Univerzita Karlova v Praze

Filozofická fakulta
katedra logiky

Bakalářská práce

Petr Švarný

Pojem malfunkce v komplexních systémech

The malfunction concept in complex systems

Praha 2009

Vedoucí práce: Doc. PhDr. Petr Jirků, CSc.

Chtěl bych také poděkovat Ing. J. Burianovi, Bc. V. Karbanovi, L. Karenové, kteří si práci přečetli a očastovali kritickými poznámkami, Doc. PhDr. P. Jirků, CSc., že mi tuto práci umožnil realizovat a za pomoc, kterou mi při tom poskytl, přednášejícím a kolegům z kurzu Budapest Semestre in Cognitive Sciences za mnohé podněty a diskuze a své rodině, bez jejíhož zázemí a podpory by tato práce nevznikla.

Prohlašuji, že jsem bakalářskou práci vypracoval samostatně a všechny použité zdroje uvádím v práci na příslušných místech.

V Praze 23. března 2009 Petr Švarný

Práce se zabývá konceptem malfunkce a nemoci, které charakterizuje jako neobvyklé či nežádoucí chování daného komplexního systému. Podobnosti a rozdílnosti mezi umělou a přirozenou myslí, jednoduchými a komplexními systémy se ukáží jako důležité s ohledem na závěr, že duševní choroby nemusí být omezeny jen na lidksou mysl.

Klíčová slova: umělá inteligence, problém mysli a těla, malfunkce, nemoc, komplexní systémy.

The paper studies the concept of malfunction or illness, which are characterized by non-standard or undesirable behaviour of the given complex systems. The parallels and differences between natural and artificial mind, simple and complex systems are shown to be crucial for the conclusion, that mental disorders must not be restricted to human minds.

Key words: artificial intelligence, mind-body problem, malfunction, complex systems.

Obsah

Seznam použitých zkratek

AI – artificial intelligence

CFAI – Creating Friendly AI

DAI – Distributed AI

DSM – Diagnostic and Statistical Manual of Mental Disorders

f MRI – functional Magnetic Resonance Imaging

GISAI – General Intelligence and Seed AI

GOFAI – Good Old Fashioned AI

NFAI – New Fashioned AI

PET – Positron Emission Tomography

RUR – Rossum's Universal Robots

1. I

2. ntroduction

Modern attempts to create a system capable of human-like or even superior than human reasoning and behaviour went through many difficult periods. Complete optimism or utter pessimism were replaced by a more pragmatic point of view. But one thing still remains the same, the source of inspiration. Research and development in the fields of artificial intelligence uses often nature as a database full with solutions and ideas. First the reasoning processes of the human consciousness were taken as a model, then the inspiration by neural architecture came or the imitations of social intelligent behaviour. This inspiration will be proved as a possible source of problems that could lead to undesired results. We shall investigate into the term malfunction and show how wide it's range is. Thereafter we shall use it as a first step for the inquiry into the possibility of mental disorder-like behaviours of artificially intelligent systems.

3. T

4. he Concept

4.1 The term of malfunction

In order to be able to investigate the possibility of malfunctions similar to human mental disorders we need to define the term of malfunction. The aim is not to create a new terminology and vocabulary but to show the width of the terms meaning and some possible realizations. First lets look up some formal definitions of this term and outline the possible meanings of the word itself before we set our own definition of the word. We need to remark the multitude of usages of the word and its foggy borders, especially it's relation with the word dysfunction.

Let us consult a few encyclopaedias and dictionaries to get a general idea, for the term malfunction we find these definitions:

> *A **malfunction** is when something functions wrongly or does not function at all. en.wikipedia.org*
>
> *To function imperfectly or badly, fail to operate normally. www.merriam-webster.com*
>
> *Faulty functioning. The Oxford English dictionary, second edition, 1989*

For the term dysfunction these:

> *In psychology, an abnormality.*
> *In mechanics, a malfunction. en.wikipedia.org*
>
> *1 : impaired or abnormal functioning*
> *2 : abnormal or unhealthy interpersonal behaviour or interaction within a group www.merriam-webster.com*
>
> *Any abnormality or impairment of function.*
> *The Oxford English dictionary, second edition, 1989*

It was important to cite these definitions because they, particularly those of dysfunction, show the duality of views upon the same problem. Thus in some cases the here used concept of malfunction will assume that there is no insuperable gap between man and machine. Then

the meanings of dysfunction and malfunction merge into one, as we can also see from the entries for dysfunction.

The intention in choosing the term malfunction instead of dysfunction is simple. On one hand malfunction tends to be more connected with mechanical entities and therefore seems more natural to use in the context of systems and machines, but dysfunction would not seem to be right in connection with just mechanical, artificial abnormalities in functioning especially if we enter the field of psychology of machines.

On the other hand, malfunction used also on human behaviour invokes the feeling of downgrading human beings to nothing more than complex machines. This actually is the main reason for it's choice. After all, it was already D. C. Dennett [Dennett 2004] that presented human beings as constructed from tiny automatons, robots and even J. Searle [Searle 1980] says that we are machines that think. It is important to realize that the scale of inspiration contemporary machines take from nature, neuron networks for example, and the development in bioengineering may even lead to a biological computing device and thus diminish the differences between the two traditionally distinct domains - machines and living forms. Also let us add that the use of an already existing term and it's slight redefinition and profounder analysis seems more reasonable then the creation of a completely new one.

The definition, to be clearly given, we shall use will be this:

> *Malfunction is any undesired and/or unpredicted behaviour of the studied system or subject that limits partially or totally it's proper functioning.*

Obviously any desired behaviour is not a malfunction because it is awaited and designed in the system. Thanks to this, events like car demolitions during crash-tests can not be regarded as malfunctions of the car (unless we supposed that the car will endure the test). A probably non trivial fact in the definition is, that a (even undesired) behaviour that has no effect on the systems functioning is not regarded as a malfunction. Therefore if for example a pen changes it's colour every time someone writes with it but it's ink remains the same and the writer just wants to use it to write, then the colour shifts are not taken as a malfunction[1].

1 Of course if he needs the pen to maintain the same colour for whatever reason, then it could already be regarded as a malfunction. This example obviously just demonstrates how problematic these matters can be if studied in depth.

4.2 Kinds of malfunctions

We shall determine the kinds of malfunctions with a multi-criteria approach using the malfunction's origin, type, and effects. All malfunctions have the, by definition given, common feature of non desired behaviour of the subject. Through this refinement of the term, we clarify the means and ways how to restore the proper functioning of the entity. This will also permit us to show the human - machine parallels in malfunctioning. The parallels are needed, as we shall see, because of the growing complexity and unpredictability of the artificial systems. However, it should be remarked that not all malfunctions prove to be useless or bad. The emphasis is put on *undesired*, because the gained reactions might be undesired in the given circumstances, but could be needed or wanted in some other context and therefore the malfunction might even be used as an inspiration for new designs[2].

4.2.1 types of malfunctions

We understand under types of malfunctions the kind of view we adopt depending on the sort and purpose of the observed entity. It would be easier to discover and classify as malfunctioning an alarm-clock that does not ring or a factory robot that instead of spraying the paint on the car sprays it on a fellow robot arm. But it would prove a lot harder to say if a human being is malfunctioning (as we seem to be the peak in intentional entities). The main idea is drawn from Dennett's [pg. 60] stances. The difference is that we regard these as possible views upon the given objects. Human beings for instance can be regarded from the physical, design and even intentional stances.

The first and the most simple type if we follow the stance hierarchy of Dennett would be the *physical view*. The system is passive and expected to act just according to the laws of physics. But these are for our study uninteresting because finding a passive object that would not obey the laws of physics would be impossible or an exciting scientific discovery. And obviously, all objects that are in the physical world act according to the laws of physics and therefore can not malfunction in this way. Therefore this type can be omitted.

In the *design type* are malfunctions that fail a given purpose of the system. Here it becomes really important to realize that types are given only as views upon the system's function. By adopting this view we presume that there is a beforehand, externally given

2 This can be demonstrated on the sickle cell disease which is caused by a mutation of the haemoglobin and causes that the blood cells are sickle shaped instead of the normal round shape. The disease can shorten life expectancy but also makes the bearer more resistant to malaria. As reference http://en.wikipedia.org/wiki/Sickle-cell_disease.

purpose of the system that should be realized. The designed purpose of a simple system is simply determinable. The already mentioned alarm-clock has a given task of making loud noises at a given time. If the alarm-clock is malfunctioning, then it is, according to the design view, not fulfilling it's task.

