounts. A third factor was the relative ease of evaluating algorithms, both formally and experimentally, which made it easier to produce and publish papers on such topics. Finally, university professors found it far simpler to teach AI as a set of unrelated algorithms than to present coherent frameworks for intelligent systems.

The results have been a greatly decreased interest in system-level accounts and the fragmentation of AI into a set of disconnected subfields. Research on cognitive architectures [11] provides some important counterexamples to this trend, but it is in the minority and intelligent systems deserve far more attention.

Heuristics and Satisficing

Another central assumption of initial AI research, also championed by Newell and Simon [4], was that intelligence involves *heuristic search*. Although not the only field to adopt the search metaphor, it was distinctive in its use of heuristics that, although not guaranteed to produce results, often made problems tractable which could not be solved otherwise. On this dimension, AI differed from other fields, such as operations research, that limited their attention to tasks for which one could find optimal solutions efficiently.

Instead, many AI researchers had the audacity to tackle more difficult problems to which such techniques did not apply. Their approach involved developing search methods that relied on heuristics to guide search down promising avenues and that *satisficed* (a term coined by Simon [12]) by finding acceptable rather than optimal solutions.

Unfortunately, recent decades have seen many AI researchers turn away from this practical attitude and adopt other fields' obsession with formal guarantees. For example, much recent work in knowledge representation has focused on constrained formalisms that promise efficient reasoning, even though this restricts the reasoning tasks they can address. Research on reinforcement learning often limits itself to methods that provably converge to an optimal control policy, even if the time required for convergence makes them completely impractical. Also, the popularity of statistical approaches has resulted largely from the belief, often mistaken, that techniques with mathematical formulations provide guarantees about their behavior.

We should certainly use nonheuristic methods when they apply to a problem, but this does not mean we should only study tasks that such techniques can handle. The original charter of AI was to address the same broad class of tasks as humans, but many now hope to redefine the field as something far more narrow. Of course, some work on heuristic approaches continues (e.g., Bridewell and Langley [13]; MacLellan [14]), but it is rare and often held in low regard by acolytes of the AI mainstream.

Links to Human Cognition

This point relates to another assumption prevalent in early AI research — that the design and construction of intelligent systems has much to learn from the study of *human* cognition. Many central ideas in knowledge representation, planning, natural language, and learning (including the importance of heuristic search) were originally motivated by insights from cognitive psychology and linguistics, and many early, influential AI systems doubled as computational models of human behavior, as did Newell, Shaw, and Simon's [15] General Problem Solver. The field also looked to human activities for likely problems that would challenge existing capabilities. Research on expert medical diagnosis [16], intelligent tutoring systems [17], artistic composition [18], and scientific discovery [18] were all motivated by a desire to support activities considered difficult for humans. In fact, the title of this publication reflects the early association between the two disciplines.

Even in AI's earliest days, few researchers attempted to model the details of human behavior, but many exhibited a genuine interest in psychology and in the ideas it offered. But as time passed, fewer and fewer adopted this perspective, preferring instead to draw their inspirations and concerns from more formal fields. Still worse, fewer chose

AI and Cognitive Systems (cont.)

to work on challenging intellectual tasks that humans can handle only with considerable effort or advanced training. Attention moved instead to problems on which computers can excel using simple techniques combined with rapid computing and large memories, like data mining and information retrieval.[1] There is no question that these efforts have had practical benefits, but they make no contact with psychology and they reveal little about the nature of intelligence in humans or machines. Again, some researchers continue to draw upon results about human cognition, but such efforts are few and far between.

The Future of Cognitive Systems

Despite these changes, I believe the assumptions and methods of the cognitive systems paradigm remain as valid now as they were over 50 years ago, in the first days of AI. They hold our best hope for achieving the original goals of our field, they have been abandoned by the mainstream for insufficient reasons, and they deserve substantially more attention than they have received in recent years. If so, then we should ask how we can resurrect interest in this approach to understanding intelligence and encourage its wider adoption within the research community.

One important avenue concerns education. Most AI courses ignore the cognitive systems perspective, and few graduate students read papers that are not available on the Web, which means they are often unfamiliar with the older literature. Instead, we must provide a broad education in AI that cuts across different topics to cover all the field's branches and their role in intelligent systems. The curriculum should incorporate ideas from cognitive psychology, linguistics, and logic, which are far more important to the AI agenda than ones from mainstream computer science. One example comes from a course on artificial intelligence and cognitive systems (http://cogsys.org/courses/langley/aicogsys11/) that I have offered at Arizona State University, but we need many more.

We should also encourage more research within the cognitive systems tradition. Funding agencies can have a major effect here, and the past decade has seen encouraging developments on this front. During this period, DARPA in the USA supported a number of large-scale programs with a cognitive systems emphasis [1], and the US Office of Naval Research has long shown a commitment to the paradigm. The European Union has also funded substantial projects (e.g., [20]) in the area.[2] Continued government support of cognitive systems research will aid progress, but we need committed people to join funding agencies as program officers to ensure that this occurs.

The field would also benefit from more audacious and visionary goals to spur the field toward greater efforts on cognitive systems. For instance, the General Game Playing competition (http://games.stanford.edu) has fostered research on general intelligent systems, and proposals for a 'cognitive decathlon' that would measure abilities on a set of well-defined cognitive tests is another good sign. But we also need demonstrations of flexible, high-level cognition in less constrained settings that require the combination of inference, problem solving, and language into more complete intelligent systems. The Turing test has many drawbacks but the right spirit, and we need more efforts toward integrated systems that support the same breadth and flexibility as humans. Challenging tasks will help excite both junior and senior researchers about the original vision of artificial intelligence.

Of course, we also need venues to publish the results of research on cognitive systems. From 2006 to 2011, the annual AAAI conference included a special track on 'integrated intelligence' that encouraged submissions on system-level results. The recent AAAI Fall Symposia on Advances in Cognitive Systems (http://www.cogsys.org/acs/2011/) attracted over 75 participants, and its organizers plan to launch a regular conference during 2012, along with an associated electronic journal. We need more alternatives along these lines to help counter the mainstream bias in favor of papers that report on narrow tasks, standalone algorithms, and incremental performance improvements. Broader criteria for scientific progress are necessary to advance the field, making room for papers that analyze challenging problems, demonstrate new functionalities, and replicate capabilities that are distinctively human.

In summary, the original vision of AI was to understand the principles that support high-level cognitive processing and to use them to construct computational systems with the same breadth of abilities as humans. As pursued within the cognitive systems paradigm, the field studied the content and representation of symbolic knowledge, the acquisition of such knowledge through learning, and the role of heuristic search in multi-step reasoning and problem solving. Much of this research focused on integrated systems rather than component algorithms, and cognitive psychology provided a source of ideas for these programs, many of which served as models of human behavior.

These ideas have lost none of their power or potential, and our field stands to benefit from their readoption by researchers and educators. Without them, AI seems likely to become a set of narrow, specialized subfields that have little to tell us about intelligence. Instead, we should use the assumptions of the cognitive systems approach as heuristics to direct our search toward true theories of the mind. This seems the only intelligent path.

Notes

1. Even for challenging problems like playing chess that require heuristic search, the vast majority of work has come to rely heavily on fast CPUs and large storage.

2. Not all work funded under these programs, in either the US or Europe, has focused on cognitive systems as we have defined them, but even researchers who hold views antithetical to those reviewed here are sometimes attracted to the movement's higher-level theme.

References

[1] Brachman, R., & Lemnios, Z. (2002). DARPA's new cognitive systems vision. *Computing Research News*, *14*, 1.
[2] Friedman, S., Forbus, K. D., & Sherin, B. (2011). Constructing and revising commonsense science explanations: A metareasoning approach. *Advances in Cognitive Systems: Papers from the 2011 AAAI Fall Symposium* (pp. 121–128). Arlington, VA: AAAI Press.
[3] Scally, J. R., Cassimatis, N. L., & Uchida, H. (2011). Worlds as a unifying element of knowledge representation. *Advances in Cognitive Systems: Papers from the 2011 AAAI Fall Symposium* (pp. 280–287). Arlington, VA: AAAI Press.
[4] Newell, A., & Simon, H. A. (1976). Computer science as empirical enquiry: Symbols and search. *Communications of the ACM*, *19*, 113–126.
[5] Fahlman, S. E. (2011). Using Scone's multiple-context mechanism to emulate human-like reasoning. *Advances in Cognitive Systems: Papers from the 2011 AAAI Fall Symposium* (pp. 98–105). Arlington, VA: AAAI Press.
[6] Schubert, L. K., Gordon, J., Stratos, K., Rubinoff, A. (2011). Towards adequate knowledge and natural inference. *Advances in Cognitive Systems: Papers from the 2011 AAAI Fall Symposium* (pp. 288–296). Arlington, VA: AAAI Press.

