Artificial Intelligence –
a Challenge for our Future

Theological and Ethical-Moral Reflections

English version of an Academic Thesis
for the attainment of the degree "Magister theologiae"

by

Martin Wan

Rheinische Friedrich-Wilhelms-Universität Bonn
Faculty of Catholic Theology
Seminary of Moral Theology

Supervisor: Prof. Dr. Gerhard Höver

Submitted: 18 June 2014
English translation: August 2023

Preface

Since the general availability of generative AI tools such as ChatGPT, Stable Diffusion and Midjourney, the debate about the capabilities and risks of AI has reached a broad public, after years of relative silence on the subject. Unfortunately, the debate is not always conducted in a objective manner – due to a lack of technological understanding of the underlying principles of AI, the media as well as the public all too often unreflectively adopt the promises and hopes of the AI community, which have already repeatedly proven to be unfounded in the past decades.

Despite the admittedly impressive functionality of current generative AI tools, even in 2023 there is no reason to assume that AI has developed a consciousness and become a social counterpart in the meantime, even though the opposite is sometimes claimed. Nine years after the submission of my work, AI remains – albeit complex – software and, like the very first computer programmes, functions according to the scheme of input-processing-output. The fact that the "processing" sub-step is less and less understood in the face of the increasing complexity of software and neural networks does not magically turn computer programs into conscious beings all of a sudden. Incidentally, Joseph Weizenbaum, to whom a large part of this work is dedicated, observed and predicted this development of increasingly complex and less understood software more than 50 years ago.

This text is a translation of my master's thesis, which I submitted to the Faculty of Catholic Theology at the University of Bonn (under my birth name Martin Rademacher) in June 2014. The text is unchanged. As far as the literature was available to me, I have used original English-language editions in the translation instead of German-language translations. I checked the availability of all the Internet references I used; where they were no longer available nine years after the initial submission, I resorted to the Internet Archive.

During the renewed intensive study of my work in the course of the translation into English, I was able to ascertain that most of the theses dealt with here remain valid, and in view of the current hype surrounding AI, are in part more up to date than ever. With one exception: although I tried to take as sober an approach to the topic as possible at the time, I allowed myself to be misled by the supposedly advanced state of development of self-driving cars back then: An example of how uncritical the reporting on supposed progress from Silicon Valley often is.

Similarly, the prophecies of techno-euphorics and transhumanists that I analysed have, for the most part, not come true or closer to being true until today. Although these groups present themselves as scientific, they are essentially practising science fiction. Worse still, their theses of salvation through technological progress, immortality through brain upload and even creating God through AI must be named for what they are: A substitute religion for those who are offended by the fact that even the greatest economic wealth can offer no way to overcome death.

With the publication of my work, I hope to contribute to a little more sobriety in the debate about artificial intelligence and at the same time to direct attention to the true dangers of AI. These dangers do not lie in the utopian sudden consciousness of machines and the submission of humanity. The real dangers arise when we use AI software carelessly and without reflecting on its technical basis and limits for taks it is not capable of.

Martin Wan
Bonn, August 2023

Web: https://wan.digital
Mail: mail@wan.digital

Table of Contents

I. Introduction

A major concern of Pope emeritus Benedict XVI was the compatibility of faith and reason, of science and theology. In his Regensburg speech in September 2006, he warned against a concept of reason narrowed to empiricism, which can no longer answer the questions of wherefrom and whereto and thus reduces the human being himself.[1] At the same time he emphasised, in contrast to religious fundamentalism, that actions contrary to reason contradict the nature of God,[2] and thus promoted a new coming together of theology and science.[3]

Theology inevitably makes itself untrustworthy when it contradicts established scientific knowledge. At the same time, natural science must always be aware of the limits of its empirical method. Pure empiricism, as Benedict acknowledged, is a useful scientific tool – but it cannot substantiate the world and man.[4] In this respect, scientific fundamentalism that puts itself in the place of God and believes in its ability to fully explain human beings in an empirical way must be critically questioned.

In artificial intelligence, we are partly dealing with such fundamentalism. The school of *strong artificial intelligence* (AI) believes that humans can be completely (!) described in the categories of natural science and that, once their brains are fully understood, they can be replicated in the form of software. The *cognitivist* school even tries to explain the human brain as a digital computer. Two of the main protagonists of strong AI, Ray Kurzweil and Hans Moravec, go so far as to claim that we must first create God with the help of artificial intelligence in order to complete the human race.

The school *weak artificial intelligence* is more moderate: for them, the simulation of the brain is always just a model of the brain. Nonetheless, this simulation of intelligence already delivers impressive results today and it is hard to imagine our everyday life without it. We now take it for granted that aeroplanes take off and land with the help of computer software. We have long since become accustomed to topic-specific search suggestions on *Google.* The introduction of personal smartphone assistants like *Siri* which

[1] Cf. Pope Benedict XVI, Address on 12 Sept 2006, 736.

[2] Cf. ibid., 731f.

[3] Cf. ibid., 738.

[4] Cf. ibid., 737: “The greatness of modern intellectual development is undiminishedly acknowledged: We are all grateful for the great possibilities it has opened up for man and for the advances in humanity that have been given to us. The ethos of scientificity [...] is, moreover, the will to obey the truth and, in this respect, the expression of a fundamental attitude that belongs to the essential decisions of Christianity. What is meant is not retraction, not negative criticism, but the expansion of our concept and use of reason.” (Translation: MW).

are able to recognise natural language and provide appropriate useful information, happened only a few years ago. Elderly and sick people hope for more independence in everyday life with the development of intelligent robots. Google is currently successfully testing driverless cars that are controlled fully autonomously with the help of software and have the potential to significantly change mobility in the coming decades.

The topic of artificial intelligence is currently virulent in the media: hardly a day goes by without new news on AI. In December 2013, it was announced that Google had taken over the military robot manufacturer Boston Dynamics.[5] In April 2014, Google announced impressive progress in the use of its autonomous car in urban traffic.[6] In mid-June 2014, it is announced that the chat robot *Eugene* is the first computer program to pass the Turing test[7].[8] *Transcendence* and *Her,* two Hollywood films dealing with the subject of conscious, intelligent software, are released simultaneously in early 2014.

Theologically, however, artificial intelligence has hardly been discussed so far. At the same time, old questions about the relationship between body and soul are being posed anew, especially by the school of strong artificial intelligence. But even apart from the question of whether artificial intelligence will ever be able to produce consciousness, a theological discussion of AI is of interest, for example when it comes to the ethical evaluation of robots in the care of the elderly or the use of drones in war.

This paper is divided into three parts. The first part will give a brief overview of the history of artificial intelligence before turning to the school of strong AI and two of its main representatives: Ray Kurzweil and Hans Moravec. Kurzweil's concept of the *singularity* bears strong traits of a surrogate religion, which is why his theses of strong AI will be compared with the criteria of classical religious systems.

The second part deals with the philosophical and humanistic criticism of strong AI. Among others, it is dedicated to Joseph Weizenbaum, an early protagonist and critic of AI research. In order to better classify the theses of strong AI, an overview of the current state of the mind-brain debate is provided. John R. Searle, Margaret Boden, Hans-Dieter Mutschler and Dirk Evers also provide a brief insight into the philosophical and theological discussion on this topic.

[5] Cf. news article "Google kauft zum Jahresende Militärroboter-Hersteller".

[6] Cf. news article "Googles autonome Autos unterwegs in der Stadt".

[7] A test proposed by Alan Turing in the 1950s to check whether computers can "think", cf. below, 21,

[8] Cf. news article "Computerprogramm ‚Eugene' besteht Turing-Test".

The third and last part deals with artificial intelligence as a challenge for Christian anthropology and practice. The ethical relevance of AI for our everyday life is shown by means of the topics of *robots in the care of the elderly*, *drones in warfare* and *autonomous cars in road traffic*.

II. Kurzweil, Moravec and the Man-Machine

"Robots will, as they do better and cheaper work,
displace humans from important functions.
Pretty soon they will even displace us from existence."[9]
Hans Moravec

"Before 2030, we will have machines proclaiming Descarte's dictum.
And it won't seem like a programmed response [...]. Should we believe them when they claim to be conscious entities [...]?"[10]
Ray Kurzweil

As mentioned at the introduction, the science of artificial intelligence research can be roughly divided into two schools. The proponents of *strong artificial intelligence* believe that any system that only implements the right computer program and that is fed the right inputs can produce consciousness in the same way as humans. In contrast, the *weak AI* school argues that the computer will only ever remain a helpful tool for understanding the human mind – computer programs that simulate the brain remain simulations, just as meteorological simulations help us understand the weather but do not become the weather itself.[11] I will refer to the two schools below as the *radical* and *moderate* schools. In this chapter I will devote myself to the presentation of the radical school and two of its main representatives: Ray Kurzweil and Hans Moravec.

The radical school sees artificial intelligence as the successor to natural, human intelligence. The representatives of this school are assuming that computers only need to become powerful enough to be able to carry out actual, human-like thought processes. That robots will eventually be able to think, feel and be aware of their existence is, for them, not a question of *if*, but of *when*. The school is characterised by an almost boundless faith in technology, which – I will show – takes on strongly religious character.

[9] Moravec, Computer übernehmen die Macht, 29, translation: MW.

[10] Kurzweil, The Age of Spiritual Machines, 60.

[11] Cf. Searle, Chinese room argument. "The contrast is that according to Strong AI, the correct simulation really is a mind. According to Weak AI, the correct simulation is a model of the mind", ibid.

This is especially true for Kurzweil's followers. In his books, Kurzweil advocates the idea of *singularity*[12]: it refers to the point in time when computers have attained a computing power that far exceeds that of the human brain. For Kurzweil, this is the point in time when machines will be more intelligent than humans. From then on, computers will be able to improve and duplicate themselves and, as a result, completely reshape the world.[13] Kurzweil promotes the idea of *transhumanism*, the fusion of human and artificial intelligence. With the help of intelligent technologies, humans would overcome their physical limitations and ultimately merge with computers to form a new species, Man 2.0.[14]

The followers of this idea call themselves *Singularians* and now form a worldwide movement. Kurzweil himself has founded a *Singularity University* in California, which is dedicated exclusively to the study of the Singularity.[15] The Berlin computer scientist and neural network specialist Raúl Rojas clearly characterises the movement as a substitute for religion: "If you pick apart the Singularians' argumentation, in the end only one thing remains: the desire to avoid death and to be able to live on forever as software."[16]

1. History of Artificial Intelligence

1.1 "Artificial intelligence" Before the 20th Century

Literary and real attempts to artificially create intelligent life date back to long before the 20th century, beginning with Greek mythology: Hephaestus, the god of fire, for example, sculpts Pandora on Zeus' behalf, whose box unleashes all the evils of the world; Pygmalion, out of aversion to women of flesh and blood, sculpts Galatea out of ivory, who is later given life by Aphrodite.[17]

Around the year 1300, the Mallorcan philosopher and missionary Ramon Llull was inspired in North Africa by the *Zairja*, a construction of Arab astrologers that produces certain answers from the combination of numerical values assigned to categories and philosophies. Llull adopted this concept for his *Ars magna*, a "logical machine" designed to produce insights through the mechanical combination of concepts.[18]

[12] On the term *singularity,* cf. below, 17.

[13] Cf. Rojas, Analoge versus Digitale Seele, section "Die Singularität", para. 1.

[14] Cf. Kurzweil, The Singularity Is Near, 300-310.

[15] Cf. Rojas, Analoge versus Digitale Seele, section "Die Singularität", para. 2.

[16] Ibid, section "Ars longa, vita brevis", para. 1.

[17] Cf. McCorduck, Denkmaschinen, 16f.

[18] Cf. ibid., 20f.

