COGS Seminar, University of Sussex, December 4th 2001

A Conceptual Overview of CogAff

OR

What I've been doing since I've been away

RON CHRISLEY

SCHOOL OF COMPUTER SCIENCE

THE UNIVERSITY OF BIRMINGHAM

http://www.cs.bham.ac.uk/~rlc/
rlc@cs.bham.ac.uk

http://www.cs.bham.ac.uk/research/cogaff/

WHAT IS CogAff?

CogAff = The Cognition and Affect Project

- **An investigation into the architectures required for intelligent agents.**
- **Centre of gravity is with Aaron Sloman at the School of Computer Science at the University of Birmingham...**
- **...But has involved a team of researchers – including Luc Beaudoin, Brian Logan, Matthias Scheutz and Ian Wright...**
- **...At several other universities in Europe and North America, including Nottingham, Sussex, Vienna and Notre Dame.**

My role: Leverhulme Research Fellow working on "Evolvable Architectures for Human-Like Minds"

OVERVIEW

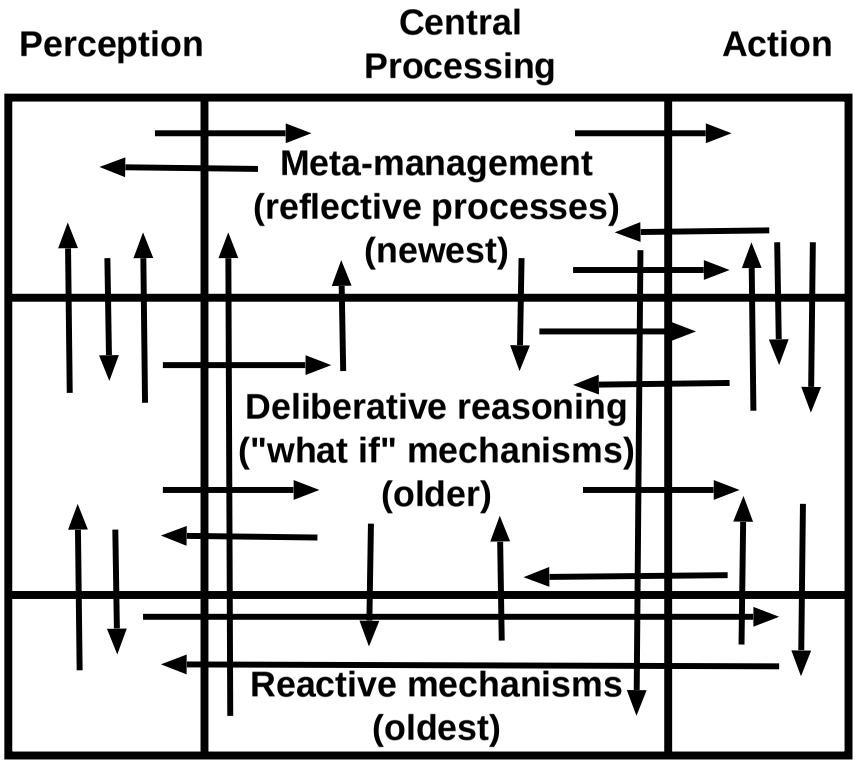
I will discuss the following aspects of CogAff:

- **Architectures and The CogAff Archiecture Schema**
- **The H-CogAff Architecture**
- **Meta-management**
- **Affect and emotion**
- **Evolvability**
- **Implemented Systems, Empirical studies, Applications**
- **Methodological and conceptual issues <– FOCUS**

