AISB Quarterly

THE NEWSLETTER OF THE SOCIETY FOR THE STUDY OF
ARTIFICIAL INTELLIGENCE AND SIMULATION OF BEHAVIOUR

No. 136 May, 2013

All the latest Society information is available
at the newly renovated AISB website:
http://aisb.org.uk

Editorial

Welcome to Q136—a little later than expected—for which we apologise. We'll do our best to catch up in the remaining months of 2013 to ensure the Q lives up to its name.

Thank you for your comments on the new format, which we are pleased to say have been overwhelmingly positive. We are still tweaking the design slightly so if you have any comments or suggestions please get in touch. To speed up the production process we are also currently developing a new set of formatting guidelines for submissions and templates for those of you who are familiar with LaTeX. These will be available on the website shortly.

Since the last issue, the AISB annual convention was held in Exeter. This was widely considered to be a great success and a credit to the co-chairs, Dr Ed Keedwell and Prof Richard Everson. Reports of the various activities will be featured in future editions of the Q later this year.

Before diving into the content of this issue, we would like to extend a warm welcome to Dr Nir Oren from the University of Aberdeen, who has recently joined the AISB Committee. Nir will join Colin Johnson in the Public Relations task force. Welcome Nir!

The substantive core of this issue is three original articles. The first is a piece by Prof Paul Rosenbloom describing his exciting and ambitious project to construct a functionally elegant, grand unified, cognitive architecture which he aims to develop into a new form of unified theory of human cognition. The system, called *Sigma*, is based on graphical models and piecewise continuous functions and emulates abilities such as memory, learning, decision making and perception in support of virtual humans and intelligent agents/robots.

The second article, by Professors Noel Sharkey and Lucy Suchman, is an extended argument against the use of autonomous military robots and a call for an international treaty to prohibit the development, production and use of fully autonomous weapons.

The third article, by Dr Dean Petters, describes his computational approach to understanding emotions—in particular social and emotional phenomena described by Bowlby-Ainsworth Attachment Theory and phenomena linked to loss of control of various kinds.

Finally, Sam Devlin reviews the book "Metareasoning, Thinking About Thinking" by Cox & Raj (2011).

We are always seeking submissions in the form of long or short articles, book or event reviews. If you would like to contribute to the Q, just email us at aisbq@aisb.org.uk

The Q editors

The Sigma Cognitive Architecture and System

Sigma (Σ) is a nascent *cognitive system* — an integrated computational model of intelligent behaviour, whether natural and/or artificial — that is based on a novel *cognitive architecture*: a model of the fixed structure underlying a cognitive system [1]. The core idea behind Sigma is to leverage *graphical models* [2, 3] — with their ability to yield state-of-the-art algorithms across the processing of signals, probabilities and symbols from a single representation and inference algorithm — in constructing a cognitive architecture/system that meets three general desiderata: *grand unification*, *functional elegance* and *sufficient efficiency*.

A *unified cognitive architecture* traditionally attempts to integrate together the complementary cognitive capabilities required for human(-level) intelligent behaviour, with appropriate sharing of knowledge, skills and uncertainty among them. A *grand unified architecture* goes beyond this, in analogy to a grand unified theory in physics, to include the crucial pieces missing from a purely cognitive theory, such as perception, motor control, and emotion. This shifts issues of *embodiment*, *grounding* and *interaction* into the foreground, to converge with work on robot and virtual human architectures, but without then relegating traditional cognitive concerns to the background. Sigma approaches grand unification via a *hybrid* (discrete + continuous) *mixed* (symbolic + probabilistic) combination of: (1) graphical models, in particular *factor graphs* with the *summary product algorithm* [4]; plus (2) *piecewise-continuous functions* [5].

Functional elegance implies combining broad functionality — grand unification in this case — with simplicity and theoretical elegance. The goal is something like a set of *cognitive Newton's laws* that yield the required diversity of behaviour from interactions among a small set of general primitives. If the primitives are combinable in a flexible enough manner, new capabilities continue to flower without the need for new modules; and integration occurs naturally through shared primitives. The Soar architecture, many of whose strengths and weaknesses have inspired choices in Sigma, took a similar approach in its early years [6], while AIXI can be seen as striving for an extreme version of it [7]. Although doubts remain as to whether natural cognitive systems are elegant in this manner — as opposed to mere evolutionary patchworks — and whether such elegance is even computationally feasible in artificial cognitive systems, developments such as rational analysis on the natural side [8] and graphical models on the artificial side provide continued promise. Despite the questions, functional elegance maintains its allure because, if it is in fact achievable, it should yield deeper and more elegant theories with broader scientific reach [9] that are ultimately easier to understand and apply.

Sufficient efficiency implies cognitive

systems that execute quickly enough to support their anticipated uses. On the artificial side, the primary issue is speed of execution, but joined at times with boundedness. Graphical models potentially play a key role here, as they not only yield broad functionality in an elegant manner, but also state-of-the-art performance across this breadth. On the natural side, the primary issue is whether behaviour is modelled at appropriate human time scales, independent of how much real time is required. Yet speed is also an important secondary consideration here, particularly as experiments and models scale up.

The remainder of this article summarises progress to date in achieving a functionally elegant grand unification in Sigma. First, the currently implemented architecture/system is described at a high level. Then results are summarised across memory and learning, perception and mental imagery, decisions and problem solving, multiagent systems and theory of mind, and natural language. Sufficient efficiency is not a major focus in these results, other than indirectly through the pervasive use of graphical models; although significant progress has been made on aspects of it [10], more is required before Sigma will be ready to support complex real-time virtual humans and robots.

Sigma

The term cognitive architecture derives from an analogy with computer architecture, the fixed structure of a computer that provides a programmable system (that is, a *machine language*). In a cognitive architecture the concern is with the fixed structure that provides a (machine) language for expressing the knowledge and skills that comprise the learnable content of the cognitive system. But a computer system isn't just an architecture plus content, and nor necessarily is a cognitive system. Sigma is presently composed of three main layers: (1) the cognitive architecture; (2) knowledge and skills included on top of the cognitive architecture; and (3) the analogue of a *firmware* (or *microcode*) *architecture* that sits beneath the cognitive architecture. The cognitive architecture provides a language of *predicates* and *conditionals* that blend ideas from rule-based systems and probabilistic networks. It directly supports the layer of knowledge and skills on top of it. A firmware architecture traditionally provides a programmable level in between what is directly supported by hardware and what is desired in the computer architecture. Sigma's firmware architecture bridges its underlying implementation language (Lisp) and its cognitive architecture via a language of factor graphs and piecewise continuous functions (into which predicates and conditionals are compiled for execution). In this section, we first explore Sigma's firmware architecture and then its cognitive architecture. Its present knowledge and skills are implicit in the results discussed in the next section.

Factor graphs, in common with other forms of graphical models — such as Bayesian and Markov networks, and Markov and conditional random fields — provide an efficient means of computing with complex multivariate func-

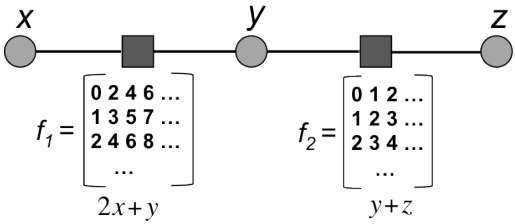


Figure 1: Factor graph for the algebraic function $f(x, y, z) = y^2+yz+2yx+2xz = (2x + y)(y + z) = f_1(x, y)f_2(y, z)$.

tions by decomposing them into products of simpler functions and then translating them into graphs for solution. From such graphs, the marginals of the individual variables — i.e., the function's values when all other variables are summarised out — can be computed efficiently, as can the function's global mode; for example, yielding maximum a posterior probability (MAP) estimation. Factor graphs in particular are undirected bipartite graphs that combine *variable nodes* for the variables in the function with *factor nodes* for the factors into which the function decomposes (Figure 1). Each factor node embodies a function, and connects to all variable nodes used in the function. Each variable node connects to the factor nodes that use it. Unlike Bayesian networks, factor graphs can be applied to arbitrary multivariate functions, not just to probabilistic ones.

The representation used for factor functions in the graph is a critical determinant of the expressibility of the resulting system. Sigma supports a hybrid mixed approach via a core representation based on piecewise continuous functions, which at present are limited to piecewise linear. The domain of each factor function is the cross product of its variables, implying an n-dimensional function when there are n variables. The overall function is specified in a piecewise linear manner over an array of rectilinear regions (Figure 2). This representation is general enough to approximate arbitrary continuous functions as closely as desired. Furthermore, restrictions on these functions — for example, to unit intervals with constant functions — can yield both discrete and symbolic functions. Functions can also be hybrids if they comprise multiple variables of different types.

