Some Thoughts on Artificial Intelligence

Guido Gloor Modjib

May 25, 2009

1 Introduction

Artificial intelligence is a subject of great interest in today's computer science, and was for decades. In fact, thinking machines are a dream that is way older than the short history of modern-day computing:

> Q: How long has the human race dreamed about thinking machines?
>
> A: Since at least the time of classical Greece, when Homer's Iliad tells us about robots that are made by the Greek god Hephaestos. Some of them are human-like, and some of them are just machines–for example, golden tripods that serve food and wine at banquets. At about the same time, the Chinese were also telling tales of human-like machines that could think.[1]

I do believe that it is not impossible to create a machine that thinks (properly), and this paper will attempt to show why. Keep in mind however that it actually started out as a part of my diploma thesis (forthcoming), and thus probably can't stand on its own without further explanations or research. I promise it will all make sense[2] once the diploma thesis is out as well.

2 The State of Artificial Intelligence

When we are talking about artificial intelligence, we immediately think of robots attempting to take over the world, or of vast distributed beings in global networks. Our expectations as to what artificial intelligences are or are not are in large parts shaped by popular culture, by ancient fears and emotional reactions. I think it makes sense to look at a few points in regards to what artificial intelligence really is or could be, what its applications are, and what the implications of these things are.

[1] Citation from [14].

[2] Or at least, more sense than it does now.

2.1 Searle: Weak and Strong Artificial Intelligence

I may not agree with the point about strong artificial intelligence being impossible that John Searle makes in his Chinese Room thought experiment.[3] But in the discussion following the article, he does put forth a very important and necessary distinction: The one between weak and strong artificial intelligence. It truly merits a closer look. While I still believe that Searle didn't successfully show that an artificial intelligence is impossible, he certainly pointed out that it is perfectly possible for a system to appear intelligent and not be intelligent at all. More on this in my chapter on the Turing Test, 2.2.

2.1.1 Weak Artificial Intelligence

Weak artificial intelligence is a mere simulation of intelligent behaviour. It might fool us, it might even pass the Turing Test, or it might be different altogether and reliably solve problems us humans can't solve. Weak artificial intelligence basically is any kind of systematic behaviour that seems as if it was intelligent to us.

Interestingly, as I will show, all (or at least, most) of today's efforts at producing artificial intelligence are purely behaviourist. We try to replicate human-like behaviour, and think that the machines that we produce will then automatically become intelligent and (more importantly) conscious. Well, they won't. They might act intelligent, and from a purely behaviourist point of view that might be enough to make them seem intelligent, but arguably there is more to consciousness than intelligent behaviour. And that is exactly the point Searle is trying to make. A chess computer might play chess really well. A pattern recognition mechanism might be able to (somewhat) reliably recognize faces, or text. But those things are merely mechanical puzzle pieces that would have to be at the disposal of something that would make sense of them, and put them into context. While these puzzle pieces are interesting on their own, can be tremendously complex, and replicate intelligent behaviour, they only simulate consciousness.

Without discussion, weak artificial intelligence is the only kind of artificial intelligence we have produced to date, and it might well be the only kind we are able to produce for years or even decades to come, until we better understand consciousness and intelligence itself.

2.1.2 Strong Artificial Intelligence

Strong artificial intelligence on the other hand is true intelligence that includes subjectivity, phenomenality and intentionality, one that can feel qualia. Essentially, it is artificial intelligence after we've solved the hard problem of consciousness, and successfully applied the solution to an artificial system. As Metzinger rightly points out[4], that doesn't necessarily have to be a purely technological

[3] See [19] for his argument, and [10] for my discussion of it.
[4] See [15], pages 206f.

system – it can well be a postbiotic one, with artificially grown neurons, or maybe with quantum computers, or a technology we haven't discovered yet.

The necessary preconditions and possible implementation details for a strong artificial intelligence are the subject of part 2. While I believe that the we will not be certain whether we can create strong artificial intelligence until the very day we manage to create it, I do think that the case is not lost.

2.1.3 Going Beyond Searle

It might become necessary to not only think in terms of strong or weak artificial intelligence, but rather, as always, the world is not only black and white. There are not just systems producing fully fleshed-out intentionality, and systems not producing intentionality at all. We might eventually have to introduce a scale ranging from "not intelligent", passing "weakly intelligent", and ending at "strongest known intelligence".

Arguably, it is very improbable that we'll sit at the top end of that scale. However, the very definition of intentionality is a priori exactly the way our mental states are intentional, so admittedly another system will probably have a hard time surpassing us in the aptitude in producing intentionality.

2.2 Turing Tests and Human Intuitions

When Alan Turing proposed the Turing Test as a measure of the intelligence of candidate machines or other systems (including aliens, I guess), he followed a deep intuition that is widely shared: Us humans are intelligent, so intelligence necessarily has to be human-like. As it turns out, the Turing Test is really good at testing said human-likeness – but the trouble is, it isn't really good at testing anything else.

The Turing Test makes several unsubstantiated claims that narrow its valid application to very few candidate intelligences, and might well classify some systems as "not intelligent" after one or two questions, when they really are – a trivial example are mentally challenged actual humans, but this also applies to other forms of intelligence.

A few things are openly human-specific, and the examiner in the Turing Test would not have to be allowed to ask about them:

- An artificial intelligence doesn't necessarily need a body, and in particular, it doesn't necessarily need a body that bears any similarity to a human body. So questions relating to body parts would necessarily give away the deviate intelligence, unless it was trained to lie, or would have a self-model and body similar enough to the one of us humans.

- A different kind of intelligence would not share a human-like childhood – so questions ranging back to the intelligence's genesis, or development, could give away this fact. Again, of course, an artificial system could be trained to pretend to have had such a childhood.

- Everyday tasks for us humans, like bathing or walking, could be out of bounds or, alternatively, very interesting for another kind of intelligence.

There are, however, a few things that aren't so openly human-specific, and would give away a non-human intelligence very quickly:

- Talking is always also empathically attempting to understand what the person we're talking with might be after – it is, as Metzinger rightly points out[5], always accompanied by us modelling our interlocutor's mind. So we are able to detect very minute changes and inconsistencies, because they don't align with how we're modeling the person we're talking to – they don't align with any way we could possibly see ourselves act or talk in their place. This is, by the way, the reason why some humans might not pass the Turing Test, particularly if they suffer from pathological mental disorders.

- A non-human intelligence could have a different memory structure. It needn't necessarily share the human distinction between episodic and semantic memory, something that probably proved of evolutional benefit for us, for economic reasons. A being that doesn't have the same memory structure would certainly appear alien to us, if not worse.

- Humans have very distinctive patterns of forgetting or suppressing things. We might be able to remember the last few days down to the minute, but our long-term memories have much sparser information, and are often actually factually wrong (often putting ourselves in the most positive light possible[6]). An artificial intelligence would have to model this perfectly, or it would be given away.

- Our mathematical prowess is very limited. Also, we are sometimes really slow. We make logical errors, we make assumptions regarding context or past history of an argument, we make spelling errors, we are irrational and let emotions guide our actions. Those things are very likely to be different in every single other system producing intelligence than our brains, even if they have errors and irrationalities of their own.

Note how for many of those things, not passing the Turing Test would actually be beneficial for at least some aspects of a candidate system.