ARCHITECTURES

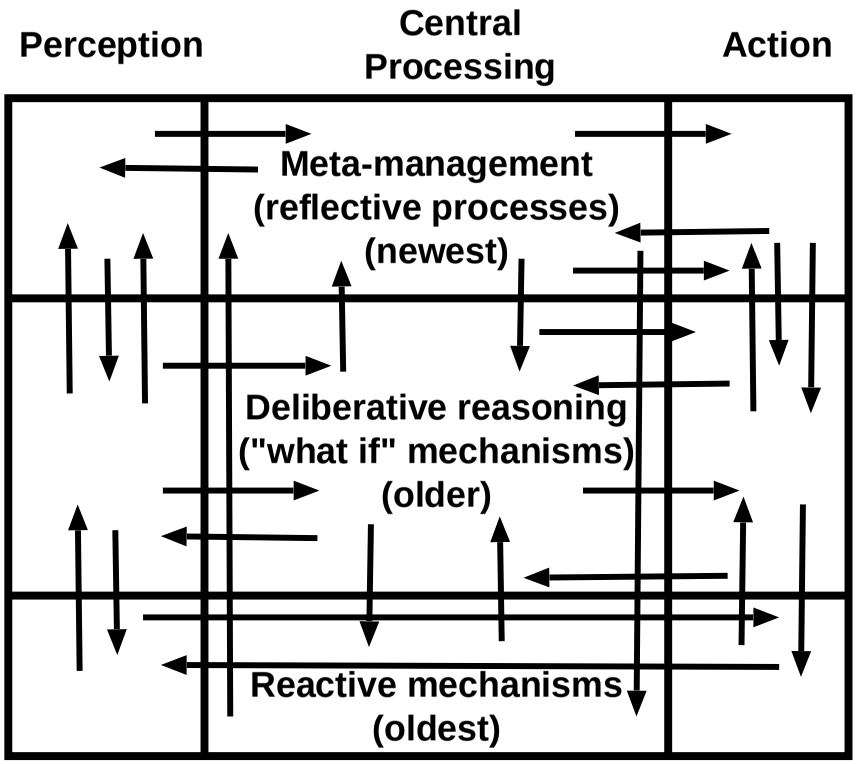
- Roughly, virtual machines: non-physical but real
- As opposed to algorithms or representations
- Functional differentiation into interacting components
- Ecology of cooperating and competing systems
- Required in order to reduce search space once one rejects behaviourism
- Requires an analysis of causation
- Investigation into both actual and possible architectures

THE CogAff ARCHITECTURE SCHEMA



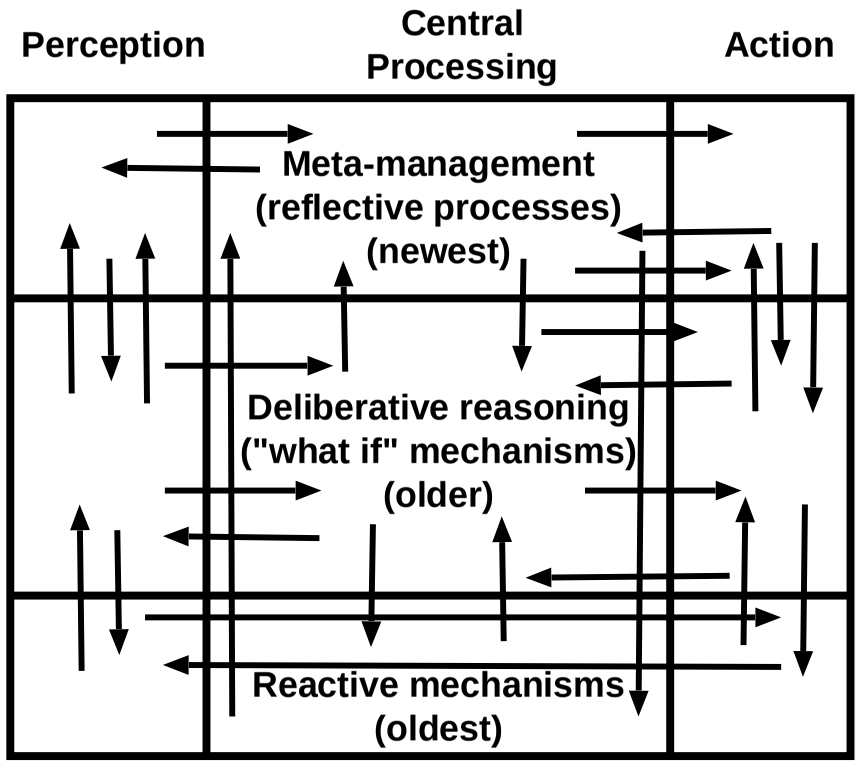
- Allows comparison of different architectures