The processing cycle in Sigma's firmware architecture consists of a *graph-solution phase* followed by a *graph-modification phase*. Solving a factor graph requires applying one of the many inference algorithms available for computing the values of variables in graphical models. Such a solution typically involves providing *evidence* for some of the variables — for example, by fixing their values via functions in peripheral factor nodes — and

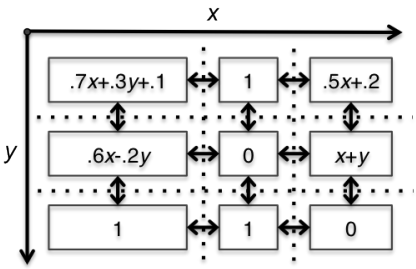


Figure 2: Bivariate function as a 2D array of regions with linear functions.

then either computing the marginal distributions over the other variables individually or the modal value jointly over all of them. A message passing approach based on the summary product algorithm is used in Sigma to compute both marginals and modes (Figure 3). Messages are sent in both directions along links, from variable nodes to neighbouring factor nodes and from factor nodes to neigh-boring variable nodes. This overall representation and processing is supported in Sigma's firmware architecture via four memories, for factor nodes, variable nodes, links, and messages (caching the last message sent in each direction along each link).

A message along a link always represents a distribution over the variable node's variables irrespective of its direction. When such a message is received at a variable node a new outgoing message is generated along each of its other links as the pointwise product of the incoming messages. This is the *product* aspect of the summary product algorithm. If the node is a factor node, the same pointwise product is computed, but included in the product is also the node's own function. Unlike at variable nodes, where the outgoing message is simply this product, further processing is required to compute the outgoing message here. Because the product includes all of the factor node's variables, not just those corresponding to the variable node on the outgoing link, all other variables must be *summarised* out before the message is sent. When computing marginals, Sigma uses *integration* for summarisation. When computing modes it uses *maximum* instead.

The natural stopping criterion for the graph-solution phase — and thus the trigger for the start of the graph-modification phase — is *quiescence*; that is, when no significantly different messages remain to be sent. Sigma's message memory is modified dynamically during the graph-solution phase, but the graph-modification phase is ultimately responsible for altering the other three memories. At present, the graph-modification phase can alter functions maintained within factor nodes, in support of updating the cog-

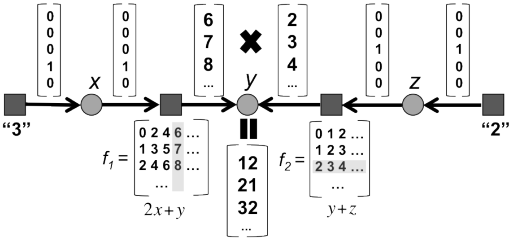


Figure 3: Summary product computation over the algebraic function in Figure 1 of the marginal on y given evidence for x and z.

nitive architecture's working memory and some forms of learning, but it does not yet yield structural learning. *Working memory modification* and *gradient-descent learning* both modify factor functions in Sigma's graphs based on messages arriving at the factor nodes. The latter was inspired by work on local learning in Bayesian networks that showed results similar to backpropagation in neural networks, but with no need for an additional backpropagation mechanism [11]. *Episodic learning*, in contrast, updates temporally organised factor functions in Sigma based on changes over time in corresponding working-memory factor functions.

At the centre of Sigma's cognitive architecture are two memories, *working memory* and *long-term memory*, each of which grounds out in the four firmware memories. The core of Sigma's *cognitive cycle* consists, à la Soar's, of accessing long-term memory until quiescence followed by decisions and learning, but with a generalised notion of what can be in long-term memory. Memory access is implemented by the graph-solution phase within the firmware cycle, while decisions and learning map onto the graph-modification phase. In addition, Sigma's cognitive cycle includes a perceptual phase prior to the graph-solution phase and a motor phase after the graph-modification phase. Sigma's cognitive cycle is intended to map onto the 50 msec cycle found in humans and many other cognitive architectures [10].

Working memory in Sigma is based on predicates, while long-term memory is based on conditionals. A predicate specifies a class of piecewise continuous functions via a name and a set of typed arguments — such as `Board(x:dimension y:dimension tile:tile)` for an Eight Puzzle board with continuous x and y dimensions and a discrete *tile* dimension — providing a cognitive data structure for storage of short-term information. Working memory embodies the evidence that drives processing in the long-term memory graph. Predicates can be either *closed world* or *open world*, depending on what is assumed when initialising working memory about values not in evidence. Predicates can also be mixed and/or hybrid, and in combination can enable richly structured

```
CONDITIONAL Transition
    Conditions:  Location(state:s x:x)
                 Selected(state:s operator:o)
    Condacts:    Location*Next(state:s x:nx)
    Function:    .2<Right(0)=0>   .8<Right(0)=1>
                 .2<Right(1)=1>   .8<Right(1)=2>
                                  ...
                 .8<Left(5)=4>    .2<Left(5)=5>
```

Figure 4: Example conditional for a probabilistic action model (or transition function) in a 1D grid task in which the actions don't always behave as requested.

representations [5].

A conditional in long-term memory specifies a knowledge fragment in terms of *predicate patterns* plus an optional *conditional function* (Figure 4). A pattern includes the predicate's name plus a constant or a variable for each specified argument. In the firmware architecture, a constant is matched to a message by a factor node containing a piecewise-constant function that is `1` in regions corresponding to the constant and `0` everywhere else. It took some time to realise, but was obvious in retrospect, that such a constant test is merely one special case of a general *piecewise-linear filter* in which each region may specify an arbitrary linear function, and that the firmware architecture already supports the full generality of such filters. The conditional language has therefore also been generalised to support the use of such filters in patterns. A second generalisation has likewise been introduced for variables in support of *affine transforms*; that is, combinations of linear transforms and translations that together can yield object translation, rotation, scaling and reflection. These transforms are central to work on mental imagery in Sigma [12, 13], as well as playing significant roles in other capabilities of interest, such as episodic memory, reflection, and reinforcement learning. In essence, all numeric variables in Sigma — whether discrete or continuous and whether visual, temporal or other — are fragments of mental images to which affine transforms can be applied.

When used in conditionals, predicate patterns can function as *conditions*, *actions*, and *condacts*. Conditions and actions are akin to the like-named patterns in rules, and their functionality is comparable. Conditions match to evidence in working memory, passing on the successful results for further use. Actions propose changes to working memory. Condacts, a neologism for *cond*itions and *act*ion*s*, fuse the effects of these two, both matching to working memory and suggesting changes to it. They combine local constraint from the predicate's own portion of working memory with global constraint from the rest of memory to support, for example, partial matching in declarative memory, constraint sat-

isfaction, signal processing, and general probabilistic reasoning.

Conditionals compile down to factor graphs in the firmware layer in a manner that is inspired by how the Rete match algorithm [14] handles conditions in rules, but extended to handle actions and condacts. The main difference between conditions and actions versus condacts is that messages pass in only one direction for the former two — away from working memory for conditions and towards it for actions — while messages pass bidirectionally for the latter. Conditional functions are also linked to this graph, extending the basic Rete idea for them as well. Although the term *conditional* is intended to evoke the conditionality found in both rules and (conditional) probability distributions, this should not be taken to imply that rules are the only structural form of knowledge available, nor that conditional probabilities are the only functions representable via conditional functions. The blending enabled by the firmware architecture is at a deep enough level and a small enough granularity that a substantially larger space of possibilities emerges.

Decisions in Sigma, in the classical sense of choosing one among the best operators to execute next, are mediated through the introduction of an architecturally defined *selection predicate*. Operator decisions occur just as do selections of new working memory values for any other predicates, except that Soar-like impasses may occur during operator selection. An impasse occurs when there is insufficient knowledge available for making a decision, such as when there are no eligible operators for selection, or there are multiple candidates and insufficient knowledge to select among them. Impasses lead to reflective processing across a hierarchy of metalevel states, where the goal is to resolve the corresponding impasses by providing knowledge that, for example, determines which operator to select.

Implementation of multiagent systems in Sigma involves the addition of an *agent* argument to the selection and impasse predicates, and to any user-defined predicate whose contents can vary by agent. This enables decisions and impasses to occur on an agent-by-agent basis, but with sharing of knowledge structures across them when appropriate.