An entirely different line of argumentation going against the Turing Test, although it overlaps the points I made so far, is the one that not all human behaviour is intelligent, and not all intelligent behaviour is human.[7] So there

[5] See [15] page 366, where he writes: "To develop successful social strategies, on has to internally model those properties of otehr beings that are not avilable [...] through sensory perception. Because the epistemic target is the content of the self-model of another agent, the task is not only simulation but also *emulation*."

[6] See [21].

[7] See [7] – although I personally believe the Venn diagram that can be found as of 2009-05-01 depicting the application of the Turing Test to the set of possible behaviours is wrong and the field "Unintelligent human behaviour" is actually tested by the Turing Test's questioning as well.

are things humans do that aren't intelligent, and there are intelligent things humans don't do. If we let the Turing Test compare a candidate's answers to a human's, we compare it to both the intelligent things humans do and the non-intelligent things humans do, and leave out all the intelligent things humans don't do. Or worse, we decide against the candidate's intelligence because it did an intelligent thing humans don't do.

2.2.1 Turing Test a Sufficient Condition for Intelligence?

Shieber makes the argument that while passing the Turing Test may not be a necessary condition, it is at least a sufficient one. His argument is as follows:

> *Premise 1:* Humans are intelligent.
>
> *Premise 2:* The conversational behavior of humans reveals that (human) intelligence.
>
> *Premise 3:* If an agent has behavior of a type that can reveal intelligence and that is indistinguishable from that of an intelligent agent, the former agent is itself intelligent.
>
> *Premise 4:* Any agent that passes the Turing Test has conversational verbal behavior indistinguishable from that of humans.
>
> *Conclusion:* Therefore, any agent that passes the Turing Test is intelligent.[8]

But there is a problem with this attempted proof: The Turing Test doesn't test what really does lie at the core of consciousness; It is perfectly thinkable that a machine is not intelligent at all, but is very well-trained or hardwired in ways that enable it to answer questions exactly the way us humans expect a human to answer. I'd argue that premise 2 in this syllogism is wrong, and consequently the conclusion is wrong as well. They display a purely behaviourist argumentation, in the true spirit of theTuring Test itself – while it's true that an agent that has behaviour that reveals intelligence is intelligent (I do agree with premise 3), it's not certain there is such a behaviour at all in the first place.

As I'll show when looking at present-day artificially intelligent systems, many early attempts at creating artificial intelligences have gone exactly the behaviourist way of least resistance: Attempting to model human-like behavior, including typing misstakes (which seems to have been very successful), but being not more than simple state machines in the end. This even coined the term "artificial stupidity".[9]

Essentially, if we want to test for intelligence, we can't really use the Turing Test as a proper guideline. As we will find out later however, a proper replacement isn't really in sight.

[8] Citation from [20], page 136.
[9] See [1].

2.2.2 Human-Like Intelligence: Why Intuitions are Wrong

There is another deeper problem lying beneath the failure of the Turing Test to reliably find out whether an artificial system in particular is intelligent or not: What's hard for a human isn't necessarily hard for a machine. In fact, what's easy for a human might just be what's hardest to copy, for two reasons:

1. Our introspective access to how exactly those lower-level things work is very limited. We have a somewhat thorough understanding of what logic is, and how it's possible to emulate the behaviour a human follows when he's playing chess. Since playing chess, for us humans, is hard, we naturally assumed for decades that as soon as a machine could play chess, it was intelligent. As recent developments have shown however, playing chess has nothing to do with being intelligent. Particularly not when all a machine can do in the end is, well, play chess. That intuition is still deeply rooted however, as popular culture shows, the idea of a machine playing chess just well enough that it might turn sentient does still exist, I was reminded of this when watching an episode of the television series Terminator SCC.[10] Philosophers are not immune to popular culture, arguably. Reality on the other hand has machines that play chess follow strictly logical patterns, predicting future behaviour of a human (or other machine) player on the basis of vast databases of past chess matches and, even more important, probability tables – but not displaying the slightest hint of phenomenality or even empathy.

2. Maybe even more importantly, evolution had millions of years to fine-tune exactly how, for example, data from nerve endings in an eye had to be processed in order to allow for pattern recognition, colour vision, velocity information of moving objects within the field of vision, and various other bits of information – each ranging from simple to complex. For example, we are able to recognize a human face at first sight, from various angles, and without problems. That is an amazing feat and one of which we've only just begun to emulate the basics, but it comes very natural to us. We're just built that way for countless generations, and inherited much of it from our ancestors. What is easy for us must not be easy for a machine, and tasks that seem complex to us can be trivial for artificial or differently built biological systems.

Keep in mind that while thoroughly fine-tuned and adequate for granting us better chances of survival as a species, our perception of the world is anything but perfect. Metzinger introduces the term of "autoepistemic closure" to mean "a closure or boundeddness of attentional processing with regard to one's own internal representational dynamics".[11] This is due to the nature of the transparency constraint I will introduce later – we never are actually in contact with the world, but merely with a representation of the world, a mental model.

[10] In this series, the artificial intelligence that eventually rises to power and attempts to destroy all human life is born out of a highly advanced chess program.

[11] Citation from [15], page 57.

This is true even while our bodies (which we perceive as and identify with as "ourselves") seem to physically interact with the physical world; the existence of a mental representation of that interaction is all that is ontologically granted. In other words, seemingly physically interacting with the world doesn't assure us that we're not merely a brain in a vat. Metzinger concludes thus:

> Phenomenal representation is that form of mental simulation, the proper function of which consists in grasping the actual state of the world with a sufficient degree of accuracy. In most cases this goal is achieved, and that is why phenomenal representation is a functionally adequate process. However, from an epistemological perspective, it is obvious that the phenomenal "presence" of conscious representational content is a fiction, which could at any time turn out to be false.[12]

2.2.3 A Replacement for the Turing Test?

I attempted to show why the Turing Test isn't really fit for finding out whether a system truly is intelligent or not in section 2.2 already, and argued against the underlying behaviourist reading of intelligence in section **??**. I think with the information we have now, it is even more apparent than ever before:

- Assuming we solve the hard problem of consciousness, with the approaches we're having now or maybe with other, radically new ones.
- Assuming we have a system that does have phenomenal experiences of qualia, from a subjective point of view.
- Assuming the system experiences its relation with the world as intentionality.

Assuming all those things, we would have to say that it is intelligent. But even such an intelligent system wouldn't necessarily pass the Turing Test. And on the other hand, nothing tells us that a system passing the Turing Test satisfies even one of those constraints. A further complication is that intelligence and consciousness are not all-or-nothing properties, but rather come in shades of gray, making them even harder to test.

But, is a replacement for the Turing Test in sight? Arguably not, not an easy one. If a system seems to show subjectivity and introspection, we can at most say that it acts like it is intelligent. But on the other hand, can we say anything more than that of fellow humans?

There is empathy, probably partially genetically hardcoded, that makes it easy for us to believe that other beings similar enough to ourselves are intelligent, by running a mental simulation of their point of view, their goals and possible actions. But we only have to look at how pretty much everything was antropomorphed in one myth or another, or how prone us humans are to talking

[12] Citation from [15], page 57.

with plants and animals; that empathy relation isn't restricted to other humans. It is nothing but a mechanism that proved to be adequate for reproduction.

Behaviourist theories do have a point: Epistemically, it is impossible to tell anything beyond whether a system seems to be intelligent from the way it seems to act.

So, going back to when I first wrote about the Turing Test and what it tests in chapter 2.2: A replacement for the Turing Test would have to look for intelligent behaviour humans do, and intelligent behaviour humans don't do, and disregard both unintelligent behaviour humans do and behaviour that is only intelligent if a human is doing it (in the sense of the behaviour being adequate for the evolutionary position us humans find ourselves in).

