A Model for Design of Machine & Human Cognition

Tom Tollenaere

Draft V4 – April 3, 2013

Abstract

This paper proposes a thought experiment to design, from a functional point of view, of a machine that can think, and which will be called an Artificial Mind (AM). The term 'thinking' for the purpose of this work is defined in the form of a minimal Mission Statement. Since the only thinking entities known today are humans, human cognition will be used as an inspiration. Architectural design principles for an AM follow from the Mission Statement. Further the required functional properties of what we consider the foundation of any AM are proposed: Associative Correlative Time Memories (ACT Memories or ACT-M).

Then we consider a further thought experiment, in which we apply the ACT-M model to human cognition, in the sense of a Philosophy of Mind (Human Mind or HM). This thought experiment leads to several interesting and unexpected (philosophical) conclusions. However, we also show that the ACT-M model does not explain human consciousness, although the differences in application between AM and HM do isolate the hard part of human cognition.

Next the ACT-M model is compared to other existing models of (human & machine) cognition. The paper concludes with a rebuttal of potential weaknesses and objections.

Version History

- August 19, 2012: first version for review, excluding discussion
- February 22, 2013: second version for review, including discussion
- March 18, 2018: third version for review, expanded section on qualia, added clarification on scope of model, excluding hard problem of human consciousness.
- April 2, 2013: restructured section on criticism, minor rephrasing in various paragraphs. Removed subsection on determinism due to space constraints. Edits to previous version are highlighted.

Introductory Remarks

The following text is a draft. There is no single line of thought in what is proposed in this text; hence I will regularly refer to concepts to be explained in further sections of the text. I apologize for the resulting inconvenience. The text uses 'I' when the author proposes elements a reader may or may not object to; the text uses the 'we' form for parts of the discussion where author and reader follow the logic of the though experiment.

1. Mission Statement

Let us start a thought experiment with the following mission statement:

> I want to design a conceivably buildable machine which is conscious and which can think on a human-like level (or above), and which we will call an Artificial Mind (AM)[1].

Definitions of 'consciousness' and 'think' will follow in section 1.2. Discussion of the term 'human-like' can be found in section 3.8.

1.1. Principles

The following principles will be adhered to:

- I want to apply Occam's razor in the sense that I want minimal design requirements. Any elements that may follow from a design requirement will not be considered design requirements
- The design should be implementation-independent and conceivably buildable with current or future technology.

Occam's razor is a principle that suggests we should tend towards simpler theories until we can trade some simplicity for increased explanatory power. Phrased differently, all other things being equal, Occam's razor is a heuristic which prefers a simple model over a more complex one. The principle is attributed to the 14th-century English logician, theologian William of Ockham, although the concept was familiar long before him.

Admittedly, and contrary to the popular summary, the simplest available theory may sometimes be less accurate explanation. I will not be too dogmatic about this, but plan to only add complexity if it turns out that the available complexity is not sufficient to achieve an AM.

For the list of principles we also need to clarify what we mean by 'design'. In terms of design we will follow principles used in software design, where, typically, when a new computer system needs to be built; an initial step is writing out functional requirements (see e.g. Jacobson, Booch & Rumbaugh,

[1] What will be called an AM in this text is an example or an instance of what is often called Artificial General Intelligence (AGI) (see e.g. Yudkowsky 2007).

1999). This is often referred to as 'functional analysis'[2]. Functional requirements define *what* the computer system should be doing, from a functional point of view. In a later phase a technical design is made, in which decisions are made as to *how* the system will be built in order for it to fulfill its functional requirements. The technical design phase deals with decisions such as which operating system to use, which programming language and database system to use and decisions regarding technical, algorithmically implementation of functionally required algorithms or processes. N this sense in what follows a functional analysis & design for an AM will be presented.

1.2. Definition of Artificial Mind

Since the result of our mission should be testable (see further), and since there is no (current) consensus on what consciousness or thinking really is (also see further), we need a definition to test against. Applying Occam's razor, I want this definition to be as simple as possible. For the purpose of this thought experiment, a machine is conscious and thinking if:

- The machine is able to learn from experience:
- The machine is able to think in the following sense: 'thinking' is
    - The ability to draw rational conclusions. By rational conclusions we mean conclusions based on reasoning that a human mind can understand, and can agree to their reasonability. This does not imply that a human mind would have to agree with the conclusions of the machine, but it should be able to understand the reasonability of the argumentation.
    - The ability to come up with original ideas, that is principles, theories that it has not learned a priori
- The machine is self-conscious, meaning it is aware of itself in the sense that it knows what it is thinking, and knows that it is itself that is doing the thinking.
- We can test the above requirements, i.e. we can inspect the machine's behaviors and/or have access to what it thinks and how it reasons, and come to the conclusion that the machine meets the requirements.

[2] In software engineering, several approaches to Software Development Life Cycles exist. One extreme is a waterfall model, in which a technical implementation only proceeds after the functional design has been completed. The other extreme are techniques such as Agile (e.g. Beck, 2001) or SCRUM (e.g. Schwaber & Beedle, 2002) which are iterative, and in which both functional analysis & design and technical implementation are progressing incrementally and coupled. Since in this paper we are mainly concerned with the design of an AM, we will ignore technical aspects. Furthermore, we will argue further in the text that an AM is not necessarily a software system; the reference to principles of software design is a reference of methodological nature; not one of a software design nature.

Any machine which fulfills the above conditions will be considered to be an AM. In the remainder of the text we will refer to the above definition as The Definition.

1.3. What Mission Statement Is Not & what The Definition does Not Imply

The Definition is not a priori a definition of how humans think, or how human consciousness works. One may or may not agree that the Definition covers human thinking and/or consciousness, given there is no wide consensus on what 'thinking' and 'consciousness' are. Of course other machine designs might be proposed based on other definitions; and such machines may be equally or more interesting as the AMs presented here. Discussion of how the definition chosen here relates to other models and theories about consciousness is deferred to section 5.

However, for how we work forward from the Definition towards a functional model of an AM, the only inspiration for AMs are Human Minds (HMs) and hence we will have to take cues from human reasoning (to the extent this is known or knowable). I will argue later that the resulting model may be a good model for modeling human cognition, but that will be a result rather than a goal.

The Mission does not cover free will. Omohundro (2007) has proposed a list of drivers that every AI (Artificially Intelligent) system (in the sense of the AMs presented here) would have, and this list includes free will. The Mission does not include this, mainly because there is no (philosophical) consensus on what free will is, and whether humans have free will. However, the AM design presented here *will* lead to what the concept might imply for AMs (and HMs), and this will be discussed in section 3.7.

The mission of this work, at least at this point, is not to actually design a technical implementation of an AM. We are working on a conceptual, architectural level. The technology to build an AM is probably not available to us humans today, though it may become available at some point in the (near?) future. We are not insisting that the architecture is capable of processing symbols. Neither are we stating that the architecture should be connectionist. Discussion of possible implementation methods is addressed in section 5.

The Mission is not to build a machine that can pass the Turing Test; Turing testability will be addressed in section 3.9.

1.4. Consequences of Mission Statement & Definition of AM

The definition of AM automatically leads to (architectural) conclusions. These are discussed in this section. The section concludes with a high level design, which follows automatically and logically from the Definition.

1.4.1. Introspection

From the statement "the AM knows what it is thinking" follows that it can introspect, i.e. it can think about its own thoughts. A definition of what we consider a 'thought' follows in section 2.8.

1.4.2. Language

The Mission Statement does not state that the machine would have be able to learn language. The need for ability to learn language (any language, human or synthetic) follows from the testability clause. If we are to test whether the machine thinks (and what it thinks) on a human-like level or above it needs to be able to tell us what it thinks. Hence, it needs to be able to learn or (a) language that humans can understand. If it can learn English then there is no reason it would be incapable of learning Chinese, ergo the machine should be able to learn any human language, at least to a level of proficiency such that it is able to rationally convey its thoughts.

1.4.3. Sensorium

If the AM is capable of learning, it will need inputs to learn from. Hence the architecture needs a sensorium. The mission statement does not specify what kind of sensorium; this could be a visual system (one camera, or multiple camera's to give it stereo vision), and auditory system, an olfactory system and so forth. But we need not limit ourselves to human sensoria; we could give an AM radar and infrared (extensions to human-type visual systems). We could give it Geiger counters, RF readers, anything goes.

We know that the human visual system does not directly project the images caught by our retinas onto the higher level parts of our brains that deal with conscious thought; there are several phases of visual pre-processing that occur in the various areas of the visual cortex. We humans have – precognitive – hardware that performs edge detection, motion detection, Fourier transforms and so forth. All this hardware can be built into an AM's sensorium.

1.4.4. Motorium

If an AM is to communicate with us, it will need to provide some sort of output; ergo, it needs a motorium. This could be a terminal output (on which it can 'type' green characters on a black background, for example), or a speech synthesizer, a printer, a modem. We may want to provide the machine with wheels so it can move about, or provide it with a remote-controllable extension in the form of a moveable agent equipped with a camera (potential relevance of this will be discussed in section 5.4)

1.4.5. Memory

If an AM needs to learn from experience, it needs a memory to store experiences. Ergo, a memory is a necessary architectural component of an AM. Note that at this point we do not define memory in terms of "storing symbols" (the GOFAI (Good Old Fashioned Artificial Intelligence) way of thinking about AI) (Artificial Intelligence), neither about modifications of synaptic weights in an artificial neural network model (Rumelhart & McClelland 1986). Using Occam's razor, all we need is a way of storing experiences.

Hence, from the mission statement follows a high level functional design:

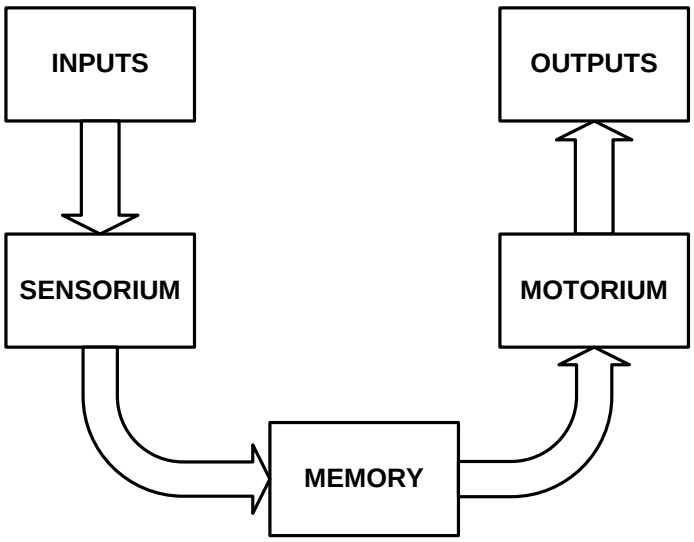

Diagram 1: Initial Functional Design of an AM

Current technology is quite capable of building sensoria, as well as motoria. Now the interesting and challenging bit is the memory. For this we propose the ACT-M model in the next section.

2. The ACT-M Model

In this section the ACT-M[3] model will be defined. We will argue that a memory for an AM needs to cover 3 crucial aspects:

- Association
- Correlation
- Time

[3] ACT here stands for Association, Correlation & Time; the acronym is not to be confused with Aleksander's Artificial Consciousness Theory (Aleksander 1996).

These will be discussed in detail in the following subsections but for the sake of clarity let us state in summary that by 'correlation' we mean a mechanism that can detect correlations in (input) data, i.e. stuff that belongs together; by 'association' we mean a mechanism that can link stuff(1) happening to stuff (2) that was earlier correlated to stuff(1). A memory that covers Association, Correlation and Time will be called an ACT Memory or ACT-M in short. In one of the follow subsections we also discuss what exactly the memory would need to 'store'. Following this section, section 3 will cover the consequences of the model proposed here for AMs.

2.1. The Importance of Time

We humans are not a system that takes static input and that produces static output. For one we have multiple sensorial inputs going on at the same time, but more importantly our inputs are continuously changing over time.

When we read, we process one character at a time. When we hear a word, we process a pattern over time and when we speak we produce a pattern over time. We are capable of processing things like movement (which is a delta in spatial position over time), speed, acceleration and so forth, all of which involve a time factor.

We would expect an AM to hear (and understand) us, and expect it to be able to 'talk' to us (whether through a speech processor or through a teletype (as in, an AM talking to us on a screen, like we communicate using an internet text-based chat box) is irrelevant for now, since both use/need the notion of time). Hence an AM needs to be able to deal with the concept of running time. (be that discretized of continuous)

Furthermore, since an AM needs to be able to learn, and learning comes from experience, and experiences come over time, a logical conclusion is that, because of the importance of time, an AM is a dynamic system, in the sense that it 'moves' (changes, evolved and thinks) over time. This notion will be further expanded in the remainder of the text.

2.2. What to Store?

We established that any design for an AM would need some sort of memory, in which to store its 'knowledge'. However, we have not defined yet what this 'knowledge' would consist of. This is the topic of this subsection.

I propose that it is functionally necessary for an AM memory to store patterns. A pattern is anything spatio-temporal, in whichever spatial domain.

This can be sensorially or motorially: sensorially when an AM hears my name 'tom' that results in a spatial pattern, over time, in the audible domain; when an AM 'sees' a car speeding on a highway,

that results in a spatial pattern over time, in the 3 dimensional space of the actual world (if the AM has the capability for stereo vision) or in a 2 dimensional representation of the actual world (in case the AM has only one 'eye' (be that a camera or a retina). For now we make abstraction of whether and how an AM would 'know' the 'concept' 'car' but we will return to this matter later.

Motorially, when an AM speaks my name, it produces (and ergo there is) some spatio-temporal pattern that results in the sound of my name in the audible; perhaps this is a spatial-temporal pattern, eventually output by a voice synthesizer, that results in a speaker affecting air, so that I can hear it uttering my name or perhaps this is a different spatio-temporal pattern within the AM that simply results in the output of the sound wave sounding like my name. When an AM 'speaks' my name on e.g. a teletype terminal, the spatial pattern is 't' 'o' 'm', over time, one after the other.

If the AM is capable of hearing itself, the audible pattern of my name is actually hears by the AM, meaning that the production of a motor pattern results in the input of an equivalent sensorial input pattern. Ditto if the AM is capable of seeing itself, the output of 't' 'o' 'm' on a teletype is seen by the AM and the production of a motor pattern results in the input of an equivalent sensorial input pattern.

For the AM to hear and talk about me, Tom, all these patterns need to be stored somewhere, hence a memory.

Note that since this is a functional exercise, no statement is made regarding how patterns would physically be stored. An auditory pattern like a voice uttering my name *could* be stored as a discretization of the sound wave of my name, but it *need* not be such. Whichever way this pattern is stored, discrete or continuous, in a digital memory, a neural network or a piece of brain tissue, encoded as synaptic connections, is for the sake of this discussion irrelevant.

2.3. Correlation

A memory that can store patterns over time, as described in 2.1 and 2.2 is not at all very exciting. In order for an AM to be able to learn, I propose that the memory needs to be able to correlate. This should be interpreted in the following sense. Suppose the AM 'sees' my mugshot, and keeps hearing my name (in the audible domain), uttered by various voices in various ways (male or female, sung, loud or whispered, clear or mumbled, slow or fast), it should be able to synthesize the information; or in other words, it should be able to discover that there is a correlation between these various voicings of my name and my picture. Instead of storing all patterns of all voicings of my name it ever heard, it should be able, by correlation, through the spatial patterns of my face, coupled with the temporal coincidence of the voicings, to correlate 'what-Tom-sounds-like' with 'what-Tom-looks-like' . 'What-Tom-sounds-like', then, is a pattern different from (but perhaps close or similar to) all these voicings.