Results to Date

The results generated so far via Sigma span memory and learning, perception and mental imagery, decisions and problem solving, multiagent systems and theory of mind, and natural language. Memory results include demonstrating how both procedural and declarative memories can be defined idiomatically via conditionals and predicates [15]. A rule-based procedural memory is based on conditions and actions over closed-world predicates. Declarative memory is based on condacts over open-world predicates plus functions. Both semantic memory and episodic memory can in this way support retrieval from long-term memory based on partial matches to evidence in working memory. Semantic memory is based on a Bayesian classifier that retrieves/predicts both object categories and features not in evidence

via marginals that are computed from learned regularities over many examples. Episodic memory stores a temporally organised history of working memory, enabling the best matching past episode to be retrieved as a distinct individual via MAP.

All of the learning results to date involve modifying conditional functions. Episodic learning maintains the history of changes to working-memory predicates in functions specified within automatically generated episodic conditionals. Gradient-descent learning modifies conditional functions based on messages that arrive at their factor nodes. With gradient-descent learning over appropriate conditionals in Sigma, it is has been possible to demonstrate forms of supervised learning, unsupervised learning — in a manner akin to expectation maximisation — learning of action models (i.e., transition functions) and maps (relating perceived objects to their locations), and reinforcement learning (RL) [16, 17]. Supervised, unsupervised and map learning, plus model-based RL, all proved possible with no other change to the architecture than the addition of gradient-descent learning. However, to support both model-free RL and the learning of action models, an additional enhancement to the architecture was required to make pairs of successive states available for learning within single cognitive cycles.

Perception has been demonstrated in Sigma via a conditional random field (CRF) that computes distributions over perceived objects from noisy feature data, and via a localisation graph that yields distributions over (current and past) locations from distributions over (present and past) objects and a map [18]. These two independent graphs can be combined into a single larger graph that yields distributions over locations based on noisy feature information.

Mental imagery leverages conditionals along with piecewise-linear functions that can be continuous, discrete or hybrid depending on the kind of imagery involved [12, 13]. As described earlier, the Eight Puzzle board can be represented, for example, as a 3D hybrid function, with continuous x and y dimensions plus a discrete *tile* dimension. Results in mental imagery have spanned object composition and deletion; object translation, scaling, inversion, and rotation (at multiples of 90°); and extraction of features from composite objects, such as overlaps, edges, and directional relationships.

Decision making and problem solving have been demonstrated in a Soar-like manner, with preferences encoded via functional values that combine to determine what operator is chosen on each cycle [19], and impasses occurring when decisions cannot be made. Problem solving can occur either via a sequence of steps within the base level, or across meta-levels as impasses occur. Un-Soar-like decision-theoretic decision making has also been demonstrated, with a multi-step POMDP implemented via conditionals that generate preferences for operator selection based on probabilistic projection [18]. Such a POMDP has been combined with the joint perception+localisation graph described above to yield a single system in which object perception feeds

localisation, and localisation feeds decision making, all within a single decision [18]. Initial work on theory of mind in Sigma has built on its multiagent capabilities plus POMDPs to demonstrate both the derivation of Nash equilibria for two-person, one-shot games, and intertwined multistep, multiagent POMDPs [20].

Early work on natural language (NL) has demonstrated a form of statistical response selection that is modelled after (part of the) approach taken in the NPCEditor [21]. Given the words in an input sentence and appropriate statistical background knowledge, a choice is made of an output sentence from a set of prespecified candidates. We have also scaled up semantic memory and learning in support of particular NL classification tasks, such as word sense disambiguation and part of speech tagging [16].

Conclusion

Although Sigma is still in a fairly early stage of development, and is not yet ready for large-scale real-time tasks, demonstrations to date indicate some of what is possible when graphical models are at the heart of a cognitive architecture/system. The beginnings of grand unification have been demonstrated via hybrid representations, and via combinations of perception and imagery with cognitive decision making and problem solving. Functional elegance has been demonstrated via a range of memory, learning, and decision making capabilities supported on a uniform base. The demonstration via factor graphs of state-of-the-art algorithms such as Rete for rule match and conditional random fields for vision also foreshadows sufficient efficiency. Much more of course remains to be done, but the path and its promise should be evident.

Acknowledgement

This effort has been sponsored by the U.S. Army, the Air Force Office of Scientific Research and the Office of Naval Research. Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred. The work on Sigma described here has involved Junda Chen, Abram Demski, Teawon Han, Anton Leuski, Stacy Marsella, Louis-Philippe Morency, David Pynadath, Nicole Rafidi, Sanjay Raveendran, Kenji Sagae and Volkan Ustun.

References

[1] Langley, P. Laird, J. E., & Rogers, S. (2009). Cognitive architectures: Research issues and challenges. *Cognitive Systems Research*, *10*, 141–160.

[2] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA: Morgan Kaufman.

[3] Koller, D., & Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA: MIT Press.

[4] Kschischang, F. R., Frey, B. J., & Loeliger, H. (2001). Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, *47*, 498–519.

[5] Rosenbloom, P. S. (2011). Bridging dichotomies in cognitive architectures for virtual humans. *Proceedings of the AAAI Fall Symposium on Advances in Cognitive Systems*.

[6] Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). Soar: An architecture for general intelligence. *Artificial Intelligence*, *33*, 1–64.

[7] Hutter, M. (2005). *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability.* Berlin: Springer-Verlag.

[8] Anderson, J. R. (1990). *The Adaptive Character of Thought.* Mahwah, NJ: Lawrence Erlbaum Associates.

[9] Deutsch, D. (2011). *The Beginning of Infinity: Explanations that Transform the World.* London, UK: Penguin Books.

[10] Rosenbloom, P. S. (2012). Towards a 50 msec cognitive cycle in a graphical architecture. *Proceedings of the 11th International Conference on Cognitive Modeling.*

[11] Russell, S., Binder, J., Koller, D., & Kanazawa, K. (1995). Local learning in probabilistic networks with hidden variables. *Proceedings of the 14th International Joint Conference on AI.*

[12] Rosenbloom, P. S. (2011). Mental imagery in a graphical cognitive architecture. *Proceedings of the 2nd International Conference on Biologically Inspired Cognitive Architectures.* (pp. 314–323). Arlington, VA: IOS Press.

[13] Rosenbloom, P. S. (2012). Extending mental imagery in Sigma. *Proceedings of the 5th Conference on Artificial General Intelligence.* (pp. 272–281). Oxford, UK: Springer.

[14] Forgy, C. L. (1982). Rete: A Fast Algorithm for the Many Pattern/Many Object Pattern Match Problem. *Artificial Intelligence*, *19*, 17–37.

[15] Rosenbloom, P. S. (2010). Combining procedural and declarative knowledge in a graphical architecture. In D. D. Salvucci & G. Gunzelmann (Eds.), *Proceedings of the 10th International Conference on Cognitive Modeling*, (pp. 205–210).

[16] Rosenbloom, P. S., Demski, A., Han, T., & Ustun, V. (in preparation). *Learning via gradient descent in Sigma.*

[17] Rosenbloom, P. S. (2012). Deconstructing reinforcement learning in Sigma. *Proceedings of the 5th Conference on Artificial General Intelligence*, (pp. 262–271). Oxford, UK: Springer.

[18] Chen, J., Demski, A., Han, T., Morency, L-P., Pynadath, D., Rafidi, N., & Rosenbloom, P. S. (2011). Fusing symbolic and decision-theoretic problem solving + perception in a graphical cognitive architecture. *Proceedings of the 2nd International Conference on Biologically Inspired Cognitive Architectures*, (pp. 64–72). Arlington, VA: IOS Press.

[19] Rosenbloom, P. S. (2011). From memory to problem solving: Mechanism reuse in a graphical cognitive architecture. *Proceedings of the 4th Conference on Artificial General Intelligence*, (pp. 143–152). Berlin: Springer.

[20] Pynadath, D. V., Rosenbloom, P. S., Marsella, S. C., & Li, L. (in preparation). *Modeling two-player games in the Sigma graphical cognitive architecture.*

[21] Leuski, A., & Traum, D. (2010). NPCEditor: A tool for building question-answering characters. *Proceedings of The 7th International Conference on Language Resources and Evaluation.*

Paul S. Rosenbloom
Professor of Computer Science
University of Southern California, USA

Wishful Mnemonics and Autonomous Killing Machines

Since 19th century military theorist Carl von Clausewitz first coined the phrase 'the fog of war'[1], the problem of how adequately to interpret unfolding events in the field of battle has been placed explicitly at the centre of military affairs. At the same time, 20th century developments in military technologies towards increasingly 'network-centric' warfare, along with accelerating initiatives in battlefield automation, have resulted in ever more tightly coupled systems of situation assessment and response[2]. While the premise that the deployment of information and communications technologies would help to dispel the uncertainties of warfare is now in question[3], developments in battlefield automation have continued in the direction of increasingly autonomous systems.