The question becomes more complicated if we take a human as an example. The reason, why humans will be used is simply because other entities are created or altered by man for some tasks, so they have clearly a goal to fulfil, and life forms are often regarded as reproduction machines which malfunction if they fail to reproduce or at least protect their close relatives. If we would manage to create an entity that would have no purpose but still have enough complexity to make choices, then we could use that as an example. Instead of this hypothetical homunculus I prefer to use the already existing biological machine - man.

A human can be regarded as having a malfunction of the design type in multiple cases depending on what function we assume he has. Herein lays the complication intentionality brings with it. With intentional and complex systems, which are able to alter their behaviour and give tasks to themselves we encounter a freedom to choose, when faced with the question of purpose. It is a matter of fact that people are animals and therefore we could point out procreation and survival as the tasks humanity and every single individual in it has to achieve. A design type malfunction would be then for example a life in celibacy. Especially in the regards of Dawkins' theory [Dawkins 1998] this matter becomes quite interesting. If man are not only gene but also meme bearers then they have two, sometimes, opposing tasks to realize, the example of celibacy is a good demonstration of this. Both the memes and genes strive to be spread and so they pose this as a function of the bearer. We can also, as exterior observants, give a man social functions, then someone who isn't fulfilling any needed function in a society or is harming it can be regarded as malfunctioning.

The tricky part is to decide, what should be regarded as a design given purpose. The best way to solve any problems connected to this seems to be a kind of a frame of reference as in physics. Depending on the chosen object and observer or user, the object can have multiple functions. It is clear that because functions of objects or systems are always, in the design type, imposed from outside, they are dependent on the observer, but this dependency can not be made absolute because there remain some more invariable purposes. The whole concept of designed purpose is supposedly a result of human culture and it's extensive use of tools. If we do not omit this origin of the view then any object, system or being has a primary designed

purpose that it was created for. In the case of any man made object it becomes then a quite simple matter how to determine the objects purpose.

The last would be the *intentional type*. It is partly covered by the design class, because we could formulate the task for the system as "to be intentional" or any similar formulation. Nevertheless not every system can have intentionality. In this view a malfunction is when a system is not able to determine it's goal[3] or purpose or was able to do so, but is not able to fulfil it. As an example of such malfunctioning system we could take someone suffering from depression or a highly indecisive person. The intentional type system, opposed to the design type, gains it's function from itself without the need of an exterior observer giving the system a function. Of course, we can use the design view even on intentional beings, in many social situations a person becomes simply a tool with a precise role for someone else.

4.2.2 origins of malfunctions

The origin of the malfunction is important to identify because it helps to find a way how to repair the system.

An obvious origin of a malfunction can lay in a mistake in the body of the system, let these be called of *material origin*. Here we understand the raw material of the system like wiring, neurons and so on. A good example of such malfunctions can be any physical damage to the system (see brain damage and it's effects or any changes in the laptop's behaviour if you drop it from the table). These are, theoretically, easy to repair if we know the right configuration (for our contemporary abilities that includes only the case of the laptop). In some cases, as in experimentally build hardware, this kind of malfunction could occur as a result of the use of non-tested architecture. Also all really brute changes on the physical body of the system are part of this origin (as an example a broken arm or DVD-ROM which limit the functioning).

The next possible origin would be a malfunction that was gained during any learning process or changes made by the system in it's patterns of functioning. Into this category would fall a neuron network that learned the wrong patterns and so does not meet out desired categories or a man who subjected to some wicked ideology becomes it's defender. The problem here lies not in the structure by herself or in the physical body of the system but in the information encoded in it. Therefore these could be named as having an *informational*

3 Of course if the system should be able to determine it's own goal(therefore using an at least seemingly non-deterministic mechanism to do so), then it is already a complex system, probably with some kind of self-awareness.

origin.

As the last origin of a malfunction we could in some cases distinguish a part of the gained malfunctions. These would be those that are done against the design of the system by the system's will. The reason, why these are just a part of the before mentioned ones, is that the system must have acquired in some stage the knowledge, why it does want to go against it's primary purpose. But opposed to the first one's, here the system is aware it is acting against some kind of purpose and it is acting willingly in this way. Because of the intentionality included in the malfunction, this could be named as having an *intentional origin*.

4.2.3 effects of malfunctions

These could be categorized simply by their harmfulness. So we get the obvious three possibilities in connection with a malfunctioning system. To these we could add the degree of the malfunction, if it is completely or just partially hindering the system from functioning. These effects would be simply called as positive, neutral and negative. We just have to keep in mind that we talk about malfunctions, therefore the system is already in some sense having a negative effect.

If the system has a *positive effect*,then it is malfunctioning and therefore does not fulfil some desired behaviour, but on the other hand his new behaviour is still beneficial to either the creator, the system or the environment[4]. It is important to take into account the whole range of the system's goals because if the malfunction makes the primary goal unfulfillable but helps with something minor instead then the effect can not be regarded as positive.

The *neutral* malfunction effects have two possible realizations. There could be a problem that only blocks the functioning in a minor way and does not bring any good nor bad side effects with it. For example a robot that has the task to move between points A and B could just become slower with a slight damage on one of it's movement units but would still manage to get to point B. The second possibility is, that the system simply does not carry any supplementary negative effects nor positive ones.

The last possibility is when the malfunction not only prevents the system from accomplishing it's task, but also has some harmful side effects. These cases are then malfunctions having a *negative effect*, because they have some additional negative influence.

4 See the already mentioned example of sickle cell disease.

4.3 Detection of malfunctions

When all the needed factors are taken into account, then determining if a given behaviour is a malfunction is easy. However we need to pay attention to some aspects of malfunctions and system functions.

In the case of simple systems, there is usually a purpose given during creation, which can be defied. This malfunction then is easily determined and found. It is important to note that many systems, even those that seem complicated, from the point of view of functioning fall into this category. For instance, most computer programs, having a precisely given function, are simply judged as malfunctioning if they fail their purpose. The more the system or it's role become complicated, the more difficult it can be to determine if it is malfunctioning and why does it do so. In these cases the notion of normal or average functioning and a range of tolerance for deviations is needed for the decision.

Also the multitude of possible realizations of malfunctions is causing trouble. Especially in cases when we can not agree on a designed purpose for the system, determining if a given behaviour is proper or not is a very difficult task. In these cases the best procedure would be determining deviations from an average system. This system does not even need to exist in order to fulfil the needed role, the reason for it's existence is simply to be a base for comparison.

Nevertheless both simple and complex systems can have a multitude of possible functions. If the system fails in one given function, it still can be given a completely new function that it will fulfil correctly (ex. using a old broken alarm clock as a paper weight or a dog that is a lousy watchdog but a great draught dog.) Therefore malfunctions are strongly function related and can not be simply generalized on the given system. In some cases the objects definition incorporates a single function, if that is left unfulfilled (ex. broken alarm clock), the object is considered as malfunctioning because of the implicit function given in its name. Other objects can have multiple functions implicitly incorporated in their name, which doesn't have to be all fulfilled in order to earn the given name (ex. dog). In these cases the object's malfunctioning is only determinable depending on the studied purpose. Here we could distinguish crucial, which are needed for the given object, and non-crucial functions, which are more optional and depend on the objects user.

However, this would probably lead to a kind of essentialist discussion or concept theory

(see [Murphy 2002]) and that is not the aim of this paper. For our purposes it is enough to use the design view. If a given system has an intended purpose (even if given just by it's name or it's intended role in the world), then it must be fulfilled in order to be properly functioning and in the case of a multitude of functions incorporated in the systems name we can regard a system as properly functioning if it fulfils at least some of the designated functions.

When looking for malfunctions the effects can not be neglected in the search because for example a positive side effect could lead us to believe that there is no malfunction at all.

4.4 The malfunction and illness relation

Let us summarize the concept and make the step needed for the the second part of the paper. We defined malfunction as any gained undesired behaviour of the studied system that limits it's proper functioning, then we analysed the terms meaning according to the view on it (design and intentional) , it's origin (physical, informational, intentional) or it's additional effects (positive, neutral, negative), then we pointed out some pitfalls of the detection of malfunctions. Let us now put this dissection to some use.

