Design of a Thinking Machine as a Supra-Human Model for Cognition

Tom Tollenaere

Draft – August 7, 2012

Abstract

This paper proposes a design, from a functional point of view, of a machine that can think, and which will be called an Artificial Mind (AM). We define the term 'thinking' for the purpose of this work, in the form of a minimal Mission Statement. Since the only thinking entities we know are humans, human cognition will be used as an inspiration. We establish architectural design principles for AM which follow from the Mission Statement. Further we propose the required functional properties of what we consider the foundation of any AM; this we call Associative Correlative Time Memories (or ACT Memories or ACT-M).

Next we propose the ACT-M model, with one small but fundamental modification, as a model for Human Mind (HM) in the sense of a Philosophy of Mind. Since the model can be used to model machine cognition as well as human cognition we consider this a supra-human model of cognition

Finally we will discuss philosophical issues such as qualia, the mind-body problem and others, and address how these apply within this model of cognition, to humans as well as machines

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.

1. Mission Statement

We want to design a conceivable 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).

Definitions of 'consciousnesses' 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 do:

- We want to apply Occam's razor (REFERENCE) in the sense that we 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. We 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 (REFERENCE TO RUP etc). This is often referred to as 'functional analysis'[1]. 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. Ergo, 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, and since there is no (current) consensus on what consciousness or thinking really is, we need a definition to test against. Applying Occam's razor, we want this definition to be as simple as possible. For the purpose of this mission, 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 I 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.

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, but there is no wide consensus on what 'thinking' and 'consciousness' are, and since this is our machine, we feel free to propose requirements for it as we see fit. This is not dogma, and of course other machine designs might be proposed based on other definitions; such machines may be equally or more interesting as the AMs presented here.

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), ergo we will have to take cues from human

[1] 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 (REFERENCE) or SCRUM (REFERENCE) 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.

reasoning (to the extent this is know or knowable). We 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 (**missing reference**) has proposed a list of drivers that every AI 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.

The Mission is not to build a machine that can pass the Turing Test (REFERENCE); 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. Ergo, 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 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. Ergo, 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 have – precognitive – hardware that performs edge detection, motion detection, Fourier transforms and so forth. (**Missing references**). 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 3.8)

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 way of thinking about AI) (REFERENCE), neither about modifications of synaptic weights in an ANN model (REFERENCES). 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 model will be defined. We will argue that a memory for an AM needs to cover 3 crucial aspects:

- Association
- Correlation
- Time

These will be discussed in the following subsections. 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 input 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). Ergo, an AM needs to be able to deal with the concept of continuously running time.

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.

We propose that it is functionally necessary for an AM memory to store patterns. A pattern is anything spatio-temporally, 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) an spatiotemporal 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. When an AM 'speaks' my name on e.g. e 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.

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 discretisation 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, we 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 my picture. 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 shaved, 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-look-like should become a pattern, which is not identical (but perhaps 'close') to what Tom happens to look like on a given day.

Again furthermore, when the AM sees me like it has never seen me before (say I ran into a door and have a block 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-toms-name-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 fall 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 sound-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-sound-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. We 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.

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 as well).

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, 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. 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. Ergo, 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 propose 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 remained 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 4 will revisit the latter comparison when comparing the ACT-M model with other AI architectures and approaches. In section 6 similarities (and dissimilarities) with theories of (human) Mind 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. We 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 we have proposed above, something must drive it. It must 'want to learn' for example. We 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 points out (Reference Maslow), humans have other needs, both physiological ones such as sex & sleep, and higher level ones such as safety, love & belonging, esteem en self-actualization.

We 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 an AM want to live, as well as negatively, e.g. I do not want the current that feeds my substrate 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 (reference Yudkowski). 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 (Yudkowski) 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. For example, if we want an (electricity based) AM to want to live, then one might consider a need for a hardware current monitor (as part of the sensorium) and a hardware 'driver' which sends an 'unpleasant' signal to the AM once a risk is detected for blackouts. Not only do we not know a priori what would constitute 'unpleasantness' to an AM, 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. In the discussion of our model in section 3 the topic of drivers will be revisited in some more detail.