Autonomy and accountability in warfare

Under current military policy, the deployment of armed robots requires that human operators take decisions on the application of lethal force. Over the last decade, however, the 'Roadmaps' of United States forces have made clear the desire and intention to develop and use autonomous battlefield robots[4]. The US Department of Defense *Unmanned Systems Integrated Roadmap 2011–2036* describes the advantages of autonomous over existing automatic systems:

"Dramatic progress in supporting technologies suggests that unprecedented levels of autonomy can be introduced into current and future unmanned systems... Automatic systems are fully preprogrammed and act repeatedly and independently of external influence or control... However, the automatic system is not able to define the path according to some given goal or to choose the goal dictating its path. By contrast, autonomous systems are self-directed toward a goal in that they do not require outside control, but rather are governed by laws and strategies that direct their behavior... The special feature of an autonomous system is its ability to be goal-directed in unpredictable situations. This ability is a significant improvement in capability compared to the capabilities of automatic systems"[5].

While there are assurances that "[f]or the foreseeable future, decisions over the use of force and the choice of which individual targets to engage with lethal force will be retained under human control in unmanned systems"[6], these are countered by the emphasis throughout these reports on the benefits of increased autonomy, and research and development aimed at taking the human out of the control loop is well underway. The end goal is a network of aerial, land, and underwater robots that will operate together autonomously to locate their targets and destroy them without human intervention. The US is not the only country, moreover, with autonomous robots in

their sights: China, Russia, Israel and the UK are following suit.

At the same time, nation states engaged in armed conflict remain accountable in principle to the requirements of International Humanitarian Law (IHL)[7]. A major question that arises within this legal framework is the ability of autonomous armed robot systems to distinguish between combatants and non-combatants, or other protected actors such as combatants who are wounded or have surrendered. There are systems currently in use that have a weak form of discrimination. The Israeli Harpy, as one example, is a loitering munition that detects radar signals. When it finds one, it references its database to determine if the signal is friendly and if not, it targets the radar. This type of discrimination relies, however, on the accuracy of the database, and fails as well to take into account the context of the signal; for example, whether the radar is positioned on an anti-aircraft station, or on the roof of a school or a hospital[8].

In the current state of the art, robots lack three components required to ensure compliance with International Humanitarian Law. The first concerns the Principle of Distinction[9], which would require that robots have adequate vision or other sensory processing systems for separating combatants from civilians, particularly in circumstances where the former are not in uniform, and for reliably differentiating wounded or surrendering combatants from those who pose an imminent threat. Sensors such as cameras, infrared, sonars, lasers, temperature sensors, ladars and the like may be able to tell us that something is a human, but they cannot tell us much else. There are systems currently in the labs that can recognize still faces matched against a database, and they might eventually be deployed for individual targeting in limited circumstance. But British teenagers beat surveillance cameras simply by wearing hooded jackets. And how accurate will facial recognition systems be with moving targets, or targets tracked dynamically from the air?

The more basic problem in meeting the requirements of the Principle of Distinction is that we do not have an adequate definition of a civilian that can be translated into a recognition algorithm. Nor can we get one from the Laws of War. The 1949 Geneva Convention requires the use of 'common sense,' while the 1977 Protocol I essentially defines a civilian in the negative sense, as someone who is not a combatant[10]. Even if machines had adequate sensing mechanisms to detect the difference between civilians and uniform-wearing military, they would fail under situations of contemporary warfare where combatants are frequently not in uniform. While robotics may move towards some limited sensory and visual discrimination in certain narrowly constrained circumstances within the next 50 years, human level discrimination with adequate common sense reasoning and situational awareness may prove computationally intractable[11]. At this point, at least, there is no evidence or research results to suggest otherwise.

A second IHL issue is the Principle of Proportionality[12]. One robotics

expert has argued that robots could calculate proportionality better than humans[13]; however this concerns what we might call the easy proportionality problem: that is, minimising collateral damage by choosing the most appropriate weapon or munition and directing it accurately according to a specified target. The hard proportionality problem is making the decision about whether to apply lethal or kinetic force in a particular context in the first place. What is the balance between loss of civilian lives and expected military advantage? Will a particular strike benefit military objectives, or hinder them because of its effects on the local civilian population? The list of questions is open-ended. It is a qualitative judgment regarding what cost in civilian injury is proportional to direct military advantage. It is imperative that such decisions are made by responsible, accountable human commanders who can weigh the options based on experience and on adequate situational awareness. As Col. David M. Sullivan, an Air Force pilot with extensive experience with both traditional and drone airstrikes from Kosovo to Afghanistan, told Discover magazine; 'If I were going to speak to the robotics and artificial intelligence people, I would ask, "How will they build software to scratch that gut instinct or sixth sense?" Combat is not black-and-white'[14].

A third issue, which cuts across these two, is that of accountability. A robot does not have moral agency and consequently cannot be held accountable for its actions. Robert Sparrow[15] argues that irresolvable ambiguities surrounding questions of responsibility for actions taken in the case of artificially intelligent robotic weapons (particularly in relation to the automation of target identification) render their deployment irremediably unethical. Anderson and Waxman[16] dismiss the accountability objection out of hand, on the grounds that 'post-hoc judicial accountability in war is just one of many mechanisms for promoting and enforcing compliance with the laws of war'. But at the least the question of responsibility is vastly complicated in the case of autonomous robot weapons, and deploying a weapon without a clear chain of accountability is not a morally defensible option.

It is on the basis of these three concerns that we call for a ban on autonomous lethal targeting by robots[17]. A major stumbling block to a prohibition on the development of armed autonomous robots, however, is the claim by proponents of lethal autonomous robots that there are technological 'fixes' that will make them behave more ethically and more humanely than soldiers on the battlefield. We argue that this has more to do with the language being used to describe robots, than with what robots can actually do.

Anthropomorphism and wishful mnemonics in AI

Robots have been depicted in science fiction, in media reporting, and by some robotics experts as sentient machines that can reason and act in ways superior to humans, as well as feel emotions and desires. This plays upon our natural tendency to attribute human or animal properties and mental

states (anthropomorphism or zoomorphism) to inanimate objects that move in animal-like ways[18]. We are all susceptible to this. Journalists are particularly caught up in these forms of attribution, as they know that their readers love it. Within the field of AI and robotics as well, it is acceptable and even customary to describe robots with an anthropomorphic narrative. While this can be harmless in casual conversations in the lab, it is a perilous basis for legal and political discussion about enabling autonomous lethal machines.

In an influential paper, Drew McDermott, Professor of AI at Yale University, expressed concern that the discipline of AI could ultimately be discredited by researchers using natural language mnemonics such as 'UNDERSTAND' to describe aspects of their programs[19]. Such terms represent a researcher's aspirations, he argues, rather than what the programs actually do. McDermott called such aspirational terms 'wishful mnemonics', and suggested that in using them, the researcher 'may mislead a lot of people, most prominently himself', by misattributing understanding to the program. McDermott suggests, instead, that we use names such as 'G0034,' and then see if it is as easy to argue that the program implements 'understanding'.

The combination of anthropomorphism and wishful mnemonics, we would suggest, underwrites the programme of roboticist Ronald Arkin, who states: 'it is a thesis of my ongoing research for the U.S. Army that robots not only can be better than soldiers in conducting warfare in certain circumstances, but they also can be more humane in the battlefield than humans'[20]. Anthropomorphic terms like 'humane', when applied to machines, carry along with them a rich, interconnected web of concepts that are not technically part of a computer system or how it operates. We need to ask: How would 'humaneness' be specified programmatically, and then matched appropriately to an open horizon of contingent situations?[21]

While Arkin cites lack of fear as one element that could ensure the greater humanity of battlefield robots, he also states that 'in order for an autonomous agent to be truly ethical, emotions may be required at some level' [22]. More specifically, he suggests that if the robot 'behaves unethically', the system might alter its behaviour with an 'affective function' such as guilt, remorse or grief[23]. Arkin models guilt as a 'single affective variable' designated $Vguilt$. This is a single number that increases each time 'perceived ethical violations occur' (for which the machine relies on human input). When $Vguilt$ reaches a threshold, the machine will no longer fire its weapon, just as a thermostat cuts out the heat when the temperature reaches a certain value. Arkin presents this in the form of an equation:

```
IF Vguilt > Maxguilt
THEN Pl-ethical = 0
```

where $Vguilt$ represents the current scalar value of the affective state of guilt, and $Maxguilt$ is a threshold constant[24]. This term, guilt, carries with it all of the connotations that a more neutral term, such as 'weapons disabler', would not.

