Multi-modal Mind Design

Jim Cunningham
Department of Computing, Imperial College of Science, Technology, and Medicine,
180, Queen's Gate, London, SW7 2BZ, UK. Email rjc@doc.ic.ac.uk

Introduction

Similar issues arise in the design of a useful agent whether it is to be realised as "intelligent" software or as an autonomous robot. Useful agents range from information gathering and trading agents to autonomous vehicles and space robots with real time physical capabilities beyond the capacity of the human mind to control. Depending on the purpose of the agent, expensive, dangerously time consuming, or otherwise unnecessary communication can be avoided if the agent has a more coherent sense of rationality, in effect a mind of its own.

Adequate encapsulations of rational behaviour and tractable engineering strategies for realisation remain an open issue for research and debate, but the considerable body of philosophical studies in the logics of practical reasoning provide some support for formal descriptions which relate actions to what an agent believes, intends and says. We do not dwell on the appropriateness of attributing mental concepts like belief and intention to a machine. It is justified by expediency as a design guide. Shoham (1992) cites McCarthy (1979), and Dennett (1987), to argue that intensional ascriptions are most useful when applied to entities whose structure is incompletely known, but that it is reasonable to require a precise theory for commonsense use of the mental property.

Here we explore a multi-modal formulation in which we seek to be reasonably faithful in style to philosophical analyses, expressively economical in its abstraction, and feasible as a technological architecture. Perhaps as we develop them, less naive mental models could become of help in diagnosing human problems. The interested reader is referred to Dennett, McCarthy, Shoham, Searle (1993) and others.

Technological Architecture

From the point of view of artefact design, the loose association of mental attitudes with an agent needs to be grounded. Beliefs, desires and intentions are not merely words for concepts disassociated from reality. When we speak of mind, it is as an abstraction of thought processes which we associate with the physical brain. If we are to build a machine with a mind, and understand what we have built, we want the abstract mental concepts which we use in design to be tied to more concrete features. Our concrete features may be representational aids like software modules, inference engines and knowledge bases; but they may equally be associative memory and emergent properties from a less representational substrata (see, e.g. Hopfield 1982). Either way these more concrete features are already quite abstract in themselves, but we have some idea how to build them as technological units. So the abstract versus concrete dichotomy is a relative scale which depends on context and evolves with technology.

In engineering it is neither necessary nor sufficient to seek fully abstract specifications, nor can we ignore the importance of structural, syntactic or intensional description in favour of the extensional mathematics of system behaviour. The reasons for this are sometimes derided as pragmatic, but it is justified by the need to reduce the complexity of comprehension, construction and verification. The traditions of language and logic in which analysis and design are expressed vary from field to field and frequently rely on graphical representation to impose structure and constrain generality. It is a measure of clarity and maturity in a field that we can graphically depict structure, and sometimes full specification detail (as may be the case functionally with a circuit diagram).

Usually a hierarchical picture is necessary. For example, real computer architecture addresses the nature of its instructions, data, and the information network. But this is predicated on a variation of a common basic idea of the computer as a collection of components: a central processor which can decode and execute instructions of finite variety, specific arithmetic and logical capabilities, a 'memory' for stored for data and instructions (which-is-which depending on the operational context), input and output devices, in effect some variation of the outline architecture of a primitive computer in figure 1. We can treat this figure a formal projection, each box depicting a component sub-system whose sum, with suitable constraints, constitute the total system. The component names constitute the elements of a simple base space onto which the components are projected and by which they are indexed. In the language of topology, each component is a fiber over its name. Constraints on the passage of data from one sub-system to another are depicted by arrows indicating information flow. Although abstract, this diagram determines the structure of a concrete system which accords with it.