Until the middle of the 14th century, complicated clocks with moving figures were erected in many towns, which many people presumably believed to be alive. The physician and mystic Paracelsus (1493-1541) is credited with a writing in which he describes a *homunculus*, a small man artificially created from human sperm.[19]

Only a few decades younger is the Jewish legend of the *golem of* Rabbi Judah ben Loew. Loew was the chief rabbi of the city of Prague and, to protect the Jewish population from pogroms, formed a human figure out of clay, which God breathed life into as soon as the tetragram יהוה was written on its forehead. Interestingly, important AI researchers such as Marvin Minsky or John von Neumann come from Jewish families who see themselves as descendants of Rabbi Loew.[20]

There are 18th century reports of a mechanical duck that Jacques de Vaucanson is said to have constructed. This duck is said to have moved its wings, drunk water, eaten grains and "digested" and excreted them. Vaucanson remained silent about the exact mode of operation throughout his life. However, an early "chess automaton" that won chess games all over Europe was later exposed as a fake: In reality, the playing table concealed a human chess player. The literature of the 19th century is full of artificially created creatures, including E.T.A. Hoffmann's *Sandman* as well as Mary Shelley's *Frankenstein.*[21]

1.2 The 20th Century until Today

During the parallel development of the first digital computers in different places at the beginning of the 20th century, the idea of creating artificial thinking was present throughout. However, no one formulated it as explicitly as the British mathematician Alan Turing. As early as the 1930s, he formulated the idea of the Turing machine, a universal computer that could perform any conceivable computer task.[22] In September 1947, he wrote a paper in which he expressed the view that man is nothing more than a machine and that electrical circuits could perform the same tasks as human nerves. In October 1950, Turing proposes a test procedure to determine whether machines can think – the so-called *Turing Test.* For him, the differences between human and machine thinking operations are only of gradual, not species-related nature.[23]

[19] Cf. ibid., 22f.

[20] Cf. ibid., 23f. and Foerst, Von Robotern, Mensch und Gott, 50.

[21] Cf. McCorduck, 25-27.

[22] At that time, no computer comparable to the Turing machine existed, cf. ibid.

[23] For more on the Turing test, see below, 21. Cf. McCorduck, 63-66.

Turing's optimism is not shared by another pioneer of computer science, the Austro-Hungarian mathematician John von Neumann. Although he sees the human nervous system as a source of ideas for the development of computers, he does not believe in a final solution to the problem of achieving human-like thinking in machines. Nevertheless, he is full of enthusiasm about the variety of tasks that computers can handle.[24]

Around 1950, the Russian-American science fiction author Isaac Asimov formulated his *Three Rules for Robotics:*

1. Robots must not harm people or leave them idle.
2. Robots must obey human commands as long as this does not put them in conflict with the first rule.
3. Robots must protect their own survival as long as this does not put them in conflict with the first or second rule.[25]

In 1956, at the suggestion of the computer scientist John McCarthy, four scientists, including Marvin Minsky, meet at the so-called Dartmouth Conference to discuss "intelligent" machines. The term *artificial intelligence* is coined at this conference, although it is controversial among those present, and later used by Marvin Minsky for his publications. However, the results of the conference are initially disillusioning for the participants: "Everyone present was quite stubborn in pursuing their own ideas [...]. Moreover, as far as I could see, there was no real exchange of ideas," McCarthy is quoted as saying.[26] The following two decades also revise the initially almost boundless expectations when it turns out that intelligent computer programs are much more difficult to produce than initially suspected.[27]

Starting in 1957, the social scientist Herbert Simon and the computer scientist Alan Newell develop a programme that is supposed to realise a simulation of human problem solving. This *General Problem Solver* codifies a whole range of problem-solving techniques and is indeed successfully tested on several puzzle problems. Nevertheless, the project is considered a failure after ten years and is discontinued.[28]

[24] Cf. ibid., 69-72.

[25] Cf. ibid., 32.

[26] Cf. ibid., 97-100, here: 99, translation: MW.

[27] Cf. ibid., 102.

[28] Cf. ibid., 203-206.

In 1963, the German-American computer scientist Joseph Weizenbaum became a visiting professor at the Massachusetts Institute of Technology. There, in 1966, he presented the computer programme ELIZA, which simulates a psychotherapeutic conversation situation and with which the user can converse in natural language. The programme demonstrates to a wide audience how information processing and artificial intelligence work; Weizenbaum himself is shocked by the overwhelming reaction to his actually quite simple programme.[29]

Convincing reports of successful Turing tests are still rare. But in select special fields, computers have gradually been able to outperform humans: As early as 1979, software succeeded in beating the acting backgammon champion, since 1986 the program *Maven* has been winning Scrabble games, and in 1997 the IBM computer *Deep Blue* beat the then-current world chess champion Gary Kasparov. In 2011, IBM succeeds in designing a computer called *Watson* with 2880 processor cores and 14 terabytes of RAM and programming it to win against human contestants in the US quiz show *Jeopardy.*[30]

Today, artificial intelligence – or at least what is referred to as it[31] – is ubiquitous in everyday life: internet companies such as Google use it to be able to place relevant advertisements for each user, digital assistants in smartphones such as *Siri* interpret spoken instructions and search appropriate answers for the user, weather forecasts are created with the help of sophisticated computer models, the military is relying more and more on unmanned robotics, high-frequency trading on stock exchanges is largely automated, and the technology for driverless cars is very advanced. Without the development of artificial intelligence, the world as we know it today would be unimaginable.

2. Kurzweil: Biographical Sketch

Raymond Kurzweil was born in 1948 as the son of Austrian secular Jews who had emigrated to the USA in 1939 shortly before the start of the war. Early on, at the age of five, he wanted to become an inventor. His parents brought him up Unitarian, i.e. pantheistic-humanistic.[32] During this time, Kurzweil noticed numerous parallels, but also contradictions between the world's religions: "It became clear to me that the basic truths

[29] Cf. Weizenbaum, Inseln der Vernunft im Cyberstrom, 89-100. For more on ELIZA, see below, 51.

[30] Cf. Rojas, Warum „Watson" ein Durchbruch ist, paras 1-3.

[31] Cf. below, 57.

[32] "We would spend six months studying one religion – going to its services, reading its books, having dialogues with its leaders - and then move on to the next. The theme was 'many paths to the truth'", Kurzweil, The Singularity Is Near, 1.

were profound enough to transcend apparent contradictions."[33] This experience feeds his philosophy that a good idea can successfully solve even seemingly insoluble challenges: "My life has been shaped by this imperative. The power of an idea – this is itself an idea. [...] This, then, was the religion that I was raised with: veneration for human creativity and the power of ideas."[34]

In 1970, Kurzweil graduated from the Massachusetts Institute of Technology with a bachelor's degree in computer science. In the years that followed, he emerged as the inventor of numerous groundbreaking technologies. From the combination of a flatbed scanner optimised by him with the technologies he developed for character recognition and speech output, he developed the first functioning reading machine for the blind in 1976, the so-called *Kurzweil Reading Machine*.[35] The development of this machine marks the beginning of a lifelong friendship with the blind musician Stevie Wonder, who buys the first Reading Machine. Inspired by him, he founded his company Kurzweil *Music Systems in* 1982, which introduced the *Kurzweil 250 in* 1984, the first synthesiser in keyboard form with the ability to accurately reproduce piano and orchestral sounds. Further inventions in the field of voice output and speech recognition followed, including the first speech recognition and control for Microsoft Windows in 1994.[36]

Since 1990, Kurzweil has published books in which he makes predictions regarding artificial intelligence, technological and civilisational progress. In particular, his books The *Singularity Is Near* (2005) and *How to Create a Mind: The Secret of Human Thought Revealed* (2012) become bestsellers, and his ideas on the Singularity and transhumanism find a large following. *Singularity University,* founded by him and the pioneer of private space flight Peter H. Diamandis in 2008 in Moffett Field, California, is dedicated to the dissemination and research of these ideas. The institution is financed by Apple, LinkedIn and Google, among others.[37]

[33] Ibid.

[34] Ibid., 2.

[35] Cf. Kurzweil, Curriculum Vitae. Kurzweil's reading machine is "considered to be the world's first consumer product to successfully incorporate artificial intelligence technology", ibid.

[36] The technology underlying *Siri*, the iPhone voice control system introduced in 2011, is based on technologies from the company Nuance, which in turn are further developments of Kurzweil's original speech recognition, cf. ibid.

[37] Cf. Meijas, Unsterblichkeit für alle.

In December 2012, it was announced that Google had hired Kurzweil as *Director of Engineering*. Since then, he has been working on machine learning and language processing projects for computer systems.[38]

Kurzweil pursues the optimisation of his own body with all available means in order to stay young and healthy as long as possible. According to his own information, he currently swallows 150 pills a day, which are specially adapted to his state of health and whose effect he controls through regular tests. His goal, he says, is to "experience the full flowering of the biotechnology revolution. I think it's pretty close – it might be another 15 years."[39] Accepting human transience is out of the question for him:

> *"Whereas some of my contemporaries may be satisfied to embrace aging gracefully as part of the cycle of life, that is not my view. It may be 'natural', but I don't see anything positive in losing my mental agility, sensory acuity, physical limberness, sexual desire, or any other human ability. I view disease and death at any age as a calamity, as problems to be overcome."* [40]

Kurzweil is married and has two children.[41]

3. Moravec: Biographical Sketch

The computer scientist, futurist and transhumanist Hans Peter Moravec was born in 1948 in Kautzen, Austria, and his family emigrated to Montreal in 1953. He became interested in science fiction literature at an early age, especially in the technical plausibility of robots and time machines. At the age of fourteen, the loner, who by his own admission still finds interpersonal obligations exhausting to this day, creates a light-controlled robot turtle, and at the age of sixteen he builds his first computer.[42]

From 1965 he studied mechanical engineering in Montreal, followed by mathematics in Nova Scotia. In 1971 he graduated from the University of Western Ontario with a master's degree in computer science, and in 1980 he completed his doctoral thesis at Stanford

[38] Cf. Kurzweil, Press release f. 14 Dec. 2012.

[39] Kurzweil, Interview f. Dec. 2012, 57, translation: MW. In his monograph published together with the physician Terry Grossman in 2004, *Fantastic Voyage. Live Long Enough to Live Forever* (New York 2004), he devotes nearly 450 pages to strategies for maintaining health until the arrival of radically life-extending and life-enhancing technologies within the next few decades.

[40] Kurzweil, The Singularity Is Near, 210.

[41] Cf. Kurzweil, Curriculum Vitae.

[42] Cf. Schult, Crazy Hans.

on a robot with the ability to move autonomously in a natural environment. Since then he has worked at Carnegie-Mellon University in Pittsburgh at the Institute of Robotics, and has been a professor since 1995.[43]

Moravec met his wife, a theologian, during a long stay in hospital.[44] He describes himself as a "physical fundamentalist" and sees religions as "invented stories".[45]

In his books, he reflects on artificial intelligence and artificial life in the future. In his book *Mind Children,* published in 1988, he predicts that in the years between 2030 and 2040 robots will develop into a species of their own, taking their further development into their own hands. Moravec sees this species as the successor to humanity, as "children of our minds"[46]. Similar to Kurzweil, he predicts that we will be able to map our personality in the form of software and transfer it to robotic computers. The consequences would be immortality and arbitrary copyability.[47]

Due to a serious illness, Moravec is infertile. However, he rejects speculation that his insistence on the idea of "robot children" is compensation for this. However, another motive, that of loneliness, could be decisive. Regarding his drive for robot research, Moravec, who describes himself as "socially handicapped", says: "I hope to get a good robot friend one day. But it's taking longer than I originally thought."[48]

4. The Concept of Singularity

4.1 Terminology

Kurzweil calls the point in time when the computer's computing power exceeds that of the human brain *singularity.* From that point on, according to Kurzweil, our future will change rapidly because computers will henceforth take over their own further development – with unforeseeable consequences for humans, our environment and the entire universe.

[43] Cf. Moravec, Curriculum Vitae.

[44] Cf. Schult, Crazy Hans.

[45] "During the last few centuries, physical science has convincingly answered so many questions about the nature of things, and so hugely increased our abilities, that many see it as the only legitimate claimant to the title of true knowledge. Other belief systems may have social utility for the groups that practice them, but ultimately they are just made-up stories. I myself am partial to such 'physical fundamentalism.'", Moravec, Robot, 191.

[46] Cf. ibid., 108ff, 125f.

[47] Cf. Schult, Crazy Hans.

[48] Moravec, quoted from: Schult, Crazy Hans, translation: MW.

Kurzweil did not invent the term singularity, but popularised it in relation to technological progress.[49] Originally, the term refers to a unique event with singular implications. It was first used by mathematicians to designate the starting point of an infinite growth on a graph of functions, such as when a finite number is divided by a number approaching zero (as in the function $y = 1 / (1-x)$ for the values $[x<1]$). Later it is adopted by astrophysicists: After a supernova, the remnant of a star decays to a point of apparently no expansion but infinite density, with a singularity at its centre – better known as a black hole.