AI and Cognitive Systems (cont.)

[7] Langley, P. (2011). The changing science of machine learning. *Machine Learning*, *82*, 275–279.
[8] Langley, P., & Simon, H. A. (1995). Applications of machine learning and rule induction. *Communications of the ACM*, *38*, November, 55–64.
[9] Clocksin, W. F., & Mellish, C. S. (1981). *Programming in Prolog*. Berlin: Springer-Verlag.
[10] Newell, A. (1990). *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.
[11] Langley, P., Laird, J. E., & Rogers, S. (2009). Cognitive architectures: Research issues and challenges. *Cognitive Systems Research*, *10*, 141–160.
[12] Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*, 129–138.
[13] Bridewell, W., & Langley, P. (2011). A computational account of everyday abductive inference. *Proceedings of the Thirty-Third Annual Meeting of the Cognitive Science Society*. Boston.
[14] MacLellan, C. J. (2011). An elaboration account of insight. *Advances in Cognitive Systems: Papers from the 2011 AAAI Fall Symposium* (pp. 194–201). Arlington, VA: AAAI Press.
[15] Newell, A., Shaw, J. C., & Simon, H. A. (1959). Report on a general problem-solving program. *Proceedings of the International Conference on Information Processing* (pp. 256–264). Paris: UNESCO House.
[16] Shortliffe, E. H., & Buchanan, B. G. (1975). A model of inexact reasoning in medicine. *Mathematical Biosciences*, *23*, 351–379.
[17] Sleeman, D., & Brown, J. S. (Eds.) (1982). *Intelligent tutoring systems*. New York: Academic Press.
[18] Cohen, H. (1975). Getting a clear picture. *Bulletin of the American Society for Information Science*, December, 10–12.
[19] Langley, P. (1981). Data-driven discovery of physical laws. *Cognitive Science*, *5*, 31–54.
[20] Christiansen, H. I., Sloman, A., Krujiff, G.-J., & Wyatt, J. (2009). *Cognitive systems*. Berlin: Springer-Verlag.

Pat Langley is Professor of Computer Science and Engineering at Arizona State University in Tempe, Arizona, and Director of the Institute for the Study of Learning and Expertise in Palo Alto, California.

AISB Convention 2012

This will exceptionally take the form of the AISB/IACAP World Congress 2012, July 2nd to 6th, 2012, University of Birmingham, Birmingham, UK.

See http://events.cs.bham.ac.uk/turing12/ (accessible via link from the society webpage, http://www.aisb.org.uk).

It is organized by AISB together with the International Association for Computing and Philosophy (IACAP) [http://www.iacap.org/], in honour of Alan Turing.

The Symposia forming the conference have issued their separate calls for papers. Please consult the calls via the above Congress page. The Symposia are as follows:

- Mathematical Practice and Cognition II
- Hypercomputation and Artificial Intelligence
- Computing, Philosophy and the Question of Bio-Machine Hybrids (4th AISB Symposium on Computing and Philosophy)
- Computational Philosophy
- Turing Arts Symposium
- History and Philosophy of Programming together with a Roundtable Discussion on Philosophy of Computer Science: PoC Meets AI & Law
- AI & Games

Symposium Group on Turing Tests and Dialogue Agents:

- Revisiting Turing and his Test: Comprehensiveness, Qualia, and the Real World
- Linguistic and Cognitive Approaches To Dialog Agents (LaCATODA 2012)

Symposium Group on Social/Collective Systems, Networks and Phenomena:

- Social Computing - Social Cognition - Social Networks and Multiagent Systems
- Understanding and Modelling Collective Phenomena (UMoCoP)

Symposium Group on Ethics, Morality, AI and Mind:

- Framework for Responsible Research and Innovation in AI
- The Machine Question: AI, Ethics, and Moral Responsibility
- Moral Cognition & Theory of Mind

Symposium Group on Natural and Unconventional Computing:

- Natural Computing/Unconventional Computing and its Philosophical Significance
- Nature-Inspired Computing and Applications: 1st Symposium (NICA)

General information about the Congress

AISB and IACAP have joined forces to run the Congress. The Congress serves both as the year's AISB Convention and the year's IACAP conference. The Congress has been inspired by a desire to honour Alan Turing and by the broad and deep significance of Turing's work to AI, to the philosophical ramifications of computing, and to philosophy and computing more generally. The Congress is one of the events forming the Alan Turing Year (http://www.mathcomp.leeds.ac.uk/turing2012/).

The intent of the Congress is to stimulate a rich interchange between AI and Philosophy on any areas of mutual interest, whether directly addressing Turing's own research output or not.

Overall Chairs:

John Barnden
School of Computer Science
University of Birmingham, UK

Anthony Beavers
Philosophy and Cognitive Science
The University of Evansville
Indiana, USA

Local Chair and Deputy Programme Chair:

Dr Manfred Kerber
School of Computer Science
University of Birmingham, UK

Musical Acts: Communication for Musical Multi-Agent Systems

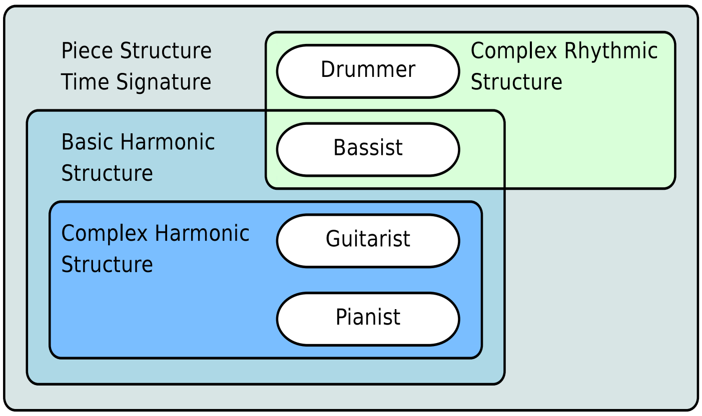


Figure 1: Different common grounds for a group of agents playing jazz

There is a long and rich history of interaction between music and computational processes; from Mozart's Musical Dice Game in which dice are cast to stochastically generate waltzes by combining pre-constructed material, through Lejaren Hillier's Illiac Suite to modern efforts such as David Cope's Experiments in Musical Intelligence (see [1] for a recent round up). There are strong links between music and formal systems — composers often use techniques which can be represented algorithmically, transforming melodies through inversion, transposition, augmentation and a host of other operators (e.g., [2, 3]). However, musicking [4, 5] is also a deeply human activity — we have emotional and physical responses to listening and performing, and individual aesthetics for judging the quality of music. It is this relationship between the formalisable and the human which makes music a fascinating and frustrating aspect of study.

There are many different points of departure for working with computers and music; is the music a score to be analysed? is the task to imitate a style or a composer? are the algorithms tools to be used under the direction of a composer or performer? are we trying to extract psychomusicologically salient features from music? It is common with many of the directions taken in the interface between machines and music to treat the music as an object to be analysed — whether a performance from which the microtimings of a particular performer are to be learnt or a score to be parsed using a formal grammar, it is the output which is analysed, rather than the processes which give rise to it. In contrast, with the MAMA system, we hope to understand what it is that happens as people play music; rather than looking at music in terms of the notes which are played, to examine it in terms of the interactions which give rise to it; in essence, to treat music as a communicative activity rather than an object of analysis. This raises a collection of issues, which we have chosen to tackle using a multi-agent systems approach.

A key quality of music is its polysemic nature [6]: there is no need for a complete agreement about what is happening; where we can expect natural language to have some grounding in an external world, this is only sometimes present in music — e.g. quoting of other pieces, copying distinctive animal calls — and does not form the bulk of the communicative impact. To make a semiotic division, the syntax of music is well studied, the semantics of music is difficult, so here we turn to the pragmatics of music [7], to understand the communicative import of music as played.

The goal of this work [8, 9], was to define a set of communicative actions, along the lines of Speech Acts [10] and existing agent communication languages [11], which are appropriate for musical activities: musical acts. These communicative acts allow computational systems to understand the interactive aspects of music, to help with analysis and performance. They are required to be: embodied through the production of music; intentional, in that they seek to change the course of the musical activity; intelligible by the other musicians involved.

An intuitive place to look for all these qualities is in difference or unexpectedness — playing which is exactly in line with what is expected can be assumed not to have any perlocutionary force. This requires the construction of some form of shared understanding of what is happening, and a way to differentiate new playing from what has gone before.

The first principle of the system is that music can be analysed for many different features — rhythmic, melodic, harmonic, spectral, structural etc. — with the constraint that the values of each of these features should be situated on a lattice. The lattice allows relations between values to be computed: are values extensions of each other, partially related or completely different? The set of features and the method of analysis is personal to the agent: for example the 'drummer' agent may have a detailed representation of rhythm, but a poorer understanding of harmony.