THE CogAff ARCHITECTURE SCHEMA



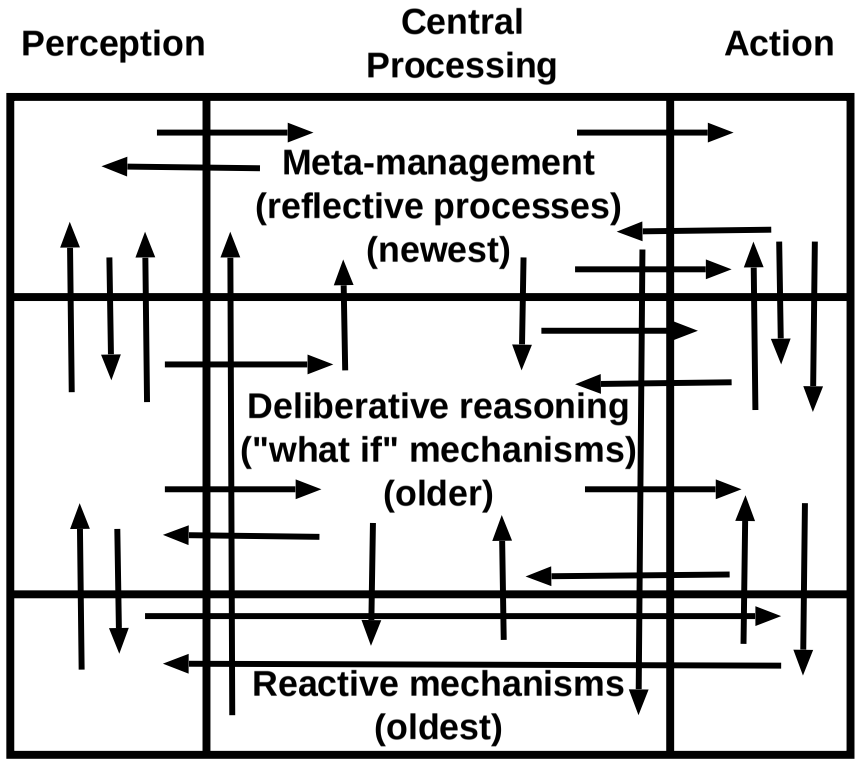
- **Vertical** divisions:
  - **Reactive**: (external) event driven
  - **Deliberative**: consideration of possibilities
  - **Reflective**: explicit control - deliberation about deliberation

THE CogAff ARCHITECTURE SCHEMA



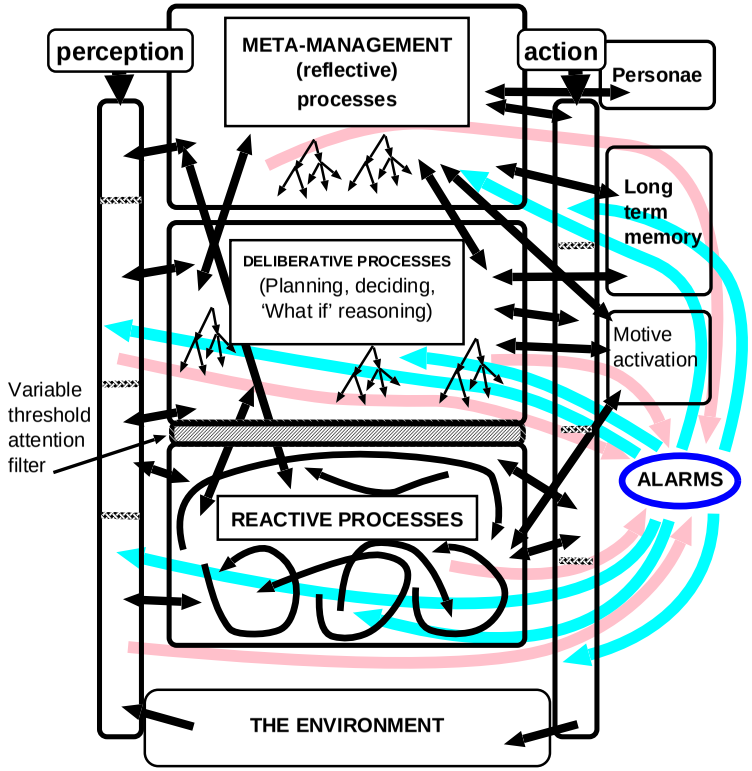
- **Horizontal** divisions:
  - perception
  - reasoning
  - action