Finally, in 1965, the British mathematician Irving John Good predicts an "intelligence explosion" as a result of intelligent machines designing their own successor generation without human interference. The American mathematician and computer scientist Vernor Vinge refers to this in 1983 and calls this event a *technological singularity*.[50]

Kurzweil dates the time of the singularity to the year 2045. According to his calculations, from this time onwards, it will be possible to acquire a computing power of 10^{26} cps (*computations per second*) for the equivalent of 1000 US dollars – according to Kurzweil, this corresponds to the combined computing power of the brains of all humans.[51]

While Moravec does not use the term singularity itself, he represents the same worldview underlying the concept of the singularity when he predicts human-like robots superior to the human mind by 2040,[52] which will self-reproduce and improve themselves from the middle of the 21st century.[53] Moravec sees this machine species as the better descendants of humans:

> *"Given fully intelligent robots, culture becomes completely independent of biology. Intelligent machines, which will grow from us, learn our skills, and initially share our goals and values, will be the children of our minds."*[54]

[49] Cf. Rojas, Analoge versus Digitale Seele, section "Die Singularität", para. 1.

[50] Cf. Kurzweil, The Singularity Is Near, 22f.

[51] Cf. ibid., 135f. On the attempt to represent the computing power of the human brain in a unit of meass urement such as *cps* or *MIPS (Million Instructions per Second)*, cf. below, 28.

[52] Cf. Moravec, Robot, 108-126.

[53] Cf. ibid., 143ff.

[54] Cf. ibid., 126. The fact that, according to his own prediction, humanity will soon be replaced by the roe bot species does not worry Moravec. Rather, he believes that "as also with biological children, we can arrange for a comfortable retirement before we fade away. Some biological children can be convinced to care for elderly parents. Similarly, 'tame' superintelligences could be created and induced to protect and support us, for a while", cf. ibid., 13.

4.2 The Belief in Exponential Progress

Underlying Kurzweil's assumptions is a belief in an exponential acceleration of technological progress. According to Kurzweil, long-term forecasts regularly dramatically underestimate the degree of future developments because they base their forecasts on an "intuitively linear" picture of past development. Even 1,000 years ago, technical development was already exponential. The progression of every exponential function, however, resembles the course of a linear function in its early phase, so the historical observer intuitively assumes a further linear progression: "The future is widely misunderstood. Our forebears expected it to be pretty much like their present, which had been pretty much like their past."[55]

According to Kurzweil, only in retrospect does it become apparent that technological progress happens at exponential rates. For meaningful forecasts of future technological development, therefore, one must use the assumption of exponential growth: "[W]e won't experience one hundred years of technological advance in the twenty-first century; we will witness on the order of twenty thousand years of progress (again, when measured by *today's* rate of progress), or about one thousand times greater than what was achieved in the twentieth century".[56]

Kurzweil sees technological evolution in continuity with biological evolution: like technological evolution, an acceleration of its development rate is already inherent in biological evolution. Based on the law named after Intel co-founder Gordon Moore, according to which the number of integrated circuits on a processor doubles on average every two years, i.e. the computing speed of processors increases exponentially,[57] Kurzweil looks at evolutionary and technological progress from prehistoric times to the present day. He finds that both biological evolution and human technology are subject to continuous acceleration, characterised by the ever shorter time between key events: Whereas it took two billion years for the first eukaryotic cells to evolve from the first life, only 14 years

[55] Kurzweil, The Singularity Is Near, 10f.

[56] Ibid., 11.

[57] For a detailed description of *Moore's Law* cf. ibid., 56f.

passed from the introduction of the personal computer to the invention of the World Wide Web.[58] Kurzweil believes this to be an overriding law, of which *Moore's Law* is only a special case. He calls this law the *law of accelerating returns*.[59]

Moravec also refers to Moore's Law and assumes a rapid increase in technological performance that even exceeds an exponential scale.[60] According to Moravec, technological developments in the past have been repeatedly disproved by reality and now exceed the wildest expectations of Jules Verne, Benjamin Franklin or Leonardo da Vinci. He is confident: "But bet that the real future will be even harder to reconcile with intuitions derived from the tiny piece of reality we've experienced thus far."[61]

Kurzweil describes how every technology reaches its limits at some point. So far, however, this has not stopped all technical progress: Whenever a technology could no longer be further developed, a new technology was already waiting in the wings, such as when computers began to be built with transistors in the 1960s instead of vacuum tubes, which had been the norm until then.[62] So, according to Kurzweil, technology will always find a way to evolve. He does not even shy away from speculating that it might be possible to accelerate the speed of light (!) in order to increase the computing power of processors,[63] or, for the same purpose, to couple processors with time machines that send executed calculations backwards through time and thus make them available immediately after the instruction is given.[64] Kurzweil emphasises that the occurrence of the Singularity is not, however, dependent on these speculations coming true.[65]

Likewise, Moravec ponders for the time when technology has reached the limits of matter. His – equally highly speculative – solutions speak of antimatter as an energy store and the belief in heavier atomic particles that could eventually replace electrons.[66]

[58] Cf. ibid., 17. Other key events that Kurzweil lists in a logarithmic chart include, in chronological order, the emergence of the first mammals, the development of the mammalian family *Hominoidea*, the first appearance of language, *Homo sapiens*, agriculture, the industrial revolution and the telephone. The thoroughly impressive chart seems to confirm Kurzweil's thesis, but the selection of key events gives the impression of a certain arbitrariness.

[59] Ibid., 7f.

[60] Cf. Moravec, Robot, 60f.

[61] Cf. ibid., 163. However, one could just as well use the same argument to defend the view that future technological development will be slower than Moravec anticipates. Moravec does not explain why the opposite will be the case.

[62] Cf. Kurzweil, The Singularity Is Near, 43-46.

[63] Cf. ibid., 139f.

[64] Cf. ibid., 140f.

[65] Cf. ibid., 139.

[66] Cf. Moravec, Robot, 159-161.

Both remarks are more reminiscent of science fiction than science and can be seen as an indication of a quasi-religious belief in technology.

4.3 Preconditions for Singularity

The belief in exponential progress is the basis for the belief in singularity. One of the technologies that Kurzweil says is currently improving exponentially is *human brain scanning*. According to Kurzweil, we already have impressive models of dozens of the several hundred brain regions in total – within the next two decades, he predicts, we will have a detailed understanding of all the brain regions. From this, he concludes, we will have developed functional software models for emulating human intelligence by the mid-2020s; the necessary powerful hardware will already be available to us by the end of the 2010s.[67] With these prerequisites, the first machine will pass the Turing test by the end of the 2020s.[68]

4.4 The Turing Test and Searle's Chinese Room Argument

The Turing Test was proposed in 1950 by the British logician and computer scientist Alan Turing to test a machine's ability to think. A machine, a human person and a human questioner take part in this test; the questioner sits spatially separated. The questioner sits in a separate room and chats alternately with the person and the machine. His task is to determine, based on the answers of the human and the machine, which of the two interlocutors is the human and which is the machine. A machine passes the Turing test if the questioner is not able to distinguish between human and machine.[69]

However, the significance of this test is based on basic philosophical premises that are not necessarily shared by every observer, since fundamental questions of the *mind-brain debate* are being affected.[70] It is true that a machine that is able to deceive a human in such a way that he does not recognise it as a machine can certainly be called "intelligent", insofar as one understands intelligence only as a rule-governed process. "The principal and continuing question that now arises, however, is whether the principles underlying these machines represent only the *necessary* or else the *sufficient* conditions

[67] Such a computer would have to be as powerful as our brain. To calculate the performance of the huuman brain, see below, 26.

[68] Cf. Kurzweil, The Singularity Is Near, 25.

[69] Cf. Oppy / Dowe, Turing Test.

[70] For an overview of the current state of the mind-brain debate, see below, 59.

of intelligence *in the first place.*"[71] Anyone who agrees with the theory of *naturalism*[72] will also see nothing else in the human brain than the special case of a machine. Under this condition, passing the Turing test would indeed fulfil the sufficient condition of intelligence.[73] But if the significance of the Turing test is so fundamentally based on basic philosophical presuppositions, it is extremely limited:

> *"The Turing test is informative in the sense that it draws attention to the fact that if there is any difference at all between artificial intelligence systems and humans, it must be that natural systems are* not completely *determined by their rule-governed and empirically ascertainable functionality."* [74]

The best-known criticism of the Turing Test is the *Chinese Room* argument by the American philosopher John R. Searle: A person who does not understand a word of Chinese sits in a room containing boxes of Chinese characters (the database) and a book with instructions on how to use these symbols (the programme). People outside the room pass questions into the room in the form of Chinese characters (input). For the person sitting in the room, these characters are meaningless, he does not understand them. However, he strictly follows the book with the instructions, which in turn is so well written that it enables him to give the correct answers to the questions asked (output). For the people outside the room, it thus looks as if the room or the person in the room understands Chinese, although this is not the case.[75] With his argument, Searle shows that computers – only because they follow sophisticated programmes – do not have any cognition. The Protestant systematic theologian Dirk Evers expresses this finding as follows:

> *"If machines made up of components that we would fundamentally deny the capacity for real thought and conscious intelligence are equal in functionality to any computer, and if, conversely, a thinking being can perform everything a computer does without developing a conscious understanding of the matter, then it is clear what distinguishes artificial intelligence from natural intelligence: Artificial intelligence follows rules, natural intelligence understands meanings."* [76]

[71] Evers, Der Mensch als Turing-Maschine? 103 (translation: MW, emphasis as in the original).

[72] *Naturalism* refers to the basic view that the behaviour of even large and complex systems can be traced back in all its diversity to the properties of their elementary components and their interactions with each other, cf. ibid.

[73] Cf. ibid., 104.

[74] Ibid., 105 (emphasis as in the original).

[75] Cf. Searle, Chinese room argument.

[76] Evers, Der Mensch als Turing-Maschine? 107 (translation: MW).

The fact that both Kurzweil and Moravec[77] instrumentalise the successful Turing test as future proof of the cognitive ability of computers thus primarily reveals their naturalistic view of the world.

4.5 Properties of Singularity

Once machines can fully simulate human intelligence by passing the Turing test, they will, according to Kurzweil, combine the strengths of human intelligence with those of artificial intelligence: In his view, the strength of human intelligence lies in recognising patterns, learning from experience and incorporating information gained from language into its knowledge.[78] The strength of artificial intelligence is to store billions of facts and recall them with pinpoint accuracy, to accurately apply skills once learned without fatigue at any time, and to share knowledge through data transmission at extremely high speed – compared to the very slow human exchange of knowledge through language. Through the internet, machines have access to and will apply all human knowledge at all times.[79]

Furthermore, computers are able to pool their resources, intelligence and memories: Any number of computers can network and thus become a supercomputer, while each of the networked computers can also act individually at any time. Kurzweil goes so far as to compare this networking of computers with human love: “Humans call this falling in love, but our biological ability to do this is fleeting and unreliable.”[80]

Equipped with human-like intelligence, machines would be able to change and improve their own hardware and software. Kurzweil sees no limits to this: By resorting to nanotechnology, their capabilities would far surpass those of biological brains.[81] Following the law of exponential progress, this repeated improvement of their own architecture would take place ever faster.[82]

The physical reality would also change along with the technical development. The machine-supported development of nanotechnology in the form of *nanobots* [83] would have an influence on the human organism, for example. Kurzweil mentions “respirocytes”,

[77] Cf. Moravec, Robot, 70-74.

[78] Cf. Kurzweil, The Singularity Is Near, 25f.

[79] Cf. ibid., 26.

[80] Ibid.

[81] Cf. ibid., 27.

[82] Cf. ibid., 28.

