From Internal Working Models to Embodied Working Models

Dean Petters [1] and **Everett Waters** [2]

Abstract.
John Bowlby introduced the 'Internal Working Model' construct into Attachment Theory to explain attachment phenomena such as an individual making plans and predictions about how attachment set-goals can be achieved. For example, when an attached individual in an anxious state considers different ways to gain proximity (and hence security) to their attachment figure before acting. According to Bowlby, an individual can possess multiple Internal Working Models, which can differ in many respects, such as representational format or how much the individual is aware of them. Existing agent based attachment simulations have implemented Internal Working Models but these are greatly simplified in comparison with Bowlby's rich and diverse conceptualisation. This paper proposes that computational attachment models can be enriched by: (i) incorporating and adapting the idea from the psychoanalytic tradition that experience can be recorded in terms of how structures of mental energy and defence are built rather than representing key attachment experiences in memory. So the particular architectural formations that develop are the residue of experience that can bias and filter future processing; and (ii) viewing cognitive architectures as more contingent and ephemeral in their structure than typically conceived. So that architectures are viewed as not just switching between configurations quickly, like the attractor states in a dynamical system, but doing so in a way that captures and brings to bear coherent biases and filters in processing laid down over long term development.

1 Introduction

John Bowlby's interest in developmental psychology started early in his career [10, 23]. After working with maladjusted children, he was training as a medical doctor when he added psychoanalysis to his studies. Melanie Klein acted as his supervisor during this psychoanalytic training. Bowlby went onto develop Attachment Theory as a theoretical vehicle to conserve some of the key insights of psychoanalysis whilst abandoning some of the aspects of the psychoanalytic framework with which he disagreed. The aspects of psychoanalytic explanation that he wished to conserve included that the cognitive and emotional life of human infants is complex and that the nature of early attachment relationships have a lasting impact, acting as prototypes of later romantic and caregiving relationships [25]. However, Bowlby disagreed with the mental energy and drive reduction models that psychoanalysis proposed to explain such internal complexity and continuity across development [25]. Other elements of psychoanalytic explanation that Bowlby wished to conserve for developmental psychology included that the phenomena of interest are bigger than the 'proxy' of behaviour. For both Psychoanalysis and Attachment Theory, overt behaviour (for example, duration of protest following separation) does not equate with strength of emotional connection [25]. In both of these frameworks, responses are guided by rich internal structures and mechanisms. However, Bowlby placed far more emphasis on the observation of current behaviour than did Melanie Klein and other psychoanalysts, who emphasised the retrospective research method of clinical reconstructions. So Bowlby viewed the immediate context an individual is in, who is around them, and their immediately prior experiences, as more important influences on their behavioural responses. The tension between the conflicting views of how research should be conducted and the relative importance of observing actual behaviour is illustrated in this passage from van der Horst:

> *"[Bowlby] was seeing an anxious, hyperactive child as a patient five days a week. The boy's mother would sit in the waiting room, and Bowlby noticed that she too seemed quite anxious and unhappy. When he told Klein he wanted to talk to the mother as well, Klein refused adamantly, dismissing the mother as a possible causal or related factor in the child's behaviour, and saying "Dr Bowlby, we are not concerned with reality, we are concerned only with the fantasy" (Kagan, 2006, p 43). When the mother was subsequently taken to a mental hospital for treatment of anxiety and depression, Klein was unaffected and untouched and only replied that is was a nuisance because now they had "to find another case" (Karen 1994, p 46). Bowlby was thoroughly annoyed - even 50 years later, in a conversation with the well-known developmental psychologist Jerome Kagan, he still become angry when relating this case (Kagan 2006) and distanced himself from Klein.". Many years later Bowlby describe his own view that:*
>
> > *"most of what goes on in the internal world is a more or less accurate reflection of what an individual has experienced recently or long ago in the external world. Of course, in addition to all that, we imagine things ... but most of the time we're concerned with ordinary events. If a child sees his mother as a very loving person, the changes are that his mother is a loving person. If he sees her as a rejecting person, she is a very rejecting person" (Bowlby, Figlio and Young, 1986, p. 43)* ([23], p 21)

So Bowlby was principally focused on how we represent our day to day experiences and use those to make predictions about future outcomes. Bowlby did want to consider how individuals imagine good and bad future outcomes. He just believed that the reality of the current moment and real past experiences around emotionally