Mental disorders could, by a behavioural approach, be defined as departures from an awaited behaviour in given circumstances or culture. Every single individual has some divergence from the standard as in physics measurements diverge from the real value. However this imagined average subject is important in the judgement of the acceptability of the behaviour of an observed individual.

The limits of acceptance are set by the observer. To demonstrate this point and connect it with malfunctions, we can even use a product instead of people: a plastic model sold in common shops can have mistakes in orders of millimetres and still be acceptable, on the other hand a highly detailed model from a specialized manufacturer does not allow this range of mistakes. In the case of humans the deviation sensitivity and observer dependency can be demonstrated on the humorous examples of an exaggerated analysis of "being in love" as a psychosis or the Koro, the fear that the subjects penis will retract into his body, both by Youngson [Youngson 2000]. Being in love, with all it's strange feelings, slight paranoia, and other symptoms, is a common and acceptable experience in our culture even if it can be viewed as having common symptoms of mental disorders. On the other hand, Koro is in Europe a strange and uncommon condition and therefore easily categorized as non-normal, but in some tribes of south-east Asia this was accepted as a valid fear.

Although the relativity of the borders of mental disorders is common knowledge, it's liaison with the average, not really existing, member of the group was necessary to emphasize for the connection between malfunctions and mental disorders.

It might come down as a kind of functionalist view, but both, malfunctions, as depicted above, and mental disorders have the same common features and vary only on the level of implementation. The only perceived difference between them being a reminiscence of the Carthesian division.

It is trivial to present injuries as malfunctions. Using the above constructed definition, (non cranial) injuries are malfunctions of the material origin with negative or no additional effects and classifiable as malfunctions from the design view. Cranial injuries, as we'll see further, can balance on the edge between simple physical injuries and mental disorders. Although they would mostly fall into the same category in our classification as other injuries. The case of illnesses is a more delicate one. Illness, in this case regarded as synonymous with the word decease, can have various causes. Depending on these causes, the decease could be classified as a malfunction of various types or even not classified as a malfunction at all (in the case of an invading organism, which simply overwhelmed the immune system).

Herein lays also the reason why this paper focuses on mental disorders and their relation to machines and malfunctions instead of deceases and injuries. They involve a great range of causes and effects and are closely connected to the mind-body problem. Malfunction in it's whole variety of possibilities finds her mirror image in human mental disorders. So in order to understand all possible malfunctions, in order to study the possibilities and pitfalls of an artificial mind, we should look into the ways how biological minds function and especially malfunction. Just to be prepared for the future.

5. T

6. he man-machine problem

6.1 The story so far

Let us have at least a brief look at the history and terminology of artificial intelligence research. Thereafter we should also regard the concisely evolution of neurology and psychology, because the thesis of the work draws from both fields.

6.1.1 the brief history of AI

AI[5] today a vast field of research with many different sub-disciplines like artificial sight, face recognition, neural networks, expert systems and others. The roots of AI could be tracked into the ancient times (see [Zelinka 2003]) but it's modern history started in the nineteen fifties preceded by a immense revolution in logic when also the creator of modern computing A. M. Turing said and asked:

> *„What we want is a machine that can learn from experience, possibility of letting the machine alter its own instructions provides the mechanism for this.“*
> *Turing, public lecture, 1947*

> *„Can a machine be made to be supercritical?“*
> *[Turing 1950]*

Where he made a metaphor between the supercritical amounts needed in nuclear physics and ideas that would in a supercritical amount lead to the mind.

6.1.1.1 good old fashioned AI[6]

The basics of many tasks and disciplines where set during these early years. C. Strachey's checkers program in 1951, the first route planning program, Shopper, from A. Oettinger in 1952, a rudiment of evolutionary computing approach using a modified Strachey checkers program by A. Samuel in 1955. A celebrity amongst programs is the Logic Theorist written in 1955-56 by A. Newell, J. C. Shaw and H. Simon, this program had the goal to prove theorems from Principia Mathematica and it did not just manage to prove all that was needed, but in one instance found a more elegant proof then there was in the book.

---

5 If not mentioned otherwise, then the informations are based on Encyclopaedia Britannica's article Artificial intelligence.
6 The GOFAI and NFAI expressions I borrowed from J. Haugeland [Haugeland 1997] .

Another star among the early programs is Eliza from 1966 by J. Weizenbaum which tried to simulate a dialogue with a therapist, not to forget Parry from the same year made by K. Colby who on the other hand simulated a human paranoiac. But both programs are clearly unintelligent and their lack of understanding of the words used can be easily proved with the right questions.

A new line of research also aimed for the AI's interaction with the world. The so called microworld approach, suggested in 1970 by M. Minsky and S. Papert, aimed to test the work on AIs firstly in simplified virtual environments. But already the first great success, T. Winograd's SHRDLU in 1972, showed the limitations of it and the actual impossibility to transfer a microworld program into the real world and failed the hopes that these programs would gain some kind of understanding of the world[Dreyfus 1979]. As the starters said:

> *„Many problems arise in experiments on machine intelligence because things obvious to any person are not represented in any program. One can pull with a string but cannot push with one... „*
> *Minsky, Papert in [Dreyfus 1979]*

However, this impasse gave birth to expert systems, one of the main AI flagships. These programs occupy a specific microworld, for example any type of database, this kind of microworld is still quite simple, but with the right information choice it is made out of clearly distinct informations that can be categorized and therefore the program can easily use them. A great instrument for building expert systems is the programming language Prolog, first implemented in 1973 by A. Colmerauer and further developed by R. Kowalski. Using a knowledge database and the right inference rules (often a simple if-then, but there can be also used basic fuzzy logics instead of classical logics) the expert system is able to give right answers to specific questions. The most important task is to choose a field of expertise that can be computerized in this way and then fill the knowledge database with the right data and give the program the right rules, this can be a long and hard work for human experts to agree on how and what should be implemented and therefore answered. This approach can be of great use in some fields as proved by E. Feigenbaum and J. Lederberg's DENDRAL (analysed spectrographic data and then suggested the structure of the molecule), MYCIN also from Stanford (1972, functioned as a specialist for blood infections and according to given answers could ask for more evidence or give a diagnosis). The closest to human mind is probably the mammoth project CYC, when a enormously huge knowledge database is manually created in order to gain a system that would reason and make inferences about the world and

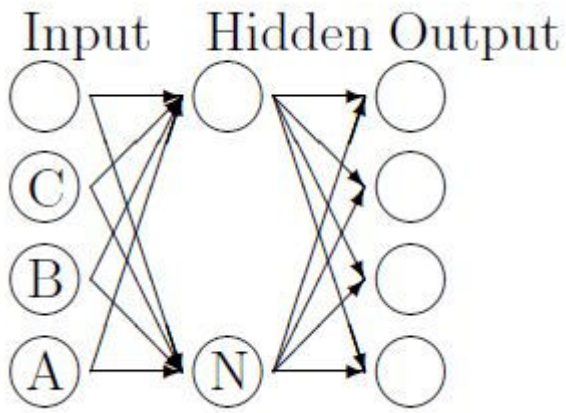
statements as humans do. This project started in 1984 and is still running [Copeland 2000].

6.1.1.2 new fashioned AI

In the meanwhile there was a different kind of research going on. The idea came with the advance of neurology and theories of E. Thorndike and especially D. Hebb's learning rule, which states that learning is made by strengthening the connection between neurons. The idea of artificial neural networks can be traced also to A. M. Turing, who came up with the idea of a trainable artificial neuron network in 1948 in his unpublished paper Intelligent Machinery [Turing 1948].The first paper however was published already in 1943 by W. McCulloh and W. Pitts, who regarded the brain as a computing machine with neurons as simple processors:

> *„What we thought we were doing (and I think we succeeded fairly well) was treating the brain as a Turing machine.“*
> *W. McCulloh*

The basic principle of an artificial neuron network are neurons connected into layers (see pic. 1), some neurons forming an input layer and others form the output layer, the layers in between those two are called hidden layers.