Furthermore, the AM might not only just see my mugshot, but actual images of me, mono or stereoscopically. And my face does not look the same every day either: I may or may not be shaven, I may or may not need a haircut, I may or may not wear my glasses, I may look sleepy, tired or awake and alert, I may be far off of nearby, I may be seen in well-lit or darkish environments, I may be lit by harsh light and hence my image may be very contrasty or very soft, I may be standing, lying or upside-down. I may move about, I may be falling, so what my face looks like depends on e.g. the angle at which I present myself to the AM (or to its visual sensorium). Again, the AM should be able to correlate all these 'images' of my face, and 'what-Tom-looks-like' should become a pattern, which is not identical (but perhaps 'close') to what Tom happens to look like on a given day. In other words, the AM should be able to detect that there is an (input) pattern that sounds like 'Tom' and that means something specific; it should be able the detect that there is something that looks like 'me' and that is invariant, and it should be able to detect that both belong together, i.e. are correlated.

Again furthermore, when the AM sees me like it has never seen me before (say I ran into a door and have a black eye), this sensorial 'pattern' of what-I-look-like-today-with-a-black-eye should automatically and immediately result in the pattern of what-I-look-like; i.e. that pattern of 'what-Tom-look-like-today-with-a-black-eye' should in the memory be correlated with the pattern of 'what-Tom-looks-like'. Ditto for 'new' audible versions of my name that the AM has never heard before – these should be auto-correlated with the pattern of 'what-Tom-sounds-like'.

2.4. Stored vs Awakened Patterns

We established that an AM memory needs the ability to store patterns, and needs the ability to detect correlations between patterns. A memory is not only capable of storing pattern, but (just like computer memories) it should be possible to retrieve patterns. The proposed mechanism is the following: a retrieved pattern is a pattern which is 'awoken'; when a pattern is awoken it is 'awake' or active with a certain amplitude, and this amplitude decays over time. Furthermore, an active pattern's amplitude get 'louder' as the pattern is re-activated. The amplitude of the pattern is an indication of 'how important' the awoken memory is. If my name happens to drop dozens of time in a 2 minute conversation, it's probably wise to infer that I (or someone or something else referred to with something that 'sounds-like-Tom') is pretty important in that conversation.

Suppose our AM is looking at a scene, perhaps participating in a conversation with myself. If John pops into our room, a (visual) pattern (or several of these patterns, as John moves about over time) of John's face is fed into the AM. This pattern awakens the pattern of 'what-John-looks-like'. Suppose John steps out of the room immediately, then the visual pattern(s) of John's appearance disappear, and the amplitude of what-John-looks-like quickly fades out. The effect would be something akin to 'hey there's John, oh he's gone again, ah well probably not important'.

If John were to stick around, or if the AM and myself were to start talking about John because of his appearance, then ‘what-John-looks-like’ and ‘what-John-sounds-like’ would remain awake with high amplitude.

Finally, just as awake patterns have an amplitude, so do stored patterns. The amplitude of a stored pattern is its relative ‘importance’, and this importance too decays over time, unless the pattern is awoken regularly.

2.5. Associativity

The mechanisms of storing patterns, and the ability to wake up these patterns, lead us to the final element of proposition for an AM memory: association. I propose that the AM memory also needs to be able to make associations, based on stored and/or awakened patterns and on the correlation between those patterns. Associations can be interpreted as re-awakening previously detected correlations.

Suppose that John has a girlfriend Mary and our AM ‘knows’ this. ‘Knowing’ that John has a girlfriend named Mary might be stored (simplistically speaking) in the memory by correlation: when John appears, very often Mary appears. When talking about John, Mary quite often comes up. Now, when John pops into the room where I was having my conversation with the AM in the previous subsection, something John-like is awoken. Since there a strong correlation between all patterns regarding John and all patterns regarding Mary, *by association* also Mary-patterns are awoken. The louder John patterns are awake, the louder Mary patterns are awake, but absent Mary (or conversation about Mary) the Mary patterns are not as loud as the John patterns. Not only the Mary patterns are awoken, but also all other patterns relating to all other stored information associated with John. How loud such patterns are awakened depends on the strength of the correlation.

Now suppose that Mary has a brother named Paul. When in the above example a Mary pattern is awoken, then by association also Paul patterns are awoken, such as perhaps the pattern of Paul’s brother Ringo, be it less loud than the Mary patterns. And not only the Paul patterns are awoken, but also all other patterns related to Mary.

This results in a chain reaction of awakening of related patterns, and patterns related to those patterns and so forth, as a kind of massively parallel association. This implies that a massive amount of patterns (potentially) irrelevant to the situation are awoken, but the further removed from the actual situation (cf. Paul who is linked to Mary who is linked to John) the more feeble the awakening. In addition, any patterns awoken and somehow linked to the situation at hand are strengthened, become louder. What the situation at hand is about is hence largely defined by the loudest patterns.

2.6. Context

The context in which the AM is 'thinking' is defined by which patterns are loudest awoken. Suppose the AM has learned about a musical instrument called a tom (a kind of drum). The concept 'tom' has been correlated with music. When the AM 'hears' the pattern 'tom' in a music-related context, the context will have awakened various music-related patterns. Then the audio pattern 'tom' comes up, the AM awakens both the pattern that refers to the drum, as well as the notion of the individual named 'Tom'. By association by the music-related patterns, the pattern coding the musical instrument is awoken 'louder' than the pattern 'individual named Tom'. (Unless of course the context involves the individual named Tom playing the tom).

2.7. Feedback

In diagram 1, an addition is now needed: a feedback arrow from memory to memory is needed. This is shown in diagram 2 below.

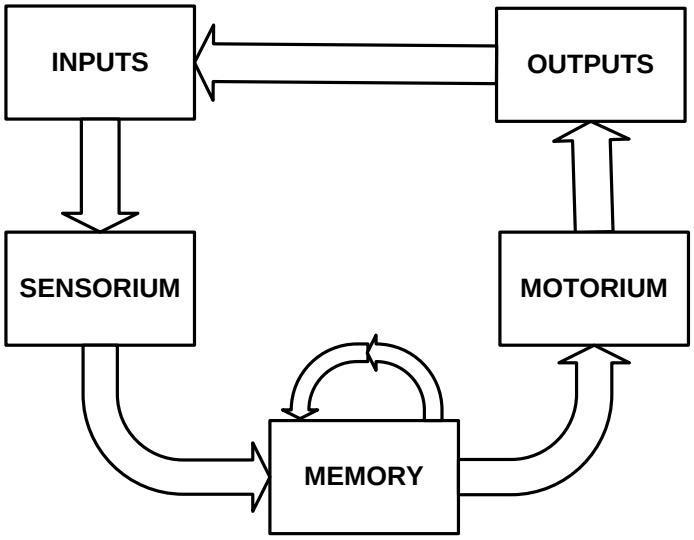

Diagram 2: Internal feedback loop in memory

There are actually various kinds of memory-to-memory feedback going on:

- Any awakened pattern may, by association, awaken other patterns, or increase amplitude of already awakened patterns
- Since we may assume that an AM can sense the effect of its own actions, such as hear itself speak, see itself move (or see the environment in which it moves change due to its movement) this also constitutes a feedback loop

- Finally, a ‘train of thoughts’ set in motion by certain sensorial input, may carry on adiabatically, by association upon association, without further sensorial input.

In the last bullet point the term ‘train of thought’ should be considered informally; the topic of ‘train of thought’ will be addressed more fundamentally in discussion on the model in section 3.6. Before this can be addressed, the material above presents sufficient basis for the definition of the term ‘thought’ in the next subsection.

2.8. Definition of Thought

In an AM based on an ACT memory as presented above, ‘thought’ is defined as any awake temporal pattern which is not solely due to sensorial input. Furthermore, ‘thinking’ is defined as the dynamic process of awake patterns, that keeps running due to chained association based upon the currently active pattern (or patterns, since a pattern of patterns is still a pattern); such chained association due to the feedback mechanism in the memory. Hence, thought is a process, and this process basically consists of ‘pattern soup’; a soup in which, once cooking, it becomes practically impossible to discern the original ingredients. The pattern soup metaphor will be revisited in section 3.1 when we discuss the consequences of the ACT-M model; more about looking at an AM as a process in section 3.3.

The remainder of this section wraps up the final aspects of the definition of ACT memories.

2.9. ACT-M as a Chaotic Dynamic System

In section 2.1 the remark was made that due to the dependency on time, an AM is a dynamic system. Based upon that material which followed section 2.1 this idea can now be expanded upon.

Dynamics occur at different levels:

- As the AM is presented with inputs, patterns are dynamically awoken.
- Awoken patterns give rise to other patterns awakening, which in turn cause other patterns to awake
- Co-incidence of awoke patterns give rise to new correlations and/or updates to the importance of stored patterns (and implicitly, the forgetting of non-awakened patterns), and new correlations cause, by association, again the waking up of other stored patterns
- Internal feedback causes further awakening of patterns; hence the dynamics of patterns waking up, causing other patterns to awaken, and subsequent changes to pattern storage in memory.
- Even should all inputs cease, the AM will keep thinking, and that act of thinking (a dynamic process) causes changes in the memory (again a dynamic process).

Because of the above proposition that an AM is a chaotic system, in the sense that an infinitely small variation in input may lead to an infinitely large difference in outcome. As a result, an AM becomes an indeterminable system: any individual emergent property of an AM cannot be traced to a unique experience. Alternatively, one may consider an AM to be a "dispredictable" system: any potential future state of 'being' of an AM cannot be predicted unless complete knowledge of all experiences is known, which, in practice will be practically impossible. We use the term dispredictable, because a certain degree of prediction is approximately possible. Dispredictability will be further explored in section 3; in order to perform such exploration additional insights presented in the remainder of this section will be needed.

Further on the dynamics of the memory: there is not necessarily a single and/or fixed pattern that defines e.g. Tom. E.g. as I grow older, my looks change, hence the patterns that define 'Tom' adapt. Also, patterns are not 'atomical' in the sense that if there is a pattern 'what-Tom-is' then that is so intrinsically linked to what-Tom-sounds like and what-Tom-looks like that one cannot identify which pattern is which. What-Tom-looks like is dynamic over time, and so is 'what-Tom-is' (in the eyes of the AM, as far as an AM can be considered to have eyes).

2.10. Memory Terms & Forgetting

An ACT memory as described in the previous subsections is definitely a memory in the sense that it 'stores stuff', be that under the form of patterns.

2.10.1. Short Term and Long Term Memory

There is a not unreasonable analogy with human short term memory and long term memory:

- Awake patterns can be thought of as short term memory
- Stored, not awake patterns can be thought of as long term memory

A similar analogy can be made with digital computer memory: awake patterns are like bits & bytes in a computer's RAM memory, whereas stored patterns are like information stored on a hard drive; these are not fetched into 'working memory' until 'needed'.

Section 5 will revisit the latter comparison when comparing the ACT-M model with other AI architectures & approaches, and similarities (and dissimilarities) with theories of (human) Cognition will be further discussed. For now the point is that an ACT memory stores information, and is able to 'forget', as argued in the following.

2.10.2. Forgetting

Based upon the previous regarding short term and long term memory, forgetting in an ACT-M happens on 2 levels.

On a short term level, one may consider that once an awoken pattern's amplitude falls below a certain threshold, the 'thought' of the pattern is forgotten, as in no longer present in active thinking. I propose this threshold is dynamic, as it depends on what other thoughts are going on, i.e. what other patterns are awake, and how loud those are awake. Short-term forgetting is simply a pattern being 'over shouted' by too many other awake patterns. An AM may at different points in time be thinking about more or less things at the same time hence the amplitude needed to be 'heard' also depends on the number of awake patterns.

On the long term level, forgetting refers to the decay of importance of stored patterns. A 'forgotten' pattern may be 'deleted' to make room for new patterns to be stored.

The correlation mechanism actually serves as a kind of lumping mechanism: when e.g. out of repeated hearing of my name and seeing of my face general patterns such as 'what-Tom-looks-like' and 'what-Tom-sounds-like' emerge, any specific previously stored instances of the sound of my name and the image of my face are no longer needed and may hence be safely forgotten.

2.11. Drivers

The discussion now turns to the last architectural element needed for a functioning AM; we call this a Driver system.

If we manage to build an AM based upon what I have proposed above, something must drive it. It must 'want to learn' for example. I propose that in an ACT-M architecture, because of the built-in correlation and association mechanisms, such a machine cannot but learn. The drive-to-learn is baked in, in a similar way as the drive to hunt or forage for food is baked into humans genetically, out of our basic physiological needs: if we do not eat and drink we die. As Maslow (1943) pointed out humans have other needs, both physiological ones such as sex & sleep, and higher level ones such as safety, love & belonging, esteem en self-actualization.

I propose that each need has a counterpart driver, an urge to fulfill the need. And vice versa each driver creates a need. Each need can be phrased positively, e.g. I as a silicon, electrically powered AM want to live, as well as negatively, e.g. I do not want the current that feeds my systems to stop flowing. Negative phrasing of a need can be considered as a fear, as in 'I as an AM fear current blackouts'.

An AM would need similar drivers, needs & fears for several reasons. Some reasons for need of drivers, needs and fears are intrinsically AM related, for example if should need to want to exist; a

machine that achieves human-like level of thoughts only to conclude its existence is futile and to commit suicide defeats the purpose. The drive to want to exist results in a need for current and a fear of blackouts (assuming our AM is based on a substrate that is electric/electronic, which is by no means a given).

Other reasons may, surprisingly, be human. If we build an AM that can reach human-level intelligence, then assuming continuing technological progress, the capacity for thinking of such a machine may at some point surpass ours. Such a machine might be potentially dangerous: it may conclude that humans are a threat to its existence and consequently decide to exterminate human life. Ergo we want to 'design' our machine such that it is inherently friendly to humans (Yudkowsky 2001). Such an AM would have a drive to do well for humanity, a fear of hurting humans and/or humanity, and a need to support humanity. There are other examples like the proposition that a machine, in order to find a proof of the Riemann Hypothesis, would consume all particles in the universe.

How exactly a driver system should be built is not the topic of this exercise. Neither is the question as to what extent a driver system for an AM would need to be elaborated. Such details do not match Occam's razor; unless there is an absolute need for such (detailed and purpose specific) hardware, we will not consider it. For now, the conclusion is that architecturally, a driver system of some sort is needed.

It is important to distinguish between drivers and sensorium: e.g. the human 'drive' to avoid pain is not what I would position in the driver system; this is sensorially induced, as humans have pain sensors. Ditto for the need for food as the unpleasant sensation of hunger is sensorial. In a 'machine that wants to live', the 'drive' to want it to maintain current to feed it might be induced by means of a sensorially unpleasant signal when current is low.

We will return to drivers in more detail and elaborate some of the above points in the following sections but for now this concludes the functional design of our AM, as shown in diagram 3 below.

