A Computational Framework for a Situated Design Agent, Part B: Constructive Memory

John S. Gero[a*] and Gregory J. Smith[a]

[a]Key Centre of Design Computing and Cognition,
Faculty of Architecture,
University of Sydney, Australia

Memory in a computational system is usually taken to be a place filled with things called "memories". These places are indexed by either knowing its location or its content. For situated design agents, however, memory is a reflection of how the system has adapted to its environment. This paper describes the characteristics required of such a memory in a situated design agent.

1. Introduction

John: How "computational" is this? Change title of this and Part A?

In (Gero and Smith 2006) we revisited (Gero and Fujii 2000) and briefly introduced an experiential model of situation design agency. In this paper we consider in greater detail the kind of memory an agent of the kind described should have. Memory in a computational system is usually taken to be a place filled with things called "memories". These places are indexed by either knowing its physical location or its content. For situated design agents, however, memory is a reflection of how the system has adapted to its environment. Recollection should be more than looking up records – it should be past experiences guiding active ones. The inspiration behind the constructive memory model described in this paper is two phrases. The first is Clancey (1997) paraphrasing Dewey:

> "Sequences of acts are composed such that subsequent experiences categorise and hence give meaning to what was experienced before"

The second is by Bartlett (1932):

> "Remembering is not the re-excitation of innumerable fixed, lifeless and fragmentary traces. It is an imaginative reconstruction, or construction, built out of the relation of our attitude towards a whole active mass of organised past reactions or experience, and to a little outstanding detail which commonly appears in image or in language form."

Our aim is a contemporary interpretation of these experiential views of agent memory. Given the risks of using the word "imaginative" in a work on design agents, we shall hereafter refer to "reconstruction" rather than "imaginative reconstruction". We will hold on to the idea of a constructive memory inventively filling-in partly recollected experiences but we refrain from calling such processes "imaginative".

The basis of the following descriptions of constructive memory is the idea of experiences. Our ideas of experience often trace back to Dewey, although we now use more contemporary descriptions of these ideas. We use Dewey and Bartlett as inspiration, not as a statement of requirements. Dewey described the quality of an experience as having two aspects:

> "There is an immediate agreeableness or disagreeableness, and there is

*This research was supported by a grant from the Australian Research Council.

> its influence upon later experiences" (Dewey 1938)

Dewey called the first aspect continuity.

> "every experience enacted and undergone modifies the one who acts and undergoes" (Dewey 1938)
>
> "The principle of continuity of experience means that every experience both takes up something from those which have gone before and modifies in some way the quality of those which come after" (Dewey 1938)
>
> "The basic characteristic of habit is that every experience enacted and undergone modifies the one who acts and undergoes, while this modification affects, whether we wish it or not, the quality of subsequent experiences " (Dewey 1938)

Dewey called the second aspect interaction.

> "Experience does not simply go on inside a person. It does go on there ... but this is not the whole of the story. Every genuine experience has an active side which changes in some degree the objective conditions under which experiences are had" (Dewey 1938)

So an experience is an interplay of continuity and interaction.

> "Taken together, or in an interplay of their interaction, they form what we call a situation" (Dewey 1938)
>
> "An experience is always what it is because of a transaction taking place between an individual and what, at the time, constitutes his environment" (Dewey 1938)

What do the Dewey (Clancey) and Bartlett inspirations mean in contemporary, computational terms? The key phrases are

- "experiences" – Section 2
- "sequences of acts" – Section 3
- "subsequent experiences ... experienced before" and "composed" – Section 4
- "reconstruction" and "categorise and hence give meaning" – Section 5

2. Experiences

If there are experiences there must be agents. In this work we take humans interacting with computational systems and artificial agents to also be agents. We shall denote agents as $\alpha_1, \alpha_2, \ldots$

An entity cannot be an agent unless it is embodied in some environment, be it of our world or of a virtual world. We shall denote the environment as ξ. As agents are embodied,

- $\alpha_1, \alpha_2, \ldots$ are a part of ξ
- $\alpha_1, \alpha_2, \ldots$ are distinct from each other and from ξ

We need not restrict the environment to only contain agents. So we say that an environment is composed of entities, some of which are agent-entities (agents) and some of which are not. We shall say that the non-agent entities of the environment are thing-entities (things) that we denote $\gamma_1, \gamma_2, \ldots$ For things,

- $\gamma_1, \gamma_2, \ldots$ are a part of ξ
- $\gamma_1, \gamma_2, \ldots$ are distinct from each other, from ξ, and from $\alpha_1, \alpha_2, \ldots$

As Dewey (1917) noted, an experience is not of a disembodied agent[2]. It is to do with interaction of the agent with an environment. An experience is also not something static; it is dynamic and is of certain kinds of entities that are coupled to their environment. This "entities ... coupled to their environment" is like Dewey's (1938) experience changing "the objective conditions under which experiences are had".

This is illustrated in Figure 1, and it shows two kinds of coupling. The "body" of an agent $\mathfrak{a}_i$ is an agent-thing, say α_i, where α_i is a part of ξ. The "nervous system" of an agent $\mathfrak{a}_i$ are

[2]Dewey would not have used the word "agent" though.

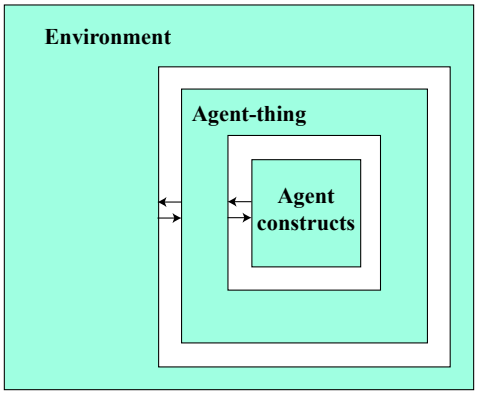

Figure 1. Agent coupled to the environment. The figure is derived from (Beer 2003).

construct-entities (constructs) $\{\beta_i^1, \beta_i^2, \ldots\}$. The scare quotes are because the words "body" and "nervous system" are those used by Beer but which we avoid. Construct-entities are parts of agent-things, so each β_i^j is a part of α_i. The agent $\mathfrak{a}_i$ is an agent-entity that is the composition of an agent-thing α_i and constructs $\{\beta_i^j\}$. That part of the environment that is not the agent is $\xi - \mathfrak{a}_i = \xi - (\alpha_i \cup \{\beta_i^j\})$.

A coupling between ξ and an agent-thing α_i is an e-experience (an exogenously generated experience) of $\mathfrak{a}_i$. An example is robot navigation experiences involving sonar sensual experiences and motion effectual experiences. A coupling between the agent-thing α_i and agent constructs $\{\beta_i^j\}$ is an a-experience (an autogenously generated experience) of $\mathfrak{a}_i$. An example is a human moving their arm, involving sensual experiences of proprioception and motor effectual experiences. Further, e-experiences and a-experiences may perturb each other directly or indirectly. An e-experience perturbs an a-experience if the agent interprets that e-experience. An e-experience fails to perturb an a-experience if the agent ignores it. An a-experience perturbs an e-experience when the agent acts on its environment – when that agent perturbs other agents or things.

3. Sequences of acts

We shall denote an experience of agent $\mathfrak{a}_i$ as $\mathfrak{e}_i^k$. If an e-experience is able to perturb an a-experience and vice versa, an agent must be able to have multiple concurrent experiences $\{\mathfrak{e}_i^1, \mathfrak{e}_i^2, \ldots\}$. An e-experience involves entities perturbing each other, where one of the entities is an α_i and the other is either another agent α_j or a thing γ_m. An a-experience also involves entities perturbing each other, where one of the entities is an α_i but where the others are constructs $\{\beta_i^j\}$. In order for one to perturb another there must be either

- Some point at which they synchronize, such as one computational process synchronously passing a message to another, or
- Some intermediary on which each synchronizes, such as one computational process asynchronously passing a message to another.

For an experience $\mathfrak{e}^x$ to perturb another experience $\mathfrak{e}^y$ there must be some part of $\mathfrak{e}^x$ that is also part of $\mathfrak{e}^y$. That common part may be so small (or temporally atomic) that we would call it an event but it must be part of each or they could not synchronize. Having a common part does not imply that $\mathfrak{e}^x$ and $\mathfrak{e}^y$ are the same. A computational process can send another a message and the actual communication event is common, but each process will have been behaving differently before and after the event. For one experience to perturb another requires two things: that experiences can have parts that are themselves experiences, and that some experiences can be parts of multiple experiences. We can describe this using mereological relations on processes (Seibt 2004, Smith 1996). These mereological relations are useful as a means to describing properties of experiences without the descriptions necessarily being reductionist.