Pic. 1

Each neuron can either fire or rest, if he fires then he emits a signal(of value 1) that propagates through the connections and is multiplied by the value of the respective connection, which is called the weight of the connection. The neuron then functions as a threshold device according to this function:

$$ni(t+1)=sgn(\Sigma jwijnj(t)-\mu i)$$

Where $n_i(t+1)$ is a given neuron (in the example the neuron N), $n_j(t)$ are the preceding

neurons that have a connection with the given neuron (in our example A,B,C), w_{ij} are the weights of the connections and μ_i is the threshold that has to be exceeded in order to make the given neuron fire. It is important to realize that the weight may even be equal to zero, an equivalent of no connection between the two neurons. If the neuron is activated then the same process continues in the next layer. The learning process of a neural network then functions on weight changes. If for a given input the output is what was desired, then the weights of participating connections are augmented and therefore raising the probability that these connections will make the succeeding neuron fire next time. Let us shrink our exemplar model to just two layers, if now an input activating the neurons A and C should give a also a signal on N, and it did, then we will reinforce the weights w_{AN} and w_{CN}, therefore their sum will have a greater chance to exceed the threshold of N. In the case the desired output of N is 0(no signal), but the network produced for the input a 1, then the weights of the active connections are lowered[7]. This process is repeated until the network gives a satisfactory response for the presented pattern(input). An interesting issue is that networks can be taught with a teacher(any supervising entity that decides if the output is as needed and adjusts the weights according to the wanted results) or without supervision when the network takes advantage of any similarities in the input data and creates it's own response pattern(example is the distinction of shapes, feeding a network with sufficient rectangles and circles leads the network to the distinction between those two shapes). After the network is trained, it is able to perform the task repeatedly based on the knowledge it got encoded from the examples that were fed to it. Already this description of the basic architecture shows the difference to symbolic approach. Any information stored in these networks is distributed in the whole network and also the process of data processing is parallel as opposed to the serial GOFAI systems.

The first, 128 neuron large, network being constructed in 1954 by B. Farley and W. Clark. This network already was able to learn simple patterns and remember them even when losing 10% of the neurons. F. Rosenblatt in 1957 started to investigate three layer neural networks that he called Perceptrons. He is also the author of the name for this branch of AI - connectionism. The work in this field was frozen, when Minsky and Papert proved in 1966 that perceptrons cannot learn any XOR problems, this proof misread as devaluating all neural networks then stopped almost all research in this field (with some the exceptions like E. Caianiello). It was until 1978 when D. Rumelhart and J. McClelland trained 920

7 For further details see for example the [Hertz 1991] or [Mařík 1993].

neurons in two layers and taught the network to conjugate English verbs' past tense. This work draw again attention to artificial neural networks and also thanks to J. J. Hopfield, who showed that connectionist networks can be studied with the mathematical theory of dynamical systems [Havel 2001] , neural networks nowadays find many applications, as an example see the Hodgkin and Huxley model for neuron activity simulations or projects like Roboneko - a robot project using artificial neuron networks[8].

6.1.1.3 a short confrontation

AI can be divided, as we saw, into two main streams, the symbolic and sub-symbolic. The symbolic AI is based more on the experience of our stream of conscious thoughts. This approach tries to study intelligence as a symbol manipulation not dependent on the structure that does the manipulation. The root of this approach lays in the base of modern computing, the Turing machine, which is inspired by the conscious thought processes of the computer, the man who computes [Havel 2001] . A. Newell and H. A. Simon postulated the physical system hypothesis that sums up the idea behind this effort:

> *A physical symbol system has the necessary and sufficient means for general intelligent action.*
> *Newell & Simon, 1976*[9]

That symbol manipulation is sufficient for a system to have also it's own intentionality is doubted, the most famous and discussed counter argument was the Chinese room thought experiment of J. Searle [Searle 1980] which essentially tried to attack the not sufficient distinction between simulation of mind and the actual mind condemning symbol manipulation as efficient in mimicking the mind as a computer simulated rainstorm is efficient in drenching us. This experiment however just showed the possible limitations of the symbol manipulation in relation to consciousness, and the so called strong AI program that aims to construct an AI that thinks the way as humans do, but did not in any way affect many other useful aspects of this approach. Therefore it is more in the field of applied AI that symbol manipulation is used nowadays: the creation of expert systems or robotics.

The other way does not incorporate any explicit symbol manipulation or usage and leaves the human cognitive processes aside and goes for the brain's design. However these networks do not aim to simulate the whole complexity of the human (or animal) brain, current

8 [Best brain boosts artificial life]
9 From [Dreyfus 1979]

state of art artificial neuron networks are not fit to simulate in the whole range even a three hundred neuron worm[10] [Copeland 2000] where the human brain has 10^{10} neurons and each one of them can have ten thousands of connections, as J. Haugeland puts it:

> *What makes connectionist system interesting in approach to AI is not the fact that their structure mimics biology at a certain level of description, but rather what they can do. After all, there are countless other levels of description at which connectionist nets are utterly unbiological. [Haugeland 1997]*

Nevertheless these networks even as primitive as they are in comparison with real neurons show analogies to mental phenomena as bistable perception, hesitation, hallucinations and others [Havel 2001] .

These two distinct methods can be also labelled as a "top-down" and "bottom-up" because of the imaginary direction they take. Symbolic AI wants to reason about the cognitive processes from the whole cognition as we live it and the sub-symbolic uses non-cognitive elements that have some emergent intelligent behaviour and both enjoy their own benefits and suffer from their respective drawbacks. Connectionist networks are really able to learn and are more capable in pattern recognition (shape or data division into different categories, image processing) on the other hand they are not able to do what the symbol systems do the best, processing syntactical patterns and inferences that is.

It is important to choose the final goal of our efforts because as we can see on both main approaches, there is a large field of application of results even though AI does not function on the same level as human intelligence and fail most of the comparative tests, as for example the famous Turing test. The engineering criticism of human similar AI, as heard from K. Warwick [Warwick 1999] for example, tries to point out the lack of purpose in imitating human cognition and processes and wants to focus on functioning models, whatever they look like or however they function. But still AI is used also as a mirror in which we can see how our cognition works. Or as J. Haugeland puts it:

> *Mind design is psychology by reverse engineering.*
> *[Haugeland 1997]*

10 Caenorhabditis elegans.

6.1.1.4 alternative AIs

We have presented two possible approaches to AI, but there are more ways in which AI is nowadays studied which often react to some of the imperfections of those two.

The **nouvelle AI** is distancing itself from the strong AI project-artificial intelligence that aims to duplicate human intellectual abilities [Copeland 2000]. Nouvelle AI, started by R. Brooks in the 1980s, does not aim for the man's mind, it tries to construct systems inspired by insects and in opposition to the classical AI- physical symbol systems [Brooks 1990]. The idea is related with the enacting approach [Havel 2001] which abandons inner representations and focuses on real world interactions and combines it with emergence theories averring that complex behaviour emerges from the interactions of simple ones. The world is left as the best model for the system to react to, instead of constructing an inner model of the world and react according to that. This approach brought some successes and even R. Brooks started a project, the Cog project, which should be a opponent to CYC and learn by himself [Copeland 2000].

DAI - distributed artificial intelligence is an accession that breaks down complex tasks into simpler tasks for a group of autonomous agents. The collaboration of these agents then produces the needed results [Štěpánková 1997]. If this seems as a unfruitful path to intelligence then simply realize that in fact, as Dennett points out [Dennett 2004], we are made out of tiny automatons and exhibit consciousness, intentionality and all these states we want to create in AI, therefore if not sooner, then with a creation of cell-like artificial automatons we could arrive to an intelligent entity[11]. Another level of approximation to the biological intelligence is the fact that the human mind is also created from multiple functional modules that communicate with each other [Koukolík 2002] and the human mind could be the result of their interaction, as for example in Dennett's suggested multiple draft theory [Cummins 1998] or [Dennett 2008].