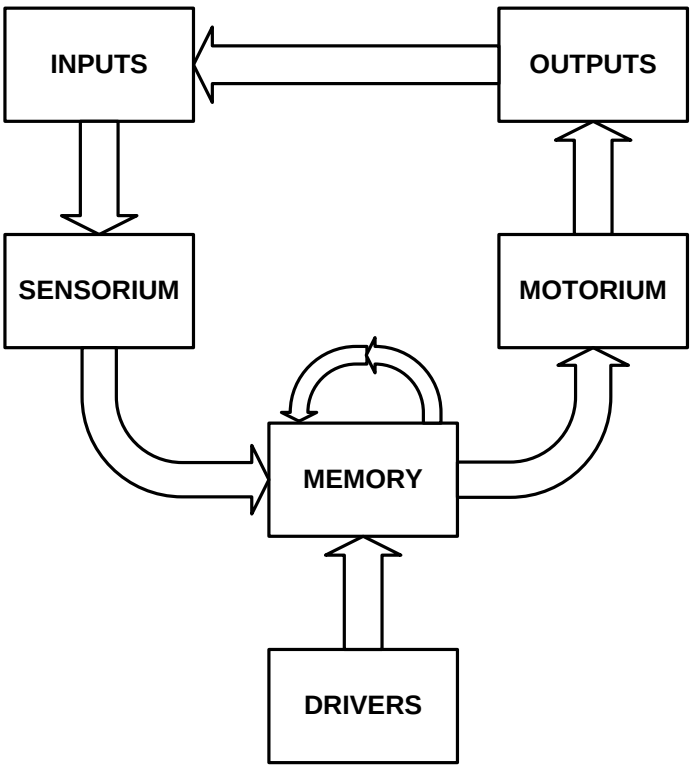

Diagram 3: Full functional design of an ACT-M based AM

Our design, from a functional, not an implementational view, includes:

- A Sensorium
- A Motorium
- An ACT memory with feedback
- A driver system

What is presented here is a model for a conscious machine mind, per our definitions in section 1. This is a functional model, centered on the concept of ACT memories. Other functional designs are obviously not excluded, and other definitions as a starting point are possible. This paper however will continue with a discussion about ACT-M based AMs in the next section.

3. Discussion of the ACT-M Model

In this section the ACT-M model as such is discussed. It will describe how in an ACT-M machine a Mind emerges, how it learns, how it thinks and whether it can be tested. In terms of testing the Turing test will be addressed as well. Comparisons of the model to other AI models and a discussion of philosophical issues is deferred to section 5.

3.1. Built for Mind to Emerge

If we manage to build an ACT-M based machine with a Mind, the mind simply will not be there once the machine is built and turned on. The Mind is the process which will 'run' on the machine. Bootstrapping the Mind comes out of the machine's (or the Mind's) experiences. A well designed and properly built ACT-M machine should be such that, once appropriately fed with stimuli, cannot but develop a Mind. I propose this happens in very similar ways that a human being develops a Mind; a well-constructed human being (barring genesis defects) subjected to the stimuli in our world, cannot be develop into a conscious, thinking human. This idea, for humans, will be revisited and expanded upon in section 4.

3.1.1. Bootstrapping

In this subsection the bootstrapping process is considered in more detail. Section 2 defined an ACT-M as a memory in which patterns are stored. When a blank machine (a machine which has not developed a Mind yet) is first subjected to stimuli, these do not a priori 'make sense' yet. In other words, the machine first needs to make some sense out of pattern soup (as was already hinted to in section 2.8. During bootstrapping, the machine would not even know what a pattern is, in the sense that it does not know where one pattern starts and ends. For example, suppose we provide the machine audiovisual input, in terms of images and language. There is no way the machine a priori knows that in a stream of 'blah blah John blah blah' and 'ladida john blahdiblah' there is something that 'sounds-like-John' – it takes time (learning) to discover that the 'john' in the pattern soup is a 'special' pattern that is e.g. always associated with John's face. The built-in correlation mechanisms in the ACT memory will figure this out, eventually, and the process of figuring this out is part of the bootstrapping process. This works in a very similar way as the development of the Self in human babies. Further discussion as to similarities and dissimilarities between Human Minds and Machine Minds is deferred to Section 5. More about learning follows in section 3.2.

3.1.2. Floating Man Thought Experiment

Avicenna, a late 10th century Persian polymath and philosopher proposed the 'Floating Man' thought experiment. Avicenna argued that a human, suspended in air and cut off from all sensoria input (including sensory awareness of the body) would still be self-conscious. We can apply this thought experiment to an AM. However, the answer would not be trivial. One might argue that a thinking AM, cut off from all sensorial input, would indeed remain conscious; it would remain thinking by virtue of the feedback cycles in the ACT Memory. It would be able to reflect on everything it knows, and even on its state of lacking sensorial input. Such a state may even be desirable from time to time; just like humans who need time to think prefer to lock themselves away so they cannot be disturbed. Humans invariably get disturbed by their primal needs such as hunger, thirst and need for sleep. An

AM would not necessarily have those needs and would ergo be able to concentrate better than a human.

However, per the above, if we boot a fresh AM, and deprive it from all sensorial input, it has been argued that a Mind will not emerge. Ergo, for AMs the Floating Man thought experiment would only work if the consciousness is already there.

3.2. The Machine Mind/Body Problem: Minds vs Substrates

At this point one may wonder what the thinking machine really is; if the 'hardware' as such, when first booted, is not a thinking conscious machine, then what is the 'consciousness' and how does that relate to the 'hardware'? This is quite similar to the human philosophical question known as the Mind Body Problem.

To clarify this discussion I propose to make a difference between the 'hardware' and the process as follows:

- Substrate: the 'grey matter' (on) which runs the machine. I prefer not to use the word 'hardware' because the substrate need not be 'hardware' in the computer (or plumbing) sense of the word. The substrate could be electric, electronic, quantumphysical, biological or other, more exotic and perhaps not yet discovered materials.
- Mind: refers to the thinking, self-aware process that runs on the substrate.

What is called a Substrate here is the 'brain' of the machine, similar to the way our brain as (merely) a collection of neurons. The substrate does not think; it is merely a substrate on which the thinking happens. Thinking is the sum of the dynamics going on on he substrate, bootstrapped by sensorial input, and equally kept 'running' by means of the feedback structure in the ACT-M. The memories, finally, is everything the AM is not currently thinking about, and those are stored in the ACT-M, in or on the substrate. The 'body' of the machine, finally, consists of the substrate, the sensorium and the motorium.

3.3. Mind as a Process

In the model proposed in section 2.9 the Mind is a chaotic process and this process runs on a substrate. In this section the idea of Mind as a process is further discussed.

Consider what a process is; take for example the business process of ordering goods, subsequently receiving the goods and finally processing invoices and payments. The process itself is not something tangible; one cannot 'see' or 'touch' this process. The process does result in tangible artifacts, such as purchase orders, invoices, and/or changes to artifacts e.g. physical delivery of goods etc. Admittedly, such a process can be formulated as a simple algorithm, or a state machine.

Now consider a chaotic dynamic process, such as our weather. Our weather has innumerable inputs; every flap of a butterfly wing is an input (and basically, everything that moves air molecules is an input). The process has outputs, e.g. thunder or rain. Again the resulting effects of the process, thunder or rain can be seen or felt, but the process itself cannot. However, the 'processing' of the inputs can be modeled, and the outputs can be predicted to some extent. Such prediction is imperfect because of the physical inability to take all possible inputs into account. It is predictable to some extent (e.g. short term weather forecasts and long term seasonal, repeatable (but not set in concrete) patterns). In the case of a weather system, the mechanism of modeling is partial differential equations, which describe how local variations/inputs affect the whole of the system. Such a system is dispredictable. The weather can be simulated on a digital computer. But such a simulation is a closed system; the simulation runs only in the/on the computer. "When it rains in a computer simulation, nobody actually gets wet". (Attributed to David Gelernter, 2006). And because a computer model cannot take every possible flap of a butterfly wing into account, the simulation does not behave like the original it simulates. The simulation in a weather forecasting computer and the actual weather on our planet are 2 different dynamic, chaotic systems. They are similar in dynamics but different in instantiation. Rain in a simulation of the weather is just as real as rain in London, but only in the sense that rain in a simulation at a location representing London is real *in that simulation*, but not in physical London. Physical rain in London is real *in our reality*, but not in the simulation. Computer models of weather do continuously receive input (of actual measurements of air pressure, temperature, humidity), but the simulations have no effect on the physical world whatsoever.

If we consider an AM as a process similar to a weather system, then the way of modeling this is similar to modeling of a weather system: we describe the (local) impact of inputs. We did this in section 2 (functionally) by describing how awakened patterns wake others, how correlations link memorized patterns together and so forth. Such a dynamic process can in principle be run or simulated in a digital computer, or run on dedicated analog hardware. There is a difference between simulations of weather in that in such a case, the system *does* affect the world it operates in: an AM has a motorium and is able to influence the world.

There is one case to consider, and that is the case where, for safety reasons, an AM is set free in a simulated world (see e.g. Yudkowsky (2001) for the concept of "Friendly AI"). Even in such a scenario, the AM would be able to influence the world it operates in, be it that such a world would be a simulation.

The bottom line of the reasoning developed in this section is: if we can define or describe the (local) operations (call them algorithms if you wish) that steer the dynamics of an AM, an AM can conceivably be built, and in such an AM the actual emerging process, driven by such operations/algorithms, constitutes thinking.

3.4. Learning

In this subsection the concept of learning is addressed. In general in the field of machine learning, 3 approaches can be distinguished[4]:

- Reinforcement learning: the machine interacts with its environment by producing actions; these affect the state of the environment, which in turn results in the machine receiving a reward or a punishment. The machine subsequently changes its behavior to act in such a way that it maximizes future rewards and/or minimizes future punishment.
- Supervised learning: the machine is given a set of sensorial inputs and it told the desired action is should take (the output). The machine learns to match outputs to inputs and is subsequentially expected that once the machine is presented a previously unseen input, it will produce an acceptable output.
- Unsupervised learning: the machine simply receives sensorial input, but is neither rewarded/punished and is neither provided with correct behavior to strive towards.

In an AM, these 3 kinds of learning come into play. We will address these in order of increasing complexity.

3.4.1. Unsupervised Learning

Unsupervised learning is what happens because of the correlation/association mechanisms in the underlying ACT memory. The ACT memory synthesizes 'structure' out of its sensorial inputs. The driver for this kind of learning is baked in by means of the correlation and association functionalities in the ACT Memory substrate.

3.4.2. Reinforcement Learning

Reinforcement learning will be needed, even if only to make sure we humans can steer an AM towards friendliness: every act towards unfriendliness will need to be discouraged. For this we need a reward/punishment system. It is however hard to imagine what would punish or reward an AM, given the minimalistic approach towards design taken in our approach. For this purpose I propose the following: every time an AM behaves undesirably, we 'shake it' in the sense that we randomly destroy or disturb its ACT-M. Larger 'mistakes' lead to more intrusive 'shaking'. Such a mechanism has similarities to simulated annealing (Kirkpatrick et al 1983).[5]

---

[4] There is one more form of machine learning, where machines are used to train one another; for the discussion here this form can be loosely covered by the other 3 kinds of learning.

[5] Simulated Annealing is an optimization method for finding the global optimum (or a good approximation thereof) of any function in a large or high-dimensional search space. The name originates from

One might speculate to what extent this would be a 'punishment' for an AM or to what extent an AM might experience this as a punishment. What is certain is that an AM would feel – in the sense of notice – the disturbance. It might realize that it now experiences or thinks differently. It might object, like humans might object to brainwashing. If we shake the AM too hard we might end up with the machine equivalent of shaken baby syndrome and end up with a totally dysfunctional AM. And finally, the AM might 'feel' the disruption and actually enjoy it, leading to more behavior that would lead to shaking, ultimately leading to self-destruction.

However, apart from 'shaking', I currently see no other alternative, and if successful, this will allow for reinforcement learning.

3.4.3. Supervised Learning

Supervised learning is, for an ACT-M based AM, a hybrid between unsupervised and reinforcement learning. Correct response to given inputs implies correlation between desired response and input and this is covered by the correlation mechanisms in the ACT-M architecture. If an AM needs prodding to accept that the desired response to a stimulus is the way to go, reinforcement (shaking) can be applied.

3.4.4. Relationship between Learning & Drivers

The need for a driver system was discussed in section 2.11. We now return to drivers, as drivers co-determine how and what humans learn, and this might provide insights in how an AM might learn. We discussed that humans have various drivers. Some are built-in, perhaps genetically, such as a particular talent for music or languages. Other drivers are cultural, or rather, are learned. One human might learn that being rich is desirable, whereas another might learn that being rich is anti-social. The final desire for richness may be a combination of genetic predisposition (need for power, e.g.) and learning. This contrasts with the only driver we have identified for AMs, namely the innate ability to correlate and associate.

I now propose the following: for most human drivers, whether a drive is innate by architecture, built in the substrate, or genetically determined (which is also built in the substrate, be it instance-dependent, not a cross-instance architectural feature), or learned, does not make much of a difference, because in the end, a drive, just like a learned aspect results in an experience. Some experiences are

---

annealing in metallurgy, which is a technique for heating and subsequent controlled cooling of a material. During simulated annealing, 'temperature' is slowly brought down, and as the system 'cools', disruptions to the system lessen. Roughly speaking the basic premise is that with infinitely slow annealing any function will statistically speaking reach its optimum.

clearly attributed to sensoria, and other humans some experiences are not. For humans this works as follows:

- Food is a basic human need. When we lack food we feel hungry; feeling hungry is an unpleasant experience. When we fill our stomachs, we feel satisfied. This is a pleasant experience. The same goes for thirst, cold and so forth. These experiences are clearly linked to primary sensory experiences: humans are equipped with sensor for hunger, thirst and cold.
- We are taught that certain behavior is not acceptable; we are punished for such behavior. The punishment is an unpleasant experience. Similarly we are taught what behavior is desirable; such experience is rewarded. A reward is a positive or pleasant experience.
- If we are taught that certain behavior is not acceptable (by reinforcement or not), then this behavior makes us feel bad – again this is an experience. The counterpart may also be true: if we violate a rule we might actually feel good, but in that case a) this feeling good is an experience and b) feeling good for breaking a rule probably indicates that one does not agree with the rule in the first place.

The point is that what is not 'learned' by innate or genetically covered drivers can be covered by learning and vice versa. Reason is that drivers lead to experiences (pleasant or not) and 'teaching' (by reinforcement) leads to experiences. This we will call the triangle of drivers, training and experience.

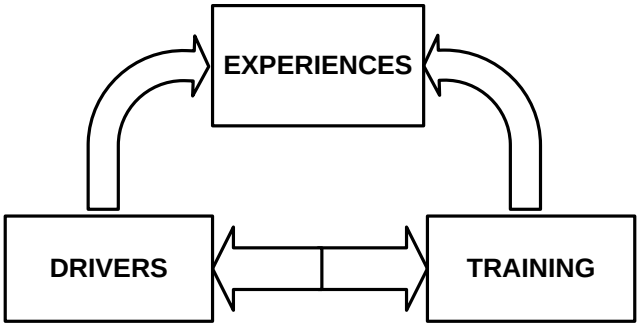

Diagram 4: Triangle of Drivers, Training and Experiences

Everything the AM learns, it learns because of its experiences. The correlation processes run per sensorial inputs; and input equals, as in leads to, experience. Unsupervised learning happens due to the stimuli the AM received, which are experience. Reinforcement learning implies punishment (and perhaps reward) which again are experienced. In that respect the AM is ultimately the sum of all experiences. Finally, not all experiences have the same importance and frequency; frequency being

relative to importance, hence we can argue that experiences are weighted by importance and frequency. As a result, I propose that an AM equals the weighted sum of its experiences.