[83] Nanobots refer to tiny robots that are produced with the help of nanotechnology and are capable of their own locomotion, for example in our bloodstream.

artificial blood cells that would enable us to survive for a long time without taking in oxygen, as well as nanobots that could reverse human ageing at the cellular level. Likewise, nanobots that interact with neurons in the human brain and thus expand our perception and intelligence would be conceivable. In this way, human and artificial intelligence would merge, with the artificial part of intelligence growing exponentially over time, while the biological part would effectively remain the same.[84] With the help of such technical developments, Kurzweil is convinced, eternal life would in principle be possible, just as a house can in principle be inhabited forever if only it is regularly renovated and repaired. The only difference between the house and the human body, according to Kurzweil, is that we have not yet fully understood the body.[85]

Moravec also believes in a fusion of human and artificial intelligence. He calls people who "transcend[] their biological humanity" by means of artificial intelligence *Exes* (short for "ex-humans").[86] Moravec warns, however:

> *"[W]hithout restrictions, transformed humans of arbitrary power and little accounte ability might routinely trample the planet, deliberately or accidentally. A good compromise, it seems to me, is to allow anyone to perfect their biology within broad biological bounds."* [87]

According to Moravec, those *Exes* who cannot or do not want to live with the imposed restrictions would have to leave the earth and relocate their existence into space.[88]

Through the interaction of machine intelligence and human neurons by means of nanobots, Kurzweil predicts that our lives will gradually shift into a virtual reality that would appear deceptively real to us. In virtual reality, we would be able to be at any desired (virtual) place in a matter of seconds, and we would also be able to choose our physical appearance ourselves or have it chosen by our counterpart - our counterpart

[84] Cf. ibid. and 296: "The *advent* [emphasis MW] of strong AI is the most important transformation this century will see. [...] It will mean that a creation of biology has finally mastered its own intelligence and discovered means to overcome its limitations. Once the principles of operation of human intelligence are understood, expanding its abilities will be conducted by human scientists and engineers whose own biological intelligence will have been greatly ampflified through an intimate merger with nonbiological intelligence. Over time, the nonbiological portion will predominate."

[85] Cf. ibid., 212.

[86] Cf. Moravec, Robot, 142f. and 144f.

[87] Ibid., 143.

[88] Cf. ibid.

would not necessarily have to perceive us in exactly the same way as we do ourselves; rather, each person would be able to choose the physical appearance that he or she likes best for each counterpart.[89]

With the help of nanobot technology in the blood vessels of human brains, it will be possible to read the brain completely, according to Kurzweil. While so far we only have low-resolution scanning techniques such as *functional magnetic resonance imaging* (fMRI) and *positron emission tomography* (PET),[90] once methods are developed for them to cross the blood-brain barrier, wirelessly networked nanobots will be able to be used for detailed imaging and observation of all neurons.[91] The idea is that we only need to understand well enough how the brain works to be able to reprogram it in the form of software. Once this form of cerebral *reverse-engineering is* complete, it will be possible to download the *mind file* from the human brain and upload it back onto a suitably powerful computer. "This process would capture a person's entire personality, memory, skills, and history."[92] The human mind would then be pure software running on this computer; our physicality could be represented in the form of virtual bodies in virtual realities. Such a software copy of the brain would then be potentially immortal.[93]

According to Moravec, the subjective perception of time of a mind transferred to computer hardware could change dramatically: On very fast hardware, one second of real time could correspond to a year of subjective thinking time – at the same time, a thousand years, if "spent" in a passive storage medium, could pass like the blink of an eye.[94]

With the ability to improve and recreate itself, artificial intelligence would spread exponentially throughout the universe, according to the *law of accelerating returns* – first in our neighbourhood, later in the entire universe: "Ultimately, the entire universe will become saturated with our intelligence. This is the destiny of the universe".[95] Kurzweil

[89] Cf. Kurzweil, The Singularity Is Near, 29 and 314f.

[90] Cf. ibid., 157f.

[91] Cf. ibid., 163-167.

[92] Ibid., 198f. See also Schanze, Plug & Pray, minute 82f.: Kurzweil: "A hundred years from now we'll think it pretty incredible that we went through the day without backing up our mind file. I mean, you wouldn't go through the day without backing up the files on your personal computer." (Transcription: MW) – So the knowledge, the thoughts, the experiences in our brain are not substantially different from the files on our computers.

[93] Cf. Kurzweil, The Singularity Is Near, 323. The idea of a mind that exists detached from the body and can move completely freely through all (virtual) worlds carries a strong dualistic character. Cf. below, 76.

[94] Cf. Moravec, Robot, 170.

[95] Cf. Kurzweil, The Singularity Is Near, 29.

equates the universe itself with God: as soon as the entire universe will be filled with (artificial) intelligence, God will "awaken". According to this view, we ourselves create God.[96]

Moravec also shares this vision:

> *"The scene may resemble the free-for-all revealed in microscopic peeks at pond water. Instead of bacteria, protozoa, and rotifers, the players will be entities of potentially planetary size, whose constantly growing intelligence greatly exceeds a human's, and whose form changes frequently through conscious design."* [97]

With reference to the cosmologists John Barrow and Frank Tipler, he assumes that the expansion of man-made intelligence will result in a coherent mind inhabiting the entire accessible universe. Because of the mass of the universe, it is predicted that its present expansion will turn into contraction. The "cosmic mind" would draw so much energy from this contraction that it would be able to "contrive to do more computation and accumulate more memories in each remaining half of the time to the final singularity than it did in the one before, thus experiencing a neverending infinity of time and thought."[98] This final singularity, to them, is the goal of the universe, the true reason for our existence. Barrow, Tipler and Moravec, with reference to the Jesuit Pierre Teilhard de Chardin, therefore call it the *Omega Point.*[99]

5. What is the Computing Power of the Brain?

Both Kurzweil and Moravec believe they can state the computational power of the human brain more or less exactly. Moravec takes the human retina, whose neuron associations he says are the best understood of all regions of the central nervous system so far, as the starting point for his calculations. At the same time he states, "retina-like operations for robotic vision" already exist in robotics, which can then be used as a conversion factor. The retina can capture a million simultaneous edge and motion detections per image; Moravec assumes that it also processes about ten images per second. A robot

[96] Cf. ibid., 375 and 390: "We can consider God to be the universe. [...] The universe is not conscious – yet. But it will be."

[97] Moravec, Robot, 145.

[98] Ibid., 202.

[99] Cf. ibid. The usage of the term is an absurd distortion of its original idea: Teilhard de Chardin refers to the "omega point" as the maximum level of complexity and consciousness towards which the universe is evolving. In contrast to Tipler and Barrow, however, Teilhard de Chardin sees the omega point as transcendent and personal – because the universe is evolving towards this point, it must exist before the universe; the growing complexity of matter in the universe has also continuously led to an ever greater expression of personhood. He identifies the omega point with Christ, cf. Teilhard de Chardin, Der Mensch im Kosmos, 250-267. For more on the subject of substitute religion, cf. below, 34.

programme would need about 100 computational instructions per edge and motion detection, corresponding to 100 million instructions for a million such detections. At a processing speed of 10 images per second, a computing unit with the speed of 1000 *MIPS (Million Instructions per Second)*[100] would therefore be necessary to map the performance of the retina.[101]

Because the human brain, at 1500 cm^3, is about 100,000 times the size of the retina, according to Moravec, one can derive by simple multiplication a total performance of the human brain of 100 million MIPS. The chess computer Deep Blue, which defeated Garry Kasparov in 1997, had a computing power of 3 million MIPS. Since it seems plausible that Kasparov could use his brain power with an efficiency of 3 per cent for the highly unnatural problems of chess, the equivalence of Deep Blue and Kasparov speaks for his extrapolation of the retinal data, Moravec says.[102]

In his calculation, Kurzweil refers to Moravec and the mathematician Lloyd Watts, who carried out a similar extrapolation of total performance based on human hearing and arrived at similar results to Moravec. In view of the relatively early state of brain research, Kurzweil conservatively increases the figure determined by Moravec by a factor of 100 and henceforth calculates with a computing power of 10^{16} cps.[103]

Neither Kurzweil nor Moravec ask themselves whether the performance of the human brain can be adequately captured at all by stating its alleged *computing power.*[104] This way of calculating the brain reveals that both *strong AI* representatives see nothing else in the brain than a highly developed calculating machine. Kurzweil does name differences between the human brain and a computer: the brain is characterised by slower circuits that are connected in parallel in large numbers, as well as by the combination of digital and analogue phenomena in computation.[105] Ultimately, however, Kurzweil sticks to his view of the brain as an "imperfect" computer since it has been shaped by evolution over thousands of years.[106]

[100] For the comparability of the units *MIPS* and *cps,* see Kurzweil, The Singularity Is Near, 531, fn. 37.

[101] Cf. Moravec, Robot, 53f.

[102] Cf. ibid., 54.

[103] Cf. Kurzweil, The Singularity Is Near, 123f.

[104] "The brain can be described as a machine, to a certain extent even with some accuracy. But if one does not realise that this description of the brain is valid only for very specific purposes and for nothing else, one is actually living in a virtual world", Weizenbaum, Inseln der Vernunft im Cyberstrom, 99 (translation: MW).

[105] Cf. ibid., 150f.

[106] Cf. ibid., 151f.

Nonetheless, given the different cell types and the non-uniform structure of the brain, to determine its overall performance by merely extrapolating a sub-area seems highly unscientific. Moravec's calculations, to which Kurzweil refers, were published in 1999. In 2004, eleven leading neuroscientists wrote in a manifesto:

> *"There is still a large knowledge gap between the upper and lower organisational levels of the brain [...]. We still know frighteningly little about the middle level - i.e. what happens within smaller and larger cell assemblies, which ultimately underlies the processes at the top level. Even about which codes individual or a few nerve cells use to communicate with each other (they probably use several such codes at the same time), there are at best plausible assumptions. What is also completely unknown is what happens when several hundred million or even several billion nerve cells 'talk' to each other."* [107]

Apart from that, even computer scientists no longer consider the unit of measurement *MIPS* (or *cps in* Kurzweil's analogue) to be a suitable indicator for the speed of today's highly developed processors. A single digit cannot describe computing power, writes processor expert Ted MacNeil. Rather, he says, there are a variety of factors such as cache size, subdivision and number of processor cores or the mix of workloads, all of which have an impact on processor performance. Among computer scientists, the acronym MIPS is now translated as "Meaningless indicator of processor speed", among other things.[108]

But if even the performance of modern processors cannot be adequately described with the MIPS indicator, doubts are warranted as to whether the performance of the human brain can be adequately qualified with a number such as 100 million MIPS or 10^{16} cps.

6. Consciousness vs. Simulation

As shown at the beginning, the radical school of *strong AI* differs from the moderate school of *weak AI* in the question of whether artificial intelligence can actually produce consciousness or remains a simulation of consciousness.[109]

Kurzweil and Moravec both describe the technical requirements for simulating the brain. Kurzweil also goes into detail about *reverse engineering* the brain, with the help of which the "software" of the brain is to be reprogrammed, because: "[A]chieving the

[107] Monyer et al, Das Manifest, 31, 33 (translation: MW).

[108] Cf. MacNeil, Don't Be Misled By MIPS.

[109] Cf. Searle, Chinese room argument.

hardware computational capacity of a […] human brain […] will not automatically produce human levels of capability."[110] But how does the simulation, however powerful and true to reality it may be, become consciousness?

The question is indeed difficult to answer because, as Kurzweil himself admits, there is no objective test for the existence of consciousness.[111] Nor is there any philosophical or scientific consensus on how to answer the qualia problem.[112] The title of Kurzweil's monograph *How to Create a Mind,* published in 2013, promises to provide an answer to this question – unfortunately, the author does not deliver on this promise:

> *"My objective prediction is that machines in the future will appear to be conscious and that they will be convincing to biological people when they speak of their qualia. […] We will come to accept that they are conscious persons. My own leap of faith is this: Once machines do succeed in being convincing when they speak of their qualia and conscious experiences, they will indeed constitute conscious persons."* [113]

If machines were to give the impression of consciousness at some point, we would therefore also have to assume that they have consciousness, says Kurzweil.[114] Of course, the impression of consciousness would also require emotions: robots would have to be able to make us laugh and cry, and they would also have to be able to laugh at our jokes or get angry with us, for example, if we do not accept them as conscious persons.[115]

Rojas is sceptical about this. Due to the subjective nature of emotions, they cannot simply be programmed into a robot. It would therefore be futile to think about emotional robots if allegedly emotional robots were not repeatedly presented with great media impact. However, at least according to the current state of technology, real emotions will not emerge – quite probably never.[116] Of course, synthetic reactions to emotions in robots would be conceivable if emotions were decoded by humans and led to the activation of certain "muscles" in the artificial robot face. Such an "emotional" robot would be able to deceive us, but would feel as little emotion as a fly in flight.[117]

[110] Kurzweil, The Singularity Is Near, 145.