An experience can be a part of another experience. An experience that is a part of, but not

identical to, another experience is a proper part. An experience with no proper parts is atomic. If the experience is temporally atomic but spatially not, we call it an event (Something spatially atomic but temporally not is an entity). Two experiences with one or more common parts are said to overlap. Experiences that perturb each other must overlap. Experiences are disjoint if they never overlap. An experience $\mathfrak{e}^x$ is emergent from experiences $\{\mathfrak{e}^y \mid y \neq x\}$ if[3]:

- $\mathfrak{e}^x$ is a part of the sum of $\{\mathfrak{e}^y \mid y \neq x\}$
- No part of $\mathfrak{e}^x$, including itself, is a part of any $\mathfrak{e}^y$ (for $y \neq x$)

We write $\mathfrak{e}^x \sqsubseteq \mathfrak{e}^y$ to mean "experience $\mathfrak{e}^x$ is a part of experience $\mathfrak{e}^y$". We write $\mathfrak{e}^x \sqsubset \mathfrak{e}^y$ to mean "experience $\mathfrak{e}^x$ is a proper part of experience $\mathfrak{e}^y$". We write $\mathfrak{e}^x \bigcirc \mathfrak{e}^y$ to mean "experience $\mathfrak{e}^x$ overlaps experience $\mathfrak{e}^y$". So

$$\mathfrak{e}^x \sqsubset \mathfrak{e}^y \mathrel{\widehat{=}} (\mathfrak{e}^x \sqsubseteq \mathfrak{e}^y \wedge \mathfrak{e}^x \neq \mathfrak{e}^y) \quad (1)$$

$$\mathfrak{e}^x \bigcirc \mathfrak{e}^y \mathrel{\widehat{=}} (\exists \mathfrak{e}^z \bullet \mathfrak{e}^z \sqsubseteq \mathfrak{e}^x \wedge \mathfrak{e}^z \sqsubseteq \mathfrak{e}^y) \quad (2)$$

$$\mathit{perturbs}(\mathfrak{e}^x, \mathfrak{e}^y) \Rightarrow \mathfrak{e}^x \bigcirc \mathfrak{e}^y \quad (3)$$

Unless an experience is infinite, something must cause it to begin. So the thing that delineates one experience from the next is a change to a different action. An experience does not begin because the environment has decided to start sending an agent data, or to change to sending it data that is somehow different. Such distinctions are up to the agent. An experience starts when the agent activates an action that is qualitatively different from active experiences. The action may be that a person changes their visual focus of attention: agent constructs β_i^j sufficiently perturb the agent-thing α_i to change what from its visual field it is looking at, triggering a new a-experience. The action may be that the robot moves in the world: agent constructs β_i^j sufficiently perturb the agent-thing α_i to trigger a new a-experience, and the new a-experience perturbs the environment by shifting the robots location in it, triggering a new e-experience.

[3]This is from (Seibt 2004), where it is defined more precisely.

So there are experiences $\{\mathfrak{e}_i^k\}$, and experiences are triggered by agent actions. Both a-experiences and e-experiences are triggered by an action. Hence if there is an experience $\mathfrak{e}_i^k$, there must have been another experience $\mathfrak{e}^l$ that overlaps $\mathfrak{e}_i^k$ at its start. Adopting the relations of Allen on the temporal extents of experiences (see Section A),

$$\mathfrak{e}_i^k \in \mathsf{E} \Rightarrow (\exists \mathfrak{e}^l \in \mathsf{E} \bullet k \neq l \wedge \mathfrak{e}_i^k \bigcirc \mathfrak{e}^l \wedge \mathit{starts}(\mathfrak{e}^l, \mathfrak{e}_i^k)) \quad (4)$$

Usually $\mathfrak{e}^l$ will be of the same agent as $\mathfrak{e}_i^k$, or $\mathfrak{e}^l \equiv \mathfrak{e}_i^l$. The cases where it is not is are those e-experiences when one agent creates another. What finishes an experience? Another action triggering another experience, although it is acceptable for an experience to autonomously extinguish. Sense-data from another agent or from a thing, and effect-data to another agent or thing, are perturbations of an experience. These sense-data and effect-data do not start new experiences. Recognising that a perturbation is qualitatively different may cause a new a-experience but this recognition is an autogenous act of this agent. The new experience isn't caused by the external agent or thing, it is caused by the agent of the experience.

What counts as qualitatively different is up to the agent. Adjusting a motor controller to remain on course may involve actions that vary quantitatively but not qualitatively, and so would be a continuation of the same experience. Deciding to top moving and start vacuuming would be qualitatively different and so would be another experience.

The experience isn't the agent taking an action followed by it sitting there waiting for something to send it data. Rather, sensations are parts of experiences and so require actions, even if those actions are internal to the agent.

> "Upon analysis, we find that we begin not with a sensory stimulus, but with a sensori-motor coördination, the optical-ocular ... the real beginning is with the act of seeing; it is looking, and not a sensation of light.

> The sensory quale gives the value of the act, just as movement furnishes its mechanism and control, but both sensation and movement lie inside, not outside the act" (Dewey 1896).
>
> "It feels distinctly uncomfortable to conceptualize people (persons) as things (substances) ... However, there is no problem with experiential access to the processes and patterns of process that characterize us personally" (Rescher 2002).
>
> "Instead of very small things (atoms) combining to produce standard processes (windstorms and such), modern physics envisions very small processes (quantum phenomena) combining in their modus operandi to produce standard things (ordinary macro-objects)" (Rescher 2002).

An agent is more than a feed-forward pipeline of sensation, conception and effection processes. Rather, an experience is an ongoing, interactive dynamic. Failing

> "to see unity of activity, no matter how much it may prate of unity, it still leaves us with sensation or peripheral stimulus; idea, or central process (the equivalent of attention); and motor response, or act, as three disconnected existences, having to be somehow adjusted to each other, whether through the intervention of an extra-experimental soul, or by mechanical push and pull" (Dewey 1896).
>
> "If one is reading a book, if one is hunting, if one is watching in a dark place on a lonely night, if one is performing a chemical experiment, in each case, the noise has a very different psychical value; it is a different experience. In any case, what proceeds the 'stimulus' is a whole act, a sensori-motor coördination" (Dewey 1896).

Some of the work done on "direct perception" in the past is of interest here. Shaw and Todd (1980) are an example. They describe interactions of an agent as a dual pair of agent-environment equations:

- The next response by the agent on environment is a function of the current stimulus on the agent by the environment and the history of the "state of affairs" of the agent concerning its "transactions" with the environment up until the current time
- The next stimulus on the agent by environment is a function of the current response by the agent on the environment and the history of the "state of affairs" of the agent concerning its "transactions" with the environment up until the current time

We are not describing "direct perception" but the resemblance of "state of affairs ..." to experiences should be clear. The important thing about these equations is that the agent is not described as a state machine. Beer also characterises agents with a dual pair of agent-environment equations:

> "an agent and its environment should be understood as two coupled dynamical systems whose mutual interaction is jointly responsible for the agent's behavior" (Beer 1997).

One difference of Shaw and Todd with Beer is that the latter's equations update an agent's internal state and this state drives future interactions, whereas the former simply requires that future interactions depend on past ones. For Beer the next *state* of the agent depends on the current state of the agent and environment. For Shaw and Todd the next *response* depends on current stimuli and past experiences, not on an abstracted representation of past experiences as state. Somewhere in the middle are Wegner and Goldin (1999): dynamic, non-deterministic but state-based but where future states depend on a stream of past actions and outputs, not only on the current state.

4. The role of past experiences

Let α_i be an agent in environment ξ and let $\{\beta_i^j\}$ be constructs of $\mathfrak{a}_i$. So α_i is a part of ξ and

each of $\{\beta_i^j\}$ is a part of α_i. $\mathfrak{e}_i^k$ is an experience of agent α_i if it is of the agent-entity that is the composition of α_i and $\{\beta_i^j\}$. $\mathfrak{e}_i^k$ is e-experience if it is perturbed by or perturbs one or more entities not part of $\mathfrak{a}_i$. $\mathfrak{e}_i^k$ is an a-experience if it is perturbed by or perturbs one or more entities from $\mathfrak{a}_i$, but is not perturbed by and does not perturb any entities not part of $\mathfrak{a}_i$.

When one entity perturbs an experience of another we say that there is an effect on that entity. An effect by α_i on $\xi - \mathfrak{a}_i$ is via an effector of agent α_i and is of an e-experience. An effect by $\xi - \mathfrak{a}_i$ on α_i is via a sensor of agent α_i and is of an e-experience. Effects that are of a-experiences are internal to the agent and are via perceptors, conceptors and action activators.

The role of past experiences on active ones is central to the what a constructive memory is about. Dewey again:

> "While this is generally admitted, it is often thought that the laws of the association of ideas, conjoined with the past experience, are enough to account for the facts of memory. We have had experiences; these exist stored up, in some unexplained way, in the mind, and when some experience occurs which is like some one of these, or has been previously contiguous with it in time or space, it calls this other up, and that constitutes memory. This, at most, solves but one half the problem. The association of ideas only accounts for the presence of the object or event. The other half is the reference of its present image to some past reality. In memory we re-cognize its presence; i.e., we know that it has been a previous element of our experience. We place the image in the train of our past experiences, we give it some temporal relation; we refer it to some real object once perceived. No idea, however it comes into the mind, certifies of itself that it has ever been experienced before, or under what circumstances it has been experienced. The mind must actively take hold of the idea and project it into time, just as in perceiving it takes hold of the sensation and projects it into space. Were it not for this projecting activity of the mind all would be a fleeting present; the range of intelligence would not extend into a past world" (Dewey 1887)

At a particular time $\mathfrak{a}_i$ will have some number of active experiences $\{\mathfrak{e}_i^k\}$. Some of these will be perturbed by other experiences from $\{\mathfrak{e}_i^k\}$, some by entities from $\xi - \mathfrak{a}_i$, and some by both. Consider a particular active experience $\mathfrak{e}_i^k$. One perturbation of $\mathfrak{e}_i^k$ was of the action that start it, another is the perturbation that finishes $\mathfrak{e}_i^k$. If $\mathfrak{e}_i^k$ is an e-experience then some other agent-entities, or thing-entities, or both, will perturb $\mathfrak{e}_i^k$. If $\mathfrak{e}_i^k$ is an a-experience then only $\alpha_i \cup \{\beta_i^j\}$ will perturb $\mathfrak{e}_i^k$.