I did mention that certain human experiences cannot be traced to basic sensors; this we will address in section 4.5.

3.5. Relation between Mind and Substrate

3.5.1. Substrate co-determines the Mind

Under the above logic the substrate is not the Mind, but the substrate is a determining factor to the Mind. This is necessarily so because the physical capabilities of the substrate are a determining factor in the Mind's experience. An AM which has infrared vision experiences the world differently than an AM which lacks infrared vision, and since the AM is the weighted sum of all experiences, infrared vision is a factor in shaping the Mind.

3.5.2. AM Levels of Cognition & Substrates

I propose that an AM can be described on 4 levels, increasing in (artificial) cognitive complexity:

- A substrate level, on which the AM runs. The substrate includes the sensoria & motoria. Nothing cognitive is going on here.
- A process definition, such as a set of partial differential equations, that define the dynamics of the system. One may consider this an algorithm, but it is not an algorithm that processes mental symbols. If the substrate is a digital computer, this algorithm may be phrased as a computer program.
- A cognitive level, which emerges from the dynamics of the system.
- Several instantiations of "computation" performed by the cognitive level.

The substrate itself may be electronic, digital, analog, biological or something else. The process definition may be coded in the substrate; this would be the case in a biological or analog substrate. Suppose we can model the sort of dynamics performed by such a biological of analog substrate. Then nothing prevents us from running the same dynamics by means of a computer program on a digital computer. In that case we consider the computer program part of the substrate: the computer program does not think, it is merely a substrate on which the dynamics of the AM thinking is executed. Bullet #3 follows from the definition of an ACT-M as described in section 2.

The fourth bullet point merits some explanation: by instances of 'computation' we imply the ability to follow algorithms (or recipes). These also include Bayesian reasoning or processes akin to fuzzy logic. These also include the ability to "think in symbols" (and then perform Boolean logic

'computations' on it). These also include the ability to perform mathematics (like a calculator can). In other words, an ACT-M is not a symbol processor by design, but as the Mind emerges that Mind should be able to process symbols. Similarly, and ACT-M based AM should be able to perform Bayesian reasoning. It should even be (or become) smart enough to come up with the 'rules' of Bayesian computation. This leads to a rather unexpected conclusion: suppose the substrate is a digital computer (which is able to manipulate symbols), and ACT-M based AM based upon that substrate should be able to (learn to) perform symbol manipulation, but not using the circuitry purpose built to manipulate symbols, but by means of the dynamic process that constitute its thinking (in combination with the memories).

3.5.3. Substrate Replacement Thought Experiment

Furthermore, the Mind and the Substrate are irrevocably linked. The thinking of an AM is the process running on the substrate, but the thinking relies heavily upon the memories stores in the ACT memory, which is part of the substrate. One cannot simply transfer a thinking process to a different substrate, unless that substrate has exactly the same memories as the original substrate. Which is, for all practical purposes, only possible if that substrate was the one of an AM identical to the original, and which has been presented to exactly the same experiences as the original.

Now consider the following thought experiment (see e.g. Chalmers 1994): suppose we replace very component in the substrate, part by part, by exactly the same component as the original, and suppose this could be done without interrupting the thinking going on. Would, once every part has been replaced, the AM still be the same as the original? The answer to this question would be 'yes' – the Mind – as a whole of the ongoing processes and the stores memories – would be identical. However, the key lies in the premise that the ongoing thoughts should not be disrupted. Any disruption (e.g. shutting down the 'machine') would be an interruption, which would be an experience, which would alter the AM.

Even further, we argue that 'shutting down' an AM irrevocably alters it, for 2 reasons:

- It is hard to conceive that shutting down an AM would be atomical; any thoughts it was thinking would be lost; how an AM thinks upon wake up, would depend on the actual waking-up process, and would inevitably be steered dramatically by the inputs it receives upon waking up.
- Furthermore, since everything the AM was thinking about influences (and hence) changes the AM itself (in the sense that during thinking patterns are re-correlated), any disrupted thought will not 'complete' and its effect will be lost. Since the AM is a chaotic system, any loss of 'effect' no matter how minor may lead to a potentially large effect in end result, meaning it may lead to a very different AM.

There is another potential effect of in-Mind substrate replacement: suppose our AM runs on a biological substrate that has a finite life expectancy. The AM knows its life span is limited because of the limits of its substrate. Knowing this or learning this is an experience which influences the AMs thinking. Suppose we now replace every living cell by a functional electronic equivalent, with infinite life span, without interrupting the thought processes. Suppose the AM is not told about the modification (the full replacement) of its brain – in that case our original premise holds – the AM would be the same as before. Suppose now we tell the AM its substrate was replaces, and it now has an infinite lifespan (barring power outages). This knowledge, the experience of this knowledge would alter the AM and it would no longer be the same. Such knowledge would actually alter the AM's look on life quite substantially, and hence alter the AM quite substantially.

3.6. Trains of Thought and Language of Thought

We established earlier that an AM is not built with an innate capability of (a particular human) language; however, it is designed such that the ability to interpret human language, and output human language (whether through a speech synthesizer or a teletype-like functionality) cannot but emerge, given the appropriate stimuli. The question addressed in this subsection is in what language an AM would think, in other words, what would be its Language of Thought (LOT) (Fodor 1975), or more radically, would it *have* a LOT? And if so, would it have (or perceive) 'Trains of Thought' or 'Streams of Consciousness'?

3.6.1. Train of Thought vs Jungle of Thought

First of all, I propose that an AM would not have *a* 'Train of Thought' or TOT. As we established in section 2, thought is not a linear phenomenon in an AM: every thought in the sense of an awakened pattern causes a multitude of other patterns to awake; hence instead of a linear awakening of patterns, we get an explosion of thoughts; rather than referring to a Train, it might be more appropriate to refer to this phenomenon as a Jungle of Thought.[6]

Consider the observation by an AM (over time) : John eats an apple.

[6] In earlier drafts of this text the author considered using the term 'Trees of Thought': every thought triggers 'child' thoughts, and each child thought triggers its own children, giving rise to a kind of tree structure. This has proved a poor metaphor, as the child of a thought reinforces the original thought; if one is correlated to the other then the other is correlated to the one. A similar kind of feedback happens with the grandchildren, which by correlation though their parents, reinforces the original thought again. Ergo, the term 'tree' is not appropriate; it is too hierarchical, and the mechanisms proposed here are all but hierarchical.

- John triggers lots of things about this person. Where the AM met him. What his wife's name is. What car he drives. How he has problems with his weight. How ill he was the last couple of days.
- Each of these things triggers other thoughts, all in parallel. E.g. 'ill' triggers the fact that when one is ill one's appetite is affected.
- At the same time the AM has the thought 'eats'. It knows it is John who is eating. 'Eating' correlated with 'weight' but also with the fact that ill people have little appetite. Of all the things the AM thinks about when the thought John comes about (the jungle of thoughts caused by thinking about John), the 'eats' makes it likely that the jungle about John's health and his weight have higher priority (more chance of staying 'alive') than jungles that branch in the direction of his wife, his car or the circumstances when the AM first met John.

An AM might however *perceive* a TOT; the TOT as perceived by the AM is then simply the linear chain of thoughts that are most active, have the highest amplitude.

3.6.2. Language of Thought

The next question is then whether a perceived (by the AM) TOT would be in an actual language or not.

Imagine an AM that speaks and reads both English and French. Suppose it utters the word CHAIR. There is a pattern CHAIR, it is the motor pattern for the AM to utter or output the word; a pattern of controlling a voice synthesizer for example. The AM need not always pronounce it exactly the same way, it can speak slow, fast, use an accent and so forth. Hence the motor pattern for outputting CHAIR is not fixed, it is 'fuzzy'[7]. It is a 'group of motor patterns such that the resulting spoken word can be reasonably recognized by a (human) listener as 'CHAIR'.

When the AM outputs the word CHAIR, it must be that for some reason it is thinking about what computationalists might call the 'symbol spoken CHAIR'. (a discussion about computationalism and its relationship to the theory proposed here is postponed section 5). The 'symbol' CHAIR is the pattern in the AM's thinking in its LOT (assuming for now there is such a thing as LOT). The 'symbol spoken' CHAIR in that LOT refers only to the AM thinking about the sound of the result of the motor pattern CHAIR. However the pattern 'symbol spoken CHAIR' is *not* the pattern for the concept CHAIR. The reason for that is simple and is elaborated below.

When the AM outputs the word CHAIZE (French for chair) there is another motor pattern.. It is

[7] 'Fuzzy here does not refer to fuzzy logic, which is a kind of probabilistic logic that deals with reasoning that is approximate rather than exact. The concept 'fuzzy' in the context of this paper is discussed further in section 5.

‘fuzzy’. That pattern is linked to the pattern ‘symbol spoken CHAIZE’. The ‘symbol’ CHAIZE is the counterpart of ‘symbol spoken CHAIR’ but in a different LOT. The AM can think in English and it can think in French. Both fuzzy patterns are linked to a pattern (another ‘symbol’) that represents the concept of ‘chair’. Actually, this pattern is not a fixed pattern either - as the concept chair may be different things, depending on the context in which the AM think about chairs. So again, this ‘symbol’ is a whole bunch of patterns, some that refer to a thing with 4 legs, some that refer to a think with wheels, some may refer to a big stone on which one might sit. Ergo, the ‘symbol concept chair’ is not a well-defined thing. It is fuzzy.

Now suppose the reason for uttering CHAIR is the fact that the AM perceives me to be tired. It may conclude I might need a chair. This reasoning is again a pattern. Not a fixed pattern, because it may refer to knowledge about me just returning from a long walk in the woods, or from an excruciating day at the Mall. On top of that, if I happen to be French the AM won’t ask ‘Do you need a chair’ but rather ‘Veux-tu une chaise?’ However the thinking, the reasoning, which started perhaps with a sensory pattern of me being tired would be the same. By the time it resulted, the same thinking might have resulted in CHAIR or CHAIZE or even both. Ergo there is thinking going on that is in a sense pre-language in the sense of public languages known by humans (and AMs).

Suppose now an AM does create conceptual, atomic symbolic representations of e.g. the concept ‘chair’, we may consider the body of all such representations the Language of Mind (LOM). This could be considered the ‘base language’ in which the system thinks. This is a language we cannot decode, if only because it is ever changing. But really, whether it exists or not is irrelevant (be it interesting) because whatever ‘thought’ in LOM will, if significant enough as explained in the subsection on TOT, automatically and by association, also be thought in a public language. Which of those languages we define as ‘the language of thought’ is not relevant and a matter of definition.

The AM then, may *perceive* the equivalent of a voice in its head, talking in a particular public language. This would be quite logical and expected, since the thinking of any word in a public language will awaken the (fuzzy) pattern of the sound of the word thought of, as we established earlier.

3.7. Testability

One of our functional requirements for an AM was testability. Let us review the requirements that need to be tested. The 2 first requirements (learning from experience and ability to think) are relatively easy to verify. The third requirement, “The machine is self-conscious, meaning it is aware of itself in the sense that it knows what it is thinking, and knows that it is itself that is doing the thinking.” is harder to test. In this section we address the possibility of testability.

3.7.1. The Turing Test

Although passing the Turing Test was not a functional requirement for AMs in the early sections of this paper, it is still an interesting question to consider. I propose that it need not, or rather that that Turing Test is not the appropriate mechanism for testing whether an AM is aware of itself in the sense that it knows what it is thinking, and know that it is itself that is doing the thinking.

The test is described as follows: a human judge engages in a natural language conversation with a human and a machine. All participants are separated from one another by means of a screen and limited to on-screen conversation. If the judge cannot reliably tell the machine from the human, the machine is considered to have passed the test, considering how closely the answer resembles typical human answers.

A full discussion of how an AM might interact with a human in such a test setting is beyond the scope of this paper, but let us briefly consider the following:

- When asked the question "Are you human?", what would one expect the AM to answer? An honest answer would be no, but then the AM would fail the test.
- As a matter of game-playing it could lie, but then it could lie about just about anything.
- As a matter of friendliness (in the Yudkowsky (2001) sense) one may not want an AM to be capable of lying.

An AM will never be human, if only for the fact that it lacks a human body and the basic drivers that result from having that body. Aeronautical engineering texts do not define the goal of their field as 'making machines that fly so exactly like pigeons that they can fool other pigeons'. (Russel & Norvig 2003). As such, then, the Turing Test ignores the fact that computer intelligence may be fundamentally different than human intelligence. The perspective an AM would have – could have – of the world would have to be different from a human's perspective. Hence I propose that a clever interrogator can always come up with questions that will cause the AM to fail the test, from which I conclude the Turing Test is not an appropriate method of testing Criterion 3.

Variations to the Turing Test exist. For example, Harnad (1989, 1990) has proposed the Total Turing Test, which adds two further requirements to the traditional Turing test. The interrogator can also test the perceptual abilities of the subject and the subject's ability to manipulate objects. More specifically this requires computer vision and robotics. Both requirements can be fulfilled by an AM's sensoria and motoria. However, merely adding those requirements does not address or alleviate the fundamental issue with the Turing test outlined above. We will return to the TTT in the discussion in section 5.

For the sake of completeness it should be noted that there exist various other arguments against the Turing test, the interested reader is referred to the Stanford Encyclopedia of Philosophy (http://www.science.uva.nl/~seop/entries/turing-test/) for a fairly complete overview.

3.7.2. AM as a P-Zombie

If the (Total) Turing test is not the appropriate method of testing Criterion #3, then what might be? Consider the following: how do I know that any other human is indeed human, and has a consciousness? How can I be sure that you aware of yourself in the sense that you know what you are thinking, and know that it is you that is doing the thinking? A solipsist may doubt this and conclude that she herself is the only conscious being in the universe. The only answer to that is that I would conclude that from what you are saying, knowing that you may be lying. For all I know you could be a P-Zombie: some agent that is indistinguishable from a normal human being except in that it lacks conscious experience, qualia or sentience. Note that – as will be the case for many more topics we will touch upon in the remainder of this paper - there is no philosophical agreement on whether P-Zombies are (logically) possible (see e.g. Chalmers (1996) for arguments pro logical possibility and Dennett (1991) for arguments contra).

Irrespective of whether P-zombies are possible, I propose that the only way to test that an AM is conscious in the sense of Criterion 3, is something akin to a Duck Test: if the AM behaves like a conscious agent, *taken its differences from a human agent into account*, one may conclude it fulfills Criterion 3. It would be for me to conclude it is, based on what it tells me. This is subjective: upon the same conversation with an AM I may conclude an AM fulfills Criterion 3 whereas you disagree.