[1] University of Northampton, UK, email: dean.petters@northampton.ac.uk
[2] SUNY, Stony Brook, USA.

valenced possible attachment outcomes like loss, separation and reunion anchor imaginative 'what if' reasoning when an individual looks ahead to possible futures. However, whilst keeping in mind attachment behaviour was a stand-in (shortcut) for the richer internal and hidden phenomena he wanted to study, Bowlby also sometimes treated behaviour as the phenomenon to be studied. Focusing on the study of behaviour itself (within an ethological framework) was methodologically valuable because it focused research on the part of the problem that is empirically accessible. It was also strategically valuable because it allowed Bowlby to break with psychoanalytic clinical reconstructions as a research methodology. So in part, Bowlby emphasised the importance of observation of actual current behaviour for pragmatic reasons, not because there was not more to consider.

2 Internal Working Models as a scientifically respectable concept

Bowlby's recognition of the tension between the need for both conservation and change in regard to psychoanalytic constructs led him to propose an alternative motivational model based on the ethological and control systems theories of the day ([7], chapter 1). At the centre of the attachment control system framework is the Internal Working Model (IWM), a more scientifically respectable construct than psychoanalytic drives and psychic energy. IWMs, in their proposed structure and operation, were compatible with the emerging information processing frameworks in Cybernetics, Artificial Intelligence and Cognitive Psychology. The conceptual ancestry of IWMs goes back indirectly to Craik's (1943) proposal of 'Working Models' [8]. In the broader sense in which Craik used the term, Working Models are not confined to attachment but apply to all representative models of the world. In his work Bowlby restricted the term Internal Working Models (IWMs) to models of self and other in attachment relationships. IWMs capture the relation-structure of attachment phenomena, not every aspect of reality but enough to make possible the evaluation of alternative actions. These include spatio-temporal causal relations among the events, actions, objects, goals and concepts represented. IWMs of attachment are what hold an infant's expectations of the levels of predicted availability and responsiveness for a given carer. These expectations are derived from the carers past performance. IWMs of self and attachment figure develop in a complementary manner. For example if the carer is responsive the self is valued. Their operation can be seen when an attached individual is in an anxious state and considers how to gain proximity to their attachment figure. IWMs allow the individual to predict the outcomes of possible actions to achieve their set-goal of proximity. They can then choose an action likely to increase security and not provoke a negative response from their attachment figure.

Although Bowlby used the IWM concept more narrowly than Craik in confining it more to just the attachment context, he also framed the IWM more broadly than Craik's working models. For Craik, the working models which a living organism might possess in their minds were comparable to the physical systems which scientists used to explain natural phenomena:

> *"By a model we thus mean any physical or chemical system which has a similar relation-structure to that of the processes it imitates. By 'relation-structure' I do not mean some obscure non-physical entity which attends the model, but the fact that it is a physical working model which works in the same way as the process it parallels, in the aspects under consideration at any moment. Thus, the model need not resemble the real object pictorially; Kelvin's tide-predictor, which consists of a number of pulleys on levers, does not resemble a tide in appearance, but it works in the same way in certain essential respects"*([8], p 51)

Craik's working models are physical systems which can act as models to explain natural phenomena because their physical operation captures key aspects of how the target system operates. When an organism holds a working model in its mind which represent its self and environment, it can configure the working model to act as memories of past events and then run this model forward in time to make predictions or imagine the results of differing actions:

> *"If the organisms carries a 'small scale model' of external reality and if its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilise the knowledge of past events in dealing with the present and future, and in every way react in a much fuller, safer, and more competent manner to the emergencies which face it"* ([8], p 61).

Bowlby himself sometimes presented IWMs in this way, for example suggesting that Internal Working Models allow *"small scale experiments"* to be conducted *"within the head"* ([7], p 80). However, it would be a mistake to go back to Craik's ideas about working models and interpret IWMs as experienced like a simulation of a visual scene, rather than the more abstract variant Craik proposed where it is the similar 'relation structure' that is key. This narrow 'visual-image' concept for what IWMs might be would be quite slow for adults for use in dynamic situations, and would not be available to sensorimotor infants. Although many attachment theorists think of IWMs as *'the'* attachment representation, a more developed treatment is to instead view IWMs as referring to a category of internal representations.