An agent will have some minimal rationality (see (Cherniak 1986)), so the way that an agent guides an experience must not be at random. The future of $\mathfrak{e}_i^k$ at some time t will depend on

- How $\mathfrak{e}_i^k$ came to be what it is at t – call this the trajectory $traj(\mathfrak{e}_i^k, t)$ of $\mathfrak{e}_i^k$ until t,
- What perturbations of $\mathfrak{e}_i^k$ there have been up until time t – call this the history $hist(\mathfrak{e}_i^k, t)$ of $\mathfrak{e}_i^k$ until t, and
- The situation at t.

Let $\mathfrak{e}^1, \mathfrak{e}^2, \ldots \in \mathsf{E}$ be particular experiences and $tproj(\mathfrak{e}^k)$ be the projection of a particular experience onto an interval of time, or

$$tproj : Expr \rightarrow (\mathcal{T} \times \mathcal{T}) \tag{5}$$

The limit of the temporal extent of a history and a trajectory is the current time. For all $e \in \mathsf{E}_i$ and all $t \in \mathcal{T}$,

$$hist(e, t) \subseteq e \tag{6}$$
$$traj(e, t) \subseteq e \tag{7}$$
$$hist(e, t) \subseteq traj(e, t) \tag{8}$$

$$\exists\, t_1, t_2 \bullet tproj(traj(e,t)) = (t_1, t_2) \\ \wedge\ max(t_1, t_2) = 2) \tag{9}$$

$$\exists\, t_1, t_2 \bullet tproj(hist(e,t)) = (t_1, t_2) \\ \wedge\ max(t_1, t_2) = 2) \tag{10}$$

An agent should recognise in an active experience something of the trajectory, or history, or both of past experiences and use these to project forward. This is continuity. Recognising the continuity of experiences is what we call memory. It is guiding an experience in a fashion similar to how past experiences progressed, and recognising that this is so. Memory is not retrieving an object from a data store. It is experiences being guided in familiar ways. If an experience is to do with continuity and interaction, what is required of an agent to facilitate this?

We denote the type of experiences of $\mathfrak{a}_i$ as E_i such that $\mathfrak{e}_i^1, \mathfrak{e}_i^2, \ldots \in \mathsf{E}_i$. Now for continuity something persists, and for interaction something changes. So experiences have a temporal aspect, but it cannot be solely temporal. Suppose we want to look closely at an experience and see what this "something" of experiences is, so we fixate on an experience at a particular time $t \in \mathcal{T}$. This fixation is a function from E_i onto some space that we denote as $\mathcal{N}_i$. Let that function be $onetime_i : \mathsf{E}_i \to \mathcal{T} \to \mathcal{N}_i$ for agent $\mathfrak{a}_i$ and time t. As we have only fixed a time, the result is a subspace of reduced dimension: $\dim(\mathcal{T}) \geqslant 1$ and $\dim(\mathcal{N}_i) = \dim(\mathsf{E}_i) - \dim(\mathcal{T})$. We shall write $\mathfrak{e}_i^k(t)$ to mean $(onetime_i\, \mathfrak{e}_i^k\, t)$ with a result $n \in \mathcal{N}_i$ that is what that particular $\mathfrak{e}_i^k \in \mathsf{E}_i$ was like at time $t \in \mathcal{T}$. If $\mathfrak{e}_i^k$ is from the space E_i, this $n \in \mathcal{N}_i$ is from a hyperplane that is a subspace of E_i. A trajectory is these n changing over time. Each hyperplane will contain "somethings" that have meaning to the agent, so we call the subspace $\mathcal{N}_i$ the space of notions of $\mathfrak{a}_i$. We use the word "notion" to maintain independence from any particular kind of agent representation.

The concepts $\mathcal{C}$, percepts $\mathcal{P}$, acts $\mathcal{A}$, sense-data $\mathcal{S}$ and effect-data $\mathcal{E}$ of an agent are all subspaces of $\mathcal{N}$. Any subset of $\mathcal{N}$ is a notion, including $\varnothing$ and $\mathcal{N}$ itself, as is the intersection of any two notions. As such, any given notion n will be a subset of $\mathcal{N}$, or $n \subseteq \mathcal{N}$. Further, that n may itself contain other notions that are themselves both a subset of n and a subset of $\mathcal{N}$.

A space $\mathcal{N}$ may be defined inductively for a particular agent system. As an example, consider wanting to implement agents with notions of "I perceive a thing x that may be a dog (probability 0.6) but it may also be a statue (probability 0.4)". There will need to be concept notions corresponding to "dog" and "statue" from $\mathcal{C}' \subseteq \mathcal{C} \subset \mathcal{N}$. There also needs to be notions of interpretation. In this case the interpretations could be of $(\mathcal{C}' \to \mathbb{R}) \subset \mathcal{N}$ such as $\{(dog, 0.6), (statue, 0.4), \ldots\}$.

If experiences were computed by a Rete engine, $\mathfrak{e}_i^k(t)$ would use the set of facts asserted at t. If experiences were computed by a neural network, $\mathfrak{e}_i^k(t)$ would use network weights and neuron activations at t. For a Beer-style agent, $\mathfrak{e}_i^k(t)$ would use the state vector and its gradient at t. Even for a Shaw-Todd-style agent, we could write an expression for $\mathfrak{e}_i^k(t)$ that would use the state of affairs until t along with the stimuli and response. The Shaw-Todd-style agent itself would not know anything of this expression. It simply keeps on with its direct perception. But we as observers could – after the fact – describe what happened in this way. This is important: $\mathfrak{e}_i^k(t)$ as a subspace of notions from $\mathfrak{e}_i^k$ at a particular time is only our observer description. This is because

1. The implemented agent may not be aware of time as such: it may just compute a new response every time there is a new stimulus.

2. Recollecting an experience differently only becomes evidently so if the original experience is available for comparison. We as observers may know that an experience has changed but the agent probably would not.

An $\mathfrak{e}_i^k$ is a mathematical object with a particular value. Imagine an agent $\mathfrak{a}_i$ at time t building the tower on model of (Gero and Smith 2006, Figure 8(a)). A number of experiences would be active: $\mathfrak{e}_i \subseteq \{$ *moving-the-hand-that-holds-the-yellow-block*, *perceiving-the-tower*, *conceiving-of-*

a-model-church, ...}. These experiences perturb each other, as do sense-data and effect-data from the thing that is aggregation of Lego blocks in (Gero and Smith 2006, Figure 8(a)). Low level experiences like the first two here will be familiar to the agent as they involve notions and processes similar to those of past experiences. The third experience is an active current experience of conceiving the model church tower, but it also may be similar to past experiences and may perturb other experiences in familiar ways. The memories here are the agent recognising having previously had a similar experience. But the current situation will not be identical with the previous one and the agent will have adapted in the meantime, so any recollected experience will be distinct from an original experience. This means that, regardless of similarity, each recollected experience is a different mathematical object. If an old experience was $\mathbf{e}_i^k$, any later recollection of it will be distinct mathematical objects $\mathbf{e}_i^{k'}, \mathbf{e}_i^{k''}, \ldots$.

5. Reconstruction

Memory of an experience as "being guided in familiar ways" is temporally twofold:

- Projecting active experiences into the future:

  > "experience in its vital form is experiental, an effort to change the given; it is characterized by projection, by reaching forward into the unknown" (Dewey 1917).

- Recognising having previously had a similar experience:

  > "reference of its present image to some past reality. In memory we re-cognize its presence; i.e., we know that it has been a previous element of our experience" (Dewey 1887).

The most important idea from the Bartlett quote of Page 1 should be remembered here – that recollections of earlier experiences may vary. Projections into the past and future are not against recordings of experiences, they are against reconstructions of experiences.

If a memory is an experience being guided in a familiar way, but the agent has adapted to other experiences since that familiar experience was active, then what the agent remembers at the later time may not be what was originally the case. Let the active experiences of agent α_i at a time t be $se = \{\mathbf{e}_i^k\}$. If the agent recollects these at some later time:

- one or more e from se may be different than it was at the original time
- the order of one or more e from se may be different than it was at the original time
- one or more e from se may not be recollected at the later time
- one or more e not from se may be recollected at the later time

So a perturbation may trigger the recall of an experience but the perturbation may be understood differently when recalled. The reasons for this include:

- The agent adapts to subsequent experiences, hence recollections of what was experienced before may be different.
- Memories of experiences are dynamic and interlinked, so a recollection is a reconstruction rather than a lookup. This is imperfect recollection accompanied by filling-in of what is missing.
- The agent adapts to subsequent experiences, hence some recollections of what was experienced before may be interpreted differently.
- The role of the current situation; see Section 6.