The 'musical now' is intended to capture the complex notion of what is happening 'now' in music, and, for a given agent, consists of the outputs of all of its analysis routines. These may be quite diverse in scope, so it might include ideas like '12 bar blues in C', 'rhumba rhythm', 'accents on the 2', 'getting louder' and so on. A set of agents will likely have divergent values for the 'musical now' — due to their differing analysis methods — so the notion of common ground is used to capture 'that which an agent can reasonably expect the other agents to understand is happening'. This musical context is formally constructed by inferring a set of features that appear to be shared with the other agents, and using this as the basis for analysis; Figure 1 gives an example of the common grounds constructed by an imaginary jazz group.

The playing of others can be analysed by comparison with the current values, and this is the basis for forming intentional actions: playing which differs from the context is assumed to be an intentional action, with the relations to other playing determining the type of action it is. The musical interaction may be collapsed to a set of states — musical contexts — which differ from each other by at least one value. The differences between these states are characterised in terms of the relations between the lattice values which comprise them, rather than the values themselves, to allow general patterns of communication to be observed (Figure 2). These abstracted sets of relations are termed 'musical action signatures', as they represent the communicative signature of the action performed. For example, one musician might introduce an idea, and then the other musicians might respond: taking up the idea — adopting the same value; embellishing it — an extension of the previous value; pointedly ignoring it — taking a disjunct value. Performative actions can be described in terms of musical action signatures, for example:

- **Propose** introduces a new value where there is no commonly agreed value — for example, the first notes played in a previously percussive piece would propose harmonic and melodic structures.
- **Confirm** agrees with something

Musical Acts (cont.)

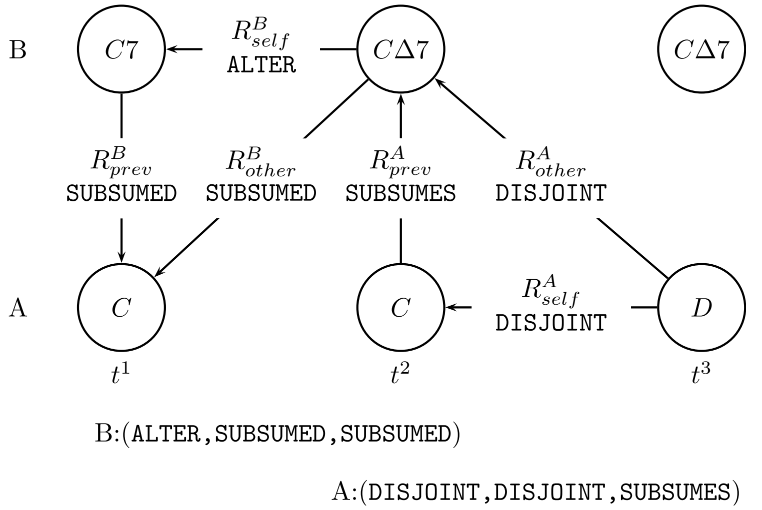


Figure 2: Action signatures extracted from the playing of two agents

that has been Propose'd, by taking it up and incorporating it in the agent's playing.

- **Alter** takes some aspect of the playing and changes it, while keeping a relationship with the original.

These are roughly at the level of the communicative actions specified in an agent communication language — they define preconditions, the form of the action and some inferences which may be drawn.

The ulitmate test of this formalism is whether it improves the interaction when playing with a human. This was tested using an experiment where pianists played several excerpts from a piece of minimalist music — constrained enough for easy analysis, but flexible enough to allow interaction —with a hidden partner. The hidden partner was one of the following: another human; a recording of a human, which has human expressiveness, but is not interactive; a straight rendering of the piece, which has no expressive or interactive qualities; the system set up to copy the expressive features of the human; and the system set up to use musical acts to inform its playing. The participant was asked to fill out a questionnaire about their partner for each excerpt, with the questions designed to separate expressive qualities — how much it sounded like good, human playing — from interactive elements — how well the partner responded to the performance as it progressed. Initial results were encouraging, implying that musical acts added to the interactivity of the system, but due to the small sample size, no statistically significant conclusions could be drawn.

Overall, Musical Acts seems to be a useful formalism for looking at musical interaction; it addresses several of the issues which we encounter when working with music — temporality, the polysemic and non-semantic nature of music, intentionality in interaction — and helps to translate useful concepts like common ground into a musical setting. The experiments performed so far are small and limited in scope, but we have shown that it is implementable in a real-time system, and we have an indication that it improves performance.

References

[1] Edwards, M. (2011). Algorithmic composition: Computational thinking in music. *Communications of the ACM*, *54(7)*:58–67.

[2] Mazzola, G. & Milmeister, G. (2006). Functors for Music: The Rubato Composer System. *Proceedings of the ICMC*. San Francisco: International Computer Music Association. 83–90.

[3] Pachet, F. (1994). The MusES system: An environment for experimenting with knowledge representation techniques in tonal harmony. *Proceedings of the 1st Brazilian Symposium on Computer Music*, Caxambu, Minas Gerais, Brazil. 195–201.

[4] Small, C. (1997). Musicking: A ritual in social space. In R. Rideout (Ed.) *On the Sociology of Music Education*, 3–4. Oklahoma: The School of Music, University of Oklahoma (1–12).

[5] Elliott, D. (1993). Musicing, listening, and musical understanding. *Contributions to Music Education*, *20*, 64–83.

[6] Cross, I. (2001). Music, cognition, culture and evolution. *Annals of the New York Academy of Sciences*, *930*, 28–42.

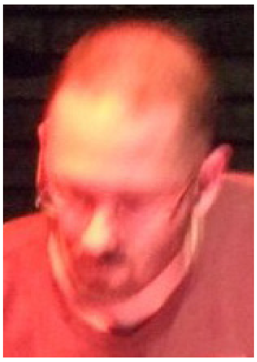
[7] Coventry, K. R. & Blackwell, T. (1994). Pragmatics in language and music. In M. Smith, A. Smaill, & G. A. Wiggins (Eds.) *Music Education: An Artificial Intelligence Approach*. Workshops in Computing, Springer, 123–142.

[8] Murray-Rust, D., & Smaill, A. (2011). Towards a model of musical interaction and communication. *Artificial Intelligence*, *175(9–10)*, 1697–1721.

[9] Murray-Rust, D. (2007). *Musical Acts and Musical Agents: theory, implementation and practice*. PhD thesis, Informatics.

[10] Searle, J. (1969). *Speech Acts*. Cambridge University Press.

[11] FIPA Specification. http://www.pa.org/repository/bysubject.html.



Dr Dave Murray-Rust
School of GeoSciences
University of Edinburgh

John McCarthy — some reminiscences

John McCarthy died aged 84 on 24th October 2011. Since then, much has been written about his life and work (e.g. search for his name and "homage", or "obituary"), and no doubt there will be much more. I shall not attempt to emulate or compete with any of the formal obituaries. Instead, I'll offer a few personal recollections and reflections.

There is also much to read on his web site[1] since he was one of the people who led the way in making everything he wrote freely available to all. It was from him that I learnt to cross out any part of a publisher's copyright agreement that restricted my right to post versions of my papers on my web site. Only one publisher has ever objected (so I withdrew the paper).

One of the most important events in my academic life occurred when Max Clowes, then the leading AI researcher at Sussex university, introduced me to AI, allowed me to attend his programming tutorials, and gave me things to read, by Simon, Newell, Minsky, McCarthy and others. It quickly became clear that AI was very relevant to old philosophical problems, especially in the papers I read by Minsky and McCarthy. One day Max suggested that I should read the 1969 paper by McCarthy and Hayes [1] and lent me his copy. I found it very interesting, especially the distinction between metaphysical, epistemological and heuristic adequacy of forms of representation of the world (echoing, but different from, the three kinds of adequacy in [2]).

However, I thought the main claim that a logical formalism would suffice for an intelligent machine was mistaken.

This (and much pushing by Max) provoked me into writing a dissenting paper presented at IJCAI 1971, subsequently reprinted in the AI Journal and elsewhere [3]. The McCarthy/Hayes formalism, first order predicate calculus enhanced with modal operators and fluents, was an example of a "Fregean" form of representation, i.e. one whose syntax used only function/argument structures, first identified as a core part of the structure of ordinary languages by Frege.

I could see that logical and other Fregean formalisms (including algebraic formulae, and other mathematical and programming notations) are very useful in many contexts, but I thought it far from obvious that the *only* form of representation required by an intelligent machine is a form of logic. My objection was that we need, and robots will need, different forms of representation for different purposes, and it is sometimes useful (both on epistemological and on heuristic grounds) to employ non-Fregean "analogical" representations (often mistakenly assumed to be isomorphic with what they represent). Examples of the latter include maps, diagrams used in proving geometric theorems, pictures of mechanisms that could be used to reason about causal connections, and 2-D pictures of 3-D scenes.