And a target system obviously would not have to show all those intelligent behaviours fully in order to be classified as intelligent – after all, if the test would require that, us humans would obviously not pass it – but rather have a scale of sorts.

Where would we draw the line between a conscious being and one that isn't conscious? How would we test for those of Metzinger's multilevel constraints that make a system minimally conscious? Can we even do that, or are we at yet another point where epistemology of a subject matter fails at fully grasping its ontology?

2.3 Today's Kinds of Artificial Intelligence

In today's implementations of artificially intelligent systems, there are quite a few approaches – the most prominent and well-known among them being scripts and neural networks. All of them can pretty reliably produce weak artificial intelligence, but arguably, additional or different structures will have to be introduced in order to create a strong artificial intelligence.

In addition to explaining how those approaches work, I will also point out some methodological specialties of artificial intelligence research and then show up where today's weak artificial intelligence is actually already used.[13]

2.3.1 Formal Approaches

The following formal approaches are currently used for artificial intelligence:

- *Scripts:* This is the most trivial approach. And although and maybe particularly because it is trivial, it is also the most widespread. Most artificial intelligence systems are simple "if this happens, do that" kinds

[13] Much of this information has be verified and researched with the help of online sources, including but not limited to Wikipedia. While admittedly the Wikipedia, unlike other online encyclopedias like the Scholarpedia, is not peer-reviewed on a professional basis, particularly for technology-heavy subjects like artificial intelligence it is arguably the most comprehensive and probably most complete resource available today. As such, I decided to use it and cite it as source despite the lack of acknowledgement it gets in some more traditional fields of philosophy and other sciences. I also used it as reference for some philosophy topics, but took great care to go beyond it in fields that are important for my argumentation.

of scripts, executing predefined orders after certain trigger events occur. Of course, this is also the least interesting and most predictable kind of artificial intelligence.

- *Decision Trees:* A decision tree is a rooted tree (also called aborescence in graph theory, which is a special kind of connected directed acyclic graph). Starting at the root of the tree, decisions are made whether to go to one or the other child node, until a state marked as "final" is reached. Essentially, decision trees are more sophisticated and multi-level scripts. Individual decisions can include thorough calculations, and access to databases of pre-calculated decisions (for example, when probabilities are involved). Noteworthy subtypes of decision trees are the minimax tree and its successor, the alpha-beta tree.[14]

- *Neural Networks:* These attempt to model a simplified animal's brain. Like the dendrites and axones of a real neuron, modelled virtual neurons have usually two inputs and one output each, and "fire" if the weighed input values exceed a certain threshold. Unlike real neurons in actual brains however, those virtual neurons aren't cross-linked multidimensionally, but rather arranged in one or few layers, or in set patterns (like it's the case with feedforward and recurrent neural networks). Before the neral network can do anything, it has to be trained, with thousands of iterations where the weights of the involved neurons are adapted.

- *Kernel Methods:* Data is mapped into a high dimensional feature space by means of complex statistical methods.[15] This produces very good models for all kinds of pattern recognition tasks, and is computationally efficient because of the use of kernel functions, which don't actually compute all the data but rather only inner products (generating hyperplanes, where the normal vectors are then compared to find similarities between different sets of data).

- *Other Approaches:* More approaches are widely used.[16] They include k-nearest neighbor algorithm (for simpler pattern recognition), Gaussian mixture models (for classification of data), or naive Bayes classifiers (also for classification of data – often used for example in automatic email spam filters).

Most actual, non-trivial artificial intelligence systems are actually a mix of the above. For example, a neural network might be used for making decisions in a decision tree. Or scrips invoke a neural networking algorithm once certain conditions have been met.

It is notable that among all those approaches, even the one that comes closest to actual brains, neural networking, doesn't fully attempt to replicate the logical

[14] See [4].
[15] See [11].
[16] See [5].

structure of the brain but rather is only inspired by actual real neurons.[17] Also, none of the approaches attempts to create intentionality or consciousness in any stricter sense in the first place – they all are purely behaviourist attempts at duplicating the graspable effects of intentionality.

2.3.2 Methodological Specialties

Some methods were developed that are mostly used in artificial intelligence.

- *Fuzzy Logic:* Inspired by truth values used by humans in everyday tasks, fuzzy logic (unlike binary or ternary logic) does not depend on a fixed set of truth values, but rather has statements use a continuous range of 0 (false) to 1 (true) as their truth values. This allows contradictory statements like "it is cold" and "it is warm" to be both partially true (like they are for us for temparature ranges that could be interpreted both ways), leaving more options for behaviour control of a system and increasing its capability to classify and process imprecise information.

- *Pattern Recognition:* This is widely seen as one of the most complex fields of today's research into artificial intelligence, and closely related with all perception-type tasks. This includes face and object recognition, speech or text recognition, but is also used in data mining for classification and clustering of data patterns. Pattern recognition is, by the way, an area us humans are particularly good at, as very simple thought experiments show: We are able to decide whether something is a tree or not within fractions of a second, whether we've ever seen this particular kind of tree (colour, shape of leaves, bark structure) before or not.

These methods will of course have to be used in the context of a strong artificial intelligence. They will however then probably not form the core of the project, but rather serve as suppliers – preprocessing data in ways that makes it easier to conceptualize on a higher-order abstraction level, like we're used to working with in our own conscious experience.

2.3.3 Fields of Application

There are few fields in our everyday life where we aren't in contact with one kind or another of artificial intelligence. Most of these applications however are such weak forms of artificial intelligence that they wouldn't fool the most innocous researcher.

- *Research:* This is trivial. Artificial intelligence is the focus of many institutes and departments of research centres and universities around the world. While experimental research is no longer a focus, directed development into specific fields (like the ones below) is one of the main focal points of today's computer science research.

[17] This is true at least in its regular form. Alternate approaches like the one of [22] do attempt to model actual neurons, but are limited to specific applications so far.

- *Science (Expert Systems):* Expert systems are programmed in ways that replace or complenent an actual human expert. They are used in a wide range of application fields, for diagnosis and evaluation, classification of a wide range of input parameters into distinct findings. Examples of fields that expert systems (along with other techniques of artificial intelligence) are used in include medicine, molecular biology, and high-energy physics.

- *Everyday Computing:* It is common in many industries to scan documents and apply optical character recognition (OCR) to them, in order to archive or further process those documents. The algorithms used for that are pattern recognition algorithms in a first step, and then a semantic analysis based on databases of word combinations and their probabilities. The exact nature of those steps is different in every software, but they are applied artificial intelligence research. Other fields of application are speech, (mouse or touch screen) gesture and handwriting recognition. Even search engines like Google order search results according to smart ranking mechanisms that include contextual and semantical analysis.

- *Mapping and Pathfinding:* A car's GPS device can find shortest paths, most fuel-efficient paths, and it can lead the driver towards his destination quite reliably with automated voice instructions.

- *Criminology:* Comparing a scanned set of fingerprints with a database of stored prints is no easy task for a man (due to the sheer volume of the stored fingerprints), and it is no easy task for a machine for completely different reasons (due to the difficulty of seeing when two prints are equal) – this is another field where pattern recognition comes into play.

- *Household Items:* Simple tasks like a rice cooker automatically switching off when the rice is done, or a water boiler switching off when the water is hot, are hardware implementations of simple scripts. A microwave's or oven's programs are more sophisticated, and can even include further sensor data in addition to user input and heat sensors. Finally, we all heard of the fictive refridgerator that's "aware" of its contents, and makes shopping lists when necessary.