[111] Cf. ibid., 378.

[112] The qualia problem is the question of the relationship between subjective-phenomenal perception and mental states. For an overview of the mind-brain debate, see below, 59.

[113] Kurzweil, How to Create a Mind, 209f.

[114] How quickly a simulation creates the illusion of a conscious counterpart surprised the AI pioneer Joseph Weizenbaum as early as the mid-1960s, cf. below, 51.

[115] Cf. Kurzweil, How to Create a Mind, 209.

[116] Cf. Rojas, Die Angst des Roboters beim Elfmeter, section "Emotionen als Entstehungsprozess", para. 3f.

[117] Cf. ibid., section "Reaktive Emotionen und Spiegelneuronen", para. 4.

Moravec also considers the simulation of consciousness to be actual consciousness. He even goes so far as to claim an equivalence of simulation and reality: according to this, actions in a virtual reality and in reality would in principle be to be judged morally the same, their only difference would be that actions in reality, unlike in the virtual one, have consequences for us.[118] Moravec also speculates on the extent to which simulating human emotions would be advantageous for robots: "Not every emotion found in humans makes sense in robots."[119] Sexual behaviour, for example, is not useful, he says, because robots cannot reproduce sexually. A feeling of "agape" towards the robot's owner could be helpful, however, if the robot were thus able to assess the effects of its actions on the feelings of affected humans. "Nice" robots would also be easier to sell.[120]

In principle, according to Moravec, depending on one's view, one can attribute consciousness and intelligence to any object, be it a stone or a human being, if one understands, for example, the thermal movements of the atoms of a rock as the workings of a complex, conscious mind.[121] The British philosopher Colin McGinn comments on this in a review for the New York Times:

> *"Where Moravec is weak is in attempts at philosophical discussion of machine consciousness and the nature of mind. He writes bizarre, confused, incomprehensible things about consciousness as [...] mere 'interpretation' of brain activity. He also loses his grip on the distinction between virtual and real reality as his speculations spiral majestically into incoherence."* [122]

Kurzweil's and Moravec's theories both assume that one only has to understand and simulate the brain well enough to create an artificial consciousness that is not just simulation but actually perceived consciousness.[123] But apart from the immense difficulty of simulating a brain made up of 100 billion neurons and many other cells, an average of about 1000 connections per neuron, and the "irritating habit of periodically questioning all the theories [about its functioning, MW] that are already finished and long since

[118] Cf. Moravec, Robot, 196-199.

[119] Cf. ibid., 118.

[120] Cf. ibid. 118f. In his novel The *Hitch Hiker's Guide to the Galaxy, the* British writer Douglas Adams describes the robot Marvin, endowed with human feelings, who becomes manic-depressive due to his limitless computing power and simultaneous chronic underchallenge by his biological owners, cf. Adams, The Hitch Hiker's Guide to the Galaxy, 72-75.

[121] Cf. ibid., 199.

[122] McGinn, Hello, HAL, 11.

[123] Cf. Kurzweil, The Singularity Is Near, 375: "If we emulate in as detailed a manner as necessary everything going on in the human brain and body and instantiate these processes in another substrate [...], why wouldn't it be conscious?"

concluded"[124] , every simulation is also always steered – consciously or unconsciously – in a certain direction by its programmer. Whether a certain behaviour of the simulated brain reflects reality or not can therefore not be determined at all.[125]

Kurzweil and Moravec's view that a sufficiently detailed simulation can achieve consciousness ignores the gulf between subjective perception and objective consideration of consciousness. In his essay *What Is It Like to Be a Bat?* [126], the American philosopher Thomas Nagel suggests that even if we had a detailed understanding of the physical workings of a bat's brain, we still cannot imagine what it feels like to be a bat.[127] While an objective view will help us understand physical and biological processes, it takes us further away from a subjective understanding.[128] In this respect, the question of how the objective-physical explanation of all brain processes and their simulation should lead to actual consciousness remains unanswered by Kurzweil and Moravec.

The Protestant theologian Eilert Herms emphasises that there is always an ontological difference between natural and artificial intelligence. Regardless of how well artificial intelligence simulates its natural archetype, it always remains an image of this archetype. The image may be more or less similar to the original image, it may even be made to resemble it to the point of confusion, but it still does not cease to be the image. "The forgery may be perfect. But what it achieves is only the perfect forgery."[129]

[124] Rojas, IBM vs. Blue Brain, section "Die Unterschätzung des Gehirns", para. 4.

[125] Cf. ibid., section "Too big to fail", para. 4.

[126] Cf. Nagel, What Is It Like to Be a Bat?

[127] Cf. also Spaemann, Schritte über uns hinaus, 134f. (emphasis as in the original): "No question of time is the answer to the question of what it is like to be a bat. We don't know that and never will, unless we ourselves have the soul of a bat, i.e. are bats. But then we would no longer be us, but bats, and would no longer know what it is like to be a human being. And we wouldn't know what it's like to be a bat either, because being a bat is kind of like being a bat, but part of being a bat, in all likelihood, is not reflecting on what it's like to be what you are. Yet we have reason to believe that it is somehow to be a living being, while it is equally probably *not* somehow to be an atom." (Translation: MW).

[128] Cf. Nagel, What Is It Like to Be a Bat, 444f. Weizenbaum argues similarly: "Complete knowledge of the physical, genetic, neurological structures of a living being is not enough to understand the living being. Whoever, for example, has all this knowledge about an ant, but does not know that the ant lives in a huge society of ants, does not understand the ant. The same applies to understanding man. In principle, it is impossible to understand humans in a purely scientific way. That is why the quest to make robots in human form is absurd. It can only arise from megalomania or uterine envy" (translation: MW), cf. Weizenbaum, Wir gegen die Gier.

[129] Herms, Künstliche Intelligenz, 291, translation: MW.

6.1 Benjamin Libet, Intentionality and Free Will

A sub-issue of the question of artificial consciousness is the question of the intentionality and free will of machines. While the Christian view of man does not work without free will,[130] the existence of free will is disputed in natural science and philosophy. Cognitivists and materialists usually reject the concept of free will when they generally regard consciousness phenomena as epiphenomena of the chemical processes in the brain.[131]

Consequently, the two terms "intentionality" and "free will" are not mentioned by Moravec. Kurzweil, on the other hand, writes about free will, but refers to the Libet experiment, which he regards as an indication that free will could be an illusion.[132]

In 1979, the physiologist Benjamin Libet conducted an experiment on the temporal sequence of action decisions. In this experiment, he asked subjects to perform simple motor actions such as finger movements at a freely selectable time. In doing so, they were asked to indicate at which point in time they consciously planned the action by looking at a clock. In fact, Libet was able to detect a readiness potential in the subjects' brains by means of an electroencephalogram between 350 and 850 milliseconds before the time at which the subjects stated they planned the action – i.e. even before the subjects were aware of their intention to act. From this, Libet concluded that free will is nothing more than an imagination of the brain.[133]

The result of the Libet experiment was the subject of much debate in brain research. Today, however, a large number of neuroscientists deny that it can be used to deduce the absence of free will. The Düsseldorf brain researcher Karl Zilles, for example, sees methodological flaws in Libet's experimental design: First of all, the test person was instructed to perform a *predefined* action – only the time of the action was freely chosen. Moreover, this action was ethically and emotionally irrelevant, but current research results suggest that emotionally relevant actions activate different brain mechanisms than emotionally irrelevant ones. The biggest flaw in Libet's experiment is conceptual: Libet equates the moment of voluntary decision with the moment of consciousness of the decision. However, the volitional decision and the awareness of a volitional decision

[130] Cf. Vat. II, GS, *No. 17.*

[131] Cf. Freeman, Intentionality, and below, 75.

[132] Cf. Kurzweil, The Singularity Is Near, 191 and Idem, How to Create a Mind, 229f.

[133] Cf. ibid.

are two different processes that require different neural mechanisms.[134] In this respect, at least on the basis of the results of the Libet experiment, one cannot seriously speak of an illusion of free will.

7. Digital Philosophy

For Kurzweil, information plays a fundamental role in the universe: from the Big Bang to humans, information has been passed on in increasingly complex forms, starting with atomic structures, later in DNA, then in brains and finally in the form of human technology.[135] To underline the importance of information for the universe, he cites the example of *digital philosophy.*[136]

"Digital philosophy" is a direction of philosophy widespread among computer scientists, mathematicians and physicists, which goes back to an idea of the computer science pioneer Konrad Zuse. Its underlying assumption that the universe is a digital computer was first formulated by Konrad Zuse in 1967.[137] In the 1980s, the American computer scientist Edward Fredkin took up this idea. According to his conviction, the entire universe is a single cellular automaton.[138]

Fredkin believes that information, along with matter and energy, makes up the universe. Information is the fundamental primal principle, of which matter and energy are only manifestations. Atoms, electrons and quarks would therefore ultimately consist of binary information units, *bits.* The behaviour of these bits – and thus of the entire universe – is determined by a simple programming rule. Fredkin calls this rule the "reason and first mover of all things":[139]

> *"I don't believe that there are objects like electrons and photons and things which are themselves and nothing else. [...] What I believe is that there's an information process, and the bits, when they're in certain configurations, behave like the thing we call the electron, or the hydrogen atom, or whatever."* [140]

[134] Cf. Zilles, Hirnforschung widerlegt nicht Freiheit.

[135] Cf. Kurzweil, The Singularity Is Near, 14-21.

[136] Cf. ibid., 85-94.

[137] Cf. Zuse, Rechnender Raum, 343.

[138] A cellular automaton is a computer based on simple mechanisms that, for example, changes the colour of individual cells according to previously defined rules depending on the status of their neighbouring cells. The results are highly complex, despite the simplicity of the rules. Cf. Kurzweil, The Singularity Is Near, 105f.

[139] Cf. Wright, Did The Universe Just Happen? 30.

[140] Fredkin, quoted from: Wright, Did The Universe Just Happen? 34.

Fredkin's universe is therefore completely deterministic. The principle of a creator is replaced by a computer, the principle of love, towards which creation is directed, by a programming rule.

Kurzweil does not go as far as Fredkin in claiming a cellular automaton as the primordial principle of the universe, but believes that further conceptions are needed to explain the progressive spread of information. Nevertheless, he considers digital philosophy to be a significant contribution to understanding the importance of information for the universe.[141]

8. Criticism

8.1 Singularity as a substitute religion?

If one considers man as *homo religiosus* and assumes the existence of religion in all peoples of all times,[142] it is not surprising that substitute religions are emerging to the same extent that traditional religions are losing importance. As mentioned at the beginning, Kurzweil's and Moravec's predictions for the future have strong religious overtones. This section argues that the concept of the Singularity and the belief in powerful artificial intelligence can indeed be seen as substitute religions.

Even if the attempt to clearly define a generally accepted concept of "religion" is difficult to impossible in view of the plurality of religious views,[143] there are elements that apply to religion and religiosity in general and that can be found in Kurzweil and Moravec. Religion conveys a sense of the *infinite,* of *transcendence,* in the broadest sense of *God.* It provides *orientation in* the world of life by offering explanatory models for the interrelationships of the world of life and is able to give people a function in this world.[144] Another aspect of religion is that of *distinction of the sacred from the profane*[145] or an idea of *redemption from suffering and death.*[146] Some forms of religion develop a *fundamentalist* character – and fundamentalist elements can also be identified in Kurzweil's and Moravec's works.

[141] Cf. Kurzweil, The Singularity Is Near, 94.

[142] Cf. Feil, Religion I, 264.

[143] Cf. ibid., 265.

[144] Cf. ibid., 264 and Zirker, Religion, 1035.

[145] Cf. Bürkle, Religion, 1040.

[146] Cf. Kornwachs, Prothese, Diener, Ebenbild, 406.

Taken by themselves, all these elements do not yet form a religion – for this, they would have to grow together into a conceptually and philosophically more cohesive unit. But as the cultural scientist Hartmut Böhme notes, it is a characteristic of the post-Enlightenment age to break individual religious motifs out of their theological and institutional bond: "[S]uch motifs do not form discourses, but the quivering base of seemingly religion-free techniques. This is the form of religion after the death of God."[147] Against this background, to speak of singularity and strong AI as a substitute religion thus seems quite appropriate.