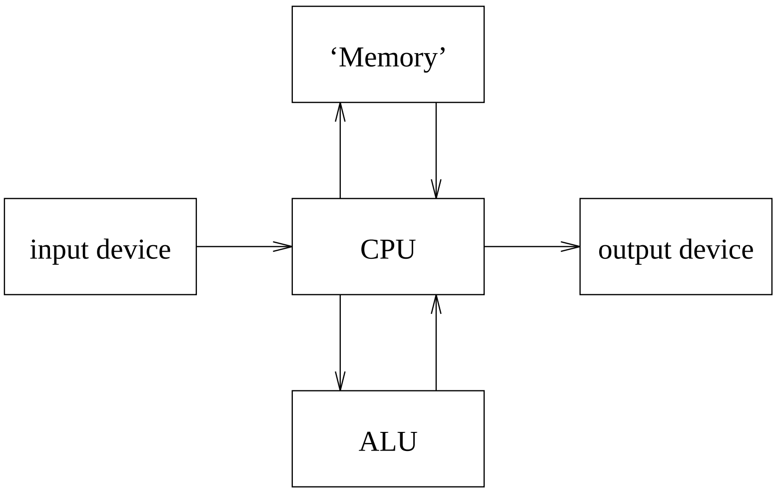
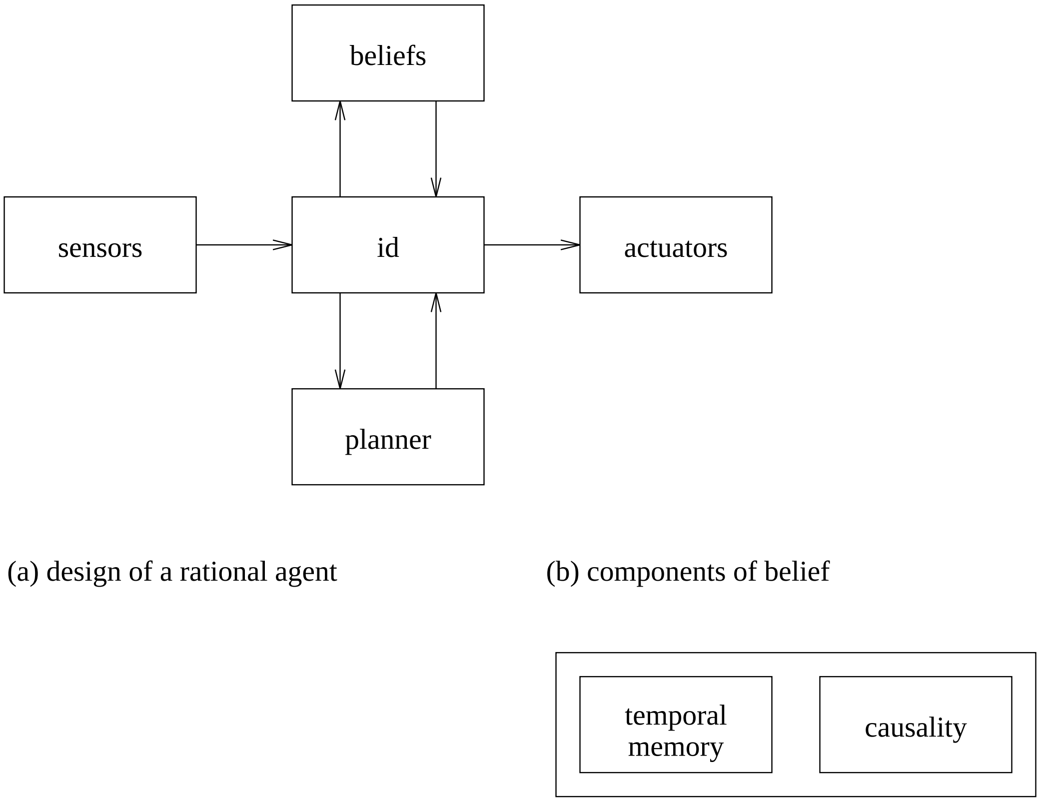


Figure 1: *Components of a Basic Computer*

What distinguishes a rational agent from a computer? There are many perspectives on both rationality and agency. If a rational agent is software, what is special about it? Perhaps it needs responsibility or self awareness. Perhaps it is merely a special sort of computer for logical inferences. It would help the design process to encapsulate distinct components in the overall design of a rational agent. For instance, we might allow figure 1 to transmogrify into figure 2(a). Just as the first figure hardly touches on serious questions of architecture the same is true of the second outline, but it gives us a framework. What information flows from a sensor or to an actuator? What is recorded from sensory information as memory of a past event and how is it recalled as a belief? Where does concept arise, is it stimulated by associating patterns? How do we relate the desires and intentions by which we might epitomise an id, the unconscious source of psychic energy, to the other components? While an attempt to describe the architecture of one artefact by modifying the architecture of the other is a recipe for producing the historical horseless carriage, we retain critical functionality and flow of information. We want rationality to be more than a metaphysical theory, even though, as inference rules, the theory may be the program which turns a suitable computer into an agent.