It is worth emphasizing that in the examples above which Craik gives for working models, although these systems can be argued to symbolize reality, it is by their physical properties rather than with abstract or arbitrary symbols that they represent other systems. This can be contrasted with Bowlby who proposed IWMs to include analogue and sensorimotor variants, but also to be represented as internal symbols and even interface with other high level processes of integration and control like natural language ([7], pp 81-82). This is an important distinction. Attachment Theory has moved away from more narrow conceptualisations of Working Models and proposes that there are multiple varieties with differing modes of representation. So a key issue for computational modelling is whether IWMs are a single type of representation or a diverse class of representations that afford prospective capabilities. If a diverse class if IWMs is required, what range of designs will these possess?

3 Internal Working Models versus Inner worlds: imagined futures and subjective experience

Bowlby compared IWMs to the Internal Worlds of psychoanalysis:

> *"The environmental and organismic models described here as necessary parts of a sophisticated biological control system are, of course, none other than the internal worlds of traditional psychoanalytic theory seen in a new perspective."* ([7], p 81)

However, Holmes, as psychoanalytically inclined Attachment Theorists observes:

> *"There is something in the kernel of psychoanalysis which Bowlby seems not to have fully assimilated. In comparison with Freud's and world of infantile sexuality, Attachment Theory appears almost bland, banal even. An appreciation of phantasy, and the complexity of the relationship with external reality, is somehow missing in his work. It is not loss alone that causes disturbance, but the phantasies stirred up by loss the lack of this appreciation makes Bowlby appear at times simplistic in his formulations."* ([10] pp 6-7).

As we noted earlier, Bowlby de-emphasised fantasy. So it is not a surprise that psychoanalytically inclined Attachment Theorists might criticise this omission in Bowlby's approach.

When Bowlby suggested an equivalence between IWMs and Inner Worlds he was minimising the importance of a kind of 'make believe' fantasy. However, even if we disregard the fantastical element of the psychoanalytic framework, the equivalence of IWMs to the psychoanalytic inner world is not very persuasive. This is because IWMs as conceived by Bowlby are a cognitive rather than affective solution to thinking about the future. The kind of 'look ahead' that IWMs suggest for attachment contexts is not suffused with the kind of subjective qualities we might expect to find when someone imagines emotionally charged outcomes like separations, loss, and reunions with attachment figures. It is merely predicting outcomes rather than giving subjective meaning to outcomes.

How should we reconcile this issue - Bowlby gives primacy to the real past and makes realistic predictions about the future, and the psychoanalytic approach emphasises a phantasy that presents outcomes that are explicitly separate from and beyond reality - outside of 'the reality principle' ([12], p 48).

According to Isaacs [13], fantasies derive from meaning from instincts but they are a psychic parallel to non-psychic instincts. If a modeller does not hold the reality principle as important for their modelling, then the importance is lessened for implementing mechanisms for phantasy (as opposed to merely modelling 'what if' reasoning with a subjectively emotional flavour).

Perhaps there are three elements that together can bring a resolution. First is to accept Bowlby's position on reality - modelling from real past experience to realistic futures. So IWMs should predict realistic outcomes with no need for imagining the fantastical. However, the imagined outcomes of an IWM should fulfil part of the role of fantasy in psychoanalysis and acts as *"the mental corollary, the psychic representative of instinct"* ([13], p 83). So impulses, instinctual urges, and responses tied to these inner experiences would be linked with related 'what if' imaginings. This new view of IWMs is that they imagine the results of achieving or not achieving the goals related to these desired states [13, 12]. Second is to accept that IWMs as currently conceptualised do not capture the emotionality of attachment interactions. So just predicting possible futures is not enough. To capture what attached individuals feel whilst they imagine possible outcomes simulations of IWMs should explain the emotional flavour of these experiences. The third bridge between the two positions that cognitive modellers need to focus on is interpretation. Should IWMs just afford an individual to operate as a scientist or also lead to the individual operating as a hermeneutician? IWMs within Attachment Theory are mainly focussed on assessing the results of actions in pursuit of set-goals linked to instincts (for example, attempting to gain proximity and hence security and so modelling what actions will achieve this); fantasy in psychoanalysis is also focussed on meanings linked to instincts (for example, fantasising 'I am adopted' as an interpretation of the meaning of feeling insecurely attached). Holmes contrasts the two approaches:

> *"Bowlby was always careful to distinguish between the scientific and therapeutic aspects of psychoanalysis. As a scientist he was struggling for simplicity and clarity and for general principles, while therapy inevitably concerns itself with complexity and concreteness of the individual case. Much of the disagreement between Bowlby and psychoanalysis appears to rest on a confusion of these two aspects. Bowlby's main concern was to find a firm scientific underpinning to the Object Relations approach, and Attachment Theory, with its marrying of ethology to the developmental ideas of psychoanalysis, can be seen in that light. Although couched in the language of science, psychoanalytic therapy has come increasingly to be seen as a hermeneutic discipline, more concerned with meaning than mechanism, in which patient and therapist collaboratively develop a coherent narrative about the patient's experience. Such objectification and coherence are themselves therapeutic, irrespective of the validity or otherwise of the meanings that are found. An extreme illustration of this come from the finding that schizophrenic patients with complex and coherent delusional systems are better able to function socially than those who lack such meanings, however idiosyncratic"* ([10], pp 8-9).

So perhaps the lesson from psychoanalysis is that IWMs should possess scientifically plausible mechanisms but at the same time have the potential to link to subsystems that allow the attached individuals to translate experiences to meanings. In this view, IWMs do not need to carry out look ahead reasoning outside of what is real, or could be real. Instead, perhaps what IWMs need to do is facilitate, within an appropriate architecture, the meaning making and interpretation of personal understanding that fantasies and dreams are possess and which are focused upon psychoanalysts.

4 Attachment Theory has three elements and an explanatory gap

Whilst Bowlby maintained that most forward thinking originated from actual events, it is clear that IWMs were not principally proposed to act as an explanatory mechanism for the subjective experience of this forward thinking. IWMs explain how individuals predict possible outcomes but not how they feel about those predictions as they ponder them. This delinking of cognition and subjective feeling is not just apparent with the Internal Working Models concept and it is not that Bowlby did not consider the subjective feelings associated with attachment interactions. Bowlby presented Attachment Theory as possessing three elements: a behavioural element that describes observable attachment behaviours in controlled laboratory procedures and in naturalistic contexts; a cognitive element that provides an underlying information processing explanatory framework for the behavioural element; and an experiential element that describes how people feel in attachment related contexts.

External behaviour and internal cognition can be linked by agent based simulations which use behavioural observations as a specification of requirements for their design process [17, 18]. Following this design based methodology for research, information processing architectures can be implemented and hence run in simulation to produce behavioural patterns which match, at an abstract level, observable behaviour.

The cognitive component of Attachment Theory postulates a variety of representational forms, from embodied aspects of sensorimotor social-interactions in infancy, (like sinking in when held), to Internal Working Models in later childhood and adulthood which may be symbolically or linguistically mediated. A simple view is that an individual progresses from relying on embodied attachment representations early in development to higher level representations in later development. However, even in adulthood very different levels of attachment representation may interact [20].

Computational attachment models implemented as robotic and agent based software simulations show a range of possibilities, from cybernetic inspired control systems [6] and behaviour based architectures [14, 19, 2, 9] to implementing Internal Working Models in agent based simulations as internal subsystems which operate over symbolic representations [19, 18]. The symbolic systems in [19, 18] (illustrated in figure 1 and implemented in pop11 using the sim-agent toolkit) deliberate about the likely future outcome of actions by systematically searching and appraising possible actions and outcomes ([18], pp 103-152).

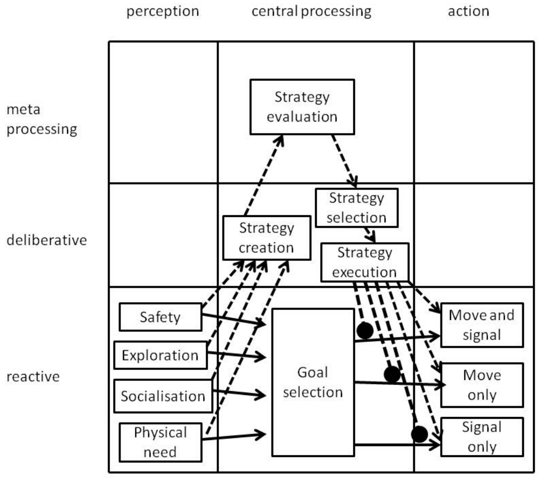

Figure 1. A hybrid attachment architecture with reactive and deliberative subsystems that has been implemented in pop11 using the sim-agent toolkit [19, 18]. The lower reactive behaviour system has a winner take all action selection system which decides which movement and signalling actions to take. In parallel, the deliberative and meta-management component in the architecture uses the same perceptual data and goals to create, evaluate, select and execute actions, which include inhibition of the actions output by the reactive behaviour based subsystem (inhibiting actions denoted with round-headed information flow). So this deliberative and meta-management component can be viewed as an Internal Working Model that 'allows small scale experiments to be conducted within the head' before external actions are taken.