Programming languages that use syntactic ordering of commands to represent the temporal sequence of processing, or use syntactic ordering of items in a data-structure to represent ordering of items in some application domain, include "analogical" representations, in which properties of and relations between parts represent properties of and relations between things represented, though they need not be isomorphic with what they represent, since e.g. a 2-D picture can represent a 3-D object despite being far from isomorphic with it.

In many cases the information in an analogical representation can be re-formulated using a Fregean representation (e.g. specifying locations and orientations in fragments of terrain in a collection of logical assertions rather than a map) yet using the information in that form will often be dreadfully inefficient, because it loses the structural correspondences between representation and what is represented, which can lead to a loss of efficiency during searching, for example. The paper also showed how the notion of a "valid inference" could be extended to include inferences represented by manipulations of spatial representations, as in mathematical reasoning with diagrams — whether in the head, or on paper, in sand, etc.[2]

I remember that JMC attended my talk and that because we ran out of time we decided to continue the discussion after the final session that day. But I cannot now remember what he said in response! However, Pat Hayes later wrote a critical response [5]. As a result of writing that paper I was later able to spend a year (1972-3) in Edinburgh, in Bernard Meltzer's group, learning about AI and having my brain rewired, which substantially changed the subsequent direction of my teaching and research. So I owe a very great personal debt to McCarthy and Hayes.

Thereafter I met JMC occasionally at conferences, e.g. a conference in Edinburgh on "Expert Systems in the Microelectronic Age" organised by Donald Michie in 1979. In the discussion of the ethics of using AI in development of weapons (e.g. Cruise missiles), I remember JMC arguing that a good (and ethical) use of AI would be to enable a missile to fly down the chimney of a munitions factory and destroy it, instead of missing the target and destroying a civilian accommodation block.

In 1980, apparently as a result of reading [6], he invited me to visit Palo Alto, where he had a collection of researchers in AI and philosophy (including Dan Dennett, Pat Hayes, John Haugeland, and possibly one or two others) funded by the Sloan Foundation, meeting and talking about philosophy and AI at the Centre for Advanced Studies in the Behavioral Sciences. The other participants came for a year, but family and other commitments meant I could visit for only a month, a very interesting and enjoyable month. Alas, I don't have any detailed recollections of our discussions (though I recall writing comments on a draft version of "Beyond Belief" by Dennett). I also recall sitting at my desk in CASBS with screen and keyboard connected to a computer in the Stanford AI lab, via a modem that made a buzzing noise while transmitting (at about 9k bits/s). I think we used a text editor implemented by Art Samuel. We had neither mouse nor graphics in those days.

I think our next meeting was at IJCAI 1981 in Vancouver, where I presented a paper on emotions in robots, jointly authored with Monica Croucher [7]. JMC, like many others since then, had misread the paper as claiming that we should try to give robots emotions. Unlike most others, he objected that that would be a bad idea. I agree with him that if we want our robots to be useful we should try to minimise their emotionality.

However, our paper did not claim that robots *should* have emotions because they are desirable in intelligent systems, a claim that is often made, usually based on fallacious arguments[3]. Instead, we argued that there are resource constraints and knowledge limitations which require mechanisms that sometimes have to react quickly on the basis of partial knowledge, including sometimes overriding other, more intelligent, mechanisms, and that emotional states could result from the operation of such mechanisms. Similar points had been made earlier, by Herbert Simon, in response to Ulric Neisser's claim that only cold cognition, not hot cognition, could be explained or modelled computationally [8]. For machines with more knowledge and much greater computational power, such mechanisms might not be necessary, and avoiding such emotional episodes would be preferable. If I ever need brain surgery, I hope I'll have a completely unemotional but highly competent surgeon.

Thereafter, I met JMC from time to time at conferences and during visits

John McCarthy (cont.)

to Palo Alto, always finding our conversations interesting and rewarding. On one occasion we discussed limitations of symbol-grounding theory, the latest incarnation of the old philosophical doctrine of concept empiricism, much discussed by past philosophers, including Hume and Berkeley who regarded it as obviously true, and Kant who refuted the theory [9], and later philosophers of science who showed that many of the deep concepts of science (e.g. "neutrino", "gene", "magnetic field") could not be derived by abstraction from experience of instances.

An alternative, summarised in [10, 11], is that such concepts are implicitly defined by their role in an explanatory and predictive theory. But it may be difficult to make the theory rich enough to exclude all unwanted models, since in general any axiomatic system with undefined symbols can have multiple models in different parts of the universe. When I said, in one of our conversations, that there might be no alternative to using what [12] called "meaning postulates, which link theoretical statements with observable evidence and measurements to help at least partially restrict the possible interpretations (or "tether the theory"[4]), JMC responded that he thought it would always be possible to avoid the need for that by enriching the axioms in the theory. For any intended portion of the world it may be in fact possible to produce a unique identifying description (not including any references to particulars) even if we can never prove that the referent is unique. I don't think he provided an argument that it was always possible: he merely thought it was true. If he is right, then philosophical discussions about the "Twin earth" problem are ill-informed.[5]

This may be one of several philosophical debates in which philosophers wrongly conclude from the fact that they *think* that they can imagine something that they really *can* imagine it, or that it could possibly exist.

On one of my visits I noticed that his car bumper had a sticker saying "More people died at Chappaquiddic[6] than Three Mile Island"[7]. He was in favour of developing use of nuclear energy. People who have not encountered his commentaries on contemporary debates may enjoy this: http://www-formal.stanford.edu/jmc/commentary.html

I think it was during a visit in 1985 that he insisted on taking me for a spin in a two-seater plane that he liked to fly, from San Francisco airport. I was concerned about insurance, but could not find a way to refuse the invitation. The flight was certainly very enjoyable — until he could not make contact when he tried requesting permission to land. Nothing he tried made the radio work, so he decided to head for the airport hoping the controllers would understand what was happening and take charge of the situation. I asked whether the problem could be that the map we had been looking at had altered a switch centrally located above the windscreen. He was sure that switch was irrelevant. I pleaded with him to try it, and he did, and it worked. I guess he had never previously had to use it because the radio was always on. We were both using common sense reasoning, but with different premises!

We have both always had a strong emphasis on the importance of trying to unpack common sense (which includes a great deal of implicit knowledge and know-how) in order to identify both what needs to be explained by theories of how minds work, and what needs to be implemented in intelligent machines. But there were differences of emphasis: JMC was mostly interested in how we represent, reason about and make use of relatively abstract logically representable, information about the environment or the constraints on some collection of actions [13], whereas much of my interest was on continuous variation in structures, e.g. surfaces with changing curvature, such as a tea-cup, processes such as rotation of nut on a thread, or straightening a string, or getting a finger through the handle of a mug. That included wanting to understand how humans or machines can discover or prove theorems in Euclidean geometry by manipulating real or imagined spatial configurations, as human mathematicians often do.

A lot of progress has been made on JMC's problems. One of the reasons for the limitations of current robots is the lack of progress on the problems concerned with spatial structures and processes, including continuous variation. It's clear that there is considerable development regarding the latter in the first few years of a human's life, but what exactly that development amounts to is far from clear. I don't think the recent emphasis on embodied AI really addresses the problems [15]. I suspect JMC would agree.

I wish I had kept records of our interactions, which were neither frequent nor extended. I enjoyed our conversations and I think he did as well, though in retrospect I also wish I had pressed him harder on our points of disagreement. I had heard others say they found him difficult to converse with. I did not notice that, possibly because we had enough disagreements to discuss and enough shared assumptions to make the disagreements fruitful. I suspect we both were incompetent at small talk and social chat.

Moreover, he had always been interested in and fairly well-read in philosophy but probably did not often meet philosophers who could actually program and were doing AI research. He immediately accepted when I invited him to join me in a two hour special session entitled "A philosophical encounter" at IJCAI 1995, in Montreal. As requested, he submitted a two page summary of his position [16]. Marvin Minsky also accepted the invitation to take part, after some uncertainty as to whether he would be able to attend, which is why there is no paper from him in the proceedings.

During the discussion session I was amazed when Herbert Simon, who had made important contributions to philosophy, and who was the recipient of the IJCAI research achievement award that year, stood up and objected strongly to the inclusion of a philosophy session at an AI conference, as did Pat Hayes. Herbert Feigenbaum noted that it was the first occasion since the Dartmouth conference that so many of the founders of AI had been in the same room at the same time.