- *Customer Relationship Management:* There are two major subgroups here; for one, the automated telephone and customer classification systems can be looked at as a simple form of scripted artificial intelligence. The more interesting part is data mining, where vast databases of customer-related data are searched for patterns and positive, often predictive analysis. Credit card companies for example use such systems to determine whether a given transaction is typical for the user in question, and bonus systems in big retail chains are often introduced for obtaining reliable customer data for these data mining tasks.[18]

[18] It is an interesting observation that customers are often very willing to part with this data in exchange for only relatively minor rewards. We are practically giving away our shopping patterns, data that is worth a huge lot to retailers, for free.

- *(Computer) Games:* These usually have one of two of the aforementioned approaches as artificial intelligence. Strategy games and massive multiplayer online games usually have nothing but mere scripts: If somebody steps on the wrong bit of floor, there will be fire. After exactly 240 seconds, the chief adversary will enrage. Or, once this enemy has been neutralized, that wave of new enemies will spawn. Chess, Go, even Tic Tac Toe and other board game adaptations on the other hand are predestined for decision trees, in fact, they were what the alpha-beta pruning method was invented for in the first place.

- *Robotics, Toys and Novelty Items:* These fields are closely related, as the most prevalent representation of artificial intelligence in toys are robotics. Robotics themselves then do have many subfields that involve findings from artificial intelligence research: locomotion, object manipulation, localization and mapping, to name but a few.

A very interesting thing is the so-called AI effect: As soon as an application of a (weak) artificial intelligence technique enters mainstream and is widely used, it is no longer regarded as "proper" artificial intelligence – although only a few decades ago, many of the machines we use daily would be considered highly intelligent, they just work, from our today's point of view.

Of course, this is partially attributable to the demystification of all things explainable. It is an interesting thought that there is a parallel to how many fields of magic and miracles made room for science a mere few centuries ago, and there might be parallels to research into consciousness as well.

2.3.4 Seemingly Strong Artificial Intelligence

There are two main directions of research that attempt to make artificial intelligences seemingly stronger:

- *Visible Emotions:* A robot that can display facial expressions similar to a human's (like Kismet could already in 1999[19]) seems more intelligent. Funny enough, this works (to some extent) even if the resulting robot is thoroughly trivial. We encountered empathy before, in chapter 2.2, and apparently relatively simple projection surfaces already work for that empathy.[20]

- *Common Sense:* When humans communicate, or even when we perceive events in the world, we rely on a large body of common sense judgements and facts. While it is well-known that the term "common sense" implies a thorough commonality that does not exist (cultural and individual differences can be rather large, which frequently is the reason for

[19] See [2].

[20] See [12], however keep in mind that the Uncanny Valley, which has first been proposed already in [18], keeps us from bonding with some human-like projection surfaces – "as robots appear more humanlike, our sense of their familiarity increases until we come to a valley," where robots look like zombies to us.

misunderstandings and disputes), those differences are usually on quite a high level – a lot of data we just know is very basic. The kind of "rocks fall down if dropped" knowledge. There are various attempts to collect and classify this data, and use them for diverse applications – among them, stronger artificial intelligence.[21]

While these approaches are nice and do not attempt to be more than behaviourist, a true strong intelligence needs more than these. Let us now try to investigate what more there is to consciousness.

2.4 The Need for a New Approach

Minsky is a proponent of an older tradition of artificial intelligence. Let's hear what he has to say about genetic algorithms and neural networks:

> I'd like to argue that if you're interested in artificial intelligence these days you're exposed to a lot of arguments that ... well, in the early days, people tried to do things with symbolic AI, sometimes contemptuously called old-fashioned AI, and that was too rigid and rule-based and mechanical, and it had to be programmed. What happened starting around 1980 was that most AI researchers tried to move in the direction of making machines smart without programming them, by using neural networks or genetic algorithms or statistical models. The value of that is that you're hoping that your machine can learn to be smart without your knowing how it does it.[22]

While neither him nor me want to diminish the value of the many achievements of artificial intelligence research in recent years, he does have a point: We cannot expect from our machines to learn to be smart without us knowing how not only it'll do that, but also how it could even learn that. We have to set some guidelines, and for that we have to find out what those guidelines will have to be first.

The rest of my thesis will revolve around two things: Finding out what recent research in the field of philosophy of mind has to tell us about how consciousness happens in animal (and in particular, human) minds, and then attempt to find ways in which those findings could be applied to research in the field of artificial intelligence.

3 Ethical Concerns

3.1 Artificial Intelligences as Agents

For weak artificial intelligences, the ethical situation couldn't be any clearer: They are not autonomous agents, so the moral responsibility for their actions

[21] A very nice overview over current efforts can be found at [3].
[22] Citation from [17], 8:40-9:45.

lies fully within the realm of their creators. As soon as an artificial intelligence is a strong artificial intelligence and satisfies those of Metzinger's multilevel constraints necessary for agency however, it becomes an agent.

As soon as that happens, there is a big difficulty in assessing the moral situation we find ourselves in. After all, drawing parallels to our own legal system, a child isn't morally responsible for its own actions until it has reached a certain age – one that has historically shown to be one in which the human can be considered an adult, and where the parents' influence is sufficiently small so the human can now be considered a full agent, one responsible for their own actions.

I will have to admit that this very problem is one that has concerned me for a few years already now. From an intuitive point of view, it is probably safe to assume that anything but the creator having moral responsibility for all of the artificial intelligence's actions will not be accepted by the society. If we assume full-blown agency for hypothetical future artificial intelligences however, that is however not really fair for these, as they would have to be considered not only agents but also persons.

This thought certainly warrants some further discussion, and the subject will eventually require normative guidelines. I am not sure if those can be anything but arbitrary at this point.

3.2 Metzinger: Negative Utilitarianism

Negative utilitarianism is the name Thomas Metzinger gives the idea that we, as moral beings, should not attempt to maximize welfare like traditional utilitarianism does, but instead minimize suffering. And since he points out how "Suffering starts on the level of PSMs", but doesn't make an equivalent point for wellbeing, I believe he has a more negative view on the world than I do. First though, this is his definition of the concept of negative utilitarism:

> People differ widely in their *positive* moral intuitions, as well as in their explicit theories about what we should actively strive for. But in terms of a fundamental solidarity of all suffering beings against suffering, something that almost all of us should be able to agree on is what I will term the "principle of negative utilitarianism": Whatever else our exact ethical commitments and specific positive goals are, we can and should certainly all agree that, in principle, and whenever possible, the overall amount of conscious suffering in all beings capable of conscious suffering should be minimized.[23]

I think what he means is this: What we can hope for when creating a postbiotic artificial intelligence, is at most a lucky strike, something that just happens to work without us knowing how or why. Indeed, the potential for a postbiotic system that is created just by growing random parts of a human brain

[23] Citation from [15], page 622, emphasis his.

and connecting them in experimental ways being a horribly misformed kind of consciousness at first is huge.

Metzinger's negative utilitarianism seems to point primarily at those systems: They suffer consciously because they are built from components meant to be in an entirely different kind of organism, they have phantom limb pains where there are no limbs, they intuitively know they should have eyes when they don't have any, they attempt to localize their heart when there is none. Indeed, we should not do any advances towards that kind of postbiotic system.