THE CogAff ARCHITECTURE SCHEMA



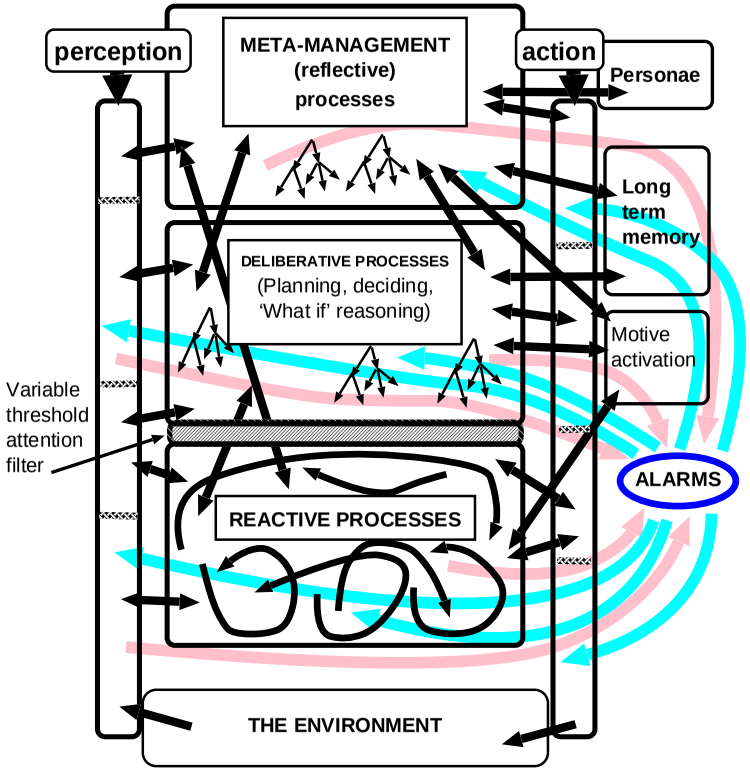
- Self-modifying, self-monitoring control system: less ambiguous than “computational system”
- Multiple interacting control loops
- Traffics in both factual and control information

THE H-CogAff ARCHITECTURE



- **Particular architecture currently under investigation**
- **Hierarchical perception, action and control**
- **Learning (reinforcement)**

THE H-CogAff ARCHITECTURE



- Multiple goals (generalised to motivators), asynchronous, vary in their insistence
- Takes resource limitation seriously; rationale for, e.g., reflection

META-MANAGEMENT

- Agent and its deliberations are in the world, so can be reasoned about
- Meta-management processes can explain qualia by explaining qualia reports (compare Dennett's heterophenomenology)
- Required for true autonomy: making self/non-self distinction.
- Required given limited resources: Nursemaid

AFFECT AND EMOTION

- Origins in alarm system: reactive layer first, then others (compare Dennett and consciousness)
- Necessary for intelligence? A side-effect of something necessary? Or just an accident?
- Three different kinds, distinguished in terms of architectural features involved:
  - Primary emotions: involves primarily the reactive layer; hedonic states lack representational content?
  - Secondary emotions: require deliberative capabilities
  - Tertiary emotions: require reflective capabilities

AFFECT AND EMOTION

- (Tertiary) Emotion as perturbance, pathological, loss of control of attention
- Intentionality: longing for one's mother requires the ability to represent one's mother
- Anti-behaviourist: not shallow
- Simulations show evolutionary advantage of affective states in some tasks: Nursemaid again

EVOLVABILITY

- An extra constraint on modelling human cognition
- But vacuous? Is anything *not* evolvable? Perhaps not, but should go with the architecture that is more evolutionarily probable.
- Tensions between design-based and evolutionary approaches?
- Trial-selectivity: not a random search