In omitting detail and providing only the sketch of a component we can appeal to some common intuition on its rough purpose and seek enlightenment in deeper analysis by philosophical logicians and psychologists. There is also a more technological justification for the omission of detail. There



(a) design of a rational agent

(b) components of belief

Figure 2: *Components of a Basic Agent*

appear to be many facets of learned behaviour where detailed logical analysis is unnecessary for implementation, where the symbolic expression is a superficial representation of a property which emerges in an artefact and which is validated by other means. We expect such components in the overall synthesis of a hybrid reasoning agent. An artificial neural net for the recognition of face, the stimulated recall of a represented pattern, and the synthesis of a robot arm movement are instances which spring to mind.

The multi-modal approach to rationality

For good reasons the use of symbolic logic dominates the representational approach to inference in artificial intelligence, with classical truth functional propositional and first order predicate logic the standard styles. Of the variety of non-standard logics used in philosophy for more acute representation of particular forms of inference, modal style logics have been most widely explored for styles of intensional inference, for instance with in reasoning about time and belief. Where a proposition can be qualified as a temporal truth or as a belief, deduction can both presume the features implicit in such qualification and exploit the context of the proposition. To say *in the past, something true in the past is still today true in the past*, asserts a global property implicit in the interpretation of the modality *past*. An analogous property of *belief* is positive introspection: *if we believe something, we believe we believe it.* So each modality implicitly carries a particular sort of inference and the use of different modalities allows the applied logician to make distinctions on the sort of inference. The benefit is economical notation, akin to natural language in the way detail is encapsulated in context, but at the cost of the functional transparency of extensional mathematics.

To give some flavour of the multi-modal formulation of rationality and how it relates to technological architecture consider the qualification of symbolic propositions in order to distinguish the components of figure 2(a) (as can also be done for figure 1). A traditional use of logical propositions is for formalising facts about the world. But an autonomous agent has no oracle for distinguishing fact from illusion. Things that are *felt, seen, heard or received* via any input medium are thus qualified by the medium of reception and, incidentally, immediately become past, albeit recent, information. If we use the more general atemporal modality *Senses* and separate it from the temporal *past* we can treat the former as a qualifier for propositions about input events and attribute the capacity for temporal qualification to some ability to perceive an order for these sensed events. (We ourselves can be observed to perceive as the present those events which are about half a second past).

If we seek to qualify the propositions of belief, there is much scope for fine distinction, for instance between belief in a temporal proposition that a light was previously on, and belief in causality, from we infer belief that a switch was on. As a starting point we will gloss over this philosophical minefield by qualifying all memorised and learned information the same way, by the modality for belief. This is very superficial, but it can be supplemented by other qualifications, in particular temporal ones. The idea of belief suggests an imperfect model of the environment and can also allow for imperfect memory, besides capacities for introspection and other features of the way information is associated. So we now have two qualifications $\mathcal{S}$ and $\mathcal{B}$ for a simple proposition like the light is on. We use these qualifications to distinguish which component the proposition arises from. They allow us to express, and reject, a first, tentative, multi-modal axiom schema linking the propositions of the sensory and belief components of mind: :

$$\text{Immediate belief:} \quad \mathcal{S}\alpha \;\rightarrow\; \mathcal{B}\alpha$$

Here α is a metavariable ranging over propositions. The inadequacy of this schema is immediate for it amounts to the aphorism *seeing is believing*. But the arrow does suggest a legitimate direct data flow between the architectural components. We may not want an agent to be so naive, but we defer discussion of more innate sense in deciding whether to believe what it senses because there is more obvious refinement of the coarse classification of beliefs. We can for instance separate a temporal memory of events as contingently sensed, from more atemporal beliefs, perhaps some acquired causal association between events, supposing divisions of belief such as that in figure 2(b). For a more appropriate temporal interaction with the sensors we now need to accommodate the progression of senses and memory, a progression which we may perceive as the passage of time. So more sophisticated, but still naive interaction might be:

$$\begin{array}{ll} \text{Updating belief axioms:} & \mathcal{S}\alpha \;\rightarrow\; \mathcal{N}\mathcal{B}\alpha \\ & \mathcal{S}\neg\alpha \;\wedge \mathcal{B}\alpha \rightarrow\; \mathcal{N}\mathcal{B}\mathcal{P}\alpha \end{array}$$

Here $\mathcal{N}$ is presumed to be a modality, now, in the sense of "bringing it about that" (Porn 1977), $\mathcal{P}$, "previously to this", some temporal ordering modality. So without being specific about the modalities, and while still describing naive belief, we are working towards a way of qualifying both propositions and their interactions to record a progression of senses.