This concludes the functional design of our AM, as shown in diagram X 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 around 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 will be dealt with in section 4, and a discussion of philosophical issues is deferred to section 6.

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. We 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 5.

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 prior knows that in a stream of 'blahbah 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 loc 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. We will return briefly to this point in section 5.

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 We propose to make a difference between the 'hardware' and the process. In what follows we define:

- Substrate: the 'hardware' which runs the machine. We 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. 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 the substrate. The 'body' of the machine, finally, consists of the substrate, the sensoria and the motoria.

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, 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 affects 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. (REFERENCE (Searle?)). 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 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 (reference Yudkowski and others). 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 distinguished[2]:

- Reinforcement learning: the machine interacts with its environment by producing actions

  These actions affect the state of the environment, which in turn results in the machine

  receiving a reward or a punishment. The machine subsequentially changes its behavior to act in such what 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 is subsequentially expected that once the machine is presented a previously unseen input, it will produce n 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 we 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 (reference & short description of what that is).

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', we currently see no other alternative, and if successful, this will allow for reinforcement learning.

[2] There is one more form of machine learning, where machines are used to train one another; for the discussion here this form is not relevant.

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. 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, sex and so forth.
- 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 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, We propose that an AM equals the weighted sum of its experiences.

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, 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

We propose that an AM can be described in 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 (ref) or processes akin to fuzzy logic (ref). 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 purpose, 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 (ref Chalmers): 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 it's 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 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 it's brain – in that case my 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 Though 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), or more radically, would it have a LOT? And if so, would it have (or perceive) Trains of Thought?

3.6.1. Train of Thought

First of all, we propose that an AM would not have a 'Train of Thought' or TOT. (reference). 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.[3]

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

[3] 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.

- 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 i*s* 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'[4]. 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 to the next section). The 'symbol' CHAIR is the pattern in the AM's thinking in it's 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 '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 an 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 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

[4] 'Fuzzy here does not refer to fuzzy logic; this is discussed further in section 5.

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 nor 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. Determinism & Free Will

Another question which needs investigation is whether an ACT-M based AM would be pre-determined and /or whether it would have free will.

(see notes on free will, as in given the same circumstances, an AL cannot be choose what it chooses)

If we consider an AM as a chaotic system, everything thought has its reasons, be it that these reasons are not simply determinable; virtually every previous experience might be, to some extent, a factor in what is being thought right now. There would be multitudes of – weighted – reasons. There is another potential source of indeterminability, and that would be chance or randomness. In an ACT-M based AM there are 2 potential sources of randomness.

First, if one chooses to believe our universe is not determined then certain experiences of the AM will be attributable to randomness, ergo that the AM is not predetermined. If one chooses to believe in a fully determined universe, in which everything that happens happens because of what happened before, a university in which the outcome of every throw of a dice is predetermined, then one might argue that an AM is predetermined and that is does not have free will.

In today's world where a quantum physical view of reality is widely accepted, a purely causal universe is hard to imagine, hence under those circumstances one cannot but conclude that an AM's thinking is not predetermined.

Second, if techniques such as shaking are used in the AM's training, and such shaking is truly random, e.g. determined by quantum phenomena, the shaking leads to experiences which are truly random and hence the AM cannot be considered determined.

Under both conditions, an AM's thinking would not be predetermined. Not being determined however does not automatically imply free will; in the above reasoning the AM would be at the whim of randomness which it does not control, hence it may not be determined but still dependent on randomness.

However we propose an AM would at least have the impression (or illusion) that it is free to think as it pleases, and has full autonomy in its decisions. We will return to this topic in section 5, in a somewhat wider context.

3.8. 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.

3.8.1. The Turing Test

Although passing the Turing Test was not a functional requirement for AMs, it is still an interesting question to consider. We 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 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 (see elsewhere on my site), 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 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. As a result, the perspective an AM would have – could have – of the world would have to be different from a human's perspective. Ergo we propose that a clever interrogator can always come up with questions that will cause the AM to fail the test. Ergo the Turing Test is not an adequate test for Criterion #3.