What does it mean to say that an agent recollects an experience at a later time? Consider an active experience $\mathbf{e}_i^k$ at the current time t. Firstly, we need to consider the trajectory and history of $\mathbf{e}_i^k$ at t. An E_i is an experience as something of

"what it feels like" to be agent i, and a trajectory to t is of an experience up until t. An agent pursuing a goal will activate many actions, and each action will overlap with another e-experience or a-experience. Histories are of perturbations between experiences and so are overlaps of experiences. As such a history is temporally ordered, is of parts of an experience, but is discontinuous. Now an abstraction is the construction of a type X' out of a type X by ignoring that which is not relevant. We wish to restrict E_i to those elements that we intuitively think of as being experiences. So let the supertype of E_i be the type P_i, or

$$\mathsf{E}_i \subset \mathsf{P}_i \tag{11}$$

A P_i may or may not be temporally or spatially continuous, and may or may not be temporally or spatially atomic. A history is therefore of a subtype of P_i that need not be temporally continuous and that satisfies Expressions (6) to (10). An experience is a subtype $\mathsf{E} \subset \mathsf{P}$ that is temporally continuous and is neither temporally nor spatially atomic. An event is a subtype that[4] is temporally atomic but not spatially atomic. An entity is a subtype $\Theta \subset \mathsf{P}$ that is spatially atomic but temporally not. Constructs $\Theta_C \subset \Theta$ and things $\Theta_T \subset \Theta$ are subtypes of Θ.

We shall denote situations of $\mathfrak{a}_i$ as $\psi_i^1, \psi_i^2, \ldots$ of type Ψ_i (see Section 6). Let us say that $recall(\mathfrak{e}_i^k, \mathfrak{e}_i^l, \psi_i^t)$ means that a recalled experience $\mathfrak{e}_i^l$ is familiar to $\mathfrak{e}_i^k$ in the current situation. That is, $\mathfrak{e}_i^l$ at some time $t' < t$ is similar in some way to $\mathfrak{e}_i^k$ at the current time t, given that the current situation is ψ_i^t. Now $\mathfrak{e}_i^k$ and $\mathfrak{e}_i^l$ have different temporal extents, so let $aproj_i(\mathfrak{e}_i^k, t, \psi_i^t)$ be $\mathfrak{e}_i^k$ projected onto a different temporal dimension: of the same time scale but with t located at time zero.

$$aproj_i : \mathsf{P}_i \rightarrow \mathcal{T} \rightarrow \Psi_i \rightarrow \mathsf{P}_i \tag{12}$$

The projection *aproj* is affected by the age of the recalled experience: the older the original experience is, the less likely it is that it should be recalled. It is also projected with respect to the current situation. See Section 9.

[4]These subtypes are inspired by (Seibt 2004).

We need to be able to judge the similarity of two projected experiences. van der Aalst, de Medeiros and Weijters (2006) consider pairs of Petri net models N_1 and N_2. They measure process equivalence is by finding those firing sequences of N_1 that also appear in N_2. In that light we define similarity of two experiences as that part of one experience that is recollected from another experience

$$similarity(h^k, h^l) = \frac{\mid h^k \cap h^l \mid}{\mid h^k \mid} \tag{13}$$

where $h^k = hist(aproj(\mathfrak{e}_i^k, t, \psi_i^t), 0)$ and $h^l = hist(aproj(\mathfrak{e}_i^l, t', \psi_i^t), 0)$ are projected histories of the experiences. $\mid h^k \mid$ is some quantitative measure of the extent of h^k. $h^k \cap h^l$ is defined using the mereology product relation (Seibt 2004):

$$h^m = h^k \cap h^l \mathrel{\widehat{=}} \forall\, h^n \bullet \\ (h^n \sqsubseteq h^k \wedge h^n \sqsubseteq h^l) \Leftrightarrow h^n \sqsubseteq h^m \tag{14}$$

So $recall(\mathfrak{e}_i^k, \mathfrak{e}_i^l, \psi_i^t)$ if $\mathfrak{e}_i^k$ at the current time t is similar to $\mathfrak{e}_i^l$ at some prior time and there are no other more similar experiences that can be recalled.

$$recall(\mathfrak{e}_i^k, \mathfrak{e}_i^l, \psi_i^t) \Rightarrow \operatorname*{argmax}_{\mathfrak{e}_i^l,\, t'} similarity(\\ hist(aproj(\mathfrak{e}_i^k, t, \psi_i^t), 0), \\ hist(aproj(\mathfrak{e}_i^l, t', \psi_i^t), 0)) \tag{15}$$

Notice that similarity is with respect to the current situation. Recall isn't looking up experiences in storage somewhere; it is reconstructing past experiences in the current situation.

If we *want* an agent to be constructive and only constructive then we would need to force it to behave only in the way described. We *could* give it a Google-like memory that just records everything that happens but using such a memory would not be constructive. It may be okay for an agent to sometimes be constructive and sometimes not. Perhaps sometimes it needs to be creative, other times it needs to recall facts. The distinction is between subjective and objective memories. Subjective memories are the kinds of reconstruction that we have been describing. They are of what it

feels like to be this agent. Examples of objective memories are "a magnetic flux density of 1 tesla is 1 weber of magnetic flux per square metre", "an action is always opposed by an equal reaction" and "I owe Mungo $2.50 for coffee yesterday". It is for objective memories that an artificial agent with a constructive memory may sometimes also want non-constructive recall. But having objective and subjective memories does not necessitate an artificial agent having distinct or separate memory mechanisms. The same memory could be used but with habituation differing in different layers (see Sections 7 and 8).

6. What does a situation look like?

A situation is a characteristic of an environment containing interacting agents and things. Recognising that there is a situation is like recognising that there is music and so requires an agent. So that which we call "the situation" is a representation by an agent, but a representation of what? The situation is an influence on how the world is viewed. Notice that the situation is not "a view of the world"; it is a process that changes how those notions behave. To illustrate, consider a story. In the first version of the story a man is standing with his back to a large tree. I start walking towards the tree, but the man is paranoid and so believes that I am after him. He takes fright and runs away. In the second version of the story a woman is standing with her back to the same tree. The tree contains a flock of Galahs[5]. The tree in the first story also contained the Galahs but the man was to self-absorbed to notice. I start walking towards the tree. The woman is confident, not paranoid, and hears the birds in the tree. She turns around to see the birds I am obviously, to her mind, looking at.

An experience may be understood differently when recalled, and part of the reason for that is the changed situation. Consider an e-experience $\mathscr{e}_i^k(t)$ as an example. As it is an e-experience it involves the perturbation of a thing-entity or of another agent. The trajectory for $\mathscr{e}_i^k(t)$ is illustrated in Figure 2. It is shown as $\mathscr{e}_i^k(t)$ drawn over a time interval (t^-, t^+) starting from a perturbation (a "query" on the experience) at t^- and ending at an equilibrium (a "result") at t^+. Shown are trajectories from two similar initially perturbed notions $n, n' \subset \mathcal{N}$ are shown.

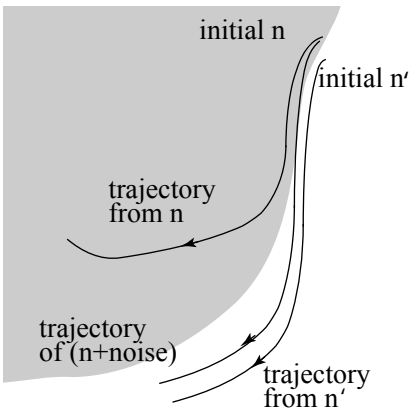

Figure 2. Trajectories of similar queries.

Each perturbation upsets the equilibrium of the experience, and the behaviour of the memory is to try to re-settle to an equilibrium. This experience will also perturb other experiences, and the equilibrium is settled with respect to the current situation. So a small change to the initial perturbation or to the current situation can result in a different eventual equilibrium, and hence an different interpretation. This idea is often described as different schemas. The trajectories shown in the example are through two similar schemas: one indicated by the grey background and one indicated by the white background. Each of these – the white and the grey – are more general notions and the trajectory arcs are through more specific notions, so the schemas are abstract interpretations of the perturbation. The first trajectory is from an initial excited notion n, settling into the grey schema. The second is from an initial n', set-

[5] A Galah is an Australian bird (a grey and pink Cockatoo) renowned for its raucous behaviour.

tling into the white schema. The third illustrates the idea that a small change to the starting point (in the example it is starting from a noisy n), a change to the situation before the query has settled, or a new perturbation before the former has settled can result in a large change in where it ends up, including being in a different schema. Such a change can change the focus of attention of an agent, or even be a "Eureka" moment. We shall, however, avoid using the word "schema" as it has been applied so variously that we risk implying something unintended to some readers. Indeed, problems with the word were recognised at least as long ago as Bartlett, so we adopt dynamical systems words like "perturb" and "equilibrium" instead.

> "I strongly dislike the term 'schema'. It is at once too definite and too sketchy ... It suggests some persistent, but fragmentary, 'form of arrangement', and it does not indicate what is very essential to the whole notion" (Bartlett 1932)

Not reaching an equilibrium within some computational bounds triggers the agent to adapt such that a similar perturbation in future will find an equilibrium. This means that influences between notions may change, or that the space of notions known to the agent may change.