On another occasion, I forget when, I told him I was trying to defend a view that all life involves *information processing*, which contrasts with the mere ability to respond to physical forces. Whereas a non-living object's movements will normally be fully explained by the resultant of all forces acting on the object, like a ball rolling down a helter-skelter (designed artefacts, like mouse-traps, excepted), a living object will typically have a store of chemical energy whose deployment can be turned on or off at least partly under the control of the organism — using not only sensors detecting external states, but also internal sensors detecting needs, etc. He immediately pointed out that that characterisation is not general enough since some animals can use external forces whose deployment they control, e.g. a bird using air-currents to control some of its flight, using only a small amount of its own energy. This required a reformulation of the distinction.

There was a period of at least 10 years, possibly more, when John McCarthy, Marvin Minsky, and other well known AI figures were regular contributors to discussions, including philosophical discussions, on Usenet — before that medium was destroyed by the combination of universal access, allowing people with no relevant prior knowledge to pontificate at great length, and worse, the rise of spamming by advertisers. Before

John McCarthy (cont.)

that, there was something very valuable about people all over the planet, who had never met, ignoring all distinctions of status, presenting questions, arguments and counter-arguments on both technical problems in AI and also philosophical problems. I presume there are online records of all those interactions. I hope someone will one day produce an edited version without the spam and without the wasteful duplication usually included by those who have not learnt email discussion etiquette. JMC's contributions (and Minsky's) will be a major feature of such an archive.

At a workshop in 2002, Marvin Minsky mentioned McCarthy's 1996 paper "The well-designed child", an early version of [17]. So I looked it up soon after, liked it very much, and started recommending it to others. It was triggered by his reading [18]. The difference between psychologists who have no experience of the problems of designing working systems and thinkers like JMC, who do have that experience, is very striking. It should especially be read by all those AI researchers working on learning, who need to be reminded that:

"Evolution solved a different problem than that of starting a baby with no *a priori* assumptions.''

"Animal behavior, including human intelligence, evolved to survive and succeed in this complex, partially observable and very slightly controllable world. The main features of this world have existed for several billion years and should not have to be learned anew by each person or animal.''

Let's hope the next 50 years of AI and cognitive science research will be more strongly influenced than the last 50 years by that viewpoint, and the implication that in order to design human-like robots we need a deep understanding of the structure of the world that shaped our evolution, including the evolution of our potential to use logic! A slightly modified version of that paper was published as [17]. (I was honoured that the journal accepted my "follow on" paper for the same issue [19]).

There are many critics of so-called classical AI, or symbolic AI, whose criticisms are based on a very superficial (and usually biased) understanding of the breadth and depth of the problems addressed by AI. For instance, criticisms that early AI systems were mainly concerned with abstract problem solving and planning, as opposed to interacting with a dynamic environment ignore the fact that in the 1960s and early 1970s CPU speeds were reported in kilocycles per second, and memories of a quarter megabyte were rare. If it takes about 20 minutes for a computer vision system to find the rim of a mug in an image, dynamic interaction with the environment is not an option. The look, think, plan, act cycle was the only kind of design that could be used: concurrent visual servoing while using a hand to manipulate an object was out of the question. However I think it is fair to say that the founders, including JMC, seriously underestimated the difficulties of the tasks, and as a result made rash predictions that seriously harmed AI. I've never understood why they did not see the complexities. When Margaret Boden and I wrote about AI we found it obvious that the problems were very deep and would take many years to address [20, 6].

There's far more to McCarthy's work than I have touched on. A taste of the breadth of his influence can be found in the recent special issue of the AI journal on his legacy [21]. I recently stumbled across an interview by William Aspray [22] that may be of interest to those who would like to know more about the early days of AI at Stanford.

I don't believe his goal of basing all of AI on logic can be achieved, and I suspect he also realised that there are problems with that approach. What's important, however, is taking something as powerful as logic and pushing it as far as possible. That will help to identify the problems that need to be solved by combining logic based AI with alternatives.

We need an AI educational system that is much less factional and produces graduates with a broad and deep knowledge of the full range of approaches, their strengths, their weaknesses, the problems solved so far, and some of the hard unsolved problems. Alas we have instead a fragmented field with factions that pontificate on the basis of incomplete knowledge of both the problems and the achievements of the various strands. I don't think I heard JMC pontificate in that way, though he did show impatience with discussions that lacked mathematical or logical rigour. Fortunately for me, that did not stop him listening to my half-baked ideas, and commenting on them.

JMC will be remembered with approval by many different researchers, including both engineers trying to solve practical problems, and scientists and philosophers, trying to understand the world and what's possible. I would say he made one huge mistake, whose consequences will go on being harmful for a long time, namely naming the new field "Artificial Intelligence", rather than, for example, "Computational Intelligence", or the more cumbersome "Natural and Artificial Intelligence". The mistake is puzzling insofar as it is clear that from the start his interests went far beyond just trying to make useful machines. He was trying to understand human intelligence as one example of a space of possible forms of intelligence, and he hoped that eventually we'll be able to produce better forms than human intelligence — e.g. intelligent machines unencumbered by emotions. Moreover he understood very well that being that sort of scientist involved also being a philosopher, as shown by the title of the 1969 paper.

However, it was sometimes hard for philosophers to take him seriously, for example when he claimed that a thermostat has desires and beliefs [23]. I think that what he was trying to say was right, namely that even in a thermostat we can distinguish what I prefer to call "belief-like" and "desire-like" states, distinguished by what some philosophers have called "direction of fit"[8]. So I was delighted to read this blog entry a couple of days ago: "Got the Nest learning thermostat installed today. Neat! Pretty easy install. I had one issue where a wire was pressing down on one of the wire mounts, and that made the Nest think there was a wire plugged in there"[9].

Many of his slide presentations are on his web site[10] but don't work because the latex source does not include [landscape] on the top line. So anyone wanting to read the slides will have to fetch the latex files, edit and run. I've reported the problem to a member of his department. The last few times I met him it was clear that his health was deteriorating, at the AAAI Spring symposium in 2004, and AAAI 2006 in Boston, and most recently at the AAAI conference in August 2011 in San Francisco, when he was in a wheel chair. Alas it was not possible in the circumstances to follow up any of our loose ends.

Notes

1. http://www-formal.stanford.edu/jmc/

2. Interestingly one of the papers in the recent special issue of AIJ on *John McCarthy's legacy* includes a paper that attempts to show how a type of 3-D spatial puzzle that humans would normally reason about spatially can also be treated in a Fregean formalism — provided that someone has worked out how to express the problem in the appropriate form, which the authors do [4]. Whether and how a machine could do that re-formulation is a hard problem.

3. As explained in http://www.cs.bham.ac.uk/research/projects/

John McCarthy (cont.)

cogaff/talks/#cafe04

4. “Tethering” was suggested later by Jackie Chappell

5. http://en.wikipedia.org/wiki/Twin_Earth_thought_experiment

6. http://en.wikipedia.org/wiki/Chappaquiddick_incident

7. http://en.wikipedia.org/wiki/Three_Mile_Island_accident

8. http://en.wikipedia.org/wiki/Direction_of_fit

9. http://plus.google.com/109662899097006452835/posts

10. http://www-formal.stanford.edu/jmc/slides.html

References

[1] McCarthy, J., & Hayes, P. (1969). Some philosophical problems from the standpoint of AI. In B. Meltzer & D. Michie (Eds.), *Machine Intelligence, 4* (pp. 463–502). Edinburgh, Scotland: Edinburgh University Press. (http://www-formal.stanford.edu/jmc/mcchay69/mcchay69.html)

[2] Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

[3] Sloman, A. (1971). Interactions between philosophy and AI: The role of intuition and non-logical reasoning in intelligence. *Proc 2nd ijcai* (pp. 209–226). London: William Kaufmann. (http://www.cs.bham.ac.uk/research/cogaff/04.html#200407)

[4] Cabalar, P., & Santos, P. E. (2011). Formalising the fisherman’s folly puzzle. Artificial Intelligence, 175 (1), 346–377. (http://www.sciencedirect.com/science/article/pii/S0004370210000408)

[5] Hayes, P. (1984). Some problems and non-problems in representation theory. In R. Brachman & H. Levesque (Eds.), *Readings in knowledge representation* (pp. 3–22). Los Altos, California: Morgan Kaufmann.

[6] Sloman, A. (1978). *The computer revolution in philosophy*. Hassocks, Sussex: Harvester Press (and Humanities Press). http://www.cs.bham.ac.uk/research/cogaff/crp

[7] Sloman, A., & Croucher, M. (1981). Why robots will have emotions. In *Proc 7th int. joint conference on AI* (pp. 197–202). Vancouver: IJCAI. (http://www.cs.bham.ac.uk/research/cogaff/81-95.html#36)

[8] Simon, H. A. (1967). Motivational and emotional controls of cognition. In H. A. Simon (Ed.), *Models of thought* (pp. 29–38). Newhaven, CT: Yale University Press.