8.1.1 Faith in technology as a sacred element

Kurzweil's and Moravec's faith in progress and technology is indeed sacred. Kurzweil's *Law of accelerating returns* and *Moore's Law*[148] can be regarded as fundamental beliefs of a religion that has made the perpetual, exponential increase in performance its foundation. Kurzweil and Moravec venerate this exponential acceleration because it is supposedly capable of solving all of humanity's problems – faster than the *profane* thinkers who persist in their world view of only linear performance growth can imagine.

This belief in progress sometimes takes on radical features, for example when there is talk of overcoming currently existing physical limits such as the speed of light or when the concept of time machines is used to further accelerate technical processes.[149]

8.1.2 God, transcendence and eternal life

Kurzweil's declared goal is the transcendence (cf. book title "*When Humans* Transcend *Biology*") of human biology through technology. This transcendence is achieved through human beings and their technical efforts alone and does not require God – humans thus put themselves in the place of God through their self-transcendence, and this transcendence takes place within the boundaries of our universe. Technology would gradually cure our diseases with the help of "nanobot doctors" and prevent and reverse the ageing of our cells.[150] Later, it would lead to a detailed understanding of the brain and thus our

[147] Böhme, The Technical Form of God.

[148] Cf. above, 19.

[149] Cf. above, 20.

[150] Cf. above, 23.

mind, which we could reprogram with the help of software. Scanning and backing up our own nervous systems would overcome death – as disembodied spirits on machines, we could effectively live on indefinitely.[151]

According to Kurzweil, we ourselves create God with the help of technological progress, namely when intelligent technology has spread throughout the universe and the whole cosmos has merged into one giant artificial brain. Because Kurzweil identifies the universe with God, God becomes conscious to the extent that the universe is filled with consciousness.[152] In a sense, this represents a reversal of the classical idea of creation.

Moravec makes a comparable reversal when he explicitly refers to Teilhard de Chardin and calls the status of an artificial intelligence encompassing the entire universe *point omega,* but reverses its original meaning.[153] While Teilhard de Chardin identifies the omega point with Christ and thus thinks of it as explicitly personal, Moravec takes an apersonal view because, as a "physical fundamentalist", he does not believe in any personal original principle.[154] For Teilhard de Chardin, the omega point must also be transcendent, since it must lie before creation as the primordial principle towards which creation is oriented. Such a form of transcendence does not exist for Moravec; his "transcendence" remains limited to the boundaries of the universe, as for Kurzweil. Infinity is subjective for Moravec: while the computational speed of the *cosmic mind* increases exponentially, the time simulated by and on it is perceived as an infinitely extended one.

For the followers of *digital philosophy,* there is nevertheless a form of transcendence outside the empirically perceptible universe: the primal principle of a computer that precedes the universe and whose calculations constitute the empirically perceptible appearance of the universe.[155] Here, the principle of a personal creator God is replaced by

[151] Cf. above, 25 as well as Rojas, Analoge vs. digitale Seele, section "Die Singularität", para. 4: "Legend has it that prancing Roman generals had 'Memento mori' whispered in their ears so that they would not forget their own transience. For Singularians, 'Memento mori' sounds more like an appeal: before it gets that far, something must be done." (Translation: MW).

[152] Cf. above, 25f and Kurzweil, The Singularity Is Near, 375.

[153] Cf. above, 26.

[154] In his description of the omega point, Teilhard de Chardin explicitly deplores modern man's desire to "depersonalise what he admires most". He sees this desire as rooted in the instrument of analysis used by scientific research, which breaks reality down into smaller and smaller parts. "A single reality seems to remain [...]: the energy – the new spirit. The energy – the new God. The impersonal for the omega of the world as for its alpha", Teilhard de Chardin, Der Mensch im Kosmos, 251, translation: MW.

[155] Cf. above, 33.

an apersonal programming rule that calculates the entire universe deterministically. However, since a "transcendent computer" by definition cannot be empirically proven, this basic assumption of digital philosophy remains a statement of faith.[156]

8.1.3 Fundamentalism

Since Kurzweil believes that technological progress is capable of solving all human problems, he sees even small delays in technological development as a great danger that could condemn millions of people to further suffering or even death. He thus fundamentally rejects cultural and ethical concerns about technological progress:

> *"[T]he reflexive, thoughtless antitechnology sentiments increasingly being voiced in the world today do have the potential to exacerbate a lot of suffering."* [157]

The promises of the singularity must therefore be fulfilled as soon as possible, technological progress must be achieved as quickly as possible – criticism of this is inadmissible for Kurzweil because it ultimately leads to human suffering.

Moravec calls himself a fundamentalist, or more precisely, a "physical fundamentalist".[158] He sees physics as the "only legitimate claimant to the title of true knowledge" and denies all other belief systems their claim to truth. These are merely "made-up stories" that may still have a social benefit for their respective adherents.[159] According to Moravec, anyone who is rational relies on natural science, and only on natural science. However, Moravec does not give his readers an explanation as to why this obvious physical fundamentalism should be more rational than belief in a religion.[160]

An example of the effects of this technical fundamentalism is given by the Protestant theologian and computer scientist Anne Foerst. She describes the reactions of some renowned AI scientists to her proposal in 1996 to offer a seminar on "God and Computers" at the Massachusetts Institute of Technology. Marvin Minsky, who coined the phrase "The brain is merely a meat machine",[161] was strongly opposed to this seminar, which

[156] "He [Fredkin] cannot give you a single line of reasoning that leads inexorably, or even very plausibly, to this conclusion", Wright, Did The Universe Just Happen? 40.

[157] Kurzweil, The Singularity Is Near, 373f.

[158] Moravec, Robot, 191.

[159] Cf. ibid., 191.

[160] Cf. also Böhme, The Technical Form of God: "They [cyberprophets like Moravec] are religious fundamentalists who long to dissolve the interconnectedness of human history and biological-evolutionary conditions. They are wild transcendental yearnings. The scrap pile of earth and the maggot bag of the human body are the sacrifice that can be made to the exit from bio-evolution all the more easily because earth and body have the stigma of sanctity attached to them." (Translation: MW).

[161] Cf. Weizenbaum, Inseln der Vernunft im Cyberstrom, 98.

he described as an “evangelical enterprise”. One student understood the class as “indoctrination”, a PhD student accused it of “suffering from the set of collective pathologies known as religious faith” (translation: MW).[162] At MIT, the bastion of objectivity and rationality, “psychologically misguided” people like theologians should not be holding seminars. Foerst calls it ironic “that so many highly intelligent people can be so religious in their rejection of religiosity.”[163]

8.1.4 When Science Transcends Religion

Analogous to the attempt to overcome the limitations of the human body with the help of technological developments and thus to “transcend” human biology, scientifically credulous researchers like Kurzweil and Moravec are engaged in replacing religious beliefs with supposedly scientific ones – presumably unconscious about the extent to which they themselves are practising religion. Physicist and philosopher Klaus Kornwachs accuses these AI researchers of a metaphysical deficit when they see a redemption of the human race in their optimisation thinking. “They play with a surrogate of salvation history, usually without even knowing the theological background.”[164]

Joseph Weizenbaum describes the substitute religion of natural science very vividly:

> *“I really believe that natural science [...] today has all the characteristics of an organised religion. There are novices, these are the students at universities. There are priests, which are the young professors, then there are the monsignori, which are the older ones. There are bishops and cardinals. There are churches and there are cathedrals. My own university, the Massachusetts Institute of Technology, is a cathedral within science. There are even popes and – this is very important – there are heretics! The heretics of natural science are punished just like the heretics of an ancient religion: they are expelled.”* [165]

8.1.5 Golems, angels and separate intelligences

Foerst, a theologian and computer scientist, draws a comparison between the attempt to build human-like robots and Jewish golem tales. According to legend, in the 16th century the Prague rabbi Judah ben Loew formed a golem out of clay[166] to protect the population of the Jewish ghetto from attacks from the non-Jewish population. After he placed

[162] Cf. Foerst, Von Robotern, Mensch und Gott, 54-56.

[163] Ibid., translation: MW.

[164] Kornwachs, Prothese, Diener, Ebenbild, 406, translation: MW.

[165] Weizenbaum, Inseln der Vernunft im Cyberstrom, 166f, translation: MW.

[166] The Hebrew root גלמ appears only twice in the Old Testament and means something like “formless thing” or “embryo”, cf. Foerst, Von Robotern, Mensch und Gott, 45.

a paper with the name of God in its mouth, the golem came to life and helped the Jews of Prague in their daily work and in the event of attacks from outside. The awakening of the golem after being named with the Tetragram means that God ultimately remains the life-giving force; without him, man cannot create anything alive.[167] According to Foerst, most golem stories show that man, as the image of God, shares in God's creative ability: "Whenever we are creative, we celebrate God and God's creativity within us."[168] Accordingly, she says, the construction of robots should be understood as an act of prayer, analogous to that of golems.[169] That many early AI researchers came from Jewish families who saw themselves as descendants of Rabbi Loew is an interesting detail that may help explain their motivation for AI research.[170]

However, some of the golem stories also deal with the danger of human hubris, such as the following: A golem has the words יהוה אלהימ אמת (God, the Lord, is truth) written on his forehead. As soon as he wakes up, he erases the second א from his forehead so that only יהוה אלהימ מת (God, the Lord, is dead) can be read. Then he explains to his builders that God is worshipped because he created us humans. If human beings themselves became creators, they would take the place of God – henceforth they, and no longer God, would be worshipped. But a God who is not worshipped is dead.[171]

Golem XIV is – surely not by chance – the name of a novel by Polish science fiction author Stanislaw Lem from 1984. In it, GOLEM[172] is a supercomputer that calls itself an "angel" and proclaims to man that it is no longer "first among the animals or over them".[173] Artificial intelligences as angelic beings? This is the comparison drawn by philosopher Rafael Capurro. He believes that the idea of artificially producible intelligences that are superior to us occupies the same position in our technological civilisation that angels and demons occupied in mythology and religion.[174]

According to Thomas Aquinas, angels are *separate intelligences* without *materia*. They are not captured by space, but can be in one place in one moment and in another in the next, without time intervening. Although an analogy is not directly possible with

[167] Cf. ibid., 45f.

[168] Ibid., 47.

[169] Cf. ibid.

[170] Cf. ibid., 50.

[171] Cf. ibid., 49.

[172] Abbreviation for "General Operator, Longrange, Ethically Stabilized, Multimodelling".

[173] Cf. Lem, Golem XIV, 137.

[174] Cf. Capurro, Leben im Informationszeitalter, 79.

regard to the hardware of artificial intelligences, it is possible with regard to the functional properties of the software. Capurro already sees in today's multimedia world networking a "decisive change in our being in space and time".[175]

Whereas, according to Thomas, man can only communicate through the barrier of sensual signs, angels can reveal themselves immediately without any barrier. For them, communication does not take place externally but internally. Whereas humans require a discourse in order to obtain information about the truth, the principles are sufficient for angels to draw immediate conclusions. Computers also exceed certain formal cognitive abilities of humans, especially in the quantity and accuracy of the stored data as well as the speed of their processing. Here, according to Capurro, "it becomes plausible why the idea of higher artificial intelligences occupies in our technical civilisation that place of a superhuman signifier which in other cultures was occupied by theological myths."[176]

According to Capurro, the effort for artificial intelligence is becoming a myth of the "technical symbolisation of the separate intelligences thought to be divine". It is absurd that the fundamentally different causal principles, God for the angels and man for artificial intelligence, and the resulting limits of artificial intelligence are no longer perceived.[177]

8.1.6 The God Machine

The Catholic theologian and philosopher Hans-Dieter Mutschler sees a fundamental change in the understanding of technology since the industrial revolution. While technology before the industrial revolution was always handicraft technology and thus comparatively inefficient and prone to failure, modern, highly efficient technology is pushing the boundaries of nature further and further.[178]

Mutschler sees this as the reason for the contrast between technology and religion that is often perceived today.[179] Pre-industrial technology was strongly dependent on nature for its functionality and thus corresponded to the religious mode of receiving. Modern

[175] Cf. ibid., 86-88, translation: MW.