As has been noted, the agent may have multiple concurrent experiences. These experiences are not hidden off in isolated compartments. They are all couplings of the same agent-thing, the same agent-constructs, and the same part of the environment. They may be modular (see Section 9) but they still need to influence each other. So if the situation is experiences influencing each other, changing how the world is viewed, then situations are processes. We need a way to describe agent processes that lets each relevant experience influence a target experience. Let a target experience be that experience if the role of the situation is ignored, and call it $\mathfrak{e}_i^{k\varnothing}$ (the un-situated experience). The experience $\mathfrak{e}_i^k$ is the target experience after the influence of the situation with respect to one or more other experiences are included. That is, in this document an experience is situated unless it is explicitly denoted otherwise such as in $\mathfrak{e}_i^{k\varnothing}$. The type of situations is Ψ such that $\psi_i^1, \psi_i^2, \ldots \in \Psi_i$. The type of experiences of $\mathfrak{a}_i$ is E, so situations Ψ_i are functions from notions $\mathcal{N}_i$ and experience E_i to a result that is also a E_i.

$$\Psi_i : \mathcal{N}_i \to \mathsf{E}_i \to \mathsf{E}_i \tag{16}$$

The current situation as seen by an experience $\mathfrak{e}_i^{k\varnothing}$ is one or more functions Ψ that each use another experience to influence this one. The idea is of the representation current situation arising from expectations, so those "other experiences" are of more abstract notions. It means that how situations change the current interpretation of an agent's world can change. To detail this requires the "layers" described in Section 7 but the idea is as follows:

- An experience $\mathfrak{e}_i^{k\varnothing}$ will be computed by one or more constructs cs
- At the current time t, cs will involve notions n
- n will be in one or more layers, and these layers will precede one or more other layers (for example, *acts* in Figure 3 precedes one layer of percepts and one layer of concepts)
- Each layer constructs a situation of type Ψ that applies to layers that precede it
- Situations that apply to $\mathfrak{e}_i^{k\varnothing}$ at time t are applied to give $\mathfrak{e}_i^k$ over temporal extent $(t, t + \delta t)$.

7. An experience as notions adapting to situations

Recall that if we fixate on experience $\mathfrak{e}_i^k$ at a particular time t we see notions from a subspace $\mathcal{N}_i$ of the space E_i, and that a trajectory is these notions changing. If we looked at all notions of all experiences of an agent at one particular time we should notice that some notions applied more generally than some others. An example is of some concepts being more general than some percepts. Suppose we partition $\mathcal{N}_i$ into regions of roughly equal generality, each of which is a subspace of $\mathcal{N}_i$. We will call these partitions "layers".

Some of these will be more general than others, so on a scale of generality some would be "less than" some others. But the "roughly equal" means that there may be some overlap. It is not necessary, for example, that all concepts be more general than all percepts. To see the effect of this, consider the example from Appendix A. Given the temporal *before* relation and Figure 4 we can say that *before*(a, c), *before*(a, d), *before*(b, e) and $c = d$, but *before* does not relate a to b. Some time intervals from $\mathcal{T} \times \mathcal{T}$ will be equal, some will be before others, and some will not be comparable with respect to *before*. Those pairs of time intervals that are not comparable with respect to *before* are those that overlap. The word "comparable" means that the pair are in the partial order.

We call a partial order on generality of $\mathcal{N}_i$ "layers". The ordering is partial because only some subspaces are related by it. Not all layers are comparable. Figure 3 illustrates an example experience at a particular time. It shows notions at that time partitioned into layers; layers are of notions and so are shown as patterned, coloured rectangles. But layers are not disconnected from each other. Notions in layers influence each other, illustrated in the figure with the black lines. An example is associative adaption of notions within and between layers (an example of non-associative adaption is habituation – see Section 9). The influence of the situation at that time on the layer *acts* is also shown (only that one situation as it influences that one layer is shown).

Conjectures are perturbations against the direction of layer precedence, and so are described as being top-down. Inquiries are either in the direction of precedence or are lateral, and so are described as being bottom-up. These directions are indicated on Figure 3. Notice (near the middle of Figure 3) the lateral effects between two layers of percepts. This illustrates that perturbations need not follow layer or level precedence. Lateral here means effects between layers with no precedence between them, but that each precede another common layer. The figure also illustrates that part of the situation that applies to the layer labelled *acts*. The situation at layer *acts* is an influence on how the world is viewed by that layer. The situations, conjectures and inquiries are what drive the memory.

Let $\mathcal{L}_i$ be the set of layers of $\mathfrak{a}_i$ and let $\mathcal{L}_i^j$ be a particular layer, where $j \in \mathbb{N}$ is an index. Consider the Figure 3 example:

- The most general layer is at the top, being the most abstract concepts and so having the most general notions.
- The most specific layers are near the bottom, being sense-data and effect-data and so having the most specific notions.
- There are 2 sensors and one effector
- There are 2 sense-data layers, say $\mathcal{L}^7$ and $\mathcal{L}^8$
- There are 2 low level percept layers – one per sensory modality, say $\mathcal{L}^4$ and $\mathcal{L}^5$
- There is one higher level percept layer, say $\mathcal{L}^3$, that combines $\mathcal{L}^4$ and $\mathcal{L}^5$
- There are two concept layers, say $\mathcal{L}^1$ and $\mathcal{L}^2$, that are separate but equally general
- There is a third concept layer, say $\mathcal{L}^0$, that has the highest level of concepts
- There is one act layer, say $\mathcal{L}^6$
- There is 1 effect-data layer, say $\mathcal{L}^9$

If we require that the layer indexes are consecutive integers starting from 0, we have a set of pairs $\{0 \mapsto \mathcal{L}^0, \ldots, j \mapsto \mathcal{L}^j\}$ for $j = \#\mathcal{L} - 1$. This is a sequence of notion spaces, where the range of the sequence is a set of non-empty, disjoint layers $\mathcal{L}^j$ that collectively partition $\mathcal{N}$. For all j and k in the domain of $\mathcal{L}$,

$$j = k \Rightarrow \mathcal{L}^j = \mathcal{L}^k \tag{17}$$

$$j \neq k \Rightarrow \mathcal{L}^j \cap \mathcal{L}^k = \varnothing \tag{18}$$

$$\mathcal{N} = \bigcup_{j \in \operatorname{dom} \mathcal{L}} \mathcal{L}^j \tag{19}$$

$$\mathcal{L} \operatorname{partitions} \mathcal{N} \tag{20}$$

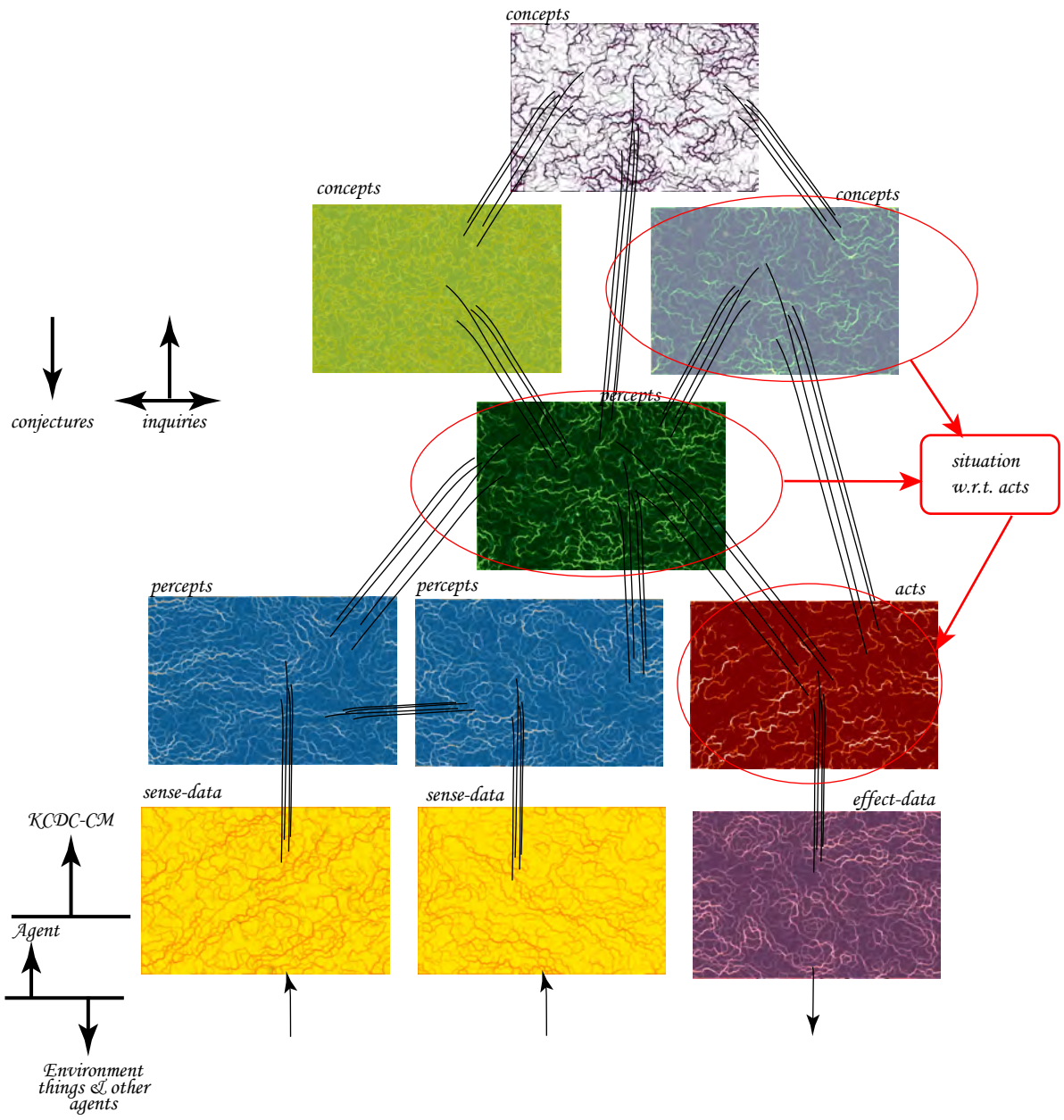

Figure 3. Example experience at a particular time.