[9] Kant, I. (1781). *Critique of pure reason*. London: Macmillan. (Translated (1929) by Norman Kemp Smith)

[10] Sloman, A. (1985). What enables a machine to understand? In Proc 9th ijcai (pp. 995–1001). Los Angeles: IJCAI. http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#4

[11] Sloman, A. (1987). Reference without causal links. In J. du Boulay, D. Hogg, & L. Steels (Eds.), *Advances in artificial intelligence - ii* (pp. 369–381). Dordrecht: North Holland. http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#5

[12] Carnap, R. (1947). *Meaning and necessity: a study in semantics and modal logic*. Chicago: Chicago University Press.

[13] McCarthy, J. (1958). Programs with Common Sense. In *Proceedings Conference on the Mechanization of Thought Processes*. Teddington. (Also in [14]) and online http://www-formal.stanford.edu/jmc/mcc59.html

[14] Minsky, M. L. (Ed.). (1968). *Semantic Information Processing*. Cambridge, MA: MIT Press.

[15] Sloman, A. (2009). Some Requirements for Human-like Robots: Why the recent over-emphasis on embodiment has held up progress. In B. Sendhoff, E. Koerner, O. Sporns, H. Ritter, & K. Doya (Eds.), *Creating Brain-like Intelligence* (pp. 248–277). Berlin: Springer-Verlag. http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0804

[16] McCarthy, J. (1995b). What has artificial intelligence in common with philosophy? In Proc. *14th International Joint Conference on Artificial Intelligence*. http://www-formal.stanford.edu/jmc/aiphil/aiphil.html

[17] McCarthy, J. (2008). The well-designed child. *Artificial Intelligence*, *172 (18)*, 2003–2014. http://www-formal.stanford.edu/jmc/child.html

[18] Spelke, E. (1994). Initial knowlege: six suggestions. *Cognition*, *50*, 431–445.

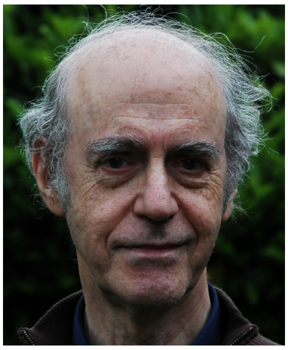
[19] Sloman, A. (2008). The Well-Designed Young Mathematician. *Artificial Intelligence*, *172 (18)*, 2015–2034. http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0807

[20] Boden, M. A. (1978). *Artificial intelligence and natural man*. Hassocks, Sussex: Harvester Press. (Second edition 1986. MIT Press)

[21] Morgenstern, L., & McIlraith, S. A. (2011). John McCarthy’s legacy. *Artif. Intell*, *175 (1)*, 1–24.

[22] Aspray, W., & McCarthy, J. (1989, March). *An Interview with John McCarthy Conducted by William Aspray*. http://purl.umn.edu/107476

[23] McCarthy, J. (1979). Ascribing mental qualities to machines. In M. Ringle (Ed.), *Philosophical perspectives in artificial intelligence* (pp. 161–195). Atlantic Highlands, NJ: Humanities Press. http://www-formal.stanford.edu/jmc/ascribing/ascribing.html



Prof. Aaron Sloman
School of Computer Science
University of Birmingham

Conference report: AISB *Convention. York, 4–7 April, 2011*

The AISB'11 Convention held this year between the 4th and 7th of April at The University of York hosted both academics and practitioners from a number of disciplines ranging from the social sciences to computer science and mathematics and was once again a great success.

I attended the Social Network Analysis and Multi-Agent Systems (SNAMAS) Symposium on the Monday and Tuesday. The first day comprised of the delegates getting to know one another and the facilitators finding out what people wanted to achieve by the end of the symposium. Some of the delegates were interested in furthering their knowledge in specific areas such as reputation, security or simulation within Multi-Agent Systems, whereas others were simply interested in the area as a whole and wanted to gain a greater understanding or to be informed of the current research and developments.

The second day consisted of back to back presentations, starting with the plenary speaker Professor Alan Baddeley discussing *Theories of Working Memory*. This presentation was extremely interesting, though of particular relevance to the delegates attending the Human Memory for Artificial Agents Symposium. Professor Baddeley spoke about how different types of information are split up within the brain for processing and storage. For example, the 'Central Executive' which is the master controller of the working memory system; the 'Visuo-Spatial Sketch Pad' which processes the visual data; the 'Phonological Loop' which is the component that holds the speech-based data; and finally the 'Episodic Buffer' which combines information from all other components into a single representation.

Professor Baddeley was followed by the SNAMAS introductory guest speaker Professor Cristiano Castelfranchi from the Institute of Cognitive Sciences and Technologies in Rome. Professor Castelfranchi works in the area of psychology and proved throughout the two days to be a great asset to the symposium; not only by discussing his paper on power networks, but also by sharing his twenty-five years of knowledge and experience in the area of artificial intelligence.

Power networks are a particular type of social network which concentrate on the power of an agent and how that agent creates power over other agents, how power is acquired and how it changes over time.

In his introductory speech Professor Castelfranchi discussed the structure and dynamics of networks along with agents and Multi-Agent Systems. He spoke of the importance of the semantics of the links within the network along with the different types of links. Professor Castelfranchi voiced his concerns on how some researchers and practitioners are creating models at an extremely high level of abstraction and omitting necessary detail. Later he discussed the different types of social networks including Dependence, Power, Negotiation and Trust Networks along with their key properties. The introductory speech was summarised by stating that there is often no one type of social network which needs to be modelled; it is the interaction between the different types that is required.

The speaker later presented his paper entitled *The Logic of Power: How my Power Becomes his Power* which built and elaborated on the power networks which had been touched upon during his introductory speech. He explained how agents can block or allow power via interference and how co-powers play an important role within the networks. He spoke of aversion and achievement, personal and role powers and of how both types are needed for an organisation to prosper.

Professor Castelfranchi's claim on Multi-Agent Systems, Social Simulation and Social Sciences is that only Agent-Based Computer Simulation can deal with these dynamics and its multi-layered nature. This is because it is necessary to model at the same time the specific mechanisms of each layer along with their feedbacks.

The specific mechanisms of each layer include: the internal mechanisms and representations controlling the behaviour of the agent; the interpersonal relations and interactions; the self-organising emergent effects and phenomena; their dynamic mechanisms; the collective actions and structure and the institutional ones. The layer feedbacks include: the emergence and bottom-up processes and the immergence top-down process that shape and reproduce the needed cognitive/behavioural condition for a given macro-process. Only Multi-Agent Systems can deal with this layered modelling. Multi-Agent Systems and Agent-Based Social Simulation will provide to the social sciences not only experimental platforms and data, and technical tools to formalise and support existing theories but new conceptual and theoretical tools, new models and theories. They will contribute to the current revolution: the birth of the "computational social sciences".

Julia Schaumeier, a PhD Student from Imperial College, London discussed her paper entitled *Adaptive Security Scheme for Open Networks*. This paper discussed the concept of "gossiping" and "forgiveness" amongst agents in a social network and the issues surrounding the transferring and dropping of packages. Other presenters included a mixture of computer and social scientists from around the world and presented their papers as follows:

Paul Chapron with "Analysis of Power Networks among the Actors of a Social Organisation"; Samuel Thiriot with "How detailed should social networks be for labor market's models?"; Sascha Holzhauer with "Considering baseline homophily when generating spatial social networks"; Enrico Franchi with "Selected Models for Agent-based Simulation of Social Networks"; and Mauricio Salgado with "Multilevel and Agent-Based Modelling in the Analysis of Differential School Effectiveness".

Below are listed some of the SNAMAS conclusions made by the delegates:

- Agent based simulation software could have a huge impact on policy makers; with the use of agent based simulation tools such as YANG, it is now possible to predict what effects policy makers actions of today will have in the future.
- The relationships between agents in a social network carry important properties which should not be omitted.
- More formal techniques are required for creating agent based systems.

Agents are either virtual or physical entities that have skills and behave autonomously across distributed environments, aware of and understanding their surroundings. In Multi-Agent Systems these intelligent agents interact with one another to solve an ever increasing host of complex problems.

Conference report: AISB *Convention 2011 (cont.)*

Social Network Analysis on the other hand looks at the mapping and measuring of social relationships in terms of network theory, providing both a visual and mathematical analysis. There are nodes and links which represent the people or groups and the relationships between them.

The idea of both Multi-Agent Systems and Social Network Analysis dates back to the mid twentieth century, though they were not seriously used in computing until the 1990s once processing power had increased adequately to cope with the highly intensive processes which these systems required. Recent application areas of Multi-Agent Systems include: modelling, planning and prediction; robotics and manufacturing; linguistics; wireless collaboration and communications; national security; property evaluation; and gaming.

These areas of application also often apply in the field of Social Network Analysis, though surprisingly recent work has taken place where Social Network Analysis has been used as an alternative technique for conducting accounting information systems related research.