Seed AI - is a new direction in the AI research propagated by the Singularity institute. The term singularity refers to a technological singularity, an immense speeding up of technological progress, which could even get out of control. The main idea is to leave the so far futile attempts to build and intelligent AI and build an AI that would be capable of self-improvement and with that help arrive to even super-human intelligence. SingInst tries

11 This extreme ending however is the same as saying for connectionist networks that it suffices to arrive to the level of human neural network's complexity. Also DAI hopes for a much sooner emergence of intelligent behaviour. Still, the advances in bioengineering lead to questions connected with DAI, as seen in [Amos 2008].

also to make sure that this AI would be human friendly and would serve the good of mankind [CFAI]. Also a lot of effort is also put into argumentation against anthropocentrism in AI and the tendencies of fearing that AI will have human needs, lusts, desires even if not programmed so and not having the human body. The name comes from the main idea:

> *The task is not to build a mighty oak tree, but a humble seed.*
> *[GISAI]*

In spite of so many approaches we still do not interact with man-built machines that would have a man-like mind but as J. Coupland sums it up:

> *(The) lack of substantial progress may simply be testimony to the difficulty of strong AI, not to its impossibility.*
> *[Copeland 2000]*

6.1.2 the story of the mind

Strong AI's goal - the human mind - is the starting point for philosophy, psychology and to some extent neurology. We shall have now at least a succinct introduction to them.

6.1.2.1 neurons

Since Aristotle's thesis that the brain is a blood cooling device and the soul resides in the heart, we made a lot of progress in the studies of these matters [Cummins 1998]. Although F. J. Gall started with his ideas the pseudo-science of phrenology he also spread the idea that the brain houses our soul (mind) and his functions can be localized. It was later that P. P. Broca found the probable centre of speech and fully opened the gate to modern neurology [Youngson 2004]. And it was already in 1909 that K. Brodman did his mapping of the brain areas that is still used in recent neurology. The scientific progress(for ex. the new imaging methods as f MRI, PET, also other new achievements in the fields of molecular chemistry, genetics and also artificial neural networks) permit modern neurology to study the brain on all possible levels, from the influence of genetics on the brain structure, over neurons and their parts to functional systems and behaviour. The base assumption for any research of the brain was that it in some way influences our mind as shown by many patients, let us mention the famous case of Phineas Gage or patient E. V. R., who both after injuries to their frontal lobes showed changes in behaviour and personality [Koukolík 2002]. An interesting information in connection with the mind is also the progress

in the identification of brain areas that are responsible for it. Experiments were and are performed that monitor with f MRI techniques the brains activity during tasks that focus the subjects attention either on the external world or to internal processes and succeeded in identifying the important locations for our self-awareness and internal speech [Koukolík 2002]. Today's neurological research also uses artificial neurons or neural networks to study real neurons behaviour on complex simulations, as an example see [Prescott 2006].

6.1.2.2 thoughts

The advances in the field of neurology largely influenced the other two disciplines. Serious psychology managed to get through the dark valley of Freudian psychoanalysis and got rid of monsters like the collective subconscious or the trinity of id, ego and superego [Youngson 2004] and uses only the valid aspects of these ideas(as an example let us take modern clinical psychology). Also a different branch of psychology started with the experiments of H. von Helmholtz and especially when W. Wundt in 1879 started his laboratory for experimental psychology [Cummins 1998]. So psychology evolved from an extensive use of introspection, over it's absolute denial in behaviourism that drew answers only from observable inputs and outputs of a black-box like subject, to a modern science which combines all these methods and any remnants of the old folk psychology must be ready to face criticism based on the newest findings, even those from AI research [Ramsey 1991]. However there is a clear feedback from psychology to philosophy of mind and the AI research connected with it as seen on theories like the Unconscious Thought Theory [Dijsterhuis 2006] which supports the thesis that human reasoning and thoughts is less direct and conscious symbol manipulation than it seemed to be. Or the question of identity is addressed by experiments with twins, who have a similar brain structure to start with and afterwards have also similar personalities [Youngson 2000].

Psychology also partly discusses the issues of mental disorders. These are, as already mentioned in the first part, mostly community defined and have various degrees. In the US the DSM-IV manual is used to determine if a given behaviour is already considered as a disorder[12]. Neurology has again a word to say in this field because many disorders have a deep neurological cause, under which we understand not habitual brain functioning or structure[13] which permits direct medical treatment.

12 See for example http://psychcentral.com/disorders/ for a on-line DSM-IV.
13 [Koukolík 2002] reports about the causes of many mental disorders in the terms of neurology.

6.1.2.3 minds

Almost all of philosophy could be in some way related to the topics discussed in AI and cognitive science, sometimes it can be surprising how ideas of long gone philosophers can contribute to the topics of AI research. As an example we can look at Dreyfus's criticism of Minsky and others, where he quotes and inspires himself with Husserl [Dreyfus 1979]. Often the start of philosophy in connection to the mind and it's relation with the world is put to R. Descartes, who in 1641 in his Meditations on the first philosophy postulated the thesis that the world, of which I can have doubt, and me, who is doubting and therefore cannot be doubted, must be two different substances. Although philosophy spoke about similar subjects from it's Greek dawn, it is a series of philosophical arguments and counter arguments started with Descartes that shaped the today's field of philosophy of mind. The categories are fuzzy and opinions can be on the frontiers of two distinct views and to demonstrate briefly the main streams in the philosophy of mind I shall use [Havel 2001] as a basis. Obviously each view is in fact an answer to a question connected with the mind, let's use the questions as a branching guide.

The first question to answer is - *"How many substances are there?"* or *"Do you want to distinguish mental and physical?"* Descartes, as we said, answered yes and therefore is a dualist, if your answer is negative, then you are some kind of a monist. There you could, as did to some extent G. Berkeley, be an idealist saying that all that is is the mental, in Berkley's case *to be* meant *to be perceived* in an extreme case you could end up being a solipsist saying that all that is is just your imagination. The other option is to say that that all mental is in fact just an effect of the material, physical, world, a view laying in the basis of modern science. Actually materialism is not the last possibility, there can be a neutral substance that is both, material and mental or neither of them. This third possibility is a kind of an intersection and it seems that slight nuances or just the sheer will of the author determines if he, a proponent of this thesis, will be judged as being a materialist, dualist or idealist.

If you accept dualism as the appropriate answer to the question then you have to decide what kind of dualism, what are the exact two substances and how they interact. Your answers also partially reveal your motivation why you chose the dualist answer. The main arguments is simply the observation that the physical and mental seem so different or the fact how lifeless, but real, matter is judged as distinct from the not materially existing thoughts. Although for example [Havel 2001] mentions under dualism also other possibilities (here they are listed in the third monist option) here we consider only one dualism and that is

the substance dualism, where mental and physical are two completely distinct substances of the world and the crucial question rises- "how do they interact?" Descartes' view, held in the twentieth century by K. Popper and J. C. Eccles, is called interactionism and claims that mind and body interact in some way and that the mental states of a person have causal effects on the world and vice-versa. We can then study the way how and where this interaction occurs[14]. C. Campbell, T. H. Huxley and others are the defendants of epiphenomenalism, the view that physical phenomena can cause mental states but these do not cause any physical changes, they are just a by-product of the underlying physical changes. Another attitude, presented by N. Malebranche, is occasionalism which asserts that all causality is done by God's intervention, in both cases: simple physical or mental-physical causes. It can also be that the mental states and material behaviour correspond, occur synchronized, but do not interact in any way, a proponent of this view was for example G. W. Leibniz with his pre-established harmony[15].

Idealism is not much represented view because it bears many problems with it's relation to the common experience of people. Even G. Berkeley's philosophy that could be placed under idealism is more based on the need of perception of things in order to exist and in a way is either similar to Descartes' division, only replacing doubt with perception, or to a even materialist opinions which take into account the limits of our perceptive abilities.

Materialist views have many branches and attitudes towards the new questions rising from the supposition that there are no two substances but only one - the physical, but we still speak in a different manner about thoughts than about physical events. The first possible attitude is the behaviourist one which reduces all mental states to behaviour as the only observable effect of the mental life, the obvious problem shows Putnam's example where a good actor can fake a headache and a tough Spartan can not show he has one. A reaction to the behaviourism's problems was the identity theory which connects mental states with brain states (the firing of certain neurons in the brain). As a next step can be regarded functionalism which abstracts from the implementation of human mind(the brain) and concentrates on the functional relations(similarly a functional heart does not have to be biological and made out of muscle tissue, it is enough if it pumps blood in functionally the same way as

14 Descartes placed this connection point into the pineal gland. Modern neural research places, a non dualist, centre of the "I" into the dorsal medial prefrontal lobe. The pineal gland is in fact a endocrine gland, for example producing the melatonin hormone which influences sleep rhythms, see [Koukolík 2002].