The symbolic subsystem in the deliberative and meta-management components of the architecture in figure 1 operates in parallel with a behaviour based subsystem which receives inputs from goal activators for safety, exploration, socialisation and physical need. In the behaviour based subsystem the goal with the highest activation directs the next movement and signalling actions. What the symbolic IWM does in this architecture is receive the same inputs and then produce plans towards the same goals as the lower behaviour based system. When the best plans from the IWM suggest actions that conflict with the actions activated by the lower behaviour system route, the upper system can intervene and inhibit the operation of the lower system. So if the lower system activates moving and signalling the upper system may inhibit the signalling and just leave the activated movement. There are many aspects of Bowlby's conceptualisation for IWMs this simulation cannot model - it only possesses one IWM rather than the multiple varieties proposed by Bowlby; it can only interact with its environment through perception and action at a reactive level (so not perceptions or actions of language or with 'short-hand' references to mental states or other 'hidden' aspects of the environment; it can only reason from goals selected by the lower behaviour based system, rather than reason about what are good goals to hold; and of most relevance for the ongoing argument in this paper, this attachment simulation does not attempt the explain the subjective experiences which are linked with the behaviour or cognitive elements it models.

So what is missing from Bowlby's account and contemporary computational models is some clear explicit mechanistic causal framework that links all these three elements of behaviour, cognition, and subjective experience together. It is clear that links exist between behaviour, cognition, and subjective experience. However, exactly what these links are is not clear. This issue is not isolated to Attachment Theory but is an example of a more common problem which has been termed: *"the explanatory gap"* between sub-personal computational cognition and subjective mental phenomena ([22], p 6).

5 Embodied Working Models

Progress towards solving the problems of modelling subjective experience and in producing systems with more diverse ways to internally model the self, environment and attachment may be made by taking an embodied view of attachment phenomena. In this approach the body can be modelled as the context or milieu of attachment structures and mechanisms. The cognitive component of Attachment Theory could then be augmented with the incorporation of bodily sensations, physiological responses, and analogue computations that rely on the physical substrate within the attachment control system. So an embodied approach might then encompass the body as a lived experiential structure ([24], p xvi) resolving questions raised by the current use of overly cognitive and internal attachment representations in computational attachment modelling. In this view, the subjective feelings associated with attachment episodes and the prospections that occur about episodes 'yet to be' could be conceptualised as being brought forth from a history of structural coupling ([24] p 205).

6 Recording experience as structure not encoded memory

This section is concerned with several related questions: how are early experiences recorded in a way that they impact later experiences?; what changes as a result of previous experience?; and what changes during different timescales, over days, weeks or months? Ainsworth explored this issue when she discussed how what occurs over previous days may affect the sub-categories found in the Strange Situation [1]. One answer is the infants persistent state. When a baby has become anxious or worried, 'on edge', this might be measured with a cortisol assessment. This is not the representation of information in memory. A second answer is that the baby has attached some kind of meaning to the experience. This 'meaning making' reaction might be mediated through information processing mechanisms such as encoding into memory that works through a longer period of time than physiological state. There is also another longer term way of recording experience which is that there is structural change in the form of the whole information processing architecture. This latter suggestion for how experience is recorded comes from the psychoanalytic tradition and is the idea that experience was recorded in terms of building structures of mental energy and defence ([11] p 17). Little dykes to keep drives from flooding off in certain directions and ditches o direct energy in other directions. So structure is the residue of experience that filters and biases future behaviour.

One of Bowlby's contributions in the 1960s in his revision of the psychoanalytic framework was to introduce the idea of information, rather than structure, as the means of recording past experience. The notion of bolting a memory or simulation module like an IWM on only really works with these information processing formulations for the attachment control system.