But Metzinger, when generalizing his stance, falls into the trap already utilitarianism itself fell: He goes too far in his normative simplification. His overreach is two fold:

- Taken by his word, we should stop reproducing now, and we should (painlessly) stop all other forms of conscious life from reproducing as well. As the enormous amount of suffering that future generations in the next tens of thousands of years (assuming we as a species survive beyond the next few hundred) will have to endure if we live on is obviously more than what little one generation is even able to suffer.

- There is not just suffering, but also happiness. If an artificial intelligence is not created by random tampering, but by careful engineering, it should perfectly be able to experience joy just as well as suffering (more on this later when I'll talk about emotions). In fact, if it is created by any means similar to today's pretty successful neural networking attempts, it will necessarily have both positive and negative feedback mechanisms built-in, which will, if translated to our own phenomenal world, be experienced as joy and suffering respectively. Should we deny that system the joy it could experience only because it would also have to experience suffering?

3.2.1 Minksy: From Pain to Suffering

A theory that Minsky presents is pretty interesting in this context. It is his theory as to why some animals (and in particular, us humans) are able to suffer. Because while it is fairly obvious how we can feel pain, suffering (so Minsky says convincingly) is not the same thing, not even a different degree of the same phenomenal experience. But what's the difference between those two and other, similar unpleasant emotional experiences like sadness, frustration, anguish or torment? And why can't we seem to be able to get suffering out of our heads?

What is it that Metzinger's negative utilitarianism wants to minimize in the world? Minsky offers a theory:

> What could make the sensation called *"Pain"* lead one into the state we call *"Suffering"*? [I propose] a theory for this: any pain will activate the goal *"Get rid of that pain"* – and achieving this will also make that goal go away. However, if that pain is intense and persistent enough, this will arouse yet other resources that tend

> to suppress your other goals – and if this grows into a large-scale "cascade," there won't be much left of the rest of your mind.[24]

That suppression was necessary from an evolutionary point of view, because it would simply be too dangerous to be able to suppress such important urges as hunger, tiredness or pain for other goals. Only on a higher evolutionary level of consciousness would it potentially make sense in some situations. Similarly, we cannot directly switch on other emotions like anger or love. As Minsky points out, "those who were able to do such dangerous things left fewer descendants than did the rest."[25]

Persistent pain then continues to interrupt and override our other goals, calling for attention (metaphorically speaking), and what actually is suffering then is not just "a lot of pain," but rather "the states that result when this [pain] escalates into a large-scale cascade that disrupts all one's usual Ways to Think."

The reason why we suffer so much however seems to be simply that our ancestors were built in ways that prioritized plans about avoiding pain – that was necessary for survival – at the price of the minor inconvenience of disrupting other thoughts and goals. And today we pay the price for that. After all, evolution never planned ahead:

> Our ancient reactions to chronic pains have not yet been adapted to be compatible with the reflective thoughts and farsighted plans that only later evolved in our brains. Evolution never had any sense of how a species might evolve next – so it did not anticipate how pain might disrupt our future high-level abilities. And thus, we came to evolve a design that protects our bodies but ruins our minds.[26]

I do believe that it should be possible to have an artificial intelligence capable of more direct control than we have, while at the same time having enough overriding goals from sources important for survival, similar to what pain and hunger would be for us humans.[27] However, that's an entirely different discussion from the one this thesis is about, and thus I will leave this subject with the following quote: "I'm so something that I can't remember what it's called."[28]

3.3 The Adaptivity Constraint and Artificial Intelligence

Metzinger also heeds another warning: He says that in order to be truly intelligent, a system has to fully satisfy the adaptivity constraint. He even goes further and quotes Davidson's "The Swampman",[29] with an argument similar to the ship of Theseus. In the thought experiment, Davidson is standing near a tree in a swamp, when lightning strikes the tree, Davidson is disintegrated, and by pure coincidence the tree is turned into an exact physical replica of Davidson.

[24] Citation from [16], page 66, emphasis his.
[25] Citation from [16], page 93.
[26] Citation from [16], page 79.
[27] See [16], page 93.
[28] Citation from [16], page 69, citing Miles Steele (age 5).
[29] See [15], page 206.

The goal of that thought experiment is showing that an exact replica of a human being might move, think, talk, and argue just as the original, it might even have the exact same kind of phenomenality. But what it lacks is the original Davidson's intentional content – Metzinger writes:

> [...] for instance, it has many false memories about its own history be they as conscious as they may. The active phenomenal representations in Swampman's brain would [...] *not* satisfy the adaptivity constraint, because these states would have the wrong kind of history. They did not originate from a process of millions of years of evolution. [...] It would [...] still be consciousness in a weaker sense, because it does not satisfy the adaptivity constraint [...].[30]

3.3.1 Adaptivity Questioned

I'll be very blunt: I do not believe this argument has much merit. If it had,[31] every subsequent generation of humans would be less conscious than the former: After all, much of what we act and live by is not genetically hardcoded, but rather learned and perceived. All those things that we take as signs of proper intelligence beyond basic empathy (being able to play chess, to talk with each other in a common language, to read and write) do not have this kind of "proper", system-internal history – they were introduced from the outside, from our parents and teachers, our environment, our experiences. And that is the very thing Metzinger tells us is what is wrong with a hypothetical postbiotic intelligence – even if it were built from actual organic tissue, and even if it had gone through an evolutionary process of its own:

> [...] it would still follow that these postbiotic phenomenal systems would only be conscious in a slightly weaker sense than human beings, because human beings were necessary to trigger the second-level evolutionary process from which these beings were able to develop their own phenomenal dynamics. From a human perspective, just like Swampman, they might not possess the right kind of history to count as maximally conscious agents.[32]

3.3.2 Memes (and Temes) are Adaptive

Essentially, the adaptivity constraint in its strict form denies the importance and evolution, maybe the very existence of memes.[33]

- Memes are a second replicator after genes. Information that is copied from person to person, "that which is imitated".
- Why do they spread? They're selfish information, they get copied if they can.

[30] Citation from [15], page 206.
[31] I do owe this point to my good friend Simon Bünzli.
[32] Citation from [15], page 207.
[33] See [9] and [8].

- Memes are using us as meme machines to get copied. Selfishly in that they can't care about anything else.
- Completely new view of human origins and what it means to be human. Cultural evolution, what makes us different from other species: There are two replicators now. People began imitating everything, not just beautiful and true things.
- Big brains are driven by the memes. Language is a parasite that we've adapted to, in a now symbiotic relationship.
- All other species are gene machines only. We alone are both gene machines and meme machines.
- Technological memes (or "temes"): Memes are outsourced from humans, replicating on their own technologically.
- Every new replicator is dangerous!

3.3.3 A Weaker Adaptivity Constraint

When Metzinger formulated his adaptivity constraint, he seems to have been mainly driven by one of his background assumptions, one that he explicitly states at the very beginning of his book: Teleofunctionalism.[34] He introduces his commitment like this:

> [...] representata have been specified by an additional *teleological* criterion: an internal state X represents a part of the world Y *for* a system S. This means that the respective physical state within the system only possesses its representational content in the context of the history, the goals, and the behavioural possibilities of this particular system. This context, for instance, can be of social or evolutionary nature. [...] It is for this reason that we can always look at mental states with representational content as instruments or weapons. If one analyzes active mental representata as internal tools, which are currently used by certain systems in order to achieve certain goals, then one has become a teleofunctionalist or a teleorepresentationalist.[35]

Teleology may not be the best approach for everything (and in fact, when Sehon used it along with the term "supervenience" that Metzinger also uses, he constructed a theory that I can't find myself agreeing with at all). But when Metzinger uses it, he seems to mean that he looks at the things our brains do in a way that assumes they must be good for something – because if they weren't, they wouldn't have been a fitness criterion for evolution, and consequently wouldn't have survived the last few million years.