[176] Cf. ibid., 89f, translation: MW.

[177] Cf. ibid., 92f, translation: MW.

[178] Cf. Mutschler, Die Gottmaschine, 109-116.

[179] Cf. ibid., 36f.

technology, on the other hand, emancipates itself from nature; a basic religious act is no longer necessary. Therefore, modern natural science became more and more an authority of appeal for atheism.[180]

To the same extent that technology and natural science displace religion, however, they themselves become a substitute religion: in the 19th century, railway stations were built in the image of ancient temples or churches.[181] At the same time, electricity companies advertised the new form of energy with posters depicting the god Helios enthroned on a generator.[182] Carl Benz's motivation for inventing the automobile was the "liberation of man" – an almost religious motif.[183] The inventor of rocket technology, Hermann von Oberth, wrote fantastic literature on the side in which he portrayed himself as a founder of religion.[184]

According to Mutschler, a very similar divinisation of technology is taking place today in the field of computer technology. He draws a direct comparison between the deification of steam and electricity, which seems bizarre to us today, and the theses of AI researchers like Minsky and Moravec. Although there is now thorough scientific and philosophical literature that doubts that computers will ever be able to simulate all human performance, reductionists like Minsky and Moravec cling to their theses quasi-religiously and mostly keep to themselves at their congresses without seeking dialogue with philosophy. Today, it is cyberspace that evokes religious categories: the creators of artificial worlds put themselves in the place of God and are themselves masters of infinity, being and non-being.[185]

According to Mutschler, the supposed opposition between religion and technology does not exist, because new technologies have been accompanied by a form of crypto-religiosity since the industrial revolution. The longing to transcend all boundaries is a human characteristic that, when it is no longer expressed religiously, is expressed in other ways.[186]

[180] Cf. ibid., 222f.

[181] Cf. ibid., 38.

[182] Cf. ibid., 56.

[183] Cf. ibid., 23.

[184] Cf. ibid., 40.

[185] Cf. ibid., 81-103 and 244.

[186] Cf. ibid., 244f.

Mutschler predicts that the phase of divinising computer technology will also come to an end at some point. He pleads for a new, far more sober attitude towards technology, which should be a means to finite ends and not convey religious content.[187]

8.2 Hubris

After Copernicus, Darwin and Freud, progress in artificial intelligence could well lead to a fourth narcissistic affront to the human self-image: namely, when more and more tasks, for the fulfilment of which genuinely human abilities were previously necessary, can be carried out by machines.[188] However, the idea of leading AI protagonists to be able to completely understand and artificially reproduce the human brain in a few decades is also accompanied by a certain disdain for humans and biology. Kornwachs suspects a "desire to 'offend' and provoke the human self-image", which is certainly also connected to the competitive struggle of modern research programmes.[189] Rojas also believes that many research programmes now have so much funding that they are "too big to fail".[190] To a certain extent, this fact explains the media-effective sensationalism practised by leading representatives of strong artificial intelligence.

8.2.1 Human Brain Project

A good example of such a mechanism is found in the *Human Brain Project*, which was selected by the EU Commission in 2013 from among six projects worthy of funding and will be funded with one billion euros over a period of ten years.[191] The Human Brain Project, led by the Israeli brain researcher Henry Markram, has set itself the goal of mapping a complete simulation of the human brain on a supercomputer that is as true to the original as possible by 2023, on the basis of which the functioning of the human brain is to be deciphered.[192]

Markram had already been working on a similar project in Lausanne for the previous eight years, apparently with little success. "Markram claims to have simulated a so-called cortical column, the smallest unit in the architecture of the cerebral cortex. However, he has not published his results in a comprehensible way",[193] writes science

[187] Cf. ibid., 246f.

[188] Cf. Kornwachs, Prothese, Diener, Ebenbild, 402.

[189] Cf. ibid., 406, translation: MW.

[190] Cf. Rojas, IBM vs. Blue Brain, section "Too Big to Fail".

[191] Cf. news article "EU wählt zwei Projekte der Spitzenforschung aus".

[192] Cf. Markram et al, Introducing the Human Brain Project, 39.

[193] Grolle, Aufruf zur Verschwendung, tranlsation: MW.

journalist Johann Grolle. So far, only one nervous system has been completely mapped, namely that of the eelworm with 302 nerve cells – and yet it is not possible to calculate the behaviour of this tiny animal from it. He therefore considers it utopian to simulate 300 million times more brain cells whose circuit diagram is not even known.[194]

The problem is homemade: while in other research projects the idea comes first and then the search for money, the EU Commission's call for proposals was the other way round: "The call to submit a one-billion-euro idea is equivalent to an open call for wastefulness," says Grolle.[195] .

8.2.2 In the Place of God

That Kurzweil and Moravec put man in the place of God has been shown before.[196] In a physical-fundamentalist worldview without a creator, man is dependent on redeeming himself. Both Kurzweil and Moravec identify the result of this redemption with the creation of God: Kurzweil by identifying the universe with God and describing a universe filled with artificial intelligence as the awakening of God, and Moravec by calling this all-encompassing cosmic spirit the omega point.[197]

The fact that in both cases, the redemption of the human race is being propagated without any knowledge of the theological background[198] can certainly be described as hubris. The idea of being able to fully understand human biology in just a few decades is also based on excessive faith in progress and overconfidence. Behind this view lies the belief that the human mind can be completely described by numbers and thus ultimately reproduce neurological processes on non-biological systems, separate from the biological-material substrate of the human being.[199] This reductionist notion of the human brain is due in large part with the fact that over the past few decades we have become so accustomed to the way computers work that we now view numerous natural processes

[194] Cf. ibid. and Fisch, Der Griff nach dem Bewusstsein: "Like many projects with high goals or visions, Markram's project sometimes triggers great scepticism among colleagues. Many, however, do not want to openly comment on it. The reason given is that they are not up to date with the latest knowledge on the project or do not feel competent enough in the field. Rodney Douglas, Kevan Martin and Richard Hahnloser from the Institute of Neuroinformatics at the ETH and the University of Zurich have nevertheless taken the risk. In a letter to the editor of the 'Tages-Anzeiger' they complained, among other things, about the waste of public money. Hahnloser told the NZZ: 'It is outrageous to spend hundreds of millions on projects that shoot into the blue.'" (Translation: MW).

[195] Grolle, Aufruf zur Verschwendung, translation: MW.

[196] Cf. below, 36.

[197] Cf. above, 25f.

[198] Cf. Kornwachs, Prothese, Diener, Ebenbild, 406.

[199] Cf. above, 26f.

in analogy with computers.[200] In the process, we increasingly lose sight of the fact that thinking is something different from calculating. The brain is not a digital computer, a separation between software and hardware does not exist biologically. "We are what we are because our cells do not calculate, but interact chemically and physically."[201]

8.2.3 False Prophecies

In his book *The Age of Spiritual Machines*, published in 1999, Kurzweil makes predictions for the period up to 2099, which he divides into four sections. The first section is particularly interesting from today's perspective, as it attempts to predict the technological development for the year 2009.[202] Today, in 2014, it can be stated that while some of Kurzweil's predictions from 1999 have actually come true, the world as a whole looks nowhere near the way Kurzweil imagined it 15 years ago.

Kurzweil was right in much of his predictions for mobile computing, which has become increasingly commonplace in the form of smartphones and tablets since 2007. Likewise, thanks in part to smartphone technology, there are now well-functioning portable readers for the blind.[203] However, the fact that we can now access the internet from almost anywhere at broadband speeds is a development that was already foreseeable in 1999 – even back then there was mobile data access, albeit considerably slower and more expensive.

Other predictions, however, have not come true at all or are at such an early stage that they are (still) unusable. No healthy person today wears "at least a dozen computers on and around their bodies".[204] The technology of so-called *smartwatches*, wristwatches connected to the smartphone, is just in its infancy in 2014 – whether it will become generally accepted is not yet foreseeable. "Computer displays built into eyeglasses"[205] also exist only as a prototype with *Google Glass* and are meeting with strong criticism due to their high price, low battery power and limited usefulness in everyday life.[206]

[200] How far this idea can lead is shown by the example of digital philosophy, cf. above, 33.

[201] Rojas, Analoge vs. digitale Seele, section "Das Gehirn arbeitet analog", para. 1, 4f.

[202] Cf. Kurzweil, The Age of Spiritual Machines, 189-201.

[203] Cf. ibid., 192 and 201.

[204] Cf. ibid., 189.

[205] Ibid, 190.

[206] Cf. Janssen, Warum Glass (noch) nicht funktioniert.

It is also not true that we create the majority of our texts with the help of dictation software. Although speech recognition systems in smartphones are now relatively usable, they still make many mistakes, especially with multilingual texts and punctuation, which the software has not yet mastered itself. In addition, written texts often have a completely different linguistic style than the spoken word. Even in 2014, it is not yet foreseeable that dictation will become established as a means of text production in the long term.[207]

“Autonomous nanoengineered machines”, as Kurzweil predicts for 2009, still do not exist in 2014, not even as prototypes.[208] Pupils do not yet learn to write and read with the help of interactive software – which is perhaps also due to the fact that people learn faster and better when they have a human counterpart.[209] Kurzweil’s prediction of a continuous economic expansion due to technological progress between 1999 and 2009 may be given a big question mark after the bursting of the *dotcom bubble in* March 2000 and the global financial crisis from 2007.[210] The same applies to privacy, which Kurzweil declared a priority policy issue in 2009.[211] However, the question of privacy became an issue for the general public at the earliest after the NSA affair came to light from June 2013.

A technology that satisfactorily translates telephone conversations in real time[212] does not exist even in 2014. In May 2014, Microsoft showcased a real-time translation function for its Skype telephony software. Immediately before the demonstration, Microsoft CEO Satya Nadella praised the translation capabilities of his software:

> *“The one fascinating, fascinating feature of this is something called transferred learning. What happens is, say, you teach it English – it learns English. Then you teach it Mandarin – it learns Mandarin, but it becomes better at English. And then you teach it Spanish – it gets good at Spanish, but it gets great at both Mandarin and English. […] It’s brain-like in the sense of its capability.”* [213]

[207] Cf. Kurzweil, The Age of Spiritual Machines, 190.

[208] Cf. ibid., 191.

[209] Cf. ibid., 191f. This phenomenon could well be explained neurologically: mirror neurons fire in the same way whether one performs an action oneself or observes it. Emotion and learning therefore seem to be neuronal processes that are significantly influenced by mirror neurons, cf. Rojas, Die Angst des Roboters beim Elfmeter, section “Reaktive Emotionen und Spiegelneuronen”.

[210] Cf. Kurzweil, The Age of Spiritual Machines, 194.

[211] Cf. ibid., 195f.

[212] Cf. ibid., 193.

[213] Nadella / Pall, Presentation on 27 May 2014 (transcription: MW).

The subsequent presentation with the language pair German-English was rather disappointing after this full-bodied announcement. Although the spoken text was translated immediately, the translations were so flawed that the comprehensibility of the conversation suffered greatly. The German interlocutor also felt compelled to speak so slowly and clearly that there was no longer any question of a natural flow of conversation[214] – another example of human overconfidence.

The same applies to the field of art, which Kurzweil foresees as fundamentally revolutionised by technology. In 2009, human musicians would routinely jam with cybernetic musicians, non-musicians would now be able to make music, automatic composition software would make it possible for any musical layman to write music.[215] A good example of the "quality" of automatically generated compositions in 2014 is offered by the software *TransProse*, which uses certain algorithms to analyse the mood of prose texts and supposedly generates music to match. The results are at best aleatory and more reminiscent of the involuntary tinkling of several toddlers at a piano than music.[216] Of course, musical amateurs can also create music on the *iPad* using software such as *GarageBand.* However, they usually use prefabricated samples recorded by professional musicians[217] – it is not the software that produces the music, it merely offers a platform for compiling music from already existing set pieces. However, there was already plenty of sampling software of this kind in 1999. The sound quality of modern software instruments and synthesizers – including those of Kurzweil *Music Systems,* a company founded by Kurzweil – also leaves much to be desired in 2014. Electronically generated piano and string sounds still sound synthetic and are in most cases no adequate substitute for real instruments. In this respect, too, the technical development has been overestimated.