15 But this already is one step closer to monism because Leibniz believed that everything is reducible to monads but it is important to distinguish the mental and material lines of interaction which do not interact in between.

the real heart does). Functionalism is very popular in AI or computer science because of the similarities with software and hardware. The possible absurdity of functionalism was shown by N. Block who suggested the Chinese computer, if we have enough people who interact in a way that is functionally the same as interactions in a brain, then they would need to have a common over-mind. We have seen also connectionist networks which are actually based on a emergentist thesis, which say that some properties are emergent from a structure if these properties are not reducible to the properties of the parts of the structure but occur only in the structure as a whole. In the minds case, mind and subjective states of mind cannot be reduced onto properties of single neurons. For example Searle presents the mind as a feature of the brain that is caused by the brain processes. But as [Havel 2001] points out, this is not really a good answer because it replaces the enigmatic mind with an enigmatic process of emergence.

The third described option are monisms that have a neutral substance or are in the intersection areas of multiple above mentioned views. Firstly it is the dual aspect theory, a kind of weakening of dualism, where mental and physical do not exist as two different and independent substances but are two irreducible proprieties of humans (and animals supposedly), thoughts are in one aspect those mental thoughts, intentions and so on, in another aspect they are brain processes. This theory has obviously many in common with for example the already mentioned identity theory or even emergentism. Defendants of this view are D. Chalmers and already W. James suggested a similar attitude [James 1904]. Also Searle makes a clear distinction between two ontologies, the first person ontology and the third person ontology, because our first person perceptions are utterly different from the observable behaviours and states. However this distinction, even if creating two irreducible stances, does not suggest dualism or as a matter of fact not even monism(as shown in [Kelemen 2001]).

6.2 The assumptions

Without any specification what are we using, aiming to prove and what are our premises for the final claims the thesis becomes foggy and inaccurate. This can be shown on Searle's Chinese room thought experiment which is a counter example for a very limited kind of AI for which it is quite convincing, but without the knowledge what is the experiments aim and what are it's premises it becomes useless [Searle 1980]. Herein lays the reason why we need to clarify what will be the foundations on which the following argumentation is built. In the next section we shall see two claims, a strong and a weak one, it is self-evident that the weak claim has more prerequisites to fulfil. Therefore we shall first look at those that are common for both.

Errors happen. It might seem superfluous to mention this commonly known fact but as Havel points out in the context of thought experiments and futurologist visions [Havel 2001] the implicit and often unpronounced assumptions are those that will significantly shape the final picture, therefore we should mention even assumptions as this one. If someone believes that there will not be any errors, mistakes, deviations, either caused by a human or by a small electrical particle, either natural or made out of lack of knowledge in a given field, then he should not be troubled by the conclusions of this paper.

Strong AI is feasible.} It can be that mankind will never arrive to the creation of a strong AI(here not necessarily a symbol manipulating system), but at least it should advance towards that goal. As we shall see, it is not the final human-like artificial mind that is needed for the argumentation, nevertheless it is necessary that AI programs do not aim just for basic design AIs that have precise tasks assigned and are relatively simple, because in these cases malfunction would still stay the same as it is in present AIs.

AI's mind is similar to human mind. Although implicitly already given in the previous point, this premise deserves it's own mentioning. In some way it must be similar to the human mind and not just for the sake of this argumentation. As already Dennett points out, minds we try to find must be similar to our minds, if its not the case then we would not call it a mind [Dennett 2004]. It seems also sufficient if the mind is at least "translatable" into the human language of mind. This does not suggest that there is an abstract and given language of the human mind, it just points to the similarities all people, on average, have concerning their ways of thinking and their mind. The AI can be working on a different principle than human minds, but if it will be still well describable by for example intentions, plans, goals, etc. it

should be still possible to draw the strong version of the conclusions. Also if the AI's mind would be utterly alien to the human mind, then it is obviously impossible to speak about it now because it could take any form.

Brains cause minds. This Searle's fourth axiom [Searle 1990] represents here a wider idea than in the original text. If brains cause minds then we admit in one sentence that substantial dualism is not accepted because brains are the origin of minds, that minds could be, at least in theory, constructed by mimicking fully, but by artificial means, the brain and also that any artificial mind will necessarily not be completely the same as the human one because it will not have all the same structures as the human brain has(with extension to hormones, nerves in the body etc.).

Malfunction as a non discriminating concept. The "discrimination" is connected to a language barrier of using distinct words for biological or human and mechanical issues that could cause needless problems in the course of argumentation. This premise is also more of a agreement on the vocabulary used than a requirement that must be fulfilled. The full explanation of the concept and reasons for it's present use are in the first part of this paper.

To the listed premises we can add more which shall already be regarded as weakening ones. These can stay unfulfilled and for some paths of AI research these are not extrapolations of their research effort. However the first listed ones seem to be necessary whatever the AI project will be if he tries in any way come closer to an artificial mind of any kind.

AI is a designed project. Therefore AIs, even with minds, will still have a purpose which they can fail and therefore malfunction. This comes from the fact that AIs must be made by someone knowingly. It is imaginable that an AI is created unwillingly during a process and therefore has no purpose, this is the reason why the premise is given as an additional weakening option to this list.

AI will be inspired by humans. The reason why this is just an additional assumption is clear. Multi-agent systems for example do not need to be inspired by human way of thinking nor with their brains (at least not directly). We shall see that any level of inspiration can carry with it not just the merits but also the flaws of source. It is here where lays the line between the above mentioned translatable minds and the human inspired artificial minds. The second case carries with it actual artificial implementations of processes we can witness in the human mind or brain but the first one is only in a functional sense translatable onto the products of these and thus represents to some extent a more behavioural stance.

A gradual transition between man and animal. This premise might seem also redundant, but although we all accept the human evolution there seem still to be a tendency to place man on a completely different level than are animals. However even in the field of mind research and especially shown by *human* neurology- the structure of the brain includes all the evolutionary older versions of a cortex. Therefore is a gradual transition which in a way supports the views of scientists like R. A. Brooks [Brooks 1990] and the Nouvelle AI. And research in experimental psychology with animal subjects is used for (correct) conclusions about human psychology[Cummins 1998] and additionally unveils that animals have also a mental life [Shekhar 2001].

We see that these assumptions are not easy to fulfil and some even seem contradictory, but these issues will be dealt with in the next section.

6.3 Marvinian leap[16]

The present still knows only robots, AIs, androids that do not seem to have any kind of mind, at least after a thorough analysis they do not have. Human cognition is habituated to look for human signs around, to attribute human behaviour not just to animals but even to inanimate objects. This can be regarded as a quite harmless and willing stance of observation [Dennett 2004] and also it might be that our cognition was created in order to analyse social situations [Dawkins 1998] and so, as optical illusions occur due to our visual system [Cummins 1998], these kinds of social illusions occur also. Therefore we must tread lightly not to spring the traps of our vivid social imagination in context with machines. Otherwise we will go for a fruitless hunt for phantoms and shadows or, if you want to slightly misuse another term, ghosts in the machine.

On the other hand, as also the concept chosen and (re)defined of malfunction tries to suggest, we should not regard man made structures as impossible of reaching natures level in complexity and capabilities and catching up with human brains. Probably even today, fifty years after Newell and Simon's attempt to list Logic Theorist as a co-author of a paper[17], we still believe there is a barrier that cannot be overcome. This barrier seems to be built only by human fear or the desire for uniqueness and superiority, which is shown by the similar attitude towards the animal kingdom.

Let us, for now, assume, there is a possible gradual transition from inanimate matter to a goal driven system and even to a mind governed system. It seems that this step is an inevitable outcome if we take into account the evidence already mentioned in this paper earlier. The arising question now is if these artificial creations should be the same as the natural creations we are familiar with and in particular, will they have the same malfunctions as the natural ones?

With the strong assumptions given above, let us formulate the strong malfunction thesis:

Intelligent systems are susceptible to malfunctions.

This seems first as an obvious statement, but let us not forget the range of definition for the concept malfunction. Here we can now draw the parallels between biological systems and

16 Marvin - a fictional character in The Hitchhiker's Guide to the Galaxy series by Douglas Adams, he is a robot suffering from depression.