Furthermore, even if I, or all of humanity, were to agree that an AM is conscious per Criterion 3, it might still be a kind of P-Zombie: a Machine-P-zombie, which is an agent that, apart from its physical substrate and lack of human drivers, is indistinguishable from a human mind but lack real conscious experience, qualia or sentience. (that is, assuming for the sake of purpose of the previous line that Machine-P-zombies were possible). I argue that the terms 'real' conscious experience, qualia and sentience in the above cannot be interpreted in the human sense, since an AM simply is not human. An AM cannot describe what the color blue feels like to a human in human terms; and a human can never feel what it is like to be an AM (and vice versa for that matter) (cf Nagel (1973) on what it feels like to be a bat).

To wrap up this section, there is one remaining item to define, and that is what I mean by the term 'human-like' consciousness. From the discussion above it follows that accepting that an AM satisfies Criterion 3 implies that the behavior of the AM is human-like. From the above argumentation that an AM simply *is* not human, we cannot expect it to behave as if it were a human. Human-likeness for an AM is human-like in the sense that you are a human like me, but an AM can only be human-like to

the extent not affected by the differences between the AM and humans. Those differences are attributed to the physical substrate it runs on, mainly its motoria and sensoria, and to the difference in underlying driver system. Furthermore, and unfortunately, human-likeness is a subjective concept.

We will return to the subject of testability of consciousness in the discussion in section 5 of this paper.

3.7.3. AMs and Qualia

Would an ACT-M based AM have qualia? (assuming for the sake of argument that qualia exist). I would propose to answer yes to this question, be it that those would not be identical to human qualia. The argument goes as follows: consider what it feels like to observe a green cucumber. Our AM has seen cucumbers before, so in principle every instance in the past of 'looking at a green cucumber' is awakened. Most of these memories are probably pretty vague and not all that important, but they are there. They are subjective in the sense that these are unique to our AM (a different AM will have different recollections of different green cucumbers). Not only the "observing the cucumber" is awakened, but by association also the circumstances in which these were observed. Not only those experiences are awakened; everything related to green cucumbers (and everything related to green without cucumbers and everything related to cucumbers without green - be it that those are awakened less 'loud') is awakened, including perhaps the anecdote our AM read in yesterday's paper about the incident of the fellow who was treated by emergency services for inserting a green cucumber in an orifice not intended for such vegetable. I propose that all these 'recollections', each with their relative importance, and many of which by no means 'loud enough' to cause something like a conscious thought about them (conscious here to be interpreted as the equivalent of 'voice in head in a human language') are part of "what it feels like seeing a green cucumber" for our AM. What becomes conscious (in the above sense) and when is largely determined by the context. For example: if the context the AM is in at the time of observing the cucumber one is where funny anecdotes are being shared, then "seeing a green cucumber" leads to consciously recalling the cucumber-in-orifice incident.

Suppose an AM is colorblind and has been since inception (Jackson 1982). It knows about color, wavelengths and the fact that sky is blue and (some) roses are red. The AM has a "feeling about what color is" in the same sense as there was something that it was like (for the AM) to observe a green cucumber in the above example. Suppose we now add (or switch on) color-vision. I would argue that this *does* change things for the AM. And this for 2 reasons:

1) Since 'being' and AM and 'thinking' as an AM is a trajectory in state space, flipping on the color switch is an 'event' which now causes input into that system which was previously not there, so this event changes these trajectories in state space (just like any experience does, and any thinking does).

2) I would argue (for an AM) that once the color switch is flipped, the sensation of color input becomes 'color-soup' which takes time for it to make sense. There is suddenly stuff there that was not there before, and there is no reason to believe all this stuff automatically makes sense; it would take time for the self-organizing systems to self-organize as to make sense of the color information.

One may of course argue whether the term "feel like" seeing a cucumber is appropriate for a machine as a machine cannot "feel" like a human does, even if only because its sensorium is different.

One may also argue as to at what point a recollection (an awakening of a previously stored pattern) is "conscious". One might consider the actual awakening as sufficient, one may insist the awakening at least results in the awakening of a 'description of what is awakened in a LOT (or the illusion thereof; whether testable or not), or that is results in conscious (in the LOT sense of the word) realization that the AM feels what it feels like to feel what it is like to see a green cucumber.

4. ACT-M as a Human Model of Cognition

In sections 2 and 3 I proposed a model for machine cognition: ACT-M based AMs are thinking, self-aware agents within this model. In this section I propose, as a thought experiment, to apply the same model to human cognition, and to investigate where this would lead us. The premise is simple and will be analyzed in the remainder of this section:

- human cognition is the sum of its experiences
- the human mind is a chaotic dynamic system
- this dynamic system emerges upon the substrate of a human brain

4.1. Differences between AM & HM

Obviously humans are not machines such as ACT-M based AMs; although one might consider the human brain a biochemical machine, there are differences between humans and AMs.

- Humans have a particular hardware, which is fixed and which has limitations that a machine body. The lifespan of a machine may be (or become) significantly higher than that of a human. An AM might be equipped with sensoria that humans simply do not have (IR vision for example).
- Humans have genetically coded drivers. Some of those follow from the limitations of our bodies: the need for food, drink and sleep. Others have been encoded by ages of evolution, such as the urge to procreate. One human might have a talent for music or mathematics pre-coded in their substrates. As discussed in section 3, in our current model AMs largely lack such specific drivers and predispositions. This has certain consequences which will be discussed in section 4.5

- An AM might be bootstrapped on a substrate of fixed capacity; a human mind emerges on a growing substrate.

4.2. The Emerging Mind

The last bullet point in the previous paragraph deserves some more discussion. If an AM can be built, it can be booted on a machine with a certain capacity for memory, association and correlation. This capacity needs to exceed a certain threshold for AM to be able to emerge. When the machine is turned on, it starts with a blank slate, an 'empty substrate', on which AM may emerge given appropriate stimuli.

For humans this is different. After conception, a human body starts to grow, including the development of the brain. If is fair to suppose that immediately upon conception, the capacity for a foetus to think is simply not there yet. As the brain develops, at some point the human brain has to be come cognition-capable. Kurzweil (2012) presents this in layman's terms: "The natal brain is a distinctly human brain with a human neocortex by the time it reaches the third trimester of pregnancy". Hence it is subject to experiences, and per our model, these experiences all contribute to shaping the mind. Ergo humans are not born with a clean slate, an "empty brain'. The shaping of the mind, by correlation and association starts prior to birth. The genesis of human mind goes hand in hand with the genesis of the human substrate for mind. And as was the case for AMs, the exact moment one may consider the Mind to have been emerged may not be clear or known, but barring defects in genesis, once conceived, a human mind will emerge. Birth itself is, in the whole of this genesis, bus a moment (be it, sensorially-wise, perhaps an influential one, and body wise, a critical one as the lungs need to start breathing). Per this model, even identical twins, who share exactly the same genetic markup, are not born identical, since surely their experiences, even as early as in the womb, were not absolutely identical.

4.3. Human Cognition as a Chaotic Dynamic System

Under the model proposed here, cognition, including human cognition, is a chaotic dynamic system, in 2 respects:

- The process of thinking is a chaotic dynamic system in which memories are awakened by either external stimuli registered by the agent's sensors, or by means of associative functional connections that awaken memories related to those already awake. In this sense, cognition or thinking is a trajectory over time in a high dimensional state space; the state space consisting of all that is thinkable.
- Furthermore, the process of correlating patterns and associating patterns with other patterns is a complex dynamic system as well. If one considers the whole of what an agent knows, its

experiences, its beliefs as the essence of what the Mind is, the Self of the Mind, then this essence itself is in constant flux as well, since every new experience invariably changes the ACT memory. In this sense, the Self is a trajectory over time in a high dimensional state space, the state space being all that is be-able.

To expand upon the last proposition: it seems reasonable to me that who I am is a changing concept. If I see a video of me moving about as a toddler, I can hardly say that it is me, the current me, that I am seeing. What I see is a body that evolved in to my current body. I can remember what I thought like when I was much younger, and the mind that thought like that then evolved into the mind I have now. Our values, principles clearly evolve over time, and they evolve due to our experiences. One's opinion about smokers and smoking may evolve from being a smoker and enjoying it at some point in life, into a fanatic anti-smoking attitude, perhaps after seeing a loved one die of lung cancer, or having to undergo a lung transplantation oneself. Every experience changes the Self, be it that the experience of yawning when being tired tonight probably has less of an impact on who & what I am than a near death experience does.

4.4 Panta Rhei

In the view presented here everything is in constant flux. This can be summarized as follows as shown in diagram 5:

- I think therefore I am (Descartes). Let's accept this for the same of argument.
- I experience sensorially therefore I change, as every sensorial experience is a(n input) pattern, which causes (changes in) memorized patterns and correlations.
- I think, by awakening stored patterns. I am also aware of what I think. Being aware of what I think is an experience.
- Hence, I think therefore I experience.
- Ergo, thinking → experiencing → changing

Within our thought experiment, the act of thinking changes the mind, and hence influences further thinking. The opposite does not hold; changing does not imply thinking; a chair does not think although it changes over time as it wears out or gets consumed by woodworms. Because I think I change more than I would have were I not to think (be it that if I were not to think I would not be what or who I am).

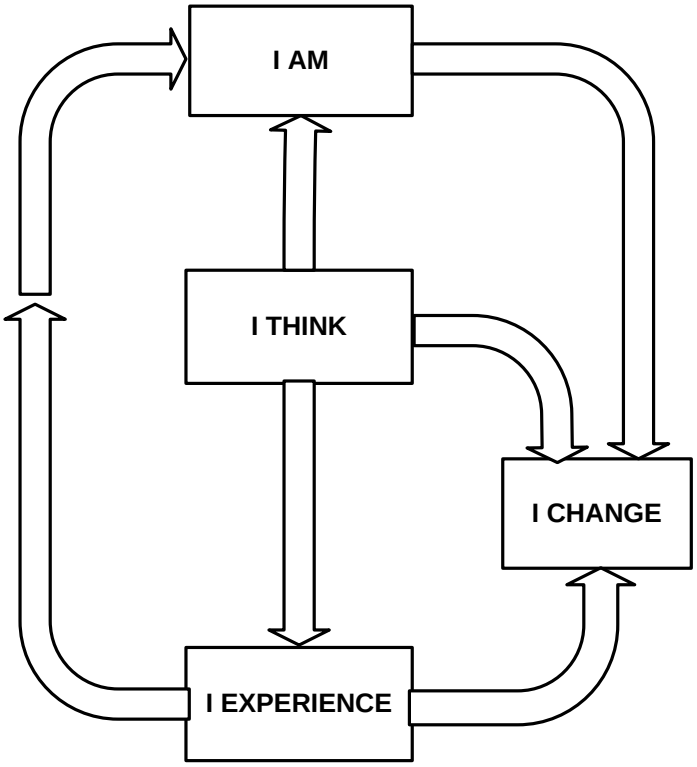

Diagram 5: Thinking, experiencing, changing, being

Phrased in the language of dynamic systems: our thinking is a dynamic process, a trajectory in state space, within the machine of being, which itself is a dynamic system too, and which itself follows a trajectory in its own meta state space. This trajectory, finally, is influenced by the trajectory of thinking. Thinking influences being.

4.5 Drivers are the Hard Part

In the above subsections we explored what the consequences would be of using the ACT-M model as a model for human cognition. As a bridge to the discussion which follows in the next section, one more issue needs to be addressed, and we already hinted to this in section 4.1: the model does not explain what human experience is, above the definition of experience in the model. It is fair to state that human experience is more, or richer, than experience defined in the model. The model delegates the mystery of why one human might enjoy Beethoven's music whereas the other might prefer Rammstein, or why one human might prefer blue over red, to the realm of drivers, without explaining why those preferences exist and how they come about.

A human might be pre-wired with a talent for music, or a preference for music (which could be translated into ' the experience of music is a positive one', and in this way would encourage more listening to or playing of music); an AM, as described here, cannot. The only explicitly built-in drivers we identified for an ACT-M based AM are the learning capacity, the capacity to detect correlations and the capacity to awaken memories by association (based on either external stimuli or

cause by its own thinking). Apart from those, and AM as proposed here does not optimize a (n explicit) 'goodness function'. As such, there is no way to design an AM such that it would have built-in talent for music or a preference for green over red. There are 2 ways to build an AM specialized in music:

- We either equip it with more and/or specialized auditory sensoria; but such hardware is 'pre-cognitive'. However, I would doubt that by adding explicit Fourier analyzers or spectrum analyzers onto an AM auditory similar to a human one would enhance the AMs talent for music; this would be like expecting a human with absolute lack of musical talent to expect te become talented when equipped with an oscilloscope.
- We focus its learning on music (rather than on, say paleontology). As for humans, learning about music, harmony, composition and so forth might enhance the AM's capability for music (as it would for a human), but that need not make music more agreeable to the subject.

Referring back to the diagram in section 3.4.4 regarding the relationship between drivers, training and experience, the bidirectional arrow between Drivers & Training should be interpreted as follows: what, in an AM, cannot be accomplished by lack of understanding of or access to the driver system needs to be covered by training. Since we cannot hardwire e.g. friendliness (as in e.g. Asimov's laws of robotics) in an AM driver system, friendliness needs to be taught.

An argument for the reasonability of this approach might be formulated as follows: the higher the capacity of learning, the less need for hardwired drivers. An insect, with its rather limited capacity for learning, largely 'knows what it needs to know' at birth; this knowledge is hard-wired. It knows, for example, how to move about. A human, given human capacity for learning, need not be born with the knowledge of how to move about, it can learn to walk (and a human is actually born before its capacity to move about has been fully developed). Given, in our thought experiment, that we might have (at some point in the future) the technological capability to build a machine powerful enough to think (within the definition of our mission statement), it is not unreasonable to envisage a machine whose capacity for (and speed of) learning would exceed human capacity; ergo I can imagine its dependency on hard-wired drivers might be less than is the case for a human.

That said, since science today does not know how subjective – in the sense of non-strictly sensorially based - preferences or dislike (or more generally, subjective experiences) in humans come about, or even why these exist, drivers are the hard part. Our thought experiment does not offer a path to answer of these questions, and as a result building a machine with e.g. an 'innate' (i.e. not learned) preference for blue over red will be impossible to build. Further discussion of this topic will be presented in section 5.1.

I proposed earlier that just like an AM, a HM is the weighted sum of its experiences. Given the above discussion I propose to rephrase this somewhat more accurately: under this model, a HM is the weighted sum of its experiences, colored by its drivers. As a further thought experiment, suppose a HM and an AM are subjected to absolutely identical (sensorial) input, the total experience for a human would be differ from the one of a machine, due to inherent preferences of the human which cannot be attributed to purely sensorial input, e.g. its preference for red over blue or its preference for Rammstein over Beethoven. The hard part of human cognition (which I delegated to the driver system) colors experience above the purely sensorial input.

5. Discussion

What remains now is a discussion of how the ACT-M model compares to (some) other models of (human and nonhuman) cognition and consciousness, as well as to currently used methods of AI.

5.1. Consciousness

In section 1 we defined as a starting point for the requirements of our AM. The requirement that it is self-conscious in the sense that "it is aware of itself in the sense that it knows what it is thinking, and knows that it is itself that is doing the thinking". There are several models and definitions of consciousness, originating in different disciplines (philosophical, psychological, cognitive, neural, or quantum mechanical); a full discussion is beyond the scope of this paper; what follows in this subsection is merely a comparison to a limited selection of such models.