IMPLEMENTED SYSTEMS

- **Cassandra: uncertainty, distinguishes information-gathering from (other) decisions, epistemic actions, opportunism.**
- **Nursemaid: puts affect and meta-management to use in real-time task**
- **NML1**
- **AIMAE: a compromise between deliberation and reaction**
- **Minder1**
- **Abbott**
- **Simagent Toolkit**

POSSIBLE APPLICATIONS

- Intelligent Software
- Believable Agents
- Education
- Therapy
- Theories Of Software Development, Etc.
- Robots
- Immune System
- Robust Text Understanding: Human Rights Violations
- Vision

OTHER ISSUES

- Agents:
  - agent taxonomy
  - multi agent systems require an economy
- Vision:
  - non-modular vision and Gibson
  - in order to understand visual representation, need to understand rep of space and motion.
- Foundations of computation:
  - Turing machines irrelevant to computation and AI
  - implementation may matter (weak strong AI)

METHODOLOGICAL AND CONCEPTUAL ISSUES

CogAff is very methodologically self-aware (even by AI or cogsci standards)

Much of the project's contribution has been to detail how one should go about the task of designing a mind, whether it be for its own sake or for the purpose of understanding natural minds.

- **Conceptual Revisionism**
- **Interactive Empiricism**
- **Pluralism**
- **Design-based**
- **Misc. Philosophical positions**

CONCEPTUAL REVISIONISM

- **Many everyday concepts relevant to the project, such as "consciousness", "emotion", "intelligence" and "free will", are unsuitable for scientific purposes.**
- **Such concepts are ill-formed, vague and indeterminate; some are cluster concepts.**
- **A central task, then, is identifying scientifically adequate concepts relevant to designing and understanding intelligent agents.**

Open questions:

- **What is the relation between the predecessor and successor concepts?**
- **Are the latter just refinements of the former?**
- **Or is some change of subject involved? If so, can it be a principled change of subject?**

INTERACTIVE EMPIRICISM

- A key component in developing these new concepts is (usually computational) modelling.
- It has already been seen in work on robots like Kismet (Brooks and Breazeal) that interaction between the researcher and the model or artefact may be required to provide the model with the form of experience necessary for learning or development.
- But such interaction may also be helpful, even necessary, for the researcher to develop or grasp an appropriate new concept; it may even be necessary that the researcher create (= code or build) a model of the phenomena under investigation.
- If so, factors which are often thought to be of marginal interest become central to both artificial intelligence and cogsci: runtime details of the model/simulation, interface/graphical display – even eyebrows.

INTERACTIVE EMPIRICISM continued

The Psychology of Cognitive Science

- **Much of recent cognitive science has emphasised the role of action, perception and experience, as opposed to disembodied inference and reasoning, in human cognition.**
- **Since cognitive scientists are humans, cognitive science itself should exploit the experiential aspect of cognition when possible**
- **First, one should acknowledge that the goal of cogsci is an explanation for experiencing agents (us), not (primarily) a set of marks on paper in a journal.**
- **Then one can ask what is required for such explanations/understanding; it may then be seen that cognitive science has been overly preoccupied with theories.**

INTERACTIVE EMPIRICISM continued

The Psychology of Cognitive Science continued

- Theories will doubtless play a crucial role, but there may be modes of understanding which only alternative forms of explanation, such as (interaction with) models and implemented virtual machines, can provide.
- Even IF all forms of understanding can, in some sense, be written down, it still seems that writing them down is not always an adequate means of transmitting the understanding.
- The idea that the experience of creating, or interacting with, a model is crucial for understanding is especially relevant when experience itself is to be explained/modelled
- We have reason to believe that there can be no purely theoretical understanding of (all aspects of) consciousness; fortunately, other modes of scientific understanding are already at hand.