How do we overcome the problem of naivety? When should an agent be sceptical of its senses, and how can scepticism be realised? Before accepting belief in sensory information a less naive agent will take into account causal expectations learnt from past events and actions. Modal qualifications for causality are well established in varieties of deontic and dynamic logic. (McCarty has described the deontic fomalisation of causality as divine obligation). But for an agent to use causality it must be either a learned or a programmed belief. When sensory data could bring about new beliefs which are incompatible with atemporal beliefs in causality, an agent may plan some investigation to choose between competing possibilities. We leave such internal learning of causality and the resolution of incompatible beliefs as further design detail.

To complete this preview of multi-modal qualifications for the propositions of a rational agent consider the remaining components of figure 2(a). Somewhere in the id reside desires, or more immediate *goals*, or *intents*. For a simple treatment we can use first the neutral qualification *wants* ($\mathcal{W}$). The treatment of goals by standard modalities have been recognised elsewhere and also criticised on technical grounds. Our justification for the modal treatment of qualified propositions is the elegance of the available formal machinery, not its use as a panacea. A rational agent should have capability to achieve certain goals by appropriate outputs, say the actions of its actuator. For formal descriptive purposes we can over-simplify by treating the execution of an action term a as a current event so that the truth of *be a-ing*, like *be running*, may be imposed directly by a parameterised actuator proposition $\mathcal{D}_a$, understood as *do* be a-ing, or more simply, *doing a*.

The main missing ingredient of rationality here is in planning what action to take, but this can be based on belief about the effect of an action. To avoid some technical difficulties we prefer a modality $\mathcal{E}_a$, denoting roughly *bring it about by action a that.* So we can now express as a multi-modal action a naive relation between goals, beliefs and actions:

$$\text{Rapaciousness: } \mathcal{W}\alpha \wedge \mathcal{B}\mathcal{E}_a\alpha \rightarrow \mathcal{D}_a$$

So with five modal style qualifiers over propositions and three multi-modal axiom schemes we have described what we may call a rapacious moron, for which sensors are believed and wants are treated like desires to be acquired continually by doing any action which it is believed to bring about the desire. As one hardly expects a rational agent to be simple we feel that this is a promising start. There is scope to refine unspecified modalities and improve their interaction. We have not provided a sensible planning component, but Cohen and Levesque (1990) summarise the philosophical analysis of intention by Bratman (1987) and others. They re-construct and justify the concept of intention by defining it as a kind of persistent goal with limited commitment for actions to bring about the desired state of affairs, thereby establishing some basic principles governing rational balance within the context of some conventional formulations of belief, action and time. Their axiomisation is naturally expressed in a multi-modal logic.

Realisation

In the introduction we asserted that this style of abstract architecture is feasible to implement. Indeed, this is critical to escape from the clutches of metaphysics. Because modal and temporal logics have attractive expressivness for formal specification in many branches of Computer Science, appropriate theorem provers, model building and various inference procedures have been investigated by many groups in recent years. Our own work, and that of Ohlbach, Gabbay and others has been linked to the Esprit Medlar project (Cunningham et al., 1991). We will therefore just assert that the time is ripe for experimental use of these logics of practical reasoning in design and hope that in a sketching some very abstract ideas there is the start of a design methodology which can be shown to be fruitful. In a companion paper (Pitt et al., 1994) we describe more immediate design objectives for cooperating software agents, where deontic (modal) logics are used for specifing protocols.

References

Bratman, M. *Intentions, Plans, and Practical Reason*, Harvard University Press 1987.

Cohen, P.R., Levesque, H.J., "Persistence, Intention, and Committment", in P.Cohen, J. Morgan, M.Pollack (eds), *Intention and Communication*, MIT 1990.

Cunningham J, Gabbay D, Ohlbach H-J, "Towards the Medlar Framework", *Proceedings Esprit '91* (CEC) pp822–841, 1991.

Dennett, D.C., *The Intentional Stance*, MIT Press, Cambridge, MA, 1987.

Hopfield, J.J., "Neural networks and Physical Systems with Emergent Collective Computational Abilities", *Proc. National Academy of Sciences*, 78, 2554-58, 1992.

Pitt, J., Anderton, M. and Cunningham J., "Autonomous Agents in Inter-Organizational Project Management", submitted to *CIKM-94 Workshop on Intelligent Information Agents.*

Porn, I., "Action Theory and Social Science. Some Formal Methods", *Synthese Library*, 120, 1977.

Searle, J.R., *Intentionality: An Essay in the Philosophy of Mind*, CUP 1983.

Shoham, Y., "Agent Oriented Programming", *Journal of Artificial Intelligence*, 1992.