[34] See [13].
[35] Citation from [15], page 26, emphasis his.

I do agree with that. I do not agree with million years of evolution being a necessary prerequisite for something being a fitness criterion however, nor do I agree with the assumption that merely because a fit system has been duplicated (as Davidson's Swampman was) it suddenly loses some kind of metaphysical fitness precondition. But that is exactly what Metzinger's reading of the adaptivity constraint presupposes: That there is some secret ingredient to proper intentionality after all, a ghost in the machine reminescent of dualist theories despite all our efforts to banish it.

I do not deny the importance of evolution for today's kind of intelligence. It was essential, both as a driving force and a shaping constraint. However, even from a purely teleofunctionalist point of view, "proper history" cannot be a criterion for the degree of intentionality a system has. "Fitness for purpose" however can be, even if we're talking about intuitively non-replicable things like emotions (that, as we saw, are nothing but "logic of survival" indicators) or qualia – and as such, I'd like to propose a weaker form of Metzinger's adaptivity constraint. While an organism may pursue the goals of its ancestors, while this may be important for all of today's biological sentient systems (including humans), it is not necessary for proper intentionality.

Essentially, I want to take Metzinger's normative approach to what good represeentata are and slightly alter it. It is:

> Phenomenal representata are *good* representata if, and only if they successfully and reliably depict those causal proerties of the interaction domain of an organism that were important for reproductive success.[36]

And, the change I propose is simple: Good representata are the ones that "depict those causal properties [...] that *are* important for *successful survival*." The two changes thus are thus:

- "Are" over "were", because while the history of a system's genesis might be interesting for evolutionary purposes, it is not required for adaptivity.
- "Survival" over "reproduction" because a postbiotic or artificial system need not necessarily be able to reproduce, particularly if it is not plagued by death of individuals and phenotypes like us humans are.

4 A Minimally Conscious Artificial Intelligence

If the Church-Turing Thesis[37] turns out to be correct (which it seems to, currently) and a Turing machine can duplicate any other logical machine's behaviour, what implements the strong artificial intelligence we are looking for can just as well be a computer with a structure very similar to what we already have today, or theoretically even one of today's computers. I have argued for this point in an earlier paper[38], but feel a need to expand on my argumentation

[36] Citation from [15], page 203, emphasis his.
[37] See [6].
[38] See [10].

back then in the light of more research on my part:

According to the Church-Turing Thesis, everything that can be expressed in any machine can be expressed in a universal Turing machine. According to any kind of remotely physicalist reading, our brain is nothing but a biological machine. Thus, there can be a turing machine reimplementation of all the things happening in our brains – even when seeing how we'd have to reduce thousands of parallel processes into only a few; calculus would be necessary, massive multitasking, but it would be possible. Theoretically, everything happening in the brain could be replicated in a Turing machine.

But: While theoretically possible, an exact replica of the human brain (and some other things that I'll point out in this chapter) is not practicable. Particularly not in a single-threaded Turing machine, and particularly not if that Turing machine would attempt to replicate the exact way a brain functions – that would require a mechanic understanding that is way beyond today's most advanced research programs.

A possible way out of this seem to be the postbiotic systems that seem to be all the rave among philosophers. There are the ethical problems with such an approach that we have already looked at, and also practical problems in that we would still require knowledge of how things interoperate in our brain. But obviously, a postbiotic approach would have the benefit that we could build on things that are already proven to be fit for enabling consciousness.

4.1 Preliminary Goals

Whatever the means, we have to know what we're aiming for when using them. After all, means are nothing but tools in this strive. A preliminary goal definition could thus be:

> We do not have to replicate all of the vehicle properties, it is enough if we replicate the content properties of a system enabling consciousness in order to create strong artificial intelligence.

There still are two shortcomings in this preliminary goal definition though:

- Metzinger rightly points out that vehicle and content, in the case of human brains, aren't easily distinguishable and interwoven to the extent of not even being practically separable.[39] A particular dificulty in this respect is that we don't even have access to all content properties, because they're not all globally available and most of them are transparent to introspection – as we have seen, the border between the conscious and the unconscious in us humans is neither clear nor rigid thanks to transparency not being an all-or-nothing property of mental states, but coming in sometimes even changeable shades of gray.
- Not all of either the vehicle or content properties will be necessary for consciousness, there are many things that might even satisfy the adaptivity constraint and are necessary for us humans to survive, but are not

[39] Reference!

> necessary for consciousness itself. Both the emotions of love and jealousy fall into this category.

So a first revised version of our goal definition:

> We do not have to replicate all of the vehicle and content properties, it is enough if we replicate an adequate subset of the vehicle and content properties of a system enabling consciousness in order to create strong artificial intelligence.

The hard work will be figuring out what subset that is. However, as Minsky rightly points out, this might not even be necessary. He says:

> I'm not trying to find out how the brain works, I'm trying to find out how to make something that can think more or less the way we do.[40]

Only once we have machines that think more or less reliably in similar ways that humans do can neurologists step in and find out how plausible those machines are in terms of explanatory force when it comes to the human brain. Of course, there is a slippery slope here: Once we start accepting systems that can think only more or less the way we do, and not exactly the way we do, there is no way of telling whether those systems actually are strong artificial intelligences or not. We arrive at mostly behaviourist argumentations again, something that we hoped to leave behind.

Furthermore, as I pointed out in chapter 2.1.3, there probably is no sharp distinction that can be drawn between weak and strong intelligence – so the best we can hope for is getting ever stronger artificially intelligent systems, until their strength eventually surpasses that of us humans.

So I would like to propose a final, again slightly changed and less ambitious goal proposition:

> We do not have to replicate all of the vehicle and content properties of a system enabling consciousness, it is enough if we construct a system with vehicle and content properties whose relative behaviour shares sufficient similarities to the vehicle and content properties of a system enabling consciousness in order to create a stepping stone towards strong artificial intelligence.

It might be necessary to emulate some vehicle properties of the constructed system by means of suitable (transparent) content properties.[41] These are merely implementatory details, although they are probably the hardest part when it comes to actually implementing such an architecture.

[40] Citation from [17], 1:16:25-1:26:33.

[41] In other words, the vehicle will probably have to be a virtual vehicle. I will talk about this in the context of virtual organs in chapter 4.3.4.

4.2 Multilevel Constraints

Metzinger already does describe a number of different strengths of phenomenal consciousness – we saw those in chapter **??**.

4.2.1 Application of Metzinger's Constraints to AI

4.3 Virtual Organs

I hereby put forth a rather important and probably controversial thesis: In order to be truly intentional, an artificial intelligence will need a special kind of virtual organs. Without these, there can be no consciousness.

4.3.1 Reasons for Virtual Organs

As Metzinger rightly points out, consciousness is not a state: It is a process. He illustrates this nicely when it comes to intentional content (which explains phenomenal content as well, as "phenomenal content is a special aspect or special form of intentional content"[42]):

> Intentionality is not a rigid abstract relation from subject toward intentional object, but a dynamical physical process pulsating across the boundaries of the system. In perception, for instance, the physical system border is briefly transgressed by coupling the currently active self-model to a perceptual object [...]. Intended cognition now means that a system actively – corresponding to its own needs and epistemic goals – changes the physical basis on which the representational content of its current mental state supervenes.[43]

This is important because like every other process, it is a transition from one (very transient) state to another, induced by input stimuli (and producing certain output stimuli). Phenomenality, and thus consciousness, is a byproduct of that process, it is not a property of starting or ending states of the process.