Samantha Dixon
School of Computing & Creative Technologies
Leeds Metropolitan University

Conference report: IJCAI. *Barcelona, 16–22 July, 2011*

IJCAI - International Joint Conference on Artificial Intelligence - is a major conference in Artificial Intelligence. It has been held biennially since 1969, and is highly selective (in the 2011 edition, held in Barcelona, only 17% of the submitted papers were accepted). With 9 research tracks, it covers all aspects of AI, from Machine Learning to Robotics, from Natural Language Processing to Multi-agent Systems.

Within the Knowledge Representation research area, an interesting trend that is emerging concerns the creation of intelligent Information Systems. One invited talk by Wendy Hall from the University of Southampton was entitled "Toward a smarter Web". The speaker chronicled the birth, the principles and the development of the Web. She started working in information management in 1987, building a system to store all digitalised documents from a specific archive (the Mountbatten archive) and, more importantly, to reveal all their relevant interconnections.

This huge amount of work did not come for free, hence to access these data everyone had to pay a fee. In 1991 Tim Berners-Lee and Robert Cailliau published a paper describing an information system that was open, free, and universal — the Web. However, in such a huge system it is impossible to systematically create conceptual links between the pages.

A first solution to this problem came up in 1997, when Sergey Brin and Larry Page introduced a ranking that takes into account the interconnection between pages rather than the times the words of interest are repeated, and implemented the search engine Google.

Despite Google being a fundamental tool to discover relevant documents on the Web, the challenge still remains to improve the quality, consistency, and breadth of linking documents, on the Web or over a more general Information System. And this is where ontologies - and AI - come into Information Management.

An ontology is a finite set of axioms, i.e. sentences encoded in a formal language, of the knowledge about the elements of a domain of interest (e.g. biomedicine) and their interrelations. Such elements — concepts, roles, and individuals — are referred to with symbols. Ontologies seek to reveal relevant logical connections between the elements of a domain. However, designing, maintaining, and debugging ontologies is a hard work, because logic infers complex interrelations between elements. An emerging trend in this research area concerns the creation of tools to support the developers of ontologies — ontology engineers — in their hard job.

In their paper "Foundations of Uniform Interpolation and Forgetting in Expressive Description Logics", Carsten Lutz and Frank Wolter take into account two operations that can be applied to ontologies: given a set of symbols {SIGMA}, one can either define a *uniform interpolant* for {SIGMA}, or *forget* the {SIGMA} symbols.

A uniform interpolant for {SIGMA} is a description that uses only symbols from {SIGMA} and that does not change their meaning. Its dual operation, i.e. forgetting, consists of removing from the ontology all references to {SIGMA} symbols preserving the meaning of the symbols that remain. These operations are not always possible; for example, let us consider the following sentences: 'a Child has a Parent who is a Person' and 'a Person has a Parent who is a Person'.

Then, if we want to obtain an equivalent description for 'Child' without using the symbol 'Person', we are forced to include all sentences of the type 'a Child has a Parent ... who is someone', for as many times as we want to nest the 'being a Parent' relation. This is clearly inexpressible in a finite set of sentences. The authors achieve 3 major results: first, they provide a model-theoretic characterisation of ontologies where defining a uniform interpolant for a set of symbols {SIGMA} is possible.

Second, in this case they define an algorithm to compute it. Third, they prove that the size of the resulting description is at most triple exponential in the ontology size, and that, in general, no shorter interpolants can be found.

A related, but distinct, operation that can be performed over an ontology is the extraction of a *module*, i.e. a subset of the ontology that preserves all knowledge concerning a given set of symbols {SIGMA}. Whilst being a well-studied issue, extracting a single module does not provide any insight in the logical interrelation between fragments of an ontology. In fact, many modules seem internally incoherent, for example because the given {SIGMA} consists of conceptually unrelated elements, like {Car, Aubergine}.

Moreover, we cannot go through all modules and discard all incoherent ones, because in general ontologies contain too many modules. In the joint paper "The Modular Structure of an Ontology: Atomic Decomposition" with Bijan Parsia, Uli Sattler, and Thomas Schneider, we define genuine modules, internally coherent modules that make up a base for all modules of an ontology. We prove that surprisingly there is only a linear number (w.r.t. the ontology size) of genuine modules, and they can be obtained after a linear number of extraction.

By comparing genuine modules we can define atoms, that are maximal sets of axioms that never split over two modules, so that modules are disjoint unions of atoms. Axioms belonging to the same atom show a strong logical interrelation, because either they all appear in a module, or none of them does. Moreover, we can define a further logical relation: we say that an atom depends on another if every module containing the former contains also the latter.

To sum up, we are able to reveal logical relations enforced between fragments of ontologies.

Chiara Del Vescovo
School of Computer Science
University of Manchester

The Making of the EPSRC Principles of Robotics

In late 2010 the EPSRC unexpectedly invited me to attend a meeting in the New Forest on the topic of Ethics and Robots. I have been writing occasional articles on that topic since 1996, in response to my experience of being on Cog (a humanoid robot project) and seeing how readily and even insistently people attributed moral obligation towards a completely non-functional (in 1993) but vaguely humanoid robot. Since 1998 I've also maintained a web page on the topic. Nevertheless this was the first time I'd been approached by a government body, and of course I said yes.

The meeting was a three-day offsite chaired (expertly) by the journalist Vivienne Parry. Besides myself, other participants included Margaret Boden, Darwin Caldwell, Kerstin Dautenhahn, Paula Duxbury, Lilian Edwards, Ann Grand, Hazel Grian, Sarah Kember, Stephen Kemp, Paul Newman, Geoff Pegman, Andrew Rose, Tom Rodden, Tom Sorell, Mick Wallis, Shearer West, Blay Whitby, and Alan Winfield, as well as able assistance from Ian Baldwin, Denise Dabbs and Paul O'Dowd. Participants came mostly from robotics, but also from the humanities, law, and social science. We were employed mostly in academia, but also by industry and the research councils (including the then-head of the AHRC, Shearer West).

I was surprised the EPSRC would splash out for so many people for so long in such a nice hotel on this topic, but they made the object of their concern very clear quite early. The EPSRC sees robotics as a critical technology for the UK, and does not want to see it face the same fate as other "futurist" technologies have, in terms of public distaste bordering on hysteria that can no longer be addressed by any amount of measured scientific assessment. The EPSRC wants to get robot ethics right from the beginning, to ensure both the safety and the acceptance of robotic technologies.

It became almost immediately apparent that they had succeeded in selecting a very pragmatic and socially-concerned group of experts. The group took a very strong line on what the moral and ethical role of robotics could be, and one that I would not say is the dominant one at typical AISB gatherings.

On the final full day, Internet Law professor Lilian Edwards and I were in a small break-out meeting together in a session intended to design deliverables as outcomes for the meeting. We decided to make a "real" set of laws for robots. Lillian was keen to have them clearly follow but correct Asimov's laws, while I was keen to include several I'd already developed while writing A Proposal for the Humanoid Agent-builders League (HAL) [1]. In the end we settled on five, the first three of which reflect and refract Asimov to the concerns of the group. The group as a whole then refined not only our "laws", but also ordinary language versions of these, and developed a further list of concepts to be communicated to you, our colleagues.

The full version of these documents can now be found by Googling the EPSRC Principles of Robotics, and they have become EPSRC policy since April of 2011. The five basic principles are reproduced at the end of this article. Below are seven high-level ideas that the group wants to communicate to our colleagues. For detailed explanations, please see the website, but I have given the highlights here.

1. *We believe robots have the potential to provide immense positive impact to society. We want to encourage responsible robot research*. We are not a bunch of luddites who "don't get" the real potential of AI. We are concerned professionals who really do want to make AI work and robots real.
2. *Bad practice hurts us all*. We can't ignore the situation if some of our colleagues do things that make all of us look bad.
3. *Addressing obvious public concerns will help us all make progress*.
4. *It is important to demonstrate that we, as roboticists, are committed to the best possible standards of practice*.
5. *To understand the context and consequences of our research we should work with experts from other disciplines including: social sciences, law, philosophy and the arts*. We were all struck by how much we learned from this multi-disciplinary working team.
6. *We should consider the ethics of transparency: are there limits to what should be openly available?* Everyone at the meeting was committed to open-source-software type solutions and approaches, but we came to realise that with robots and AI more generally we do have the obligation to make sure that every "script kiddy" couldn't hack into a system that has information or memory about the private lives of humans.
7. *When we see erroneous accounts in the press, we commit to take the time to contact the reporting journalists*. Most science reporters really don't want to be made to look silly by reporting an "expert" who turns out to be self-promoting or sensationalist. A quiet word or email can often damp hysteria being generated by irresponsible statements.