17 Logic Theorist proved a part of the theorems from Whitehead, Russell's Principia Mathematica but was not allowed by the Journal of Symbolic Logic as a coauthor, http://www.cs.swarthmore.edu/~eroberts/cs91/projects/ethics-of-ai/sec1_2.html.

artificial ones[18]. We will not take up the futuristic visions and reasoning like in Creating a friendly AI, let us simply deduce from the given premises and with the given definitions the necessary conclusions.

First, what happens if we omit some of the assumptions. Obviously not agreeing on the use or range of the word malfunction would cause confusion and different understanding of the thesis. The next easily explainable assumption is that if we are not able to create a strong AI, as described above, then we could not draw any similarities between human and AI minds, because the AI would not have one, and the same goes for creating an AI that is utterly different from us, because then the two minds would not have any similarities to compare or at least it could be difficult to do so. As described in [Horáková 2007] it seems that communicating AIs will have to be able to communicate and exist in human societies and therefore have some common ground, but still could evolve their own culture which could be then, if completely AI based, hardly judged by humans. The premise speaking about the necessity of errors is more of a friendly reminder not to assume that the AI created by man will be flawless, but in the same moment it does not say that these errors need to be the cause of any malfunctions. If we do not accept that brains cause minds(and remember, it does not need to be in Searle's original meaning) then we would need to explain in a very peculiar, probably too complicated way, how it happened, that a machine got a mind (already the biological machine of man). It is completely irrelevant which theory of those many that explain how mind raises from the gelatin tissue of our brains we prefer, but if brains do not cause minds, where do they come from?

Now let us explain a bit the thesis and it's consequences. As already mentioned during the explanation of the concept of malfunction, some instantiations are trivial. If we ram a lead pipe into the AIs system or pour a bucket of water on a not waterproof AI then it will malfunction as humans do when we ram a lead pipe into them or pour acid on the not acidproof ones. But even in human society mental disorders are a category of malfunctions that are not (always) trivial to identify. An aspect, that cannot be omitted is that human minds are a result of evolution, as also pointed out in CFAI and other publications, and not of design or perfection. Evolution forced only the minimal needed changes to survive, not the optimal or best and also had to react to many various factors that do not play today any role and it

18 The division of biological and artificial probably is not correct because as [Amos 2008] shows, biology based computing could lead to biological man-made machines. Nevertheless this distinction seems for now as usable with the implicit understanding of biological as evolved by natural selection without the direct interference of man.

could not react to those that were not present in the evolutionary history of our species[19]. We cannot simply, in the human context, judge all mental disorders as a fault, see the example of the dissociative identity disorder(also known as multiple personality disorder). This disorder, one of the commonly known and misunderstood, has from an evolutionary point of view an important role for the subject in the case he was confronted with a trauma, especially during early years, when creating another personality allows the subject to function and survive in some way [Cummins 1998]. In human (and supposedly animal) cases these disorders are judged then on a case by case basis or depending on neurological findings.

Our argumentation now steps on thin ice because we do not know precisely what AI are we dealing with. However all the models of AI mentioned above can be in a way affected by a malfunction that would in a (human) biological system be regarded as mental disorders. Because of the lack of a precise implementation it is unnecessary to go into details, but let us just point out some examples to show the idea. A symbol manipulating system, if able to create our AI, would be obviously a great candidate for systematized delusions - delusions based on a false premise but with completely correct deductions. In the case of neural networks epilepsy and learning input deprivation seem as possible and primal choices to demonstrate mental illness parallels. Actually the neuronal network based AI has the most possible mental disorder parallels simply because of it's similarity to the human brain architecture that being also the reason why connectionist networks are used in studies of schizophrenia and similar disorders. In the case of DAI we can take into account mental disorders that are connected to problems in communication between functional systems in the brain (starting with amnesias and ending with neglecting syndromes). The case of a hypothetical seed AI is complicated because of the attitude of their proponents, who claim to be able to create one day a perfect "seed". If we accept this premise, then it is an AI not usable in these parallels. But the case of a seed AI as depicted by the Singularity Institute demonstrates also the point of needed similarities between AI and biological intelligence. Therefore it does not fulfil one of the assumptions, because it is, by definition, a perfect AI. A question of similarities is also interesting in the view of other human mental disorders and states, as for example idiot savants (would we call a mind bearing expert system a idiot savant?). Obviously the more different the AI would become, the more difficult would it be to search for parallels. Although if there is maintained the demand of a translatable mind into human terms, human mental illnesses could be, based on more psychological than

19 See for example Dawkins opinion on aging [Dawkins 1998]

neurological(therefore architecture dependent) description, translated back into the AI mind's terms. The analysis of the term of malfunction shows a spare method of determining mental disorders in the hypothetical mind bearing machine, a method human disorders were defined in the first place, the statistical comparison with the average member or with the (assumed) needs and goals of the individual.

By adding the first of the additional assumptions we can weaken the thesis by limiting it to a certain type of AIs. It could happen that the given AI was a result of some aleatory, not metagoal directed process(as we are results of the evolution). In that case the AI has no given design which can be used as judging basis. Actually a aleatory made AI could be even more free of purpose than humans are (we are still the machines of survival for our genes or memes [Dawkins 1998]). But it seems more probable that the AI will be a result of some human effort which carries with it a goal that is, at least in the vagueness as we have from our genes/memes, imposed on the AI. Interestingly in this point seed AI can again come into play because seed AI's project is strongly goal driven of creating a human friendly AI, which would obviously malfunction if being unfriendly. Any other AIs' success, if we add this assumption, depends on her final goal and if the malfunction also influences that goal. However, these malfunctions would be already a kind of different malfunctions than are often in the case of people, at least according to the traditional view. Mental disorders are not regraded in human society mainly as causing troubles with some imaginary or real goals imposed on us, more they are taken as obstacles in possible goals that we impose on ourselves.

If we take into account also the premise that the AI will be strongly inspired by humans, then the source of possible mental disorders is clear. The more similar these two systems are, the more similar their problems will be. If we would assume a similar way of creating artificial intelligence as in some of the popular culture examples[20], then these AI would be susceptible to the same mental disorders as humans are. Which would obviously make it easier to draw parallels.

The last not tested assumption is connected to a gradual evolution on the tree of life. If we admit that this is true, then AI would be confronted on a much lower level with possibilities of less trivial malfunctions. Because for example the Nouvelle AI program tried in a way to focus on these lower levels of cognition, then results in animal psychology prove

20 As in K. čapek's R. U. R. or the TV series Battlestar Galactica.

that the program is not safe from similar problems as the others have. This is also important to take into account because of a effect of closing scissors - experimental psychology tests hypotheses about mental life in the case of animals (and also infants) and shows how little is needed to gain quite complex mental life and also develops methods that could be in the end used up in the field of AI to test AI systems, in the meantime AI is trying to work herself up on the imaginary ladder of complexity of minds. By adding this assumption we admit the possibility that these two scissor blades will reach each other one day and cut a hole into both fields of research.

If we want to gain a final, weakest thesis, we can add all the assumptions up and make an intersection of their effects. This weakest thesis seems to be an inevitable conclusion of them, let us formulate it clearly:

> *Any system that will be based on human architecture or ways of functioning, will proportionally to the amount of inspiration also suffer from human-like problems, including mental disorders.*

7. C

8. onclusion

Hopefully this paper has successfully shown the point postulated as the weakest thesis and also the dependency of AI on it's architecture. If we leave the unnecessary clinging to the exceptional status of the mind, then we can see that problems that man had with his mind would occur again if we use these minds as a basis for the construction of the artificial ones. Reminding ourselves of this risk can only be useful. What still stays an open subject for research and investigation is the mind. Would it be possible to construct an abstract theory of mind as when Nils Nilsson speaks about the ideal AI as aerodynamics, which works as good for the study of flying of planes and birds [Kelemen 2001]? Can we really uncover the processes that constitute our thoughts even if neurology, psychology, philosophy and AI have show the immense number of pitfalls and illusions that surround us? Probably a more directed research similar to the efforts of creating the minimal living organism [Amos 2008] could be started with the goal of finding the minimal amount of components needed for a sentient being by eliminating one by one human systems, groups of neurons till we arrive to a necessary minimum. The combined effort of sciences approaching a common subject slowly solves one question after another and so it is already a pressing issue if we are prepared to get the final answer. So the exhausting siege by empirical science of one of the few bastions of philosophy that are left continues.