INTERACTIVE EMPIRICISM

Dynamics of system development: Put built systems to use

- De-bugging and testing
- Interaction: as discussed earlier, especially teaching
- Bootstrapping: intelligent software helps build more intelligent software (compare Cyc)
- Acquire data: allows direct modelling and emulation of behaviour
- Synthetic metaphysics: more instances mean better concepts and accounts

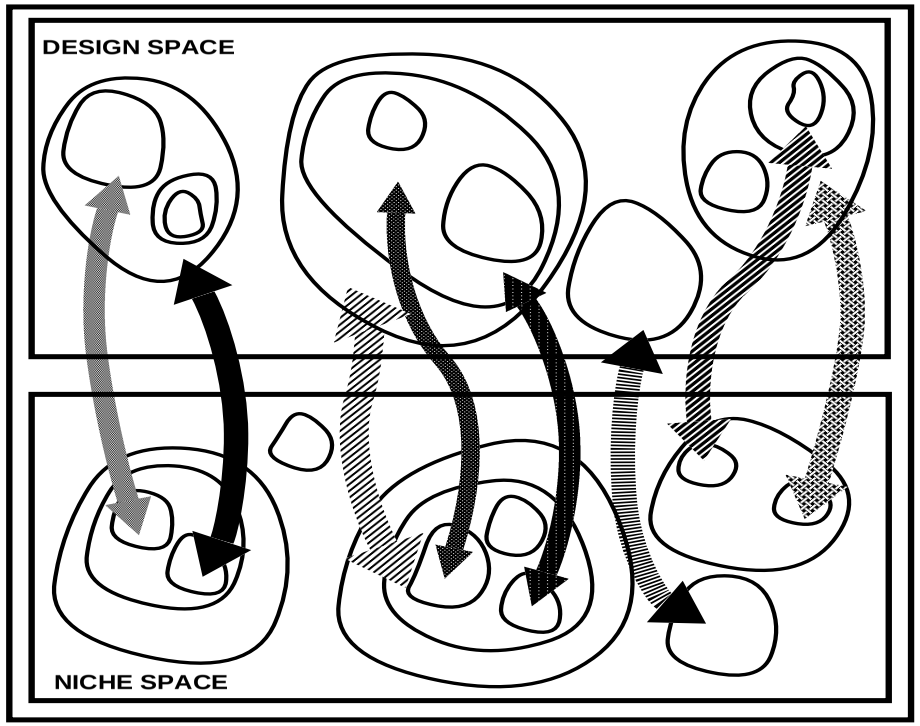
PLURALISM: Forms of Open-Mindedness

- **Of Method: The "NO IDEOLOGIES!" Ideology ("Let a thousand flowers bloom)**
- **Of Capacities: Not just a model of this or that ability, but entire working "broad but shallow" systems (compare ALife, Brooks, and Dennett's "Whole Iguana"); not just intelligence, but, e.g., emotion as well.**
- **Of Mechanism: evolved nature of cognition makes it unlikely that there will be a single representational scheme or architecture type. True, simpler accounts are preferable, however: "A theory should be as simple as possible – but no simpler" (Einstein)**
- **Of Scope: not just actual architectures, but possible architectures (cf synthetic metaphysics, above)**

A DESIGN-BASED APPROACH

- Not just the emphasis on the design of working systems, as discussed above
- Also a rejection of Dennett's intentional stance, an interpretive scheme in terms of idealised rationality
  - Much of cognition is not rational – evolution and satisficing
  - Intentional stance places few constraints on, and is hardly constrained by, underlying mechanism
- Instead, adopt the design stance: viewing a system as composed of mechanisms, each designed to perform some function, of likely use to the system as a whole
- Not a historical notion of function: different past is not enough for difference of function

DESIGN-BASED APPROACH, continued



- **Evolution best understood as trajectories (not shown) in design space vs niche space**
- **Design-based approach as opposed to semantics-based or phenomena-based approaches**
- **Conceptualisation of design: 6 types of design decision**

OTHER PHILOSOPHICAL POSITIONS

- **Anti-reductionist, but not dualist**
- **Embodiment not required, except to provide causal basis**
- **Internalist in the sense that brain in a vat cognition is possible; compatible with empirical claim that some forms of human cognition exploit the environment**
- **Non-causal theory of reference**