We denote the partial order by $\preceq$. For the example, the partial order is $\preceq = \{\mathcal{L}^1 \mapsto \mathcal{L}^0, \mathcal{L}^2 \mapsto \mathcal{L}^0, \mathcal{L}^3 \mapsto \mathcal{L}^1, \mathcal{L}^3 \mapsto \mathcal{L}^2, \mathcal{L}^3 \mapsto \mathcal{L}^0, \mathcal{L}^4 \mapsto \mathcal{L}^3, \mathcal{L}^5 \mapsto \mathcal{L}^3, \mathcal{L}^6 \mapsto \mathcal{L}^3, \mathcal{L}^6 \mapsto \mathcal{L}^2, \mathcal{L}^7 \mapsto \mathcal{L}^4, \mathcal{L}^8 \mapsto \mathcal{L}^5, \mathcal{L}^9 \mapsto \mathcal{L}^6\}$. If $\mathcal{L}^j \preceq \mathcal{L}^k$ and $j \neq k$, then we say that layer j precedes layer k and that $\mathcal{L}^j \prec \mathcal{L}^k$.

Layer precedence is according to notion generality, so percepts and acts are in layers that precede layers containing concepts. Sense-data are in layers that precede layers containing percepts, and effect-data are in layers that precede layers containing acts. Consequently, all percepts in $\mathcal{P}$ precede all concepts in $\mathcal{C}$. All acts in $\mathcal{A}$ precede all concepts in $\mathcal{C}$. All sense-data in $\mathcal{S}$ precede all percepts in $\mathcal{P}$. All effect-data in $\mathcal{E}$ precede all acts in $\mathcal{A}$. Let $(layerof\ n)$ be the layer that contains notion n.

- *clayers* is the set of layers that contain notions from the space of concepts $\mathcal{C}$.

$$clayers = \{c \in \mathcal{C} \bullet (layerof\ c)\} \tag{21}$$

- *players* is the set of layers that contain notions from the space of percepts $\mathcal{P}$.

$$players = \{p \in \mathcal{P} \bullet (layerof\ p)\} \tag{22}$$

- *slayers* is the set of layers that contain notions from the space of sense-data $\mathcal{S}$.

$$slayers = \{s \in \mathcal{S} \bullet (layerof\ s)\} \tag{23}$$

- *alayers* is the set of layers that contain notions from the space $\mathcal{A}$ of acts.

$$alayers = \{a \in \mathcal{A} \bullet (layerof\ a)\} \tag{24}$$

- *elayers* is the set of layers that contain notions from the space of effect-data $\mathcal{E}$.

$$elayers \mathrel{\widehat{=}} \{e \in \mathcal{E} \bullet (layerof\ e)\} \tag{25}$$

$$\forall\, c \in clayers;\ p \in players;\ a \in alayers \bullet p \prec c \wedge a \prec c \tag{26}$$

$$\forall\, s \in slayers;\ p \in players \bullet s \prec p \tag{27}$$

$$\forall\, e \in elayers;\ a \in alayers \bullet e \prec a \tag{28}$$

The set of layers, the notions in a layer, and effects by layers on each other need not be static. Notice that we are not prescribing any particular architecture of perceptors, conceptors and so on. If there are conceptors then they must be within the layers $\mathcal{C}$, but it may be that $\mathcal{C}$ is empty. But if there is at least one concept notion then $\mathcal{C} \neq \varnothing$, and if there is a concept then something must compute it, implying that there must be at least one conceptor. The same applies to sense-data, pecepts, acts and effect-data.

For a construct $x \in \Theta_C$, let $exp(x,t) \in \mathbb{P}\,\mathsf{E}$ find the active situated experiences that construct x is involved in computing at time t:

$$exp : \Theta_C \to \mathcal{T} \to \mathbb{P}\,\mathsf{E} \tag{29}$$

By "situated experience" we mean an experience $\mathfrak{e}_i^k$ after situations from Ψ_i have operated on $\mathfrak{e}_i^{k\varnothing}$. For a construct $x \in \Theta_C$ of agent $\mathfrak{a}_i$, the notions of interest to construct x at time t are $exp(x,t)(t)$.

Sensors, effectors, perceptors, activators and conceptors are contructs that compute sense-data, effect-data, percepts, act and concepts respectively. It computes sense-data (effect-data, percepts, act, concepts) if it is involved in an experience that is of sense-data (effect-data, percepts, act, concepts) notions except if it is only involved in a perturbation. That is, $(traj(e,t) - hist(e,t))(t)$ for some t contains sense-data. The reason for this expression form is that an effector may be perturbed by a sensor but that does not make the effector a sensor.

$$sensor(x) \Rightarrow \exists\, t \in \mathcal{T} \bullet \bigcup \Big(\{e \in exp(x,t) \bullet traj(e,t) - hist(e,t)\}(t)\Big) \subseteq \mathcal{S} \tag{30}$$

$$effector(x) \Rightarrow \exists\, t \in \mathcal{T} \bullet \bigcup \Big(\{e \in exp(x,t) \bullet traj(e,t) - hist(e,t)\}(t)\Big) \subseteq \mathcal{E} \tag{31}$$

$$perceptor(x) \Rightarrow \exists\, t \in \mathcal{T} \bullet \bigcup \Big(\{e \in exp(x,t) \bullet traj(e,t) - hist(e,t)\}(t)\Big) \subseteq \mathcal{P} \tag{32}$$

$$activator(x) \Rightarrow \exists\, t \in \mathcal{T} \bullet \bigcup \big(\{e \in exp(x,t) \bullet traj(e,t) - hist(e,t)\}(t)\big) \subseteq \mathcal{A} \quad (33)$$

$$conceptor(x) \Rightarrow \exists\, t \in \mathcal{T} \bullet \bigcup \big(\{e \in exp(x,t) \bullet traj(e,t) - hist(e,t)\}(t)\big) \subseteq \mathcal{C} \quad (34)$$

In the above, difference is defined as follows. For $\mathfrak{e}^1, \mathfrak{e}^2, \ldots \in \mathsf{E}$,

$$\mathfrak{e}^3 = \mathfrak{e}^1 - \mathfrak{e}^2 \mathrel{\widehat{=}} \mathfrak{e}^3 \sqsubseteq \mathfrak{e}^1 \wedge \neg\, (\forall\, \mathfrak{e}^4 \bullet \mathfrak{e}^4 \sqsubseteq \mathfrak{e}^3 \wedge \mathfrak{e}^4 \sqsubseteq \mathfrak{e}^2) \quad (35)$$

Given these descriptions of layers and constructs we can reconsider the description of situations from Section 6.

- An experience $\mathfrak{e}_i^{k\varnothing}$ is computed by one or more constructsat current time t.

  $$cs = \{c \in \Theta \mid \mathfrak{e}_i^{k\varnothing} \in exp(c,t) \bullet c\} \quad (36)$$

- At the current time t, cs will involve notions.

  $$n = \Big(\bigcup\nolimits_{c \in cs} exp(c,t)\Big)(t) \quad (37)$$

- n will be in one or more layers.

  $$ln = \bigcup\nolimits_{x \in n} (layerof\ x) \quad (38)$$

- These layers will precede one or more other layers.

  $$pln = \{p, pl \in \mathcal{L}_i \mid pl \in ln \wedge pl \prec p \bullet p\} \quad (39)$$

- Each layer constructs a situation of type Ψ that applies to layers that precede it. Let the situation constructed at time t for a layer be *layersitn*.

  $$layersitn : \mathcal{N} \to \mathcal{T} \to \Psi \quad (40)$$

- Situations that apply to $\mathfrak{e}_i^{k\varnothing}$ at current time t are applied to give $\mathfrak{e}_i^k$ at t.