I would like to thank the EPSRC and also our colleagues who advocated for this meeting. Two of these latter were Alan Winfield and Tom Rodden. Personally I feel extremely proud and happy for my profession and nation that the UK now has an official set of robotics principles that address such important matters. But we are only one country, and there is still much work and advocacy to be done to ensure that intelligent robotics are used appropriately in our society.

Five principles of robotics

1. *Robots are multi-use tools. Robots should not be designed solely or primarily to kill or harm humans, except in the interests of national security*. While acknowledging that even dead fish can be used as weapons by creative individuals, we were concerned to ban the creation and use of autonomous robots as weapons. Although we pragmatically acknowledged this is already happening in the context of the military, we do not want to see these used in other contexts.
2. *Humans, not robots, are responsible agents. Robots should be designed & operated as far as is practicable to comply with existing laws & fundamental rights & freedoms, including privacy*. We were very concerned that any discussion of "robot ethics" could lead individuals, companies or governments to

abrogate their own responsibility as the builders, purchasers and deployers of robots. We felt the consequences of this concern vastly outweigh any "advantage" to the pleasure of creating something society deigns sentient and responsible. This was the law we knew would most offend some of our colleagues in AISB — consequently (with David Gunkel) I am running a symposium at AISB 2012 to examine whether this is a reasonable rule. The symposium is called "The Machine Question: AI, Ethics and Moral Responsibility".

3. *Robots are products. They should be designed using processes which assure their safety and security*. This principle again reminds us that the onus is on us, as robot creators, not on the robots themselves, to ensure that robots do no damage.
4. *Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent*. This was the most difficult rule to agree on phrasing for. The idea is that everyone who owns a robot should know that it is not "alive" or "suffering", yet the deception of life and emotional engagement is precisely the goal of many therapy or toy robots. We decided that so long as the responsible individual making the purchase of a robot has even indirect (e.g. Internet documentation) access to information about how its "mind" works, that would provide enough of an informed population to keep people from being exploited.
5. *The person with legal responsibility for a robot should be attributed*. It should always be possible to find out who owns a robot, just like it is always possible to find out who owns a car. This again reminds us that whatever a robot does, some human or human institution (e.g. a company) is liable for its actions.

References and additional resources

[1] Bryson, J. J. (2000). A proposal for the Humanoid Agent-builders League (HAL). In J. Barnden, (ed.) AISB'00 Symposium on Artificial Intelligence, Ethics and (Quasi-)Human Rights. 1–6.

http://www.epsrc.ac.uk/ourportfolio/themes/engineering/activities/Pages/principlesofrobotics.aspx

http://machinequestion.org/symposium/

http://www.cs.bath.ac.uk/~jjb/web/ai.html

Wilks, Y., (Ed.) (2010). *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues*. John Benjamins, Amsterdam.

Dr Joanna J Bryson
Department of Computer Science
University of Bath

Society News

AISB Committee membership 2012

Following recent elections there will be some changes in the composition of the AISB committee, with effect from January 2012.

- Aladdin Ayesh, Savas Konur and Fiona McNeill are standing down.
- Etienne Roesch, Floriana Grasso, Colin Johnson (currently co-opted) and Mark Bishop (previously elected Jan 2009 - Jan 2012) are elected unopposed.
- Dimitar Kazakov's term as a co-opted member is extended until January 2013.
- Yasemin Erden joins the committee as a co-opted member.

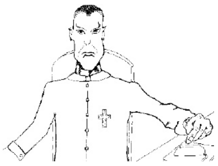
The AISB constitution sets a maximum of 13 elected members, though the committee may co-opt an unspecified number of non-voting members in addition. Both co-opted and elected members must stand for (re-)election after three years. The current committee members are listed at www.aisb.org.uk/secretary/committee.shtml

On behalf of the committee I would like to thank the retiring members for their contributions to the Society over the years, and to welcome the new members.

Rodger Kibble
Secretary and Returning Officer, AISB.

AISB Convention 2012

See page 4



Dear Aloysius...

Fr. Aloysius Hacker answers your questions

Cognitive Divinity
Programme
Institute of Applied
Epistemology

About the Society
The Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB) is the UK's largest and foremost Artificial Intelligence society. It is also one of the oldest-established such organisations in the world.

The Society has an international membership of hundreds drawn from academia and industry. We invite anyone with interests in artificial intelligence or cognitive science to become a member

AISB membership includes the following benefits:

- Quarterly newsletter
- Student travel grants to attend conferences
- Discounted rates at AISB events and conventions
- Discounted rates on various publications
- A weekly e-mail bulletin and web search engine for AI-related events and opportunities

You can join the AISB online via:

http://www.aisb.org.uk

Dear Aloysius,
Some strange things have been happening to me recently and I'm wondering whether you, with your wealth of experience of the shadier side of life, can shed any light on them. About a year ago, I got an anonymised request for a draft of a paper I was working on. I thought nothing of it, but when the paper was eventually published, I got a flaming letter from the same source accusing me of altering the draft and threatening me with unspeakable consequences if I ever did this again. What's going on?
Yours, Scared

Dear Scared,
You'll recall the recent court case against some Pakistani bowlers over 'spot betting' on 'no balls' during cricket matches. All this adverse publicity means that sports matches are no longer a lucrative hunting ground for the gamblers who ran this industry. They have, therefore, sought alternative outlets for their particular skills. What more innocuous target than academic research publications. Early sight of your paper enabled them to predict accurately such things as the location of the first typo, the first use of a particular technical term, etc. Unfortunately for them, your subsequent editing of your paper rendered their prediction false – and lost them a lot of money. Hence, the threats.

If want to live a quiet and safe life, I suggest you refuse to reveal the drafts of your papers to anyone not personally known to and trusted by you. On the other hand, if you like adventure, then you could cash in on this new opportunity. Subscribe to SWINDLE™ (Society for Wagers Insured to Nullify Damaging Losing Events), the Institute's spot-betting consortium. Your academic connections will provide ample opportunity to make reliable spot bet predictions. Convey these opportunities to SWINDLE™ and bets will be placed anonymously on behalf of the whole consortium, with the winnings divided proportionately. SWINDLE™ also takes spot bets from gamblers outside the consortium but, again, our inside knowledge and influence ensures that we rarely pay out. Journal editors in the consortium are especially effective in ensuring that minor corrections can be made to papers even after the final copy has been submitted to the publishers.
Yours, Aloysius

Dear Aloysius,
The 2012 Olympics in London has provided my robotics research group with an unparalleled opportunity to demonstrate our running robot, R...RUSH™ (Robot Runner Undertakes Sprints and Hurdles). Our experiments show that R...RUSH™ can outrun Usain Bolt over 100 metres. Our challenge is going to be persuading the Olympic officials to allow him to enter the competition. Can you offer us any advice?
Yours, Coach

Dear Coach,
What a wonderful opportunity. Your first problem will be getting R...RUSH™ into a national team. I'd suggest a small country with so much enthusiasm to reverse a poor track record that they will not ask awkward questions. Myanmar (Burma), for instance, did not compete in the 100 metres in 2008. I have excellent contacts there so, for a small consideration, may be able to help oil the wheels for you. Assuming that R...RUSH™ is a convincing humanoid, your next problem is going to be passing the random drugs test. Have you equipped R...RUSH™ with the capacity to pass urine and give blood? If not, our PISS™ (Pass Inspection via a Source of Solutions) attachment may help you, as it has helped numerous human athletes.
Yours, Aloysius

Dear Aloysius,
I am the leader of a large Eurozone country. You may be aware of the difficulties that the Eurozone has been facing due to sovereign debt. We've tried treaties, entreaties and threats, but nothing seems to work. If we are to avoid a meltdown, we need a miracle. I've long been impressed by the miraculous results that you and your Institute have achieved with your Artificial Intelligence research. Can you help us?
Yours, Leader

Dear Leader,
Whether you own them or owe them, large sums of money are no longer recorded in bars of gold or even wads of bank notes, but in numbers in a bank's database. For a relatively small fee, our new service, ERODE™ (Economic Rectification Of Debt European), can exploit this fact to help solve your debt problem. Banks carry out billions of financial transactions a second. Each such transaction provides an opportunity to round down a credit and transfer the difference into another account.

ERODE™ can hack into a bank and transfer several tiny, undetectable sums into a Eurozone sovereign debt account each nanosecond. Billions of such tiny reductions add up to one quite large reduction each second - and a billion of these large reductions, spread over a large number of banks, will solve your sovereign debt problem in a few years. For an extra consideration we can exact your revenge on the bankers by eroding their bonuses.

Yours, Aloysius

Agony Uncle Aloysius, will answer your most intimate AI questions or hear your most embarrassing confessions. Please address your questions to fr.hacker@yahoo.co.uk. Note that we are unable to engage in email correspondence and reserve the right to select those questions to which we will respond. All correspondence will be anonymised before publication.