5.1.1. P-, E- and S-consciousness

There is no widely accepted definition of consciousness; there is not even consensus on whether consciousness exists at all. Block (1998) has proposed a distinction between two types of consciousness: Phenomenal (P-consciousness) and Access (A-consciousness). P-consciousness is raw, subjective experience; A-consciousness is the accessibility of information in our minds, for verbal report, reasoning, and the control of behavior. (Readers should be aware Block's proposal is not universally accepted (see e.g. Dennett (2004) for a different opinion). In the ACT model design, A-consciousness is covered by the feedback loop from memory to memory; P-consciousness is not specifically covered here and will be topic of a separate paper. We mention Block's proposal because it has lead to another, different but related view on consciousness. Floridi (2005) distinguishes 3 kinds of consciousness:

- An AM is Environmental Conscious if it is not switched off and/or able to process information about, and hence to interact with, its surroundings, its features and stimuli effectively

- An AM is P-conscious if it experiences the qualitative, subjective, personal or phenomenological properties of a state it is in. This is the sense in which Nagel (1974) famously speaks of being conscious of a certain state as having the experience of "what it is like to be" in that state;
- An AM is Self Conscious (s-conscious) if the AM has a sense of, or is (introspectively) aware of its personal identity (including its knowledge that it thinks) as well as its perceptual or mental experiences (including its knowledge of what it is thinking).

Our definition in section 1 fits Floridi's concept of s-consciousness. Interestingly, Floridi proposes a test for S-consciousness; a test I propose our ACT based AMs would pass. In summary the test is a game played as follows: 3 prisoners are offered one of 5 pills, 3 of which are placebo but 2 of which turn the subject totally dumb. Prisoner A is subsequentially asked the question of which kind of pill he received. *A* cannot know which tablet he has taken *A* hears the question and is then allowed to answer. Since he has no way of knowing or inferring whether he is in a dumb state, he answers by reporting his state of ignorance. Now, *whatever A* says to communicate his state of ignorance, e.g. "Heaven knows", either a) his verbal report about his state of ignorance triggers no further reaction; or b) his verbal report about his state of ignorance triggers a counterfactual reasoning of the following kind: "had I taken the dumbing tablet I would not have been able to report orally my state of ignorance about my dumb/non-dumb state, but I have been and I know that I have been, as I have heard myself speaking and saw the guard reacting to my speaking, but this (my oral report) is possible only if I did not take the dumbing tablet. I know that I am in a non-dumb state, hence I know that I have not taken the dumbing tablet, and I know that I know all this. In case b), A passes the test. In case a) A fails the test (for full details see (Floridi 2005)). Per our requirement for an ACT based AM to it be aware of itself in the sense that it knows what it is thinking, and knows that it is itself that is doing the thinking, by design our AM will pass Floridi's challenge.

Bringsjord (2010) has argued that (in the foreseeable future) a Turing machine-like AI robot can be built that would pass Floridi's test, but that it nonetheless would lack s-consciousness and p-consciousness. Bringsjord states that one might know such robots lack these attributes by showing that s-consciousness and p-consciousness are more than computation; we address computationalism in the next subsection.

The author tends to agree with Bringsjord that Floridi's challenge is an inventive, unprecedentedly difficult challenge to (logic-based) AI, but I also side with Bringsjord on this topic that Floridi's challenge – philosophically speaking – is not the one and ultimate test. By lack of an ultimate test, we unfortunately need to revert to the subjective duck test proposed in section 3.7.

5.1.2. Global Workspace Theory

Among the various theories about human consciousness, Global Workspace Theory GWT is one of the more prominent psychological/cognitive ones. It was proposed by Baars (1988, 1997, 2002, 2005) and introduces the metaphor as a theatre. GWT be thought of as a theater of mental functioning. Consciousness in this metaphor resembles a bright spot on the stage of immediate memory, directed there by a spotlight of attention under executive guidance. Only the bright spot is conscious, while the rest of the theater is dark and unconscious. The sensory ‘‘bright spot’’ of consciousness involves a selective attention system, the ability of the theater spotlight to shine on different actors on the stage. (Baars 2005).

The ACT model presented here bears resemblance to GWT, in the sense that the awakening of patterns is similar (but not identical) to the spotlight of consciousness. However, as we have argued earlier, an awake pattern need not necessarily be equivalent to a ‘symbolic’ thought in a Language of Thought or inner speech, but may remain unconscious. Apart from this, the ACT model and GWT do not contradict one another.

5.1.3. Cartesian Materialism

The term ‘Cartesian Theater’ has been coined by Dennett (1991), if only to immediately reject the concept. Cartesian Materialism proposes a special, specific brain area (or areas) that stores the contents of conscious experience. Dennet argues that it is impossible to precisely determine when something enters conscious human experience, based on experiments which reveal timing anomalies in conscious human experience, hence Cartesian Materialism must be false. Experiments Dennett refers to are e.g. Libet’s W-time (Libet et al 1983) , which refers to an experiment which shows that that the unconscious brain activity leading up to the conscious decision by the subject to flick his or her wrist began approximately half a second before the subject consciously felt that she had decided to move. The ACT-M model seems incompatible with Carthesian Materialism.

Baars argues that GWT is distinct from the concept of the Cartesian Theater, since it is not based on the implicit dualistic assumption of "someone" viewing the theater, and is not located in a single place in the mind. Similarly, in an ACT based AM there is no clear and specific place where consciousness is located, so the ACT model fits Baars’ arguments in this respect.

5.1.4. Multiple Drafts Model

The most prominent philosophical model of consciousness is probably Dennett’s Multiple Drafts Model (MDM) (Dennett 1991). Using his arguments against a Cartesian Theater, Dennett proposes a model with a variety of features, among which the non-existence of qualia and P-zombies. MDM understands conscious experience as taking time to occur, such that percepts do not instantaneously

arise in the mind in their full richness; the model denies any clear and unambiguous boundary separating conscious experiences from all other processing. Different parts of human neural processing assert more or less control at different times and for something to reach consciousness is comparable to becoming famous, in that it must leave behind consequences by which it is remembered. Which inputs are "edited" into our drafts is not an exogenous act of supervision, but part of the self-organizing functioning of the brain, and at the same level as the circuitry that conveys information bottom-up. In this sense the ACT model does not contradict MDM. As to the existence of qualia and P-zombies, the reader will remember that this paper does not take a particular stance on this issue (section 3.7) so in that respect the ACT model does not contradict MDM either.

As is the case for any model of consciousness, MDM is not without critics, notably philosophers who maintain the notion of possible existence of qualia (including aforementioned Block (1998) and Floridi (2005)). Chalmers (1996) proposes that Dennett has produced no more than a theory of how subjects report events. I do not take a position pro or contra MDM here but merely state that the ACT model does not contradict MDM.

5.1.5. The Hard Problem of Consciousness

Chalmers has argued that most if not all current models of human consciousness focus on accessibility and reportability; this he considers the "easy problems", which can be explained in terms of computational or neural mechanisms. The "hard problem" is the problem of subjective experience: "Why is it that when our cognitive systems engage in visual and auditory information-processing, we have visual or auditory experience: the quality of deep blue, the sensation of middle C? How can we explain why there is something it is like to entertain a mental image, or to experience an emotion?" (Chalmers 1996a). The ACT-M model does not provide answers to Chalmers' hard problem (neither do MDM of GWT for that matter); this is not a weakness as the model derives from the Mission Statement, which did not set out aiming to explain all of human consciousness. The ACT-M model, applied to humans, does isolate the hard problem *to an extent*. This matches well with the statement that "drivers are the hard problem" in section 4.5. The last but one sentence ("to an extent") deserves some further attention.

Although ACT-M cannot entirely explain subjective experience, the argument made for subjectivity in experience for AM in section 3.7.3, also holds for humans: i.e. I do propose that part of the subjectivity of an experience is due to (sensorial) experience and hence learning, and another due to the particularities of the driver system of the human in question. A (very unscientific) example would be "learning to drink beer", or phrased in a less limited fashion, acquiring a taste.

Furthermore, our thought experiment of using ACT-M as a model for human cognition may offer explanations to certain questions (though I make no claims that those answers are absolutely correct; this remains a thought experiment after all). This subsection offers 2 examples of such explanations.

Firstly, let us reconsider the colorblindness example we mentioned in 3.7.3, but now for humans. This thought experiment is known as "Mary's box" (Jackson 1982) and serves to demonstrate the non-physical nature of mental states. A human (Mary) who knows everything science knows about color, about how light works to create different color wavelengths and about the perception of color in human brains, but has been living in a box without color. If she were to be freed from the room and visually perceive color for the first time, would she learn something new about it, namely what color looks like?

If we use ACT-M as a model for human cognition the answer would be yes and explanation would be similar to the one provided for AMs: I can imagine[8] that once a human's color sensors were to be switched on for the very first time one would have very strange 'experiences' (perhaps like hallucinations), and that it would take time for our neural tissues to get to know how to process this information, and to make sense out of the information. I do not believe that what you & I experience as 'blue sky' (even apart from the fact that our experiences are not the same and perhaps even incomparable) would be what Mary would experience after switch-on, I even doubt that 'red' would uniformly be 'experienced' as red for quite a while, until our pattern detector systems have settled into detecting that some stuff is always the same wavelength and hence can be associated with the name of a color. In other words, Mary would need to learn to feel what it is like to see blue, and that would be a new experience.

Another example of (potential) usefulness of using the ACT-M model for human cognition is the following. Block (2009) poses the following: "If you have seen and heard a circumcision, you may

[8] I personally actually can imagine: in 1996 I lost a contact lens, and for a variety of reasons I was not immediately able to order a replacement. As a result I walked around for 9 months with basically one missing eye (I am extremely shortsighted). Initially this posed a problem because of lack of stereo sensorial input, my depth sight was missing and I would end up 'missing' my cup of coffee when reaching out for it. Over time I learned to compensate by continuously moving my head in order to receive sufficient input through one eye to be able to estimate depth. When in early 1997 I did manage to get a replacement contact lens, my brain had for all practical purposes shut down processing of the input of my left eye and I was not even missing my second contact lens. When I put in the missing contact, my brain had to re-learn processing stereo input; I can attest I did feel something: I felt nauseated for a couple of days and had splitting headaches for 2 weeks, which my eye doctor attributed to 'getting used to seeing stereo'. This is admittedly a very unscientific argument, it is based on introspection (and not all will accept introspection as a valid approach to science) but for me the experience of having to (re-) learn to process stereo vision was (or at leased seemed) very real (and literally painful) to me.

find it difficult to doubt that it hurts. Relevant evidence: newborns who are circumcised without anesthesia or analgesia are more stressed by later vaccination even 6 months later. […] My point is not that you should be totally convinced of phenomenal consciousness in early infancy, but rather that you should be convinced that there is a better case for phenomenal consciousness in infancy than there is for those instances of phenomenal consciousness being accompanied by higher order thought."

Irrespective of whether I believe an infant is conscious, and/or whether one needs higher order thought before one can use the term consciousness at all, the ACT-M does offer an explanation of why a circumcised infant might be more stressed by later vaccination than uncircumcised ones: the sensation of pain of the experience (which is controlled by the human pain receptors, not by a "hard to grasp" driver system) in the ACT-M model makes the experience very "loud", and hence it is "remembered" very "loudly", and all associated memories (knifes, doctors etc) become pretty significant as well.

5.2. Computationalism

Computationalism considers the mind to function like a computer or symbol manipulator (Putnam (1961), Fodor (1975)). Newell & Simon (1976) postulated that a physical symbol system, i.e. a system that considers physical patterns (symbols), combines these into structures (expressions) and manipulates them (using processes and algorithms) to produce new expressions, has the necessary and sufficient for general intelligent action. The model proposed here is not a computational one: in our model, the capacity to manipulate symbols is an emergent property of the processes running on the substrate; the substrate need not be a symbol processing machine. Our model also disagrees with Newell & Simon, in the sense that a symbol manipulator by itself would not lead to a thinking machine. As described in 3.5.2, we do conclude that a digital computer (as a symbol processor) may be a suitable substrate, if running the appropriate program, but even in that case the thinking agent is not the symbol processor of the substrate, nor the program that manipulates symbols on the substrate, but the dynamic process ran by the program on the substrate. An emergent property of this dynamic process would be the ability to reason on symbols, or the illusion that the process is doing so.

5.2.1. Symbols and Symbolic Logic

When we state that the ability to perform symbolic logic would be an emergent property, the question arises as to how symbols (or even more complex, mental states) would be represented in an AM. Given the extreme dynamic nature of Minds in this model, I propose it will be impossible to absolutely determine which dynamic pattern in an ACT-M or a human mind represents which symbol (or mental state), we referred to this earlier in 3.5 as being 'fuzzy' At this point in our discussion we can define the concept of fuzzy in this context:

- Representations are not fixed, they evolve over time. What I consider to be a chair may change. The day we invent the technology to suspend a seat, without legs, in air, we will consider a floating seat a chair. Or perhaps, just considering this possibility, my brain just added this possibility to the concept of chair and hence changed the concept of chair. Every time we perceive a particular, not previously perceives instance of what-a-chair-is (or might be), our internal representation of chair changes. Once we have learned, as a child, the notion of the word chair, our internal representation might not change dramatically, but it does not remain set in stone either.
- Representations are not fixed because they depend on context. 'John' might be Wayne or Lennon. A chair might be an instance of a design object, or it might be something to sit on when tired. In the context of this paragraph, a chair is neither; it is a concept used to illustrate a point, and in the context of the logic presented here interchangeable with a horse, a car, a cat.
- When we talk about thinking about 'chair', the definition of 'thinking about chair' is undefined as well; since I can think about the concept 'chair' (whatever that means given the context I am thinking in) in both English & French, i.e. about 'chair' (the English word) and 'chaize') the French word.
- Since patterns of patterns are patterns, and thinking is a massively parallel activity, it would be impossible to separate the patterns of 'chair' from the other thoughts going on in either a human mind or an ACT-M. Neither is it possible to separate the English word 'chair' from the concept of 'chair' (and adding insult to injury that concept itself is not atomically discernible). To humans it is impossible to think only of 'concept chair' and nothing else at all (the reader is invited to try this; such an exercise would be similar to meditation techniques using a mantra; the difference being that a mantra should not mean anything). Humans (and I propose AMs based on ACT-memories) are incapable of not associating; hence thinking of a chair automatically leads to thinking of other things, and as a result, the 'atomic' pattern that would encode 'chair' becomes indiscernible.

One might argue that the pattern that causes an AM to output, on a computer screen, the word 'CHAIR' in a particular font would be fixed, since the output is fixed and could be determinable. That would be correct, but that pattern would merely be a motor pattern; it would not be the pattern of thinking of a chair. Actually I propose there is no such thing as 'the pattern' that is isomorphic with thinking of 'chair', even if 'thinking of chair' were perfectly mathematically defined and in addition, 'thinking of chair and nothing but chair' were humanly (or 'machinably') possible.

When thinking of 'chair' in whatever instance of chairness, depending on context, the 'pattern' of CHAIR pops up in our minds, what actually pops up is a crystallization, a projection, of what is being thought, into humans language, in this case English. It is not an isomorphism, and the fact that I can equally seemingly think 'the same thing' in terms of 'chaizes' indicates it is merely *a* projection. In this respect Language of Thought is an illusion in the sense that what we 'hear' we are thinking is not what we are actually thinking; it is merely an imperfect projection of what is being thought, formulated in a human (public) language (see also Blackmore 2002).