  A Ψ can use how an experience came to be what it is at t to change what it or another experience is next. This applies to the Section 3 described models of Shaw and Todd, Beer, Wegner and Goldin, and to the experiential descriptions of Dewey. A Ψ cannot use how an experience came to be what it is at t to change what it or another experience was before t. The situation can affect the present, it can affect the future, it can affect recall of the past, but it cannot rewrite the actual past. So there will be some function f of type $(\mathsf{E}_i \to \mathsf{E}_i)$ such that the part of $\mathfrak{e}_i^k$ that extends from time t until time $t + \delta t$ is f applied to $\mathfrak{e}_i^{k\varnothing}$ and $layersitn(t, pln)$. To this end, let

  $$\mathfrak{e}_i^k = \mathfrak{e}_i^{k-} \cup \mathfrak{e}_i^{kt} \cup \mathfrak{e}_i^{k+} \quad (41)$$

  where

  $$\mathfrak{e}_i^{k-} \rrbracket \mathfrak{e}_i^{kt} \rrbracket \mathfrak{e}_i^{k+} \quad (42)$$

  $$tproj(\mathfrak{e}_i^{kt}) = (t, t + \delta t) \quad (43)$$

  Given these, the current situation applies as:

  $$\exists \mathsf{f} \in (\mathsf{E}_i \to \mathsf{E}_i) \bullet \mathfrak{e}_i^{kt} = \mathsf{f}(\mathfrak{e}_i^{k\varnothing}, layersitn(t, pln)) \quad (44)$$

8. Kinds of reasoning

Maher and Gero (2002) described three kinds of agent reasoning: *reflexive*, *reactive* and *reflective*. We can now say what these three are. Note that, as Rosenschein and Kaelbling (1995) start by describing "pure action" and "pure perception" agents, we need at least two more kinds that we call *senseless* and *effectless*.

$$senseless \Leftrightarrow \mathcal{S} = \varnothing \quad (45)$$

$$effectless \Leftrightarrow \mathcal{E} = \varnothing \quad (46)$$

$$interactive \Leftrightarrow \neg\ senseless \wedge \neg\ effectless \quad (47)$$

$(\neg\ \mathit{senseless} \vee \neg\ \mathit{effectless})$
$\wedge$ at least one agent process is perturbed by
at least one external process, or
at least one external process is perturbed by
at least one agent process
$\Leftrightarrow \mathit{isagent}$
(48)

$\mathit{interactive}$
$\wedge$ at least one effection is perturbed by
at least one sensation
$\Leftrightarrow \mathit{reflexive}$ (49)

$(\mathit{interactive} \wedge \mathcal{A} \neq \varnothing)$
$\wedge$ at least one action is perturbed by
at least one perception
$\Leftrightarrow \mathit{reactive}$ (50)

$(\mathit{interactive} \wedge \mathcal{A} \neq \varnothing \wedge \mathcal{C} \neq \varnothing)$
$\wedge$ at least one action is perturbed by
at least one conception
$\Leftrightarrow \mathit{reflective}$ (51)

The reason that *reactive* is not described as $(\mathit{reflexive} \wedge \mathcal{A} \neq \varnothing)$ is because an agent may be both reflexive and reactive, or it may be reactive but not reflexive. Similarly, a reflective agent may or may not be reactive and it may or may not be reflexive. Why is an agent with percepts but not acts called reflexive? Because it is what drives the effectors that matters. So why have percepts in a reflexive agent? To determine the current situation and the conjectures for the sensors.

Notice that these conditions are independent of any kind of implementation. Systems described by this document could be implemented in an imperative or declarative or object-oriented or functional language, or with neural networks, or even in hardware: anything that satisfies the descriptions here.

Now *reflexive* reasoning is described as being of an *interactive* agent where "at least one effection is perturbed by at least one sensation". Effection and sensation are of effectors and sensors respectively. So this condition will hold when the behaviours of an effector x without any sensors are different to the behaviours with at least one sensor y present. That is, when there is a sensor y that perturbs that x.

Let $\mathit{perturbs}(x, y)$ be true if construct x is perturbed by construct y. That is, if an experience of x overlaps an experience of y.

$$\mathit{perturbs}(x, y) \Leftrightarrow \exists\, \mathscr{e}^x, \mathscr{e}^y \in \mathsf{E};\ t \in \mathcal{T} \bullet$$
$$e^x \in \mathit{exp}(x, t) \wedge e^y \in \mathit{exp}(y, t) \wedge \mathscr{e}^x \bigcirc \mathscr{e}^y \quad (52)$$

Given $\mathit{perturbs}(x, y)$,

$$\mathit{reflexive} \Leftrightarrow \mathit{interactive} \wedge (\exists\, x, y : \Theta_C \bullet$$
$$\mathit{effector}(x) \wedge \mathit{sensor}(y) \wedge \mathit{perturbs}(x, y)) \quad (53)$$

$$\mathit{reactive} \Leftrightarrow \mathit{interactive} \wedge \mathcal{A} \neq \varnothing \wedge (\exists\, x, y : \Theta_C \bullet$$
$$\mathit{activator}(x) \wedge \mathit{perceptor}(y) \wedge \mathit{perturbs}(x, y)) \quad (54)$$

$$\mathit{reflective} \Leftrightarrow \mathit{interactive} \wedge \mathcal{A} \neq \varnothing \wedge \mathcal{C} \neq \varnothing \wedge$$
$$(\exists\, x, y : \Theta_C \bullet$$
$$\mathit{activator}(x) \wedge \mathit{conceptor}(y) \wedge \mathit{perturbs}(x, y)) \quad (55)$$

The space of concepts is $\mathcal{C} \subset \mathcal{N}$. The space of percepts is $\mathcal{P} \subset \mathcal{N}$. The space of acts is $\mathcal{A} \subset \mathcal{N}$. The space of sense-data is $\mathcal{S} \subseteq \mathcal{N}$. The space of effect-data is $\mathcal{E} \subseteq \mathcal{N}$. Why are concepts $\mathcal{C} \subset \mathcal{N}$ but sense-data $\mathcal{S} \subseteq \mathcal{N}$? An *effectless* agent may only have sense-data, meaning that $\mathcal{S} \subseteq \mathcal{N}$. An agent with concepts will be *interactive*, meaning that $\mathcal{S} \cup \mathcal{E} \neq \varnothing$. Hence it is possible to have an agent with only sense-data but it is not possible to have an agent with only concepts.

Reflective construction of particular notions may eventually lead to reactive reasoning over similar notions if that refection recurs sufficiently. This is non-associative adaption to continuity of experiences, and it is often called habituation. Reactive reasoning may similarly lead to new reflexive reasoning. The idea of habituation is not new; James (1890), for instance[6]:

[6]Emphasis in the following is by James

> "When we are learning to walk, to ride, to swim, skate, fence, write, play, or sing, we interrupt ourselves at every step by unnecessary movements and false notes. When we are proficients, on the contrary, the results not only follow with the very minimum of muscular action requisite to bring them forth, they also follow from a single instantaneous 'cue.' "
>
> "In action grown habitual, what instigates each new muscular contraction to take place in its appointed order is not a thought or a perception, but the *sensation occasioned by the muscular contraction just finished*."

As experience is an interplay of continuity and interaction, habituation of interactive experiences leads to grounding of notions in such interactions.

9. The finitary predicament

The "finitary predicament" is Cherniak's (1986) name for the boundedness, spatially and temporally, of computation in an agent. Cherniak is actually only discussing humans but any agent acting in real time is computationally bounded. Focusing the resources of the agent on continuity and interactivity of its experiences is an advantage here. It need only focus on projecting its own interactions into the future and past, rather than trying to learn a model of its world.

Even with this focus, computational bounds remain. Noticing a failure to reach equilibrium after a perturbation must be bounded in time; updating an experience following such notice must likewise be bounded. In both cases it is not always possible to consider all implications, agent-wide, of all possible changes. Hence the modularity inherent in our descriptions of both experiences and constructs.

The advantages of such modularity are seen in Carpenter and Grossberg (1987) preventing the "relentless degradation of its learned codes by the "blooming, buzzing confusion of irrelevant experience" ". The solution is to make habituation subject to an attention mechanism and to the novelty of events and have it subject to. That is, the learner is distributed and localised, or modular. This also helps when learner stability and plasticity are pulling a learner in different directions.

Not only should there be modularity but memories should age. A particular experience should eventually be forgotten even if no similar recent experiences replace it. This is referred to as explicit ageing. Implicit ageing means replacing of parts of old memories by newer experiences. This takes advantage of the interactive nature of designing: as new memories get absorbed, generalisations change and so a later reconstruction of the same query results in different memories being retrieved.

10. Conclusions

In this paper we considered what kind of memory a situated design agent should have. We described an experiential view of memory in a situated agent.

John: obviously more is needed here. I thought that some of the stuff from your ARC proposal (on why we would want to build such an agent) should go here and in the introduction of part A?

REFERENCES

Allen, JF: 1991, Time and again: The many ways to represent time, *International Journal of Intelligent Systems* **6**: 341–355.

Bartlett, FC: 1932, *Remembering: A Study in Experimental and Social Psychology*, Cambridge University Press.

Beer, RD: 1997, The dynamics of adaptive behavior: A research program, *Robotics and Autonomous Systems* **20**: 257–289.

Beer, RD: 2003, The dynamics of active categorical perception in an evolved model agent, *Adaptive Behavior* **11**(4): 209–243.

Carpenter, GA and Grossberg, S: 1987, A massively parallel architecture for a self-organizing neural pattern recognition machine, *Computer Vision, Graphics, and Image Processing* **37**: 54–115.

Cherniak, C: 1986, *Minimal Rationality*, MIT Press.

Clancey, WJ: 1997, *Situated Cognition*, Cambridge University Press.

Dewey, J: 1887, *Psychology*, Harpers & Brothers. Reprinted in *John Dewey: The Early, 1882-1898*, Vol. 2: 1887, Southern Illinois University, 1969, pp. 3–21.

Dewey, J: 1896, The reflex arc concept in psychology, *Psychological Review* **3**(4): 357–370.