9. S

10. eznam použité literatury

AMOS, Martyn. *Na Úsvitu živých strojů* (Genesis Machines). Praha: Mladá fronta, 2008. ISBN 978-80-204-1674-2.

Artificial intelligence. In *Encyclopædia Britannica*. [online] Chicago(Illinois): Encyclopædia Britannica, Inc., 2009- [cit. 2009-23-03] Dostupný z WWW: http://www.britannica.com/EBchecked/topic/37146/artificial-intelligence

Best brain boosts artificial life. In *BBC News.* [online] London: British Broadcasting Corporation ,2007- [cit. 2009-23-03] Dostupný z WWW: http://news.bbc.co.uk/2/hi/science/nature/250343.stm

BROOKS, Rodney A. Elephants Don't Play Chess. In *Robotics and Autonomous Systems*, vol. 6, no. 1\&2., 1990,s. 3-15.

COPELAND, Jack. What is Artificial Intelligence. In *Alanturing.net* [online]. University of San Francisco. Dostupný z WWW: http://www.alanturing.net/turing_archive/pages/Reference%20Articles/What%20is%20AI.html

Creating friendly AI [online] Palo Alto, CA: Singularity Institute for Artificial Intelligence, Inc. 2001- [cit. 2009-23-03] Dostupný z WWW: http://www.singinst.org/upload/CFAI.html

CUMMINS, Denise D. *Záhady experimentální psychologie* (The other side of psychology). Praha: Portál, 1998. ISBN 80-7178-186-X

DAWKINS, Richard. *Sobecký gen* (The selfish gene). Praha: Mladá fronta, 1998. ISBN 80-204-0730-8.

DENNETT, Daniel C. True believers: The intenional strategy and why it works. In HAUGELAND, John (ed.). *Mind design II.* 1997, s. 57-80.

DENNETT, Daniel C. *Druhy myslí* (Kinds of minds). Praha: Academia, 2004. ISBN 80-200-1177-3

DENNETT, Daniel C. Can we close the carthesian theatre. In *Vienna Conference on Consciousness, 26 September, 2008* [online] Vienna: Department für Verhaltensbiologie Universität Wien, 2008 [cit. 2009-23-03] Dostupný z WWW: http://vcc.univie.ac.at/index.php?id=28426.

DIJKSTERHUIS, Ap; NORDGREN, Loran F. *A* Theory of Unconscious Thought. *Perspectives on Psychological Science*, 2006, roč. 1, číslo 2, str. 95-109.

DREYFUS, Hubert L. From Micro-Worlds to knowledge representation: AI at an impasse. In HAUGELAND, John (ed.). *Mind design II.* 1997, s. 143-182.

General intelligence and seed AI [online] Palo Alto, CA: Singularity Institute for Artificial Intelligence, Inc. 2001- [cit. 2009-23-03] Dostupný z WWW: http://www.singinst.org/ourresearch/publications/GISAI/GISAI.html

HAUGELAND, John (ed.). *Mind design II : philosophy, psychology, artificial intelligence* . Cambridge, Mass.: MIT Press, 1997. ISBN 0-262-58153-1

HAUGELAND, John. What is mind design. In HAUGELAND, John (ed.). *Mind design II.* 1997, s. 1-28.

HAVEL, Ivan M.: Přirozené a umělé myšlení jako filozofický problém. In MAŘÍK, V. *Umělá inteligence 3.* 2001, s. 17-75.

HERTZ John A.; KROGH, A.; PALMER R. *Introduction to the Theory of Neural Computation.* Addison-Wesley, 1991. ISBN 0-201-50395-6

HORÁKOVÁ, Jana; KELEMEN, Jozef. Robot - stroj a metafora 20.století. In MAŘÍK, V. *Umělá inteligence 5.* 2007, s. 43-74.

JAMES William. Does 'Consciousness' Exist? *Journal of Philosophy, Psychology, and Scientific Methods,* 1904, vol. 1,no. 18, s. 477-491.

KELEMEN, Jozef. *Kybergolem.* Olomouc: Votobia, 2001. ISBN 80-7198-504-X

KONOLIGE, Kurt G. *Experimental Robot Psychology.* Technical Note 363. AI Center, SRI International, 333 Ravenswood Ave., Menlo Park, CA 94025, Nov 1985.

KOUKOLÍK, František: *Lidský mozek.* Praha: Portál, 2002. ISBN 80-7178-632-2

MAŘÍK, Vladimír; ŠTĚPÁNKOVÁ, Olga; LAŽANSKÝ, Jiří. *Umělá inteligence 1.* Praha: Academia, 1993. ISBN 80-200-0496-3

MAŘÍK, Vladimír; ŠTĚPÁNKOVÁ, Olga; LAŽANSKÝ, Jiří. *Umělá inteligence 2.* Praha: Academia, 1997. ISBN 80-200-0504-8

MAŘÍK, Vladimír; ŠTĚPÁNKOVÁ, Olga; LAŽANSKÝ, Jiří. *Umělá inteligence 3.* Praha: Academia, 2001. ISBN 80-200-0472-6

MAŘÍK, Vladimír; ŠTĚPÁNKOVÁ, Olga; LAŽANSKÝ, Jiří. *Umělá inteligence 5.* Praha: Academia, 2007. ISBN 978-80-200-1470-2

Merriam-Webster [online] Chicago(Illinois): Encyclopædia Britannica, Inc. 1996- [cit. 2009-23-03]. Dostupný z WWW: http://www3.merriam-webster.com/opendictionary/

MURPHY, Gregory L. *The big book of concepts.* Cambridge, Mass.: MIT Press, 2002. ISBN-10: 0-262-13409-8

OXFORD. *The Oxford English dictionary.* Oxford: Clarendon Press, 1989. ISBN 0-19-861186-2

PRESCOTT, S. A., Sejnowski T. J., de Konick Y.: *Research of anion reversal potential subverts the inhibitory contorl of firing rate in spinal lamina I neurons: towards the biophysical basis for neuropathic pain.* http://www.molecularpain.com/content/2/1/32, 2006

RAMSEY, William; STICH, Stephen; GARON, Joseph. Connectionism, Eliminationism, and the Future of Folk Psychology. In HAUGELAND, John (ed.). *Mind design II.* 1997, s. 351-376.

SEARLE, John. Minds,brains, and programs. In HAUGELAND, John (ed.). *Mind design II.* 1997, s. 183-204.

SEARLE, John. Is the Brain's Mind a Computer Program? *Scientific American*, 1990, vol. 262, no. 1, p26-32.

SHEKHAR, A.; McCANN, U. D.; MEANEY, M. J. Summary of a National Institue of Mental Health workshop: developing animal models of anxiety disordes. *Psychopharmacologia,* 2001, vol. 157, no. 4, s. 327-339.

ŠTĚPÁNKOVÁ, Olga; MAŘÍK, Vladimír; LHOTSKÁ, Lenka. Distribuovaná umělá

inteligence. In MAŘÍK, V. *Umělá inteligence 2.* 1997, s. 142-177.

TURING, Alan M. Intelligent machinery. In *Turing's Anticipation of Connectionism* [cit. 2009-23-03] Dostupný z WWW: http://www.alanturing.net/turing_archive/pages/Reference%20Articles/connectionism/Turing%27santicipation.html

TURING, Alan M. Computing machinery and Intelligence. In HAUGELAND, John (ed.). *Mind design II.* 1997, s. 29-56.

WARWICK, Kevin. *Úsvit robotů- soumrak lidstva.* Praha: Vesmír, 1999. ISBN 80-85977-16-8

Wikipedia : the free encyclopedia [online]. St. Petersburg (Florida) : Wikimedia Foundation, 2001- [cit. 2009-23-09]. Dostupný z WWW: en.wikipedia.org.

YOUNGSON, Robert M.: *O šílenství, podivínství a genialitě* (The madness of Prince Hamlet). Praha: Portál, 2000. ISBN 80-7178-401-X

YOUNGSON, Robert M. *Vědecké omyly, bludy a podvrhy* (Scientific Blunders). Jinočany: H&H, 2004. ISBN 80-86022-84-6

ZELINKA, Ivan: *Umělá inteligence: Hrozba nebo naděje?* Praha: BEN - technická literatura, 2003. ISBN 80-7300-068-7