In this 'Back to the Future' view proposed here (that draws upon old psychoanalytic ideas in a new context), the attachment control system grows in a way that records what it experiences during its development. The elements (building blocks) for an attachment control system, as a result of experience, become coordinated into a system to give a particular structural form. In the same way that a tree planted on a hill with a strong directed wind will grow leaning and so record the prevalent wind direction. In this new view, the word architecture is a little misleading as a label because it suggests a structure (like a building) that is designed towards a blue-print, or some design in the mind of the architect. Instead we might think about landscapes with a 3D structure that changes over time. So like a building completed over many generations. Or perhaps the analogy of experience being like water running over a landscape and eroding gulleys nicely captures psychoanalytic view of how structure can gets formed by existing within its environment. Or another analogy is a cognitive architecture like a living, flexible organ like a heart, with blood being pumped around. In this analogy, experience is like triggers for valves opening and closing and blood pushed one way or the other and some routes becoming preferred/more likely than others.

What does this mean for the imagining of possible futures? Perhaps with particular experiences an architecture takes a form that prospects just 'pop-out' in a way that does not require fully deliberative mechanisms. So the future that comes to be imagined does so because of architectural structure as much as the contents of memory. Such a component would be situated at the deliberative level but in the perceptual column of figure 1. The same may occur with mental state ascription - the architectural structure as well as the contents of memory promoting particular interpretations of what others are thinking and feeling. Mind-reading due to the right architectural formations rather than the right algorithmic operations on the contents of memory. Such a component would be situated at the meta-management level but in the perceptual column of figure 1. This 'architecture based imagination' concept has similiarities with Sloman's idea of 'architecture based motivation' [21], where the architectural organisation rather than some response to reinforcement or reward signals can decides motives.

How does this view relate to the scientific objectives of cognitive modelling? When an attachment modeller wants to implement a system that encodes experience in memory they are on familiar territory as computers naturally use memory encoding. Having an architecture that changes in the way that a landscape or living organ does is much further from traditional digital computing. So there is a bias in people who do modelling. However, we should not have theories of development get produced because they seem manageable from an AI techniques perspective. What is required is an unbiased view of the modelling outcome and then try and fit to what can currently be modelled, aware of any possible shortfall. Otherwise what will be produced is a mere simulation rather than a very deep one.

7 Viewing architectures as ephemeral structures with affordances that change moment to moment

Newell (1990) defined a cognitive architecture as:

> *"the fixed (or slowly varying) structure that forms the framework for the immediate processes of cognitive performance and learning."* (Newell 1990, p 12)

This section will present an argument that a more fine grained view of architectural temporality has some benefits when we move from a model that has a central processor and passive memory stores to a model where experience and biases are recorded in architectural structure. To unwrap this issue we will first consider a high level cognitive architecture (of a generic type similar to ACT-R [4], or EPIC [15, 16]) which processes information with productions rules. Some production rules will be involved in deliberative processes and others in meta-management processes. In dual task activities the meta-management processes may be very busy, working out when to swap from one deliberative task to another. However, in terms of resource constraints, both the meta process and the deliberative process both use attention requiring resources, they are both resource constrained. So the architecture cannot deliberate and meta-process at the same time. How does the architecture 'decide' which process to run next? At an implementational level, this means deciding which production rule to fire next when only one can fire. Older versions of cognitive architectures based upon production rules might have a goal stack to mandate the order in which goals and sub-goals should be processed [3, 5]. With a goal stack, everything that would happen would be predictable. The control structures would just go down the goal stack and deal with whatever process needs to be carried out as its turn arrive. However, goal stacks are not very biologically plausible and are absent in contemporary production system architectures that aspire for biological plausibility (such as ACT-R 6, [4]). Instead, the 'correct' production in some sequence is primed or activated rather than strongly directed. The goals are dynamically constructed. This is 'soft power' rather than 'hard power'. However, because of this processes can get swapped about. From the perspective of one single expected/desired chain of processes in a particular task, if all the particular parts were not quite right the 'wrong' thing may happen. Priming and spreading activation help form the context to trigger the next correct action. What priming and spreading activation do is change what the architecture is - in terms of not only what states are active, but in terms of what states are accessible from the current states. They change what the architecture affords, at that moment, in terms of it possibilities.