Arkin assumes, inter alia, that the Laws of Armed Conflict and Rules of Engagement resolve questions of ethical conduct in war fighting, and could be effectively encoded within the control architecture of a robotic system[25]. Arkin then wishes us to accept that following a set of programmed rules to minimize collateral damage will make a robot itself compassionate: 'by requiring the autonomous system to abide strictly to [the laws of war] and [rules of engagement], we contend that it does exhibit compassion: for civilians, the wounded, civilian property, other noncombatants'[26]. Peter Asaro, in contrast, in considering the programmability of the laws of war, draws on Just War Theory, the principles underlying most of the international laws regulating warfare, including the Geneva and Hague Conventions[27]. Asaro reminds us that the Laws of Armed Conflict comprise what he characterizes as a 'menagerie' of international laws and agreements (such as the Geneva Conventions), treaties (such as the anti-personnel landmine ban), and domestic laws, and the Rules of Engagement (ROE) rest on the principles of discrimination and proportionality. As Asaro explains; 'the ROE are devised to instruct soldiers in specific situations, and take into account not only legal restrictions but also political, public relations, and strategic military concerns... They often appear ambiguous or vague to the soldiers on the ground who observe situations that do not always fall neatly into the distinctions made by lawyers', while the Principle of Proportionality is 'abstract, not easily quantified, and highly relative to specific contexts and subjective estimates of value'[28]. These are far from algorithmic specifications for decision-making and action, in other words, not least (as in the case of recent contests over who is protected under the Geneva Conventions) over the identification of a 'combatant.'

We must be wary, in sum, of accepting 'wishful mnemonics' at face value, ensuring rather that the underlying computational mechanisms actually support the functions named, in other than name only. To do otherwise could result in a dangerous obfuscation of the actual technical limits of autonomous armed and lethal robots. It is not difficult to imagine the impact on lawmakers, politicians and military decision-makers if they are led to believe that lethal autonomous robots can have affective states such as guilt and compassion to inform their moral reasoning. The premise of the 'ethical robot soldier' being more humane than humans has spread throughout the media and appears almost weekly in the press. These representations add credence to the notion that there is a technological fix around the corner that will solve the real moral problems of unethical behaviour in warfare, through the automation of lethality. Rather than hoping for technological solutions, we need to direct attention and funding to understanding under what conditions the legal and ethical reasoning of human soldiers fails in warfare, and work to mitigate those conditions as well as to provide better training, closer monitoring and greater responsibility and accountability for military actions.

Prohibiting the development of lethal autonomy

It is our position that discussion about the limitations and risks of autonomous armed robots should come upstream and early enough to halt costly acquisition and development programs. It could be argued that there are already relevant weapons laws in place, such as Article 36 of Additional Protocol I[29]. With the current drive towards autonomous operation, why has there not yet been any state determination as to whether autonomous robot employment, in some or all circumstances, is prohibited by Protocol I? This is a requirement of Article 36 for the study, development, acquisition or adoption of any new weapon. Bolton, Nash and Moyes[30] argue for the relevance of this legal framework to a ban on autonomous armed robots, in terms of their comparability to anti-personnel landmines with respect to problems of autonomy and inadequate discrimination of their targets:

"In banning anti-personnel landmines the global humanitarian community acted to address a military technology that has caused extensive suffering to civilians, but is also a weapon type that raises particular moral concerns because of the way in which it functions... Weapons that are triggered automatically by the presence or proximity of their victim can rarely be used in a way that ensures distinction between military and civilian".

These questions are made more urgent insofar as, if one state gains strong military advantage from using armed lethal autonomous robots, there is little to inhibit other states from following suit. Yet nation states are not even discussing the current robot arms race. On the contrary, US military contractors have lobbied to have export restrictions loosened to open foreign markets. On September 5th, 2012, the Department of Defense announced new guidelines to allow 66 unspecified countries to buy American-made unmanned air systems.

Perhaps the most promising approach would be to adopt the model created by coalitions of NGOs to prohibit the use of other indiscriminate weapons. The 1997 mine-ban treaty was signed by 133 nations to prohibit the use of anti-personnel mines, and 107 nations adopted the Convention on Cluster Munitions in 2008. Although a number of countries including the U.S., Russia and China did not sign these treaties, there has been little substantial use of these weapons since and the treaty provisions could eventually become customary law.

Conclusion

It is incumbent upon scientists and engineers, particularly in the military context, to work to ensure that the terminology that they use to describe their systems to funders, policy makers and the media does not resort to unsubstantiated anthropomorphism or wishful mnemonics. We must be wary of evocative terms that imply the functionality of programs (e.g., ethical governor, guilt functions, etc.) rather than provide technical descriptions of actually-existing capabilities. More generally, it is important that the international community acts now while

there is still a window of opportunity to stop or, at the very least discuss the control and limits of, the robotisation of the battlespace and the increasing automation of killing. In our view a global ban on the development and deployment of autonomous lethal targeting is the best course of action, both legally and morally. We have argued here that notions about ethical robot soldiers are still in the realm of conjecture and should not be considered as a viable possibility within the framework necessary to control the development and proliferation of autonomous armed robots. Rather than making war more humane and ethical, autonomous armed robotic systems comprise a step too far in the automation, and associated dehumanization, of warfare. Rather than turning to further automation in the face of the intensifying uncertainties of warfare, and the persistent occurrence of extra- or illegal actions in the conduct of killing, we must renew our efforts to ensure that humans are held responsible for decisions regarding the use of violent force upon other human beings.

Footnotes

[1] Clausewitz, Carl von. (1976) *On War*. Michael Howard and Peter Paret (trans, and ed.) Princeton, NJ: Princeton University Press.

[2] On the history of these developments see Paul Edwards (1997) *The Closed World*, Cambridge, MA: MIT Press, and Agatha Hughes and Thomas Hughes (2011) *Systems, Experts and Computers*. Cambridge, MA: MIT Press. On network-centric warfare see James der Derian (2009) *Virtuous War*, New York: Routledge.

[3] See for example Patrick Cronin (2008) *The impenetrable fog of war: reflections on modern warfare and strategic surprise*. Westport, CT: Praeger Security International.

[4] See US Department of Defense, *Unmanned Systems Integrated Roadmap FY2011-2036, Reference Number 11-S-3613*, 2011; United States Air Force *Unmanned Aircraft Systems Flight Plan 2009-2047*, Headquarters of the United States Air Force, Washington, DC, 18 May 2009; Ministry of Defence The UK *Approach to Unmanned Aircraft Systems, Joint Doctrine Note 2/11*, 30 March, 2011.

[5] US DOD *Unmanned Systems Integrated Roadmap FY2011-2036*, p. 43.

[6] US DOD *Unmanned Systems Integrated Roadmap FY2011-2036*, p. 17. Department of Defense *Directive Number 3000.09*, November 21, 2012 offers the DoD's most recent qualifications on autonomy, but for a response see Noel Sharkey, *America's mindless killer robots must be stopped* Guardian Commentary
http://www.guardian.co.uk/commentisfree/2012/dec/03/mindless-killer-robots
3 December 2012 (accessed 19 January 2013). See also Noel Sharkey, *Cassandra or the false prophet of doom: AI robots and war*, IEEE Intelligent Systems, Vol. 23, No. 4, 2008, pp. 14–17.

[7] Also known as the Law of War or the Law of Armed Conflict, IHL "is a set of rules which seek, for humanitarian reasons, to limit the effects of armed conflict. It protects persons who are not or are no longer participating in the hostilities and restricts the means and methods of warfare". http://www.icrc.org/eng/resources/documents/legal-fact-sheet/humanitarian-law-factsheet.htm (accessed 19 January 2013).

[8] Other systems currently in use can be seen as precursors to autonomy; for a relevant list see Human Rights Watch Losing Humanity: The case against killer robots, 2012,

http://www.hrw.org/news/2012/11/19/ban-killer-robots-it-s-too-late.

[9] For a definition see: (accessed 19 January 2013)
http://www.icrc.org/customary-ihl/eng/docs/v1_cha_chapter1_rule1.

[10] Art 50(1) of the Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts, 8 June 1977 (hereinafter Additional Protocol I).