And those input stimuli do come, in the case of human consciousness, from our organs. They come largely from the ones we know as senses of course. In the case of humans, those mainly include sight, hearing, taste, smell, touch, balance, temparature, pain, but also proprioception (the kinesthetic sense) and input we receive from internal organs such as the lungs or the bladder. Other organisms have an even wider range of senses: Electroception (direct perception of electric fields), echolocation (sonar-like capabilities, for example in bats), pressure and current detection, among others.

Notably, all these stimuli start out external to consciousness, although they might be internal for the body anyway. They are perceived and made into consciousness-internal stimuli through organs: Eyes, skin, ears, nose, tongue, aforementioned lungs and bladder.

[42] Citation from [15], page 111.
[43] Citation from [15], page 114.

In that sense, any artificial intelligence will require such organs: Means for external stimuli to reach the places where they can initiate processes that produce consciousness.[44]

4.3.2 Consciousness With Only One Virtual Organ

Of course, a consciousness without a wide array of virtual organs like the one I just suggested is thinkable: Assuming that the physical requirements for all the processes that will produce consciousness are in place, including those requirements for the supervening "mental content" structures. And assume that there are stimuli without a plethora of virtual organs anyway.

Examples for this could include clock ticks that are generated with the only one remaining virtual organ, a clock. There need be at least this minimal form of virtual organ, I'd argue: As without something like this, there won't be any input stimuli, and consequently there won't be any processes, and thus no consciousness.

If we were to produce such a minimal one-virtual-organ configuration, life for the resulting artificial intelligence would be incredibly dull:

- If it was created with the aid of virtual organs before, and its phenomenal self model would include those virtual organs and expect input from them, it might start making up inputs from (more or less) random fluctuations and hallucinate, like us humans do in sensory deprivation situations.
- If it was created with just that one virtual organ in the first place, it would probably never form any interesting form of consciousness at all, as it would not be able to relate to anything, and wouldn't even be able to form some minimal kind of self-world boundary that is so important for the phenomenal self model.

Keep in mind that these thoughts are purely speculative, but I do think they are interesting nonetheless.

4.3.3 Parallels to Metzinger's Virtual Organs

Metzinger already does postulate virtual organs, in various places. One of them is particularly distinct:

> There are two kinds of organs: permanently realized organs like the liver or the heart, and "virtual organs." Virtual organs are coherent assemblies of functional properties, only *transiently* realized, typically by the central nervous system. Classes of integrated forms of phenomenal content [like a book in your hand] are classes of virtual organs. [...]
>
> But not only simple presentata, phenomenal *representata* are virtual organs as well: Consciously experienced objects [...] are distinct,

[44] I want to thank my good friend Andreas Hunziker for a very interesting discussion on this subject.

> functionally active parts of the organism currently making global stimulus properties and the high internal correlation strength, that is, the *coherence* of a perceptually given set of properties, globally available. In doing so, these object emulators form a functional cluster, that is, a casually dense, discrete subregion within the *global* functional cluster constituting the organism's world model [...]. Phenomenal scene segmentation [...] too is a dynamic property of the transient organ we call our world-model.[45]

When talking about virtual organs in the context of artificial intelligence, I actually go beyond the virtual organs that Metzinger postulates, in one simple aspect: In artificial intelligence, there aren't necessarily any permanently realized organs, so depending on the embodiment of the artificial intelligence all organs might even be virtual. In all other aspects, I would like to follow suit with Metzinger's definition of what can be such virtual organs:

- *Presentata:* Perceived objects that are part of the world model can be virtual organs if we perceive them as a part that we can influence, and thus consciously extend our phenotype (or rather, our phenomenal self model) into including them. They are even strictly embodied for us humans, in the sense that the neural correlates representing them are part of the body. For artificial intelligences, this will have to be similar in that a model representing the perceived object will have to be formed as part of the intelligence's mental content.
- *Representata:* Since we are able to manipulate not only objects in the world (and consequently have to model their presentata as virtual organs), but also our representations of these, the representations themselves can become virtual organs – can be embodied in the above sense and included in the phenomenal self model. Of course, this does not stop at this level, meta-representations can become virtual organs as well.
- *The World-Model:* This virtual organ is only special in that it is the outermost layer of virtual organs that make up Metzinger's convulted holism of nested, sometimes overlapping structures.

4.3.4 Virtual Organs as Virtual Vehicle Properties

4.4 Virtual Emotions and Hormones

4.4.1 Learning through "Punishments and Rewards"

4.4.2 Emotions

- Logic of survival: Damasio 1999 p. 54ff, Metzinger 198
- Minsky's Emotion Machine, Critics and Selectors
- Genuine emotions: Metzinger 199

[45] Citation from [15], page 201, emphasis his.

4.4.3 Critics, Encouragers and Selectors

For Minsky, emotions are just one kind of process[46] that the body (and with it, the brain) uses to make us think about problems in different ways. Because that really is one of the core things that make us intelligent: All the different ways to think that we do have at our disposal, that we use situationally or even within the same situation in order to look at something from a different angle.

Now these ways to think are generated through what Minsky calls critics and experts.

The brain consists of hundreds of different realatively distinct functional centers (According to [17], there are around 400 of those), similar to organs, that enabled evolutionary processes to optimize one of them at a time – similar to how organs are locally distinct, or how (in the technological and man-built domain) we neatly arrange different functionalities in a program within different code files while programming.

Like either of those two examples, such functional brain centers will rely on specific forms of input from other functional brain centers, and are functionally intermingled with one another – probably more so than a freshly designed program, similar to how an older program quickly becomes a maintenance nightmare after a few years, when structures from one functional center (function, method, class or file, in programmer's terms) are used in another in a completely different way than it was first built for. Evolution does not care about strict boundaries or neat-looking arrangements, as we've seen in chapter **??** it only cares about survival.

Minsky postulates that some of these functional centers are critics, of which there are four (or three, not counting the encouragers) kinds:

- Correctors, who declare that we are doing something dangerous.
- Suppresors, who interrupt before we begin the action we are now planning to take.
- Censors, who prevent ideas from even occurring to us (by knowing which steps usually lead to the ideas the censor wants to avoid).
- Encouragers – who are not really critics, but rather, reinforce actions that will likely lead to success.[47]

Those critics aren't necessarily active all the time. Rather, the brain switches them on and off repeatedly, as the situation calls for. Minsky suggests that this is what can happen (and indeed often does so "on timescales of one or two

[46] There is a reason why he called his book "The Emotion Machine" and not "The Thinking Machine". As he says in [17], 49:36-49:53: "The reason I called the book The Emotion Machine was to fool people into reading it. Because people are very excited and want to know about emotions, but if you said The Thinking Machine, they'd say "Oh I hate thinking, it hurts."" This is in line with Metzinger when he writes in [15], page 406 (in a different context), that "Thinking clearly is hard."