What this means is that if we are coming up with metaphors for what a cognitive architecture is, we have think about more dynamic metaphors than an architecture which is similar to a building like a castle. Cognitive architectures are more contingent on the particular context. If a cognitive architectures is compared to building with rooms, then it is a strange building where rooms that rise or lower so you cannot get in some rooms until they have appeared (perhaps by fast-acting hydraulics!). So there is an ephemeral nature to the organisation of a cognitive architecture - perhaps better termed 'ephemeral structures'. Instead of thinking about a cognitive architecture like a castle we should perhaps think if it like a special children's bouncy castle. This gets blown up and collapses and blown up again slightly differently. Perhaps the turrets have all moved around affording a completely different set of processes to possibly become active.

In the context of 'what if' reasoning, and explaining the imagined outcomes that can arise from this, it is not just your upbringing that changes your predispositions, it is what someone said to you five minutes before. Or the aroma in the room or the temperature. These things may change the particular nature of your accessible architecture. Of course, this kind of dynamic contingency will occur more in some contexts than others. Doing well practised routines like arithmetic, the architecture carries out the next goal stage is as if it were the next one on a goal stack. However, when we are considered something outside of the well-canalised routines of formal mathematical training then context will be more important and the next production to fire will be much less certain. A traditional view of a cognitive architecture suggests we should be concerned with all the structures, mechanisms and processes that may be possible on some occasion (in logical terms all the processes that may exist). These are the structures, processes and mechanisms that find their way on to traditional architecture diagrams. This traditional view also emphasises the processes that are active at any given moment in time. The concept of a cognitive architecture viewed as ephemeral structure suggests we should not only be concerned with these two dimensions, but also be more focused on the space of processes that are possible in the immediate future. This view of an ephemerally structured architecture is comparable to evolution, in that can a system can only reach the evolutionary solutions at any point in time that are accessible from where the system is that precise point. In evolution, if you have a paddle you can evolve to having fingers, but otherwise not.

How might we visualise these ephemeral possibilities within an architecture. If a spatial representation were formed of the components of an architecture through time with the activated parts in bright colour and the uninvolvable parts were in black and white you would see streams of colour running spatially through time. To this animation you could add some highlight (such as a tinge of colour) at each moment in time to the areas which were not accessible but which could have been. These highlighted structures, neither active nor inaccessible, would form a dynamically changing and contingent affordance structure. There would be a spatial 3D structure to this but it is completely fluid. The fixed architecture would be represented as the base page.

How does the ephemeral nature of a cognitive architecture relate to the idea, presented in the previous section, of an architecture as a lived experiential structure that forms over weeks, months and years? The long term formations will provide some strong predispositions for particular ephemeral affordance states to repeatedly appear. Or alternatively, to perhaps very rarely appear. We might say that the overall architecture is sculpted by previous experience but the idea of sculpting is not quite right because of short term changes that occur in what an architecture affords.

8 Conclusion

The central theme of this paper is that the IWMs that Bowlby proposed ([7], p 79-81), and those that have been implemented in computational attachment models [19, 18], emphasise how attached individuals predict future attachment outcomes but do a less good job on explaining the subjective experience and meaning making related to such processes. This paper therefore proposes that Internal Working Models might be adapted to become Embodied Working Models and record experiences not just in memory encodings, but in the very structure of a developing cognitive architecture. In this view, there is more to how experience is recorded than just another slot in an index for a memory. Structural recordings of experience can have an impact on the kind of outcomes that individuals will imagine because of biases in the processes that are afforded by architectures on a moment to moment timescale.

This structural underpinning to experience can also be engaged by a meaning making process. Another person looking at the same environment on the basis of a different set of experiences is going to make meaning differently because of different interconnections and structural organisation. Then they can interpret whatever they do by way of making meanings, within the filter of their existing architecture and memories. Tying the making of meaning to architectural structure in addition to memory encoding means the individual is engaged in meaning making rather than just meaning mapping or otherwise representing meaning. Embodying Internal Working Models allows for a structural underpinning to meaning which is read off of the environment rather than just represented.

REFERENCES

[1] M. Ainsworth, M. Blehar, E. Waters, and S. Wall, *Patterns of Attachment: a psychological study of the strange situation*, Erlbaum, Hillsdale, NJ, 1978.