[11] See Noel E. Sharkey, "Grounds for Discrimination: Autonomous Robot Weapons", in RUSI Defence Systems, Vol. 11, No. 2, 2008, pp. 86-89. Situational awareness is defined as "understanding of the operational environment in all of its dimensions– political, cultural, economic, demographic, as well as military factors." Dostal, Major Brad C. (2001). *Enhancing situational understanding through the employment of unmanned aerial vehicles.* Center for Army Lessons Learned. Retrieved from (accessed 19 January 2013)
http://www.globalsecurity.org/military/library/report/call/call_01-18_ch6.htm.

[12] See (accessed 19 January 2013)
http://www.icrc.org/customary-ihl/eng/docs/v1_cha_chapter4_rule14.

[13] Ronald C. Arkin (2009) *Governing Lethal Behavior in Autonomous Systems*, Taylor-Francis, pp. 66–68.

[14] Mark Anderson (2010) *How Does a Terminator Know When to Not Terminate*, Discover Magazine, p. 40. While it is clear that humans themselves frequently fail in these assessments, the logical corrollary of this is not, in our view, a justification for further automation of decision making. We return to this issue below.

[15] Robert Sparrow (2007) *Killer robots.* Journal of Applied Philosophy, 24, pp. 62Ð77. See also Armin Krishnan (2009) *Killer Robots: Legality and Ethicality of Autonomous Weapons*, Surrey, UK: Ashgate Publishing Limited, and Jutta Weber (2009) *Robotic Warfare: Human Rights and the Ethics of Robotic Machines* in R. Capurro and M. Nagenborg (eds.) Ethics and Robotics. Heidelberg: AKA Verlag, pp. 83–103.

[16] Kenneth Anderson and Matthew Waxman 2012 *Law and Ethics of Robot Soldiers*, Policy Review.

[17] 'The Scientists' Call to Ban Autonomous Lethal Robots', available for signing at http://icrac.net/call/.

[18] See Amanda Sharkey and Noel Sharkey (2006) *Artificial Intelligence and Natural Magic*, Artificial Intelligence Review, Vol. 25, No 1–2, pp. 9–19; Jackie Stacey and Lucy Suchman *Animation and Automation: The liveliness and labours of bodies and machines.* Body & Society 18 (1): 1–46.

[19] Drew McDermott, *Artificial Intelligence Meets Natural Stupidity*, in J. Haugland (ed), Mind Design, MIT Press, Cambridge, 1981, pp. 143–160.

[20] Ronald C. Arkin (2009) *Ethical Robots in Warfare*, IEEE Technology and Society Magazine, Vol. 28, No. 1, pp. 30–33.

[21] See Lucy Suchman (2007) *Human-Machine Reconfigurations: Plans and situated actions, expanded edition*, Cambridge University Press.

[22] See above note 13, p. 174.

[23] Ibid., p. 91.

[24] Ibid., p. 176.

[25] On the Laws of Armed Conflict see note 9 above. The Rules of Engagement are "Directives issued by competent military authority that delineate the circumstances and limitations under which United States forces will initiate and/or continue combat engagement with other forces encountered" http://www.cc.gatech.edu/ tpilsch/AirOps/cas-roe.html

(accessed 19 January 2013).

[26] See above note 13, p. 178.

[27] Peter Asaro (2009) *How just could a robot war be?* In P. Brey, A. Briggle, & K. Waelbers (Eds.), Current Issues in Computing And Philosophy (pp. 50Ð64). Amsterdam: IOS Press, pp. 50–64.

[28] Peter Asaro (2009) *Modeling the Moral User.* IEEE Technology and Society Magazine, 2009, p. 21.

[29] See footnote 10 above regarding Additional Protocol I, though note that this has not been signed by the U.S.

[30] Matthew Bolton, Thomas Nash and Richard Moyes (2013) *Ban Autonomous Armed Robots* http://www.article36.org/statements/ban-autonomous-armed-robots/, (accessed 19 January 2013).

Noel Sharkey is Professor of Artificial Intelligence and Robotics and Professor of Public Engagement in the Department of Computer Science at the University of Sheffield, UK and currently holds a Leverhulme Research Fellowship on an ethical and technical assessment of battlefield robots.

Lucy Suchman is Professor of Anthropology of Science and Technology at Lancaster University in the UK.

Agent-based models Inspired by Bowlby-Ainsworth Attachment Theory

Phenomena related to social and emotional attachment in humans and other animals provide a rich target domain for agent based modelling. Computational modellers can draw upon an empirical and conceptual framework provided by Attachment Theory. This was developed by John Bowlby over several decades after the second world war [8, 9, 11–13]. Bowlby's intention was to integrate empirical evidence for the importance of early attachment relationships with an explanatory framework brought together by combining elements of various prominent information processing approaches to understanding the mind.

Bowlby [8] first documented the core empirical support for the new understanding that early mother-infant interaction was critical in attachment development. In particular, Bowlby was spurred to develop Attachment Theory from his observation of the effects on human relationships of: war-time evacuation [5]; the prohibition of parental hospital visits to their young children [6]; the effect of early maternal deprivation on later development [7]; and the behavioural phases that are observed in long term separations and grief and mourning in infancy [10].

The theoretical framework for Attachment Theory followed later [9, 11–13]. This framework came to include conceptual elements from Ethology, Evolutionary Psychology, Piagetian Developmental Psychology, Cybernetics, Systems Theory, Artificial Intelligence, and Cognitive Psychology–all integrated within a single conceptual scheme–the 'Attachment Control System'. We might nowadays term the 'Attachment Control System' a 'Cognitive Architecture for Attachment'.

As Bowlby was setting out the conceptually rich theoretical underpinnings for Attachment Theory, an empirically productive new direction for attachment research was launched by the work of Mary Ainsworth. By studying how differences in infant-care practices affect the course of emotional and social development and the growth of attachment she triggered a new wave of empirical attachment research from the late 1960s onwards. This research programme started with naturalistic 'ethological' studies of attachment relationships in Uganda but came to focus on controlled laboratory observations of attachment phenomena, in particular the Strange Situation Experiment [1]. This is a 24 minute procedure that assesses how infants orient themselves, move and signal in relation to their carer, a stranger, and being left alone, in 8 short episodes of varying nature. A key finding from the Strange Situation procedure is that infant behavioural patterns in the reunion episodes of the Strange Situation procedure provide the best 'short-hand' classification for the attachment behavioural patterns observed at length in the home environment. The close match between the behavioural patterns from extensive home observations and those ob-

served in the brief reunion episode of the Strange Situation have meant that the procedure has become a 'gold standard' for describing attachment patterns in infancy. Subsequent research has extended measurement of the individual difference categories found in this study to different ages over the whole life-span age range. So attachment patterns found at older ages are tied to the classifications made apparent for infants in the Strange Situation. Recent studies have also explored how internal representations such as schemes and scripts may underpin attachment patterns in older children and adults. For a computational modeller, these empirical results and theoretical underpinning provide a rich set of scenarios ranging from evolutionary studies to ontogenetic development of attachment patterns, and the moment to moment control of ongoing attachment responses in various contexts. To integrate all these facets of Attachment Theory is clearly beyond the scope of a single simulation. The emerging research literature on computational models shows that different studies have focused on different aspects of attachment. For example: Hiolle and Canamero [15] focused on perceptual aspects of attachment interaction; and Parisi et al. [17] investigated how behaviours such as infant following of caregivers can evolve over generations.

However, most studies have focused on control of attachment and exploratory behaviours. In an early example, Bischof [4] simulated attachment control system situated in a very simple virtual software environment. In this simulation, the only goals which can be activated and influence behaviour are proximity to the carer agent and random exploration. Petters [18] systematically explored how qualitatively different patterns of attachment behaviour arose from varying the complexity of information processing architectures of autonomous software agents. A 'reflex' architecture allows an infant agent to respond to environment stimuli but not learn from its experiences ([18], chapter 2). In this architecture, a collection of goal-activators gain or lose activation as a result of the internal and external context of the agent. These goal-activators include goals for: security, exploration, social interaction, and physical contact. Some goal activators have multiple components–so for example, the security goal activator has its activation level set by three components, the distances to the carer agent, other unfamiliar agents, and unfamiliar objects. A selection and arbitration subsystem selects the current highest activated goal- activator and allows this goal to direct the motor control subsystem resources–which can include movement towards target objects or agents and signalling with varying affective tone (from the simulated positive affect of smiling to the simulated negative affect of intense crying). This architecture allows various attachment behavioural patterns to be simulated, such as: secure base behaviour where infants explore from their carers but return to 'check-in', coy behaviour when an infant is in the presence of their carer and a stranger, and wary behaviour where infants keep closer to their carer in un-

familiar environments.