I also propose that the perceived mechanics of logically (in the Boolean sense of the word) processing symbols, is equally a projection, in English or in Boolean algebra, of what is really being thought.

5.2.2. Chinese Rooms & China Minds

With respect to Searle's Chinese Room argument (Searle 1980): if a human were to function as (part of) the substrate, executing the program for an AM (the program being part of the substrate per 3.5.2), then indeed the human would not know what the dynamic process it is running is thinking, in the same way that the transistors in a digital computer substrate would know this, let along that these transistors themselves could or would think. If a human were executing the program, it would be absurd to state the program itself is thinking. If every neuron in my brain had a brain, neither of these brains would know what the brain they are part of is thinking. Similarly, consider Block's China Brain thought experiment (Block, 1978), in which the whole population of China is to simulate the workings of a single brain and each Chinese person acts as a neuron, and communicates by two-way radio in the corresponding way to the other people. Such China Brain would have all the elements of a functional description of mind: sensory inputs, behavioral outputs, and internal mental states causally connected to other mental states. However, no Chinaman would know what the China brain is thinking (though each Chinaman might be aware they are participating in the thought experiment). In an ACT-M based AM, running on a digital silicon-computer, neither the substrate (the chips) neither the computer program, would be thinking. The thinking would be the dynamic process, not the substrate and neither the program. In this sense I agree with Searle; but at the same time I would argue I do have a thinking machine; the thinking just does not happen where Searle per his argument would claim I would put it (assuming he were to read this paper).

We will briefly return to Block & Searle's thought experiments in section 5.4.

5.3. Perdurantism & Dynamicism

The view presented here is perdurantistic, in the sense that as pointed out in 4.3.1 the Mind, the Self is constantly in flux. Dynamicism in our model occurs on 2 levels: the first level is dynamicism in the

memory, which constantly adapts stored patterns, finds new correlations or modifies existing ones, and the second level is the awakening of patterns by association and sensorial stimulation.

Dynamicism as a basis for cognition has been proposed before, although the work presented here was performed independently (see e.g. Beer (1990, 2000, 2003), Thelen & Smith (1994). Barsalou et al (2003), Barsalou (1999, 2003)).

Van Gelderen argued in a paper titled 'What might cognition be if not computation' (1995) that a Watts Governor – a dynamic gearing system for a rotative steam engine, patented by Scottish engineer James Watts around 1788) is a dynamic process, but it performs as if it executes an algorithm and as if it manipulates symbols. However these symbols are nowhere represented; Van Gelderen shows in his paper that a dynamic system can behave as if functionally equivalent to a symbolic algorithm processor, or in other words, Van Gelderen outlines a possible computer program, operating on symbols, which would perform exactly like a Watts Governor. Where Van Gelderen proposes dynamicism as a potential, theoretical basis for a model of cognition, the ACT model takes this idea one step further towards implementation.

The work presented here can also be related to Barsalou's Perceptual Symbol Theory (PST) (Barsalou 1999, 2003). Quoting Barsalou (1999): "The basic assumption underlying perceptual symbol systems: Subsets of perceptual states in sensory-motor systems are extracted and stored in long-term memory to function as symbols. As a result, the internal structure of these symbols is modal, and they are analogically related to the perceptual states that produced them". In PST, thinking of a color depends upon the same neural system that is recruited when the color is actually perceived. Contrast this with classic symbolically computational approaches, where the symbols are amodal, i.e. disconnected from the actual perceptions. PST is in this respect similar to the mechanisms in an ACT-M where e.g. thinking of a color (e.g. the concept of the color yellow), by association, wakes up the same patterns that are awoken when seeing the color yellow. PST is a model for human cognition from a psychological point of view; it is not a priori a model for machine intelligence, whereas ACT-M is.

Again, as with just about any model of cognition, dynamicist approaches such as Barsalou's are not without critics (see e.g. (Rupert 2006), (Machery 2007)). Without passing judgment about aforementioned dynamicist models, we merely point to the similarities in approach.

5.4. Grounding

In section 3 I proposed that an AM (and a human alike) not subject to any sensorial input would never evolve a thinking mind. This line of thought is similar to Harnad's concept of grounding (Harnad 1990, 1992). For Harnad, in a symbolic system, "the symbols, despite their systematic interpretability, are ungrounded; their meanings are parasitic on the mind of an interpreter. So the

symbol grounding problem concerns how the meanings of the symbols in a system can be grounded (in something other than just more ungrounded symbols) so they can have meaning independently of any external interpreter." (1992). A symbol is merely a token, and any token serves as well as any other: there is nothing about the label CHAIR that makes it serve any better as a label for the atomic representation of the CHAIR concept than as a label for the CAT concept (Chalmers 1992).

Since an AM builds its representation of chair based on real instances of chairs or things that can be used as chairs, it is by definition grounded. An ungrounded AM would and could simply not be thinking, and the thinking emerges based on the grounding. This should be seen unrelated to Harnad's grounding for symbolic systems, since an AM, in our model is not a symbolic system, but a dynamic one.

We mentioned Perceptual Symbols Theory in the subsection about dynamicism; it is interesting to note that PST is inherently grounded (Barsalou 2008), and again, this is a point of similarity between our proposal and PST.

We already mentioned Harnad's Total Turing Test in section 3.7. For a machine to pass the TTT requires indistinguishability between man and machine in both symbolic and robotic capacity, where robotic capacity is defined as sensorimotor capacity to discriminate, recognize, identify, manipulate and describe the objects, events and states of affairs in the world. An AM as proposed here would not be an AM without sensorimotor capacity; the fact that it depends on sensorimotor capacities for its grounding may seem like a good step towards passing Harnad's TTT, but as described in section 3.7 that our ACT based AM would and should fail the TTT, since it is simply not human.

In 5.1 we claimed that with reference to Searle's Chinese Room, if a human is executing the program, it would be absurd to state the program itself is thinking. Similarly, with respect to Block's China Brain thought experiment, no Chinaman would know what the China brain is thinking. Given our reasoning about grounding we might add that simply a standalone program, without sensoria & motoria, would never be able to evolve the complex dynamic behavior needed for a Mind to emerge. It would not be grounded. Hence, a computer program, be that executed by a digital computer, a single human, or the population of China, without sufficiently rich sensorial input, would never amount to anything intelligent.

5.5. Connectionism

The attentive reader familiar with the matter may have wondered by now why connectionism was not addressed earlier in this discussion. The reason for this is that I wanted the discussion about dynamicism out of the way.

5.5.1. Connectionism & Dynamicism

According to Van Gelderen (1995) connectionist systems are dynamical systems, just like a Watts governor. Not everyone would agree with this statement: Beer (2000), who just like Van Gelderen is a proponent of dynamicism, has stated that "dynamical models can be represented as connectionist networks, and at least some connectionist networks are dynamical systems".

This author would clear up the difference between both opinions as follows. Some connectionist models, e.g. Hopfield's models (Hopfield 1982) are dynamic systems in the sense that they can be described by means of differential equations (or difference equations, when simulated on a symbolic computer). Other models, such as backpropagation networks (Rumelhart, Hinton & Williams 1986) are not. Backpropagation networks are simple input/output networks in which a static input is transformed into a static output. This matches Beer's statement: some connectionist architectures, but not all, are dynamic systems in this sense. However, the training of a backpropagation network, in which perceived errors in output, given an a priori known input/output pattern, lead to successive modifications to the 'synaptic weights', can be considered a dynamic process. Extensions of the standard backpropagation model can deal with patterns over time (e.g. Mozer, Chauvin & Rumelhart 1995); those are dynamic in the sense that they deal with patterns over time. At the same time models such as Hopfield networks are not dynamic in the sense that they deal with fixed input patterns, after which they – dynamically – reach a fixed, stable state. Hence, what makes a system 'dynamic' depends on one's point of view.

To return to our subject, the ACT model proposed in this paper is implementation-independent. The question that remains is then whether it could be implemented as a connectionist system. Connectionist input/output models (such as backpropagation networks) cannot be used to implement an ACT-M, as they lack dynamicity in their operation. Dynamic networks like Hopfield networks cannot be used either, because they do not deal with dynamic input (patterns). However, more complex architectures, which behave dynamically (i.e. according to differential/difference equations) and which can deal with input that changes over time (patterns) may be suitable candidates for *implementation* of our model. An example of such a complex connectionist architecture is Hawkins' Hierarchical Temporal Memory (HTM) Cortical Learning Algorithm (CLA) (Hawkins 2004). Interestingly, Hawkins stresses the importance of time in a very similar way as found in the early sections of this text.

That said, from a comparison point of view, connectionist models do share similarities with what I propose here: they offer non-trivial 'representations' as patterns of activity in 'artificial neurons', i.e. highly interconnected processing units. Some connectionist models offer the kind of unsupervised learning that might match our requirements for the system to correlate and associate. Others, such as backpropagation and its derivatives, offer algorithms for supervised training. Such algorithms

however need an overall ‘fitness’ function to be optimized (or conversely an ‘error function’ to be minimized) and such algorithms do not make for a good fit with the ACT model, since there is no overall measure of desired behavior (apart from a human correcting the system). Finally, connectionist architectures offer parallel and distributed processing, which also fits the ACT model.

The grandfather of connectionist models is probably Hebb’s rule, which is a form of reinforcement learning which postulates “that any two cells or systems of cells that are repeatedly active at the same time will tend to become 'associated', so that activity in one facilitates activity in the other.” (Hebb 1949). In this respect, the model proposed here definitely has a Hebbian flavor.

As a side note it has been shown (Franklin & Garzon 1990) that Von Neumann architectures (such as digital computers and Turing Machines) and neural-network connectionist architectures are both universal; anything the former can do, the latter can as well. Most implementations of connectionist architectures are even implemented as ‘simulations’ on digital computers. Hence, if a connectionist architecture is a good model for implementing an ACT-M based AM, then so is a digital computer.

Again, connectionist models as a basis for artificial intelligence (and by extension conscience) are not without critics; the most influential body of criticism can be found in (Fodor & Pylyshyn 1988) who seem to defend a thoroughly computational point of view.

5.5.2. Subsymbolism

Smolensky (1988) considers connectionism not to be symbolic computation, but rather subsymbolical. To quote Smolensky: (in a connectionist architecture) “The intuitive processor is a subconceptual connectionist dynamical system that does not admit a complete, formal, and precise conceptual level description.” And further “Do the formal principles of cognition lie at the conceptual level? The answer offered by the subsymbolic paradigm is: No – they lie at the subconceptual level”. This matches to some extent our observation that the core ‘representations’ of what is going on in an ACT, the awakening of patterns, is not isomorph to symbolic thinking, it is subsymbolic in the Smolensky sense.

However, where we disagree with Smolensky is where he proposes “… states of a subsymbolic model can be approximately analyzed as superpositions of vectors with individual conceptual-level semantics. … the subconceptual and conceptual levels are isomorphic”. I differ in opinion in 2 ways. One is that Smolensky seems to consider encoding of ‘symbols’ as vectors whereas the ACT model proposes that ‘symbols’ would be fuzzy-isomorphic to a trajectory in state space, i.e. not a fixed vector. Secondly, I use the word ‘fuzzy’ because I propose there is no discernible and fixed match between ‘symbols’ and patterns in an ACT.

5.6. Closing Implementational Remarks vs Functional AM Design

In previous subsections ad hoc references and hints were made to implementational issues, even though the goal of this text is merely the description of a functional design of an AM, not an implementational one. It was mentioned that in principle this functional design might be implemented on a standard digital computer, which is a symbol processor. It was also mentioned that connectionist models may fit the ACT-M AM design. For the sake of completeness this subsection briefly addresses some other AI techniques, and their potential relevance in this context.

Several statistics-based techniques and algorithms are currently used in what is considered applied AI today. This ranges from consumer product features such as the face recognition software in our digital cameras, the speech processing software in our smartphones, to highly specialized functions such as automated medical diagnosis aids. Techniques used for these applications include Bayesian decision networks, (Hierarchical) Hidden Markov Models, spatial & temporal clustering algorithms of a wide variety of backgrounds.

All these, along with e.g. complex dynamic neural network-like architectures mentioned in this text, such as HTM, may be useable for actually building an ACT-M based AM. I do propose that none of these techniques are actual models of (human or machine) cognition; these are potential implementational tools, and doubtlessly valuable at that.

5.7 Potential Criticism

The author can think of several avenues of criticism to the model presented here. Whereas a full analysis is beyond the scope of this paper, I briefly address the main issues.

The main argument will be that all this is too vague, too simple and perhaps too philosophical, and hence is not enough of a basis to actually build a working thinking machine. The naivety argument can be broken down – and subsequently dealt with in the following subsections. Further subsections deal with other criticisms.

5.7.1 Too vague & Missing Components

One might argue this is too vague, and that surely the model is missing functional components.

I would rebut this in two ways. First of all the goal described in section 1 was to functionally design an AM, not technically design one. Technically speaking, yes one would need explicit mechanisms to e.g. decide on similarity of patterns, and clearly defined mechanisms/algorithms for e.g. pattern creation (the creation of (fuzzy) patterns that somehow encode ‘chairness’ out of many sensorially ‘seen’ chairs). The argument made here is that aspects of Association, Correlation and Time are necessary, not necessarily sufficient; the mechanisms to encode time, create associations and detect

correlations were not subject of this text. These are important topics and are subjects for further research.

In addition one might argue that even if and when technology is ready to build and ACT-M based AM, it would be naive to expect something like human intelligence to emerge based on so few principles (association, correlation and time coding + feedback, motorium and sensorium). There are 2 different paths of rebuttal to this criticism.

First of all, there is evidence that the human brain, though it is massive in terms of components (number of neurons and number of connections between those neurons), is structurally not all that complicated (relatively speaking). Kurzweil (2012) provides tis argument in layman's terms along with scientific references, and his argument is along the following lines: there are on the order of a quadrillion (10^{15}) connections in the human neocortex, but the relevant genome that encodes the design information for the neocortex comprises only about 25 million bytes, which is only double the memory of the laptop I write this paper on. Hence the connections themselves cannot be determined genetically (Kurzweil goes on to argue they are formed by experiences, which matches the proposal presented in this paper). The conclusion is then that the neocortex, organizationally and structurally is composed of fairly uniform modules or columns, which are all based on the same design. Markram & Perrin (2011) found the neocortex to be "elusive assemblies [whose] connectivity and synaptic weights are highly predictable and constrained. […] They serve as innate Lego-like building blocks of knowledge for perception and that the acquisition of memories involves the combination of these building blocks into complex constructs" (cited in Kurzweil 2012).

Admittedly, the above view is not shared universally. Yudkowsky (2007) for example argues that "Simplicity is the grail of physics, not AI" and subsequently coins the term "Physics envy in AI" as the search for a single, simple underlying process, with the expectation that this one discovery will lay bare all the secrets of intelligence. Whether Yudkowsky is right and the ACT-M model does or does not suffer from physics envy, the future will tell.