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

[3] J.R. Anderson, *Rules of the Mind*, Lawrence Erlbaum Associates, Mahwah, NJ, 1993.

[4] J.R. Anderson, *How Can the Human Mind Occur in the Physical Universe?*, OUP, New York, 2009.

[5] J.R. Anderson and C. Lebiere, *The atomic components of thought*, Lawrence Erlbaum Associates, Mahwah, NJ, 1999.

[6] N. Bischof, 'A systems approach toward the functional connections of attachment and fear', *Child Development*, **48**(4), 1167–1183, (1977).

[7] J. Bowlby, *Attachment and loss: volume 1 attachment*, Basic books, New York, 1969. (Second edition 1982).

[8] K. Craik, *The Nature of Explanation*, Cambridge University Press, London, New York, 1943.

[9] A. Hiolle, L. Canamero, M. Davila-Ross, and K.A. Bard, 'Eliciting caregiving behavior in dyadic human-robot attachment-like interactions', *ACM Trans. Interact. Intell. Syst*, **2**, 3, (2012).

[10] J. Holmes, *John Bowlby and Attachment Theory*, Routledge, 1993. (revised edition).

[11] M. Horowitz, *An Introduction to Psychodynamics: A New Synthesis*, Basic books, New York, 1988.

[12] J. Hughes, *Reshaping the Psychoanalytic Domain: The Work of Melanie Klein, W.R.D. Fairbairn, and D.W. Winnicott*, University of California Press, Berkeley, 1989.

[13] S. Isaacs, 'The nature and function of phantasy', *International Journal of Psychoanalysis*, **29**, 79–97, (1948).

[14] M. Likhachev and R.C. Arkin, 'Robotic comfort zones', in *Proceedings of SPIE: Sensor Fusion and Decentralized Control in Robotic Systems*, pp. 27–41, (2000).

[15] D. E. Meyer and D. E. Kieras, 'A computational theory of executive control processes and human multiple-task performance: Part 1. Basic Mechanisms', *Psychological Review*, **104, (1)**, 3–65, (1997).

[16] D. E. Meyer and D. E. Kieras, 'A computational theory of executive control processes and human multiple-task performance: Part 2. Accounts of Psychological Refractory- Period Phenomena', *Psychological Review*, **104, (4)**, 749–791, (1997).

[17] D. Petters, 'Simulating infant-carer relationship dynamics', in *Proc AAAI Spring Symposium 2004: Architectures for Modeling Emotion - Cross-Disciplinary Foundations*, number SS-04-02 in AAAI Technical reports, pp. 114–122, Menlo Park, CA, (2004).

[18] D. Petters, *Designing Agents to Understand Infants*, Ph.D. dissertation, School of Computer Science, The University of Birmingham, 2006. (Available online at http://www.cs.bham.ac.uk/research/cogaff/).

[19] D. Petters, 'Implementing a theory of attachment: A simulation of the strange situation with autonomous agents', in *Proceedings of the Seventh International Conference on Cognitive Modelling*, 226–231, Edizioni Golardiche, Trieste, (2006).

[20] D. Petters and E. Waters, 'A.I., Attachment Theory, and Simulating Secure Base Behaviour: Dr. Bowlby meet the Reverend Bayes', in *Proceedings of the International Symposium on 'AI-Inspired Biol-*

ogy', AISB Convention 2010, 51–58, AISB Press, University of Sussex, Brighton, (2010).

[21] A. Sloman, 'Architecture-Based Motivation vs Reward-Based Motivation', *Newsletter on Philosophy and Computers*, **09**(1), 10–13, (2009).

[22] E. Thompson, *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*, MIT Press,, Cambridge, Mass, 2007.

[23] F. van der Horst, *John Bowlby - From Psychoanalysis to Ethology: Unravelling the Roots of Attachment Theory*, Wiley-Blackwell, Chichester, 2011.

[24] F. Varela, E. Thompson, and E. Rosch, *The Embodied Mind: Cognitive Science and Human Experience.*, MIT Press,, Cambridge, Mass, 1991.

[25] E. Waters, K. Kondo-Ikemura, G. Posada, and J. Richters, 'Learning to love: Mechanisms and milestones', in Minnesota Symposium on Child Psychology (Vol. 23: Self Processes and Development)*, eds. M. Gunner & Alan Sroufe*, 217–255, Psychology Press, Florence, KY, (1991).