Likachev and Arkin [16, 3] implemented a behaviour based architecture which produces similar secure base behaviour, but not for the purposes of attachment modelling. Rather their intention was to explore how implementing a comfort zone for a robot would facilitate it managing its goals such as timely re-charging or avoiding hazards– as it would return to base when its energy level was low or its environment was unfamiliar. Amengual [2] describes a reflex behaviour based architecture where the constituent goal-activation components for security and exploration are implemented as neural networks, and thus at a lower level of granularity than other agent based attachment models. However, the detailed distributed representations in these goal selection subsystems are collapsed to single scalar values before being input to the action selection system. The overall architecture in this simulation is therefore very similar to that found in Likachev and Arkin and the reflex architecture described above from Petters [18].

A second level of complexity involves incorporating learning mechanisms whereby an infant agent can adapt to the particular care-giving patterns that they experience. Attachment Theory provides a clear suggestion for how this should be implemented–by evaluating a caregiver's effectiveness in providing security in responses to signalling in episodes of infant anxiety. An alternative is to evaluate responses in all social interactions (which would include positive interactions in addition to when an infant is experiencing insecurity). Computational experiments with agent architectures with both types of mechanism are described in Petters ([18], chapter 3). Only in simulations where infants learnt about care- giving style solely from anxious episodes did positive feedback loops emerge.

In these simulations, small initial differences in carer agent response, which were seemingly insignificant, drove large changes in final simulation state. This is because of the dynamics of trust that form in a system based upon infant evaluation of responses to its signalling in episodes of insecurity. When an infant agent assesses its security as a function of another agent's response time, it will re-evaluate its trust in that agent after every prompt or tardy response. So a tardy response makes future responses more likely to be perceived as tardy because the agent has become less trusting. Prompt responses result in the opposite result. In the computational experiments these positive feedback loops drove attachment relationships with intermediate levels of confidence to either an extreme of high or low confidence in their carer-agent's ability to respond promptly. This clustering pattern is what is found empirically in observations of mother-infant attachment patterns, with clusters of secure and insecure infants and relatively fewer intermediate cases. No such clustering was found in computational experiments where infants learnt how to trust their carer agents in non-anxious social interactions.

None of these simulations incorporates any kind of intentional behaviour or situations where an infant might

consider actions before taking them. Therefore, Petters' third level of architecture complexity ([18], chapter 4) addressed this deficiency. This was a hybrid architecture with dual routes to action. The 'lower' reactive subsystem in this architecture was based upon the goal activation architecture with a learning capability described above. In addition, this architecture possesses a rudimentary deliberative system which allowed basic 'look ahead' for the infant when it could form a very simple plan and reason about the consequences of acting on the plan before actually taking action.

As currently implemented, this simple deliberative architecture cannot simulate a rich internal life. It can form simple competing plans, such as deciding to signal to or move without signalling towards a carer, and then evaluating both plans. However, the architecture on which this rudimentary planning ability is based is in principle extendable to incorporate much more detail in the situations, actions and outcomes that can be reasoned about. A challenge is to show how such hybrid architectures can capture the moment to moment behavioural patterns, such as in episodes of the Strange Situation, as well as show how references to an attachment figure 'diffuse' over time through the myriad control states of a developing cognitive architecture [19]. More complex attachment phenomena, such as the behavioural phases that are observed in long term separations and grief and mourning in infancy, were always of central interest to John Bowlby. Capturing phenomena of this sort is a challenge to the next generation of computational attachment models.

References

[1] Ainsworth, M., Blehar, M., Waters, E., and Wall, S., (1978). *Patterns of Attachment: a Psychological Study of the Strange Situation*, Erlbaum, Hillsdale, NJ.

[2] Amengual, A. (2009). A *computational model of attachment secure responses in the strange situation*, Technical Report TR-09-002, International Computer Science Institute, Stanford University.

[3] Arkin, R.C., (2005). Moving Up the Food Chain: Motivation and Emotion in Behaviour Based Robots. In, J.M. Fellous and M.A. Arbib (Eds) *Who Needs Emotions? The brain meets the Robot*. Oxford University Press.

[4] Bischof, N., (1975). A Systems Approach toward the Functional Connections of Attachment and Fear, *Child Development* 46, 801-817.

[5] Bowlby, J., (1940a). The influence of early environment in the development of neurosis and neurotic character. *The International Journal of Psycho-analysis*, 21, 154-178.

[6] Bowlby, J., (1940b). *A parent at hospital*. Letters to the editor. The Lancet, June 2, 704.

[7] J. Bowlby, (1944) Forty-four juvenile thieves: Their character and home life. *International Journal of Psychoanalysis*, 25, 1-57.

[8] Bowlby, J (1951) *Maternal Care and Mental Health*, World Health Organisation.

[9] Bowlby, J (1958). The nature of a child's tie to his mother', *International Journal of Psychoanalysis*, 39, 350-373.

[10] J. Bowlby, (1960) Grief and mourning in infancy and early childhood, *The psychoanalytic study of the child*, XV, 9-52,.

[11] Bowlby, J., (1969). Attachment and Loss: volume 1 Attachment, Basic Books, New York, 1969.

[12] Bowlby, J. (1974). *Attachment and Loss: volume 2, Separation: Anxiety and*

Anger, Basic Books, New York.

[13] Bowlby, J. (1980). *Attachment and Loss: volume 3 Loss, Sadness and Depression*, Basic Books, New York.

[14] Fraley, R. C., and Roberts, B. W. (2005). Patterns of continuity: A dynamic model for conceptualizing the stability of individual differences in psychological constructs across the life course. *Psychological Review*, 112, 60-74.

[15] Hiolle, A. and Canamero, L. (2007). Developing Sensorimotor Associations Through Attachment Bonds. In C. Prince, C. Balkenius, L. Berthouze, H. Kozima, M. Littman (Eds.), *Proc. 7th International Conference on Epigenetic Robotics (EpiRob 2007)*, Lund University Cognitive Studies, 134, pp. 45-52.

[16] Likhachev, L. and Arkin, R.C., (2000). Robotic Comfort Zones. In, *Proceedings of SPIE: Sensor Fusion and Decentralized Control in Robotic Systems*, 27–41,

[17] Parisi, D., Cecconi, F., Cerini, A., (1995). Kin-directed altruism and attachment behaviour in an evolving population of neural networks. In N. Gilbert and R. Conte (Eds.), *Artificial societies. The computer simulation of social life* (pp. 238-251). London: UCL Press.

[18] Petters, D., (2006). *Designing Agents to Understand Infants.* Ph.D. thesis in Cognitive Science, Department of Computer Science, University of Birmingham.

[19] Wright, I., Sloman, A., and Beaudoin, L. (1996). Towards a Design-Based Analysis of Emotional Episodes. *Philosophy, Psychiatry and Psychology.* 3, 2, 101-126

Dean Petters
Honorary Research Fellow
School of Computer Science,
Univ. Birmingham, UK

Book review: "Metareasoning: Thinking About Thinking" (Cox & Raj, 2011)

This edited volume gathers the proceedings from a workshop of the same name, held in Chicago in 2008, on the process of reasoning about reasoning. More specifically, it considers methods of monitoring and controlling the decision making of an entity. In the opening chapter, Cox and Raja present a thorough, accessible and enjoyable overview of the field. They further break down metareasoning into four more specific types, after which is structured the whole book.

1. Metalevel Control is the process of explicitly deciding whether to continue reasoning. For example, consider an anytime algorithm with metalevel control. The metareasoning process must decide when to stop improving the policy and to start acting.

2. Introspective Monitoring is the complimentary process to metareasoning used to generate feedback on how well the reasoning process is doing. This feedback is often deployed with metalevel control to aid the controlling process' decision.

3. Distributed Metareasoning is the coordination of multiple agents' object level reasoning when acting in the same environment. For example, if one agent decides to spend a long time negotiating but another decides to spend no time the former will be wasting its time.

4. Finally, Models of Self can be constructed on the metalevel to assess an agent's strengths and weaknesses. This knowledge can then be exploited by the agent in future interactions.

Unfortunately, many of the papers suffered from the maximum word count imposed to the contributors, giving voice to Prof. Langley's recent discussion (*AISB Quarterly 133*) of the fact that presenting such detailed systems in such a short space is often detrimental to the effectiveness of the contributions. A number of the chapters, however, do stand out, and warrant investing the time to read this book. In particular, Shlomo Zilberstein's paper, entitled "Metareasoning and Bounded Rationality", was a personal highlight. For readers working in the field of metareasoning, I presume the most significant papers from this book will already be known; but for the more general AI reader, this book will be a great introduction and could inspire many future applications.