Dewey, J: 1917, The need for a recovery of philosophy, *Creative Intelligence: Essays in the Pragmatic Attitude*, Henry Holt and Company, pp. 3–69. Reprinted in *John Dewey: The Middle Works, 1899-1924*, Vol. 10: 1916-1917, Southern Illinois University, 1985, pp. 3–49.

Dewey, J: 1938, *Experience and Education*, Collier. Reprinted in 1963.

Gero, JS and Fujii, H: 2000, A computational framework for concept formation for a situated design agent, *Knowledge-based Systems* **13**(6): 361–368.

Gero, JS and Smith, GJ: 2006, A computational framework for concept formation for a situated design agent, Part A: Gero and Fujii revisited. To submit.

James, W: 1890, *The Principles of Psychology*, Vol. I, Dover. Republished 1950.

Maher, ML and Gero, JS: 2002, Agent models of 3d virtual worlds, *in* G Proctor (ed), *ACADIA 2002: Thresholds, Pamona, CA*, ACADIA, pp. 127–138.

Öhrstöm, P and Hasle, PVV: 1995, *Temporal Logic: From Anceint Ideas to Artificial Intelligence*, Kluwer Academic.

Rescher, N: 2002, Process philosophy, *in* EN Zalta (ed), *Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Centre for the Study of Language and Information, Stanford University. `http://plato.stanford.edu/entries/process-philosophy/`.

Rosenschein, SJ and Kaelbling, LP: 1995, A situated view of representation and control, *Artificial Intelligence* **73**: 149–173.

Seibt, J: 2004, Free process theory: Towards a typology of occurrings, *Axiomathes* **14**: 23–55.

Shaw, R and Todd, J: 1980, Abstract machine theory and direct perception, *Behavioral and Brain Sciences* pp. 400–401.

Smith, B: 1996, Mereotopology: A theory of parts and boundaries, *Data and Knowledge Engineering* **20**: 287–303.

van der Aalst, WMP, de Medeiros, AKA and Weijters, AJMM: 2006, Process equivalence: Comparing two process models based on observed behavior, *in* S Dustdar, JL Faideiro and A Sheth (eds), *International Conference on Business Process Management (BPM 2006)*, Vol. 4102 of *Lecture Notes in Computer Science*, Springer-Verlag. `http://is.tm.tue.nl/staff/wvdaalst/publications/z8.pdf`.

Wegner, P and Goldin, D: 1999, Models of interaction. ECOOP'99 Tutorial.

A. Temporal relations on experiences

We can use the function *tproj* from Section 4 to define three primitive relations for temporal interval projections of experiences: sum, less than (a temporal partial order that excludes temporal overlap) and follows (temporally immediately follows). These are adapted from (Seibt 2004, Öhrstöm and Hasle 1995).

$$\mathfrak{e}^3 = \mathfrak{e}^1 + \mathfrak{e}^2 \mathrel{\widehat{=}} \forall\, \mathfrak{e}^4 \bullet (\mathfrak{e}^4 \bigcirc \mathfrak{e}^3 \Rightarrow \mathfrak{e}^4 \bigcirc \mathfrak{e}^1 \vee \mathfrak{e}^4 \bigcirc \mathfrak{e}^2) \tag{56}$$

$$\begin{aligned} \mathfrak{e}^1 <_\tau \mathfrak{e}^2 \mathrel{\widehat{=}} \neg\, (\mathfrak{e}^1 \bigcirc \mathfrak{e}^2) \wedge (\exists\, t_1, t_1', t_2, t_2' \bullet \\ tproj(\mathfrak{e}^1) = (t_1, t_1') \wedge \\ tproj(\mathfrak{e}^2) = (t_2, t_2') \wedge \\ max(t_1, t_1') < min(t_2, t_2')) \end{aligned} \tag{57}$$

$$\mathfrak{e}^1]\!] \mathfrak{e}^2 \mathrel{\widehat{=}} \mathfrak{e}^1 <_\tau \mathfrak{e}^2 \wedge \neg\, (\exists\, \mathfrak{e}^3 \bullet \mathfrak{e}^1 <_\tau \mathfrak{e}^3 \wedge \mathfrak{e}^3 <_\tau \mathfrak{e}^1) \tag{58}$$

These can be used to define Allen's (1991) interval relations on the temporal extent of experiences. The following is derived from (Öhrstöm

and Hasle 1995).

$$meets(\mathfrak{e}^1, \mathfrak{e}^2) \mathrel{\widehat{=}} \mathfrak{e}^1] \mathfrak{e}^2 \tag{59}$$

$$metby(\mathfrak{e}^1, \mathfrak{e}^2) \mathrel{\widehat{=}} \mathfrak{e}^2] \mathfrak{e}^1 \tag{60}$$

$$before(\mathfrak{e}^1, \mathfrak{e}^2) \mathrel{\widehat{=}} \exists\, \mathfrak{e}^3 \bullet \mathfrak{e}^1] \mathfrak{e}^3] \mathfrak{e}^2 \tag{61}$$

$$after(\mathfrak{e}^1, \mathfrak{e}^2) \mathrel{\widehat{=}} \exists\, \mathfrak{e}^3 \bullet \mathfrak{e}^2] \mathfrak{e}^3] \mathfrak{e}^1 \tag{62}$$

$$equals(\mathfrak{e}^1, \mathfrak{e}^2) \mathrel{\widehat{=}} tproj(\mathfrak{e}^1) = tproj(\mathfrak{e}^2) \tag{63}$$

$$starts(\mathfrak{e}^1, \mathfrak{e}^2) \mathrel{\widehat{=}} \exists\, \mathfrak{e}^3 \bullet tproj(\mathfrak{e}^2) = tproj(\mathfrak{e}^1) + tproj(\mathfrak{e}^3) \tag{64}$$

$$startedby(\mathfrak{e}^1, \mathfrak{e}^2) \mathrel{\widehat{=}} \exists\, \mathfrak{e}^3 \bullet tproj(\mathfrak{e}^1) = tproj(\mathfrak{e}^2) + tproj(\mathfrak{e}^3) \tag{65}$$

$$finishes(\mathfrak{e}^1, \mathfrak{e}^2) \mathrel{\widehat{=}} \exists\, \mathfrak{e}^3 \bullet tproj(\mathfrak{e}^2) = tproj(\mathfrak{e}^3) + tproj(\mathfrak{e}^1) \tag{66}$$

$$finishedby(\mathfrak{e}^1, \mathfrak{e}^2) \mathrel{\widehat{=}} \exists\, \mathfrak{e}^3 \bullet tproj(\mathfrak{e}^1) = tproj(\mathfrak{e}^3) + tproj(\mathfrak{e}^2) \tag{67}$$

$$overlaps(\mathfrak{e}^1, \mathfrak{e}^2) \mathrel{\widehat{=}} \exists\, \mathfrak{e}^3, \mathfrak{e}^4, \mathfrak{e}^5 \bullet tproj(\mathfrak{e}^1) = tproj(\mathfrak{e}^3) + tproj(\mathfrak{e}^4) \wedge tproj(\mathfrak{e}^2) = tproj(\mathfrak{e}^4) + tproj(\mathfrak{e}^5) \tag{68}$$

$$overlappedby(\mathfrak{e}^1, \mathfrak{e}^2) \mathrel{\widehat{=}} \exists\, \mathfrak{e}^3, \mathfrak{e}^4, \mathfrak{e}^5 \bullet tproj(\mathfrak{e}^2) = tproj(\mathfrak{e}^3) + tproj(\mathfrak{e}^4) \wedge tproj(\mathfrak{e}^1) = tproj(\mathfrak{e}^4) + tproj(\mathfrak{e}^5) \tag{69}$$

$$during(\mathfrak{e}^1, \mathfrak{e}^2) \mathrel{\widehat{=}} \exists\, \mathfrak{e}^3, \mathfrak{e}^4 \bullet tproj(\mathfrak{e}^2) = tproj(\mathfrak{e}^3) + tproj(\mathfrak{e}^1) + tproj(\mathfrak{e}^4) \tag{70}$$

$$contains(\mathfrak{e}^1, \mathfrak{e}^2) \mathrel{\widehat{=}} \exists\, \mathfrak{e}^3, \mathfrak{e}^4 \bullet tproj(\mathfrak{e}^1) = tproj(\mathfrak{e}^3) + tproj(\mathfrak{e}^2) + tproj(\mathfrak{e}^4) \tag{71}$$

We do not attempt to define equivalent relations on the notions of an experience. Temporal overlap does not imply experiential overlap:

$$overlaps(\mathfrak{e}^1, \mathfrak{e}^2) \not\Rightarrow \mathfrak{e}^1 \bigcirc \mathfrak{e}^2 \tag{72}$$

$$\mathfrak{e}^1 \bigcirc \mathfrak{e}^2 \Rightarrow overlaps(\mathfrak{e}^1, \mathfrak{e}^2) \tag{73}$$

These interval relations on the temporal extent of experiences are partial orders on experiences. Given the example of Figure 4, we can say that *before*(a, c), *before*(a, d), *before*(b, e), but *before* does not relate a to b.

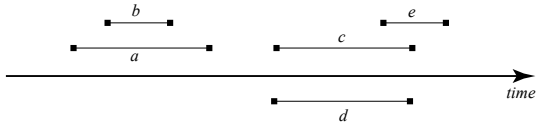

Figure 4. Example time intervals