Secondly, because is system might be based on relatively simple principles does not imply it would be easy to actually build. Current (super-)computers do not match the computational power of a human brain; even assuming that the advances in technology and Kurzweil's Law of Accelerating Return hold (which is reasonable, barring some disaster that wipes out humanity or sets human civilization back (Chalmers 2010)) it is not unreasonable that expect that within the next decade or 2 machines will be available whose computational abilities exceed the human brain. Kurzweil (2005) estimates this to happen around 2023. Suppose we build an ACT-M based AM and boot it in 2023 and suppose we have all parameters right, and this first attempt turns out to be equally efficient at developing into human level intelligence as an average human; in that case it will about 2 years for the system to learn how to talk back to us coherently. Those assumptions are probably unreasonable, so allowing for

years of trial and error, as well as even further advances in technology which would lead to 16-fold-ish increase in computational power over the next 10 years, any 'trial' run up to equivalent age of 2 would still take 45 days. I propose that even with the advances in technology (reasonably) predicted, and even if a model such as ACT-M were to be correct (which I hope and expect, but do dare claim here), we are still decades away from actually building and ACT-M based AM. This should, however, not prevent us from thinking about their design.

5.7.2 Inconceivably Buildable

One might argue that, even if technical details were available, it has not been shown why this would be conceivably buildable. To rebut this criticism I would argue that humanity today has technology to largely accomplish what is needed: we know how to build motoria & sensoria; we have techniques to detect patterns. If e.g. we have clear differential equations for such a system, we know at the very least how to discretize and simulate these on a digital computer.

In addition, we do have technology to detect correlations and implement associativity in memories. We may not have exactly the right tools to implement an ACT-M, but I see no fundamental reasons why we might not come up with appropriate such tools.

5.7.3 Dangerous

Proponents of Friendly AI (Yudkowsky 2001) will probably argue that the approach described here is dangerous, in the sense that it might lead to artificially intelligent systems, with an intelligence perhaps surpassing human-level (Kurzweil 2005), that are hostile to humans, or that behave (involuntarily) dangerously to humans (see examples in section 2.11), because the driver system described here lacks goals that are clear, distinguishable and controllable. This is a valid argument, but I would argue that the danger with ACT-M based AMs is no greater than for other artificially intelligent systems. One would always need to be careful when training (or 'programming' in Yudkowsky's terminology) such a system, and as Yudkowsky suggested, we can always set AMs loose initially in simulated worlds, until we are relatively certain hey are friendly.

Finally, please note that friendliness, although admittedly important, was not part of the mission statement with which we started our thought experiment.

5.7.4 Need for Chaos

In 2.9 I proposed an ACT-M would be chaotic. A potential criticism would be that there is no clear reason why this would be the case. I agree this is a valid remark; it may or may not be possible to build an ACT-M based AM without chaotic mechanisms. However, I would argue, admittedly more based on my personal intuition than on scientific evidence; that the richess of a human-equivalent

mind would almost have to imply chaos. Whether this proposition is correct future research will have to conclude.

5.7.4 Inefficiency

One might argue that the approach described here would be terribly inefficient: surely when we know that e.g. humans are capable of performing symbolic/Boolean logic it would be more efficient to build this into the system, rather than to expect the system to either discover Boolean logic, or expect it to learn this?

This argument too can be dealt with in 2 ways. First of all inefficiency is relative. A 1960s computer scientist would probably have ridiculed the practicality of the idea of running mission-critical enterprise functions on computer by means of interpreted languages. Today this is common practice. If Moore's law holds, at some point in the next couple of decades we should have enough raw power to make this argument go away.

Secondly, I would argue that the more we choose to hardcode or hardwire in a general AI, the more room for error: how would we a priori know what absolutely must be hardwired and what not? Conversely I would argue that less hard-coding, and more reliance on learning/discovery/self-organization in an AI is better and even desirable. An ACT-M based AM would be a truly general AI, capable of dealing with (perhaps imaginary or simulated) realities that differ wildly from the world as we know it. Given appropriate sensorium and appropriate stimuli, there should be no reason why and ACT-M based AM would not be able to learn to navigate e.g. a spatially 4-dimensional world (or any higher dimensionality) or deal with realities with laws of nature very different from ours; feats which humans brains would probably find mind-blowingly difficult.

An average chess program is specialized, it has the rules of the game baked into it. On today's hardware it performs for all practical purposes pretty efficiently, it works as intended when first booted (it does not have to learn the game) and can easily beat an average human chess player (be it that the way a chess program plays chess may not be considered efficient in approach as it basically performs a brute search in game state-space). However, it is not general; it can do one thing well, and do nothing else at all. Specialization may lead to efficiency, but only in what specialized in. The mission statement this text started with is particularly unspecialized, but rather very general purpose; we made a reference to AGI – Artificial General Intelligence – early on in this paper). It is then not such a stretch do to away with specialization as much as possible, as indeed has been done in the approach chosen here.

6 Summary & Conclusions

This paper presented several ideas and propositions, admittedly sometimes hidden in the text. These are summarized below:

- Proposition of a thought experiment which leads to a functional model for a thinking machine, which is conceivably buildable with current or future technology. The model is centered around a memory which:
    - Deals with patterns over time, and can store and detect patterns, and forgets stored patterns over time if such patterns are not (regularly) awoken
    - Is able to correlate occurrences of patterns, in real-time and in memory
    - Is associative in the sense that real time occurrence of patterns awakens previously correlated patterns

  Such a memory is called an ACT memory. An AM based on an ACT memory needs a sensorium and a motorium, needs feedback from memory to memory, and needs a driver system.

- Proposition that in a machine based on the above model, an intelligent mind would emerge, provided the machine is presented adequately rich stimuli.
- Proposition that in such a machine the mind is not equal to the substrate, and that what is considered as thinking is a (dynamic) process rather than something akin to a computer program. The actual running instance of a computer program might be able to exhibit such a dynamic behavior.
- Proposition that testability of existence of such a mind is philosophically difficult, and that a 'duck test' seems the only viable alternative. Unfortunately this does not exclude the possibility of such a mind being a P-zombie.
- Proposition of a second thought experiment: to use the same model as a model for human cognition. One big difference between AMs and HMs is the driver system; humans have biologically/genetically controlled drivers whereas in this model a machine does not. The model does not fully explain the richess of human consciousness, but does isolate Chalmers' (1996) hard part of human consciousness. A mechanism similar to simulated annealing is proposed to help 'drive' an ACT-M based AM.
- Proposition that under above model, human as well as machine minds are to a large extent the weighted sum of their experiences, colored by their drivers.

- Proposition that ‘being a mind’ is dynamic processes. In this perdurantistic view, thinking is a trajectory in state space, and at the same time being constitutes a trajectory in state space, where thinking and being mutually influence one another’s trajectories.
- Proposition that in this model, a mind is not a symbol processor, and that ‘patterns’ cannot be absolutely linked to symbols. Symbolic thinking and a language of thought are illusions in this model.

Furthermore the ACT-M model has been compared to other theories of mind; the model is incompatible with computationalism but need not be incompatible with other models such as Barsalou's Perceptual Symbol Theory, Baars’ Global Workspace Theory and Dennett’s Multiple Drafts Model.

Finally I have argued that an ACT-M based machine is conceivably buildable, although probably not with current state of technology, but perhaps with current state of (scientific) knowledge.

Further work will be done on exploration of the model to address further philosophical questions. In addition further work will be done towards translating the functional design of a machine-based ACT-M into a more mathematical formulation and a technical specification.

7 Acknowledgments

The author thanks David Chalmers, Peter Dayan, Wouter Soudan and Katrijn Van Bouwel for discussions on earlier drafts of this text. Koen Thewissen pointed out the relevance of Maslow’s work.

8 References

- Aleksander, I (1996) Artificial Neuoconsciousness, an Update, Neurocomputing (12) 91-111
- Baars, B. J. (1988), *A Cognitive Theory of Consciousness* (Cambridge, MA: Cambridge University Press)
- Baars, B. J.(1997), *In the Theater of Consciousness* (New York, NY: Oxford University Press)
- Baars, B. J. (2002), The conscious access hypothesis: Origins and recent evidence. Trends in
- Baars, B. J. (2005) Towards a cognitive neuroscience of human experience?, Cognitive Sciences, 6 (1), 47-52.
- Barsalou, L.W. (1999) Perceptual Symbol systems, *Behavioral & Brain Sciences*, **22,** 577–660
- Barsalou, L. W. (2003) Abstraction in perceptual symbol systems, *Phil. Trans. R. Soc. Lond.* B **358**, 1177–1187

- Barsalou, L.W., Simmons, W.K., Barbey, A.K., and C.D. Wilson, (2003), “Grounding conceptual knowledge in modality-specific systems,” *Trends in Cognitive Sciences*, 7: 84–91.
- Barsalou, L.W., (2008), “Grounded Cognition,” *Annual Review of Psychology* 59: 617–645.
- Beck, K. et al. (2001), “Manifesto for Agie Software Development”, Agile Alliance.
- Beer, R.D., (2003), “The Dynamics of active categorical perception in an evolved model agent,” *Adaptive Behavior*, 11: 209–243.
- Beer, R.D., (2000), “Dynamical approaches to cognitive science”, *Trends in Cognitive Sciences*, 4: 91–99.
- Beer, R.D, (1990), *Intelligence as Adaptive Behavior*, New York: Academic Press.
- Blackmore, S. (2002) “There is no stream of consciousness”, Journal of Consciousness Studies((5): 17-28
- Block, N. (1998), "On a Confusion About a Function of Consciousness", *Behavioral and Brain Sciences*, 18, 227-247
- Block, N. (2009) Comparing the Major Theories of Consciousness, in Gazzaniga, M. (ed), MIT Press 2009, 1111-1122
- Bringsjord, S. (2010) Meeting Floridi’s challenge to artificial intelligence from the knowledge-game test for self-consciousness, Metaphilosophy 41, no 3, April 2010
- Chalmers, D. (1992) Subsymbolic Computation and the Chinese Room, in The Symbolic and Connectionist Paradigms: Closing the Gap (J Dinsmore, ed.). Lawrence Erlbaum.
- Chalmers, D (1994) A Computational Foundation for the Study of Cognition
- Chalmers, D (1996). *The Conscious Mind*. Oxford University Press.
- Chalmers, D. (1996a) “Facing up to the problem of consciousness”, Journql of Consciousness Studies (2) 3, 200-219.
- Chalmers, D. (2010) The Singularity: a Philosophical Analysis, Journal of Consciousness Studies, 17:7-65
- Dennett, D. (1991) *Consciousness Explained*, Penguin books.
- Dretske, F. 2003, "How Do You Know You Are Not a Zombie?" in *Privileged Access and First-Person Authority*, edited by B. Gertler (Burlington: Ashgate)
- Floridi, L. (2005) Consciousness, Agents and the Knowledge Game, Minds and Machines 15, 3-4, 415-444
- Fodor, J. A. (1975). The language of thought. Cambridge, MA: Harvard University Press.

- Fodor, J & Pylyshyn, Z.W. (1988). “Connectionism and Cognitive Architecture: A Critical Analysis,” In *Connections and Symbols*, ed. Pinker, Steven and Jacques Mehler, MIT Press, 1988.
- Franklin, S., & Garzon, M. (1990). Neural computability. In O. Omidvar (Ed.), *Progress in Neural Networks*, Vol. 1., pp. 127-145. Norwood, NJ: Ablex.
- Hebb, D. (1949) *The Organization of Behavior*, New York: Wiley & Sons.
- Harnad, S. (1989) Minds, Machines and Searle. Journal of Theoretical and Experimental Artificial Intelligence 1: 5-25.
- Harnad, S. (1990). The symbol grounding problem. Physica D, 42, 335–46
- Harnad, S. (1993) Grounding Symbols in the Analog World with Neural Nets. *Think* 2: 12-78 (Special Issue on "Connectionism versus Symbolism" D.M.W. Powers & P.A. Flach, eds.).
- Hawkins, J. & Blakeslee, S. (2004) *On Intelligence*, Time Books
- Hopfield, J.J. (1982) Neural networks and physical systems with emergent collective computational abilities, PNAS, 79, 2554-2558.
- Jacobson, I, Booch, G., Rumbaugh, J (1999) *The Unified Software Development Process*, Addison Wesley
- Jackson, Frank (1982). "Epiphenomenal Qualia". *Philosophical Quarterly* (32): 127–136.
- Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. (1983). Optimization by Simulated Annealing. *Science* **220** (4598): 671–680.
- Kurzweil, R. (2005) The Singularity is near, Viking Books
- Kurzweil, R. (2012) How to Create a Mind – the secret of human thought revealed, Penguin Books
- Libet, B., Gleason, C.A., Wright, E.W., Pearl, D.K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). The unconscious initiation of a freely voluntary act. *Brain*. 106 (3):623–642.
- Machery, E., (2007), “Concept empiricism: A methodological critique,” *Cognition*, 104: 19–46.
- Maslow, A.H. (1943) A Theory of human motivation, *Psychological Review*, 50(4), 370-96
- Markram, H & Perrin, R. (2011) Innate neural assemblies for Lego memory, *Frontiers in Neural Circuits* 5(6)
- Mozer, M. C. (1995). Y. Chauvin and D. Rumelhart. ed. *A Focused Backpropagation Algorithm for Temporal Pattern Recognition*. Hillsdale, NJ: Lawrence Erlbaum Associates. pp. 137–169.
- Nagel, T. (1974), "What Is It Like to Be a Bat?" *Philosophical Review*, 83(4), 435-450.
- Newell, A. & Simon, H.A. (1976) (1976) Computer Science as Empirical Inquiry: Symbols and Search , *Communications of the ACM* **19** (3): 113–126

- Omohundro, S. M. (2007), The Basic AI Drives. Self-Aware Systems Inc. (http://selfawaresystems.files.wordpress.com)
- Putnam, H. (1961). "Brains and Behavior", originally read as part of the program of the American Association for the Advancement of Science, Section L (History and Philosophy of Science), December 27, 1961.
- Rupert, R., (2009), "Innateness and the Situated Mind," in *The Cambridge Handbook of Situated Cognition*, P. Robbins and M. Aydede (eds.), Cambridge University Press, pp. 96–116.
- Rumelhart, D. E. & McClelland, J. L (1986). *Parallel distributed processing: Explorations in the microstructure of cognition.Volumes I & II.* Cambridge, MA: MIT Press.
- Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986). "Learning internal representations by error propagation". *Mit Press Computational Models Of Cognition And Perception Series* (MIT Press Cambridge, MA, USA): pp. 318–362.
- Russel, S. J & Norvig, P. (2003). Artificial Intelligence: A Modern Approach. New Jersey:Prentice Hall
- Schwaber, K. & Beedle, M. (2002), *Agile software development with Scrum*. Prentice Hall.
- Smolensky, P. (1988). On the proper treatment of connectionism. Behavioral and Brain Sciences, 11, 1–74.
- Thelen, E., and L.B., Smith, (1994), *A dynamic systems approach to the development of cognition and action*, Cambridge, MA: MIT Press.
- Van Gelderen, T. (1995) What Might Cognition Be, If Not Computation? The Journal of Philosophy, Vol. 92, No. 7 (Jul., 1995), pp. 345-38
- Yudkowsky, E. (2001) Creating Friendly AI, Singularity Institute
- Yudkowsky, E. (2007) "Levels of Organization in General Intelligence." In Artificial General Intelligence, edited by Ben Goertzel and Cassio Pennachin, 389–501