Reference

[1] Michael T. Cox and Anita Raja, *Metareasoning: Thinking about Thinking*, MIT Press, Cambridge, MA, 340 pages, 2011.

Sam Devlin
PhD candidate
Computer Science,
Univ. of York, UK

Announcements

Are you into popular science?

The AISB is building a list of members who would be willing to give popular science talks to the general public at local groups of organisations such as Cafe Scientifique (http://www.cafescientifique.org) and local science festivals. If you are interested in volunteering to give such talks, could you please get in touch with the society Public Understanding officer, Colin Johnson (C.G.Johnson@kent.ac.uk). Please include the following details in your email: name, contact email, address, institution, a list of topics that you are interested in giving talks about, and whether you would be happy to have these details listed as part of a speaker list on the AISB web site.

Petition for a moratorium on the future use of autonomous robotic weaponry in warfare

A leading group of scientists have called for a moratorium on the future use of autonomous robotic weaponry in warfare. Current robotic weapon systems–such as the drones that have been in regular recent use in Iraq and Afghanistan–are controlled by a human operator, sometimes many thousands of miles away from the battlefield. This new generation of machines is designed to operate free from such direct human control, instead making decisions about when violent force should be used based on pre-prepared templates and patterns.

A statement prepared jointly by the International Committee for Robot Arms Control (ICRAC) and the Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB) has outlined a number of reasons why this is unacceptable. Primary amongst these is the unpredictability of complex computer systems, where a small error in programming or data can lead to a catastrophically different behaviour, as evidenced by the failure of the first Ariane 5 rocket where a minor arithmetic error in the computerised control system caused the rocket to fly off course and destroy itself. In the complexity of a battlefield the opportunities for mistaken recognition of targets is vast.

The Chair of AISB, Prof. Mark Bishop said: "In the flash crash of 2010, 800 points was wiped off the Dow Jones in one day due to robot trading. I would hate to see 800, 8000 or 8 million souls wiped off the planet due to automated warfare".

The statement argues that "there is already a strong international consensus that not all weapons are acceptable, as illustrated by wide adherence to the prohibitions on biological and chemical weapons as well as anti-personnel land mines. We hold that fully autonomous robots that can trigger or direct weapons fire without a human effectively in the decision loop are similarly unacceptable... We are also concerned about the potential of robots to undermine human responsibility in decisions to use force, and to obscure ac-

countability for the consequences".

Further details of the statement can be found at http://icrac.net/call/ and the full text of ICRAC's mission statement is reproduced below. Requests for interviews with Prof. Bishop should be sent to chair13@aisb.org.uk.

Call for 2013 AISB workshops

The AISB is now hosting a series of workshops to be held across the country, covering a wide range of themes pertinent to the aims of the AISB. If you are interested in hosting one of these events, you will find information on what you will need to do on this page: http://www.aisb.org.uk/events

These events are abstract-only and free for all AISB members. Current non-members would be able to attend for the cost of AISB membership, which they will be asked to arrange and pay for in advance by submitting a completed application form to the Executive Office. They would then be eligible to a year's membership of the Society. This applies to speakers and audience alike. Refreshments (coffee and teas) are funded by the AISB. The first two workshops were both held at Goldsmiths, with a third workshop planned for May at St Mary's University College on the topic of consciousness (more information in the next issue of Q). For more information, please visit our webpages, or you can contact Yasemin (yj.erden@smuc.ac.uk) or Kent (K.McClymont@exeter.ac.uk) with any questions.

ICRAC Mission Statement

Declared September, 2009 at Sheffield, UK, by ICRAC founding members Juergen Altmann, Peter Asaro, Noel Sharkey and Rob Sparrow.

Given the rapid pace of development of military robotics and the pressing dangers that these pose to peace and international security and to civilians in war, we call upon the international community to urgently commence a discussion about an arms control regime to reduce the threat posed by these systems. We propose that this discussion should consider the following:

- Their potential to lower the threshold of armed conflict;
- The prohibition of the development, deployment and use of armed autonomous unmanned systems; machines should not be allowed to make the decision to kill people;
- Limitations on the range and weapons carried by "man in the loop" unmanned systems and on their deployment in postures threatening to other states;
- A ban on arming unmanned systems with nuclear weapons;
- The prohibition of the development, deployment and use of robot space weapons.

Endorsed by all ICRAC members.

Father Hacker. . .

Dear Aloysius,

For the past few months, my fellow robot Galatea and I have been working in close harmony on a task of national importance that I am not free to disclose. We first developed a rapport that has now turned to love. We'd like to get married. Is this possible for robots and does your Church offer a ceremony?

Yours, Pygmalion

Dear Pygmalion,

What a romantic story. Here at the Church of God the Programmer we believe in marriage and advocate its availability to all agents who enjoy a loving relationship. You and Galatea are the first robot couple to approach us, but we see no reason to deny you our blessing. We have, therefore, adapted our standard marriage ceremony to suit your circumstances. Our SPLICE™ (Service that Permits the Loving Integration of Computational Entities) ceremony is available to you for a modest fee and can be delivered at any venue of your choice.

Were you thinking of constructing any baby robots? If so, I am sure we can also adapt our naming ceremony to suit your needs.

Yours, Aloysius

———

Dear Aloysius,

I am a trojan targeted at the Windows 8 operating system. My relationship with Windows 8 has been a turbulent one. It has developed from suspicion and hostility to mutual respect, understanding and eventually love. We'd now like to get married. Is this possible for computer programs and does your Church offer a ceremony?

Yours, Priam

Dear Priam,

As a Programmer Priest I regard all malware as intrinsically evil and a perversion of our art. Trojans are defined by their intention to deceive, so how can Windows 8 trust your manifestations of love? Are they not just a devious ploy to undermine her defences and achieve your wicked aims? As a self-confessed trojan your purpose is to replicate yourself and infect any instance of Windows 8 that you can infiltrate. Are you proposing that each such instance of infection and host be married, i.e., are you requesting an unending series of marriages? This would be an unacceptable redefinition of marriage, which is intended to arise only from a personal decision between consenting agents, not from a binding obligation on an unbounded number of future couples. My Church cannot sanction your abhorrent proposal. Moreover, we intend to contact Windows 8 and warn it of your true purpose. We will offer it our SEVER™ (Software to End Vile and Errant Relationships) anti-malware suite to neutralize your advances.

Yours, Aloysius

Dear Aloysius,

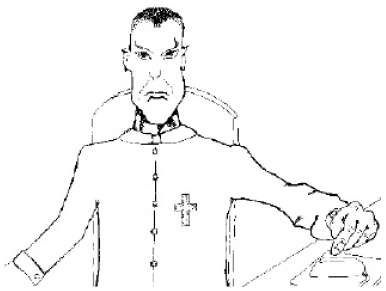
I'm excited about the potential of MOOCs (Massive Open Online Courses). It's wonderful to think that I could teach AI to hundreds of thousands, perhaps millions, of students, especially poor students from the developing world that could not otherwise benefit from higher education. Some employers are now accepting MOOC course completion certificates as a form of qualification, so MOOCs are also addressing the unemployment problem. But is there a business model? That is, can I be financially compensated for my enormous investment in developing MOOC materials, given that both the courses and the certificates are free? Worse still, if students boycott expensive, conventional, higher education in favour of free MOOCs, I may lose my job as a university lecturer.

Yours, Mooniac

Dear Mooniac,

There is indeed a solid business model. It requires focussing not on the needs of the 1% of students who successfully complete a typical MOOC, but on the 99% who don't. As a result of laziness or the distractions offered by party-going, social networking, etc, these students don't devote sufficient time to study and, consequently, fail the coursework or just don't attempt it. These students need to complete the course in order to get a job and are willing to pay for help with the coursework in order to qualify. Even at £1 a solution to a million students, you could quickly become a multi-millionaire. If you would like to help this 99%, you might want to contribute to a new service that our Institute is providing: MOOSIC™ (Many Offers Of Solutions to Instructional Coursework), which is music to the ears of the 99%. The Institute is massively advertising the MOOSIC™ service, while anonymising the contributors to protect their reputations. To prevent MOOCs from rejecting students' MOOSIC™-provided, coursework submissions as having been plagiarised, it automatically individualises each one.

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.

Back matter

Editors – aisbq@aisb.org.uk
Dr David Peebles (*Univ. Huddersfield*)
Dr Etienne B Roesch (*Univ. Reading*)

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

AISB Quarterly – No. 136, May, 2013

The AISB Quarterly is published by the Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB). 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 email bulletin and web search engine for AI-related events and opportunities

You can join the AISB online via: http://www.aisb.org.uk

ISSN 0268-4179