Variations to the Turing Test exist. For example, Hamad (reference) 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 nt address or alleviate that fundamental issue with the Turing test outlined above.

For the sake of completeness there are 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.

3.8.2. 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 youself 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 (reference): some agent that is indistinguishable from a normal human being except in that it lacks conscious experience, qualia, or sentience.

Given this, we 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. We 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) (see Nagel what it feels like to be a bat).

To wrap up this section, there is one remaining item to define, and that is what we 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.

4. ACT-M as a Supra-Human Model of Cognition

In sections 2 and 3 we proposed a model for machine cognition: ACT-M based AMs are thinking, self-aware agents within this model. In this section we propose to apply the same model to human cognition. 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

State that part of the discussion will hold for AM & HM!

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.
- 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. (zoek referenties – vanaf wanneer is een brein voldoende groot?). If is fair to state that a foetus is to some extent – perhaps not consciously- aware. It hears sound. It feels. (references)Hence it is subject to experiences, an per out 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 in the womb were not absolutely identical.

(referenties naar wanneer een foetus wat voelt. Referenties naar onderzoek wanneer een foetus 'conscious is). http://www.scientificamerican.com/article.cfm?id=when-does-consciousness-arise (foetus zou 'slapen')

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 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 the author 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 though 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, to fanatic anti)-smoking, 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.

4.3.1. Panta Rhei

In the view of body & mind presented here everything is in constant flux. This can be summarized as follows:

- I think therefore I am (Descartes)
- I think, therefore I change. In our model, 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.
- I am therefore I change, as everything changes over time

- Hence I change, because I am and because I think. 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, being, changing

5. Discussion

In section 4 the ACT-M model was compared to other models of artificial intelligence. In section 4 we extended the scope of the model from machine cognition to human cognition. What remains is a discussion of how the ACT-M model compares to other models of (human) cognition.

5.1. Computationalism

Computationalism considers the mind to function like a computer or symbol manipulator (reference: Putnam, Fodor). Newell & Simon (reference) 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 means 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 believe 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.

5.1.1. 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, we propose is 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.

- 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 if not atomically discernable). 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 using a mantra; the difference being that a mantra should not mean anything. Humans (and we 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 undiscernable.

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 only be an output pattern; it would not be the pattern of thinking of a chair. Actually we propose there is no such thing as 'the pattern' that is isomorph 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 (reference) 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.

We 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.

5.1.2. 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 is 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, with respect to Block's China Brain thought experiment (Block, 1978), no chinaman would know what the China brain is thinking (though each Chinaman might be aware they are participating in the thought experiment). We will

briefly return to Block & Searle's thought experiments in section X.Y (nadat we iets over grounding gezegd hebben).

5.2. 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. Van Gelderen (195) XXX Thelen & Smith (1994) XXX Beer. Boursalou.

Van Gelderen argued that a Watts Governor is a dynamic process, performs as if it executes an algorithm and as if it manipulates symbols. But these symbols are nowhere represented. Point being that a dynamic system is able to perform as if it processes symbolic logic. That is an illusion. Similar to our proposition that Loth is an illusion.

Van Gelder also argued about the importance of time.

5.3. Grounding

In section 3 we 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 and AM, in our model is not a symbolic system, but a dynamic one.

As a side note Harnad (1992) has also proposed an extension of the Turing Test, which he calls the Total Turing Test (TTT). In addition 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 X.Y our 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. Ergo, 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. Searle & Block were right in this respect.

5.4. Connectionism

5.5. Qualia

5.6. Global Workspace Theory

5.7. Multiple Realisability

5.8. Multiple Drafts Model

5.9. Private Language Hypothesis

5.10. Potential Criticism

5.10.1. Naïve

5.10.2. Dangerous

If internal representations are not decodeable, we cannot inspect what a mind thinks and how it goes about doing it.