[47] See [16], page 82.

seconds, or less, in the course of our everyday commonsense thinking"[48]) when we try to solve new kinds of problems:

> First, briefly shut most of your Critics off. This helps you to think of some things you could do – with little concern about whether they'll work – as though you were in a brief "manic" state.
>
> Next, turn many Critics on, to examine these options more sceptically – as though you were having a mild depression.
>
> Finally, choose an option that seems promising, and then proceed to pursue it, until one of your Critics starts to complain that you have stopped making progress.[49]

Now, what those critics also do, is activate other functional centers in the brain, through the means of what Minsky calls selectors. Essentially, every critic is linked to one or more selectors, and the selectors then enable other functional centers of the brain, the ways to think – who invoke reactions to sensory data and system-internal feedback loops. The ways to think then in turn activate yet again different critics (of different levels).[50] In analogy, the critics enable different pathways that shape and form Metzinger's mental processes. This does imply that not all of our brains is active all the time – rather, most of it is suppressed most of the time, and only activated when it is actually needed. Exceptions are only these functional centers of the brain that are necessary for survival; the ones that control respiration or the beating of the heart.[51]

Coming back to emotions: They are nothing but one such kind of way to think, one that is merely special in that it can only rarely be activated on purpose.

Minsky proposes an extensive (albeit of course not exhaustive) list of higher-level critics in [16], pages 228ff. Similar to his list of critics, Minsky proposes an interesting list of ways to think in [16], pages 226ff. Particularly the list of ways to think includes entries that correspond to processes, emotions, memory retreival, offline activation of contents, motor behaviour and others. This is plausible however, since neither reactions to sensory data nor system-internal feedback loops are constrained to just one kind of response. That is what makes us so resourceful, and that in turn is what we perceive as being intelligent:

> [I] argue that emotional states are not especially different from the processes that we call "thinking"; instead, emotions are certain ways to think that we use to increase our resourcefulness – that is when our passions don't grow till they handicap us – and this variety of Ways to Think must be a substantial part of what we call "intelligence" – although [I] call it "resourcefulness."[52]

[48] Citation from [16], page 85.
[49] Citation from [16], page 84.
[50] See [16], page 222 for an illustration of this.
[51] See [16], page 4 for an illustration of this (that he reuses and expands throughout the book).
[52] Citation from [16], page 6.

4.5 Memory Structures

4.5.1 Neural Correlates of Supervenience

- Global avilability, functional clusters: 122
- Global integration functions: 139

4.5.2 Kinds of Knowledge Representations

While autobiographic and factual memory are fully covered with the distinction between episodic and semantic memory, both of these can be about many things. Quite a few kinds of knowledge representations have to be classified, and most of them can be part of either episodic or semantic memory structures. These kinds include objects, properties, categories, situations, events, states, time, causes and effects.[53] Relations between all those things (within the same category or crossing category borders) have to be represented as well. All these things are important to keep in mind if we are to create a strong artificial intelligence.

4.5.3 Short-Term and Long-Term Memory

- Episodic vs. Semantic Memory
- Declarative vs. Nondeclarative Memory (Nägele, Patrizia Hasler)

4.5.4 Associativity and Learning

4.5.5 Minsky's Panalogies

4.5.6 Common Sense

- Everything has exceptions
- Arguing by Analogy (Panalogy)

The following things at least are needed for each fragment of common-sense knowledge (in the sense of the fragments having to be embedded in these things), if they are to be similar to human reasoning:[54]

- Types of problems it might help to solve
- Types of goals that it might serve
- Other ideas it is similar to
- Typical cases in which it is useful
- Story-like narratives depicting its use

[53] I owe this non-exhaustive list to the Wikipedia article on artificial intelligence, http://en.wikipedia.org/wiki/Artificial_intelligence

[54] Slide from [17], 1:02:40. Remember that we talked about common sense before, in chapter 2.3.4.

- Contextual cues to suggest when it's relevant
- Parallel interpretations in other realms

4.6 Reasoning and Mental Resourcefulness

4.6.1 Minsky's Panalogies

5 Putting it all Together

> In fact, all of these methods are very useful for certain problems. What we don't know in general is: For what kind of problem is it good to use a statistical inference system? What kinds of problems are ones that can be learned and handled by neural networks or by genetic algorithms?
>
> What I'm proposing along with Gerry Sussman and Hal Abelson is to develop a new kind of AI system that has places in which we can insert all of the useful results that tens of thousands of AI researchers have made.[55]

References

[1] Artificial stupidity (editorial). *The Economist (US)*, 1992. Available online at http://www.highbeam.com/doc/1G1-12504583.html.

[2] *Sociable machines - Facial expressions.* 1999. http://www.ai.mit.edu/projects/sociable/facial-expression.html.

[3] *Commonsense Computing Initiative.* 2008. http://xnet.media.mit.edu/.

[4] *Wikipedia: Alpha-beta pruning.* 2009. http://en.wikipedia.org/wiki/Alpha-beta_pruning.

[5] *Wikipedia: Artificial Intelligence; Classifiers and statistical learning methods.* 2009. http://en.wikipedia.org/wiki/Artificial_intelligence#Classifiers_and_statistical_learning_methods.

[6] *Wikipedia: Church-Turing thesis.* 2009. http://en.wikipedia.org/wiki/Church_Turing_Thesis.

[7] *Wikipedia: Turing Test.* 2009. http://en.wikipedia.org/wiki/Turing_test.

[8] Susan Blackmore. *Memes and "temes".* TED Talks, 2008. http://www.ted.com/index.php/talks/susan_blackmore_on_memes_and_temes.html.

[9] Richard Dawkins. *The Selfish Gene.* Oxford University Press, 1976.

[55] Citation from [17], 30:08-30:55.

[10] Guido Gloor. *The Chinese Chatroom*. 2007. Available online at http://www.haslo.ch/philosophy/chinesechatroom.pdf.

[11] Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171–1220, 2008. Available online at http://arxiv.org/PS_cache/math/pdf/0701/0701907v3.pdf.

[12] Kacie Kinzer. *Tweenbots - Robot/People Art*. 2009. http://www.tweenbots.com/.

[13] W. G. Lycan and K. Neander. *Scholarpedia: Teleofunctionalism*. 2008. http://www.scholarpedia.org/article/Teleofunctionalism.

[14] Pamela McCorduck. *Machines Who Think*. 2004. http://www.pamelamc.com/html/machines_who_think.html.

[15] Thomas Metzinger. *Being No One: The Self-Model Theory of Subjectivity*. The MIT Press, 2003.

[16] Marvin Minsky. *The Emotion Machine*. New York, London, Toronto and Sydney: Simon & Schuster Paperbacks, 2006. Available online (in a prelimary version) at http://web.media.mit.edu/m̃insky/.

[17] Marvin Minsky. *Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. 2007. Available online at http://mitworld.mit.edu/video/484.

[18] Masahiro Mori. The uncanny valley. *Energy*, 7(4):33–35, 1970. Available online (translated by Karl F. MacDorman and Takashi Minato) at http://www.androidscience.com/theuncannyvalley/proceedings2005/uncannyvalley.html.

[19] John R. Searle. Minds, brains, and programs. *Behavioural and Brain Sciences*, 3(3):417–457, 1980. Available online at http://www.bbsonline.org/Preprints/OldArchive/bbs.searle2.html.

[20] Stuart Shieber. *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*. The MIT Press, 2004.

[21] Susan Turk Charles, Mara Mather, and Laura L. Carstensen. Aging and emotional memory: The forgettable nature of negative images for older adults. *Journal of Experimental Psychology*, 132:310–324, 2003. Available online at http://www.apa.org/journals/releases/xge1322310.pdf.

[22] Lloyd Watts. *Lloyd Watts: Neuroscience*. 2009. Available online at http://www.lloydwatts.com/neuroscience.shtml.