A similar conclusion was reached by Forbes editor Alex Knapp, who subjected Kurzweil's prophecies to a reality check in March 2012:

> *"Out of 12 key predictions that Kurzweil highlighted for the year 2009, only one has come completely true. Four were partially true (score them a half-point each) and eight failed to come true by the end of 2011. That's a score of 3/12 – or 25% accurate.*

[214] Cf. ibid.

[215] Cf. Kurzweil, The Age of Spiritual Machines, 196.

[216] Cf. https://transprose.bandcamp.com/album/first-iteration (accessed 20 August 2023).

[217] A sample here refers to a pre-produced sound recording of a musical excerpt that is characterised by a simple harmonic and rhythmic structure and whose integral multiple comprises a classical period duration. Entire pieces can be composed from the combination of several samples from libraries sorted thematically and by genre.

> *This is actually being somewhat generous, because if you go and read the chapter that provides a fuller explication of the world Kurzweil predicted, the picture he paints of the culture and society in general were pretty far off."* [218]

[218] Knapp, Ray Kurzweil's Predictions For 2009 Were Mostly Inaccurate.

III. The Human Being - More than a Machine?

"Most of the damage that the computer could potentially result in depends less on what the computer can or cannot actually do, and more on the characteristics that the public ascribes to the computer. The non-specialist has no choice at all but to attribute to the computer the properties that come to him through the propaganda of the computer community amplified by the press. Therefore, the computer scientist has an enormous responsibility to be modest in his claims."[219]

Joseph Weizenbaum

What is the human being? The previous chapter was devoted to the school of *strong artificial intelligence* and its answer to this question. It became clear that its representatives, such as Kurzweil, Moravec or Minsky, see no more in humans than the product of their physical properties. According to the naturalistic view of strong AI, humans are completely reducible to physical processes and, once these processes are understood, can be easily replicated in the form of hardware and software.

This chapter attempts to look at the question from the perspective of philosophy, theology and humanism. With Joseph Weizenbaum, it is dedicated to a representative of *weak artificial intelligence*, the moderate of the two AI schools, who advocates such a humanistic approach to technology and the natural sciences and warns against a naïve faith in science. Although the views presented here are based on different religious and philosophical views, what they have in common is that they speak out against reducing humans to their physical processes.

The debate about artificial intelligence is also a debate about the relationship between body and soul, brain and mind. For this reason, this chapter provides an overview of the current state of the mind-brain debate.

[219] Weizenbaum, Albtraum Computer, translation: MW.

1. Joseph Weizenbaum

One of the important protagonists of AI research is also one of its strongest critics: with his ELIZA programme, Joseph Weizenbaum presented one of the first language-analytical programmes for human-computer interaction in 1966. The reactions he observed to this programme made him an early sceptic of IT technology, which is reflected in his major work *Computer Power and Human Reason. From Judgment to Calculation*[220] .

1.1 Biographical Sketch

Joseph Weizenbaum was born in Berlin in 1923 to Jewish parents.[221] His father is a master furrier, Weizenbaum describes him as a "strict man without human warmth".[222] His mother, on the other hand, is very affective; Weizenbaum says in retrospect that he "feared suffocating from her love".[223] The parents put emphasis on a religious upbringing for their children; together with his brother Heinz, Joseph attended a Torah school in Berlin.[224]

Immediately after his 13th birthday, the family fled from the National Socialists to the USA in 1936[225] and settled in Detroit.[226] As a Jew who experiences antisemitism both in Berlin and in Detroit, but who is able to "very well stand up to it", he feels a sense of "otherness" early on. Along with this "otherness" – unlike his classmates, he does not enjoy sports and ball games, for example – he discovers his love of mathematics early on.[227]

After his school years, he therefore decided to study mathematics at Wayne University in Detroit. There he was involved in the design and construction of the first computer for the university in the late 1940s.[228] His enthusiasm for technology earns him a call as a visiting professor at the Massachusetts Institute of Technology in 1963, where he

[220] Cf. Weizenbaum, Computer Power and Human Reason.

[221] Cf. Brandt-Herrmann, Typische Biographien untypischer Informatiker, 91.

[222] Weizenbaum, Inseln der Vernunft im Cyberstrom, 40, translation: MW.

[223] Ibid., 41.

[224] Cf. ibid., 40.

[225] Cf. ibid., 43. Despite his relatively young age, Weizenbaum was already aware of the necessity of ess cape: "When we [...] left Germany in 1936, I knew that we were now escaping something evil. It was clear to me that it was an escape, a real necessary escape", ibid., 45f, translation: MW.

[226] Cf. ibid., 52.

[227] Cf. ibid. Weizenbaum says in 2006 about his "otherness": "It meant establishing my identity for myself. And has remained so throughout my life. Later I became a member of the Scientific Establishment, that is, the scientific elite in America, but at the same time a dissident. I was different and I am different", ibid, 53, translation: MW.

[228] Cf. Brandt-Herrmann, Typische Biographien untypischer Informatiker, 91.

and colleagues develop a time-sharing system, a system that allows several users to use a computer at the same time.[229] A few years later, Weizenbaum develops ELIZA: while computers of the time were indirectly controlled by punched cards, his team develops a typewriter input for computers. With ELIZA, Weizenbaum now presents a programme with which one can have a "conversation" in natural language via the keyboard for the first time.[230]

After the presentation of ELIZA, Weizenbaum is increasingly irritated that even colleagues and staff who worked on the development of the programme and should therefore know exactly about its limitations take it seriously as a "conversation partner".[231] Furthermore, Weizenbaum is increasingly critical of MIT's financial dependence on the Pentagon, especially during the time of the Vietnam War: "It was no problem at all then to push through all kinds of research projects. They were funded by the Pentagon."[232] His moral qualms about collaborating on projects that could potentially be used for military purposes contribute to his scepticism about blind faith in technological progress.[233] An early testimony to this scepticism is his essay *Albtraum Computer. Ist das menschliche Gehirn nur eine Maschine aus Fleisch?* (Nightmare Computer. Is the Human Brain Just a Machine Made of Flesh?), published in January 1972 in the German weekly newspaper *Die Zeit*.[234] His 1976 book *Computer Power and Human Reason. From Judgment to Calculation* becomes a standard work of technological social criticism.

After his retirement in 1988, Weizenbaum moved back to Germany. He was a sought-after lecturer and continued to publish until he died of a stroke in his hometown of Berlin on 5 March 2008.[235] Shortly before his death, Weizenbaum wrote in an e-mail:

> *"Our death is the last service we can render to the world: if we did not get out of the way, the generations that follow us would not have to recreate human culture fresh. It would become rigid, unchanging, in other words, it would die. And with the death of culture, everything human would also perish."* [236]

[229] Cf. Weizenbaum, Inseln der Vernunft im Cyberstrom, 89.

[230] Cf. ibid.; for more on ELIZA see below, 51.

[231] Cf. Brandt-Herrmann, Typische Biographien untypischer Informatiker, 92f.

[232] Weizenbaum, Inseln der Vernunft im Cyberstrom, 54, translation: MW.

[233] "I naturally thought of the German scientists who cooperated with the military in the First and Second World Wars. I did my best not to participate", Weizenbaum, quoted in: Loos, Pionier und Pazifist, translation: MW.

[234] Cf. Weizenbaum, Albtraum Computer.

[235] Cf. Loos, Pionier und Pazifist.

[236] Weizenbaum, quoted after: Borchers / Ziegler, Der letzte Service: zum Tode von Joseph Weizenbaum, translation: MW.

1.2 ELIZA – Is the Computer Awakening?

Shortly after taking up his visiting professorship at MIT in 1964, Joseph Weizenbaum began work on the programme that is still closely associated with his name today: ELIZA. He chose the name in reference to Eliza Doolittle from George Bernard Shaw's *Pygmalion*.[237] The programme, completed in 1966, is based on simple speech analysis and simulates, very effectively for its time, a conversation situation with a psychotherapist. Weizenbaum gives an example of such a "conversation" between a human user and ELIZA (in small caps):

> *„Men are all alike.*
> IN WHAT WAY
> *They're always bugging us about something or other.*
> CAN YOU THINK OF A SPECIFIC EXAMPLE
> *Well, my boyfriend made me come here.*
> YOUR BOYFRIEND MADE YOU COME HERE
> *He says I'm depressed much of the time.*
> I AM SORRY TO HEAR YOU ARE DEPRESSED
> *It's true. I am unhappy.*
> DO YOU THINK COMING HERE WILL HELP YOU NOT TO BE UNHAPPY
> *I need some help, that much seems certain.*
> WHAT WOULD IT MEAN TO YOU IF YOU GOT SOME HELP
> *Perhaps I could learn to get along with my mother.*
> TELL ME MORE ABOUT YOUR FAMILY
> *[...]"*[238]

The example makes it clear that the programme is based on a fairly simple algorithm. It tries to isolate individual terms from the user's line of text and convert them into a counter question. Where this fails, the programme asks a general question to keep the conversation going. Weizenbaum's motivation for the programme is to demonstrate the general functioning of a computer to a wide audience.[239]

What shocks Weizenbaum after the publication of ELIZA, however, are the reactions to his programme. The psychiatrist Kenneth Colby, for example, suggests using the programme for therapeutic purposes:

[237] Cf. Weizenbaum, Computer Power and Human Reason, 3. In Shaw's comedy, the self-important linguist Henry Higgins bets that he can make the flower seller Eliza Doolittle a duchess if only he teaches her the dialect of upper-class London society.

[238] Ibid., 3f.

[239] Cf. ibid., 4f.

> *"The human therapist, involved in the design and operation of this system, would not be replaced, but would become a much more efficient man since his efforts would no longer be limited to the one-to-one patient therapist ratio. [...] A human therapist can be viewed as an information processor and decision maker with a set of decision rules [...]."* [240]

The fact that a psychiatrist no longer sees himself as a person who mediates therapy, but as a mechanical "information processor", and could thus come up with the idea of being able to delegate his work to a computer programme, is a mechanistic reduction of the human being that is simply incomprehensible to Weizenbaum.[241]

He finds it frightening how quickly human users are prepared to perceive the computer as an actual interlocutor when talking to ELIZA. His secretary, who has followed the development of the programme for months and is therefore well informed about how it works, asks Weizenbaum to leave the room during a "conversation" with ELIZA – as if it were an actual conversation partner with whom one is discussing intimate details. Weizenbaum is concerned that people seem to be willingly fooled by the illusion of computers after only a short period of use.[242]

Finally, Weizenbaum is irritated that his programme is regarded as a solution to the problem of machine speech understanding. With ELIZA, he had actually wanted to illustrate the opposite, namely that speech understanding only works, if at all, in contextually very narrowly defined areas – and even there only with considerable limitations, as the conversation with ELIZA quoted above shows.[243]

From these experiences, he concludes that even an educated public generally attributes far higher capabilities to technology than it actually possesses.[244]

[240] Colby, quoted after: Weizenbaum, Computer Power and Human Reason, 5f.

[241] Cf. ibid., 5f. as well as Idem, Inseln der Vernunft im Cyberstream, 97: "Today there are many variants of 'Eliza' on the net, all doing roughly the same thing. Only the purposes are different. There is even a variant in which the programme no longer plays the role of the psychiatrist but that of the priest and, so to speak, receives confessions via computer. Although I am not a Catholic, this idea appalls me. If one really believes that a machine can forgive one's sins and give absolution, then I really wonder what meaning faith or priestly ordination still have." (Translation: MW).

[242] Cf. Idem, Computer Power and Human Reason, 6f.

[243] Cf. ibid., 7.

[244] Cf. ibid.

1.3 The Compulsive Programmer

Weizenbaum describes a phenomenon that was new for his time: that of the *compulsive* programmer. In the computer centres of the universities he notices talented young men[245] of unkempt appearance, fixated on their computer consoles. Their fingers are always ready to make the next input, they work to the point of extreme fatigue, sometimes twenty to thirty hours at a stretch. They eat and sleep in close proximity to the computer, their unkempt appearance suggests that they are barely aware of their bodies and the outside world.[246] Unlike the regular professional programmer who wants to solve specific problems and achieve pre-determined goals, the compulsive programmer sees every problem as an opportunity to interact with the computer – his programming becomes an end in itself. While the professional programmer uses the time between programming wisely, for example to document his work so far, the compulsive programmer spends as much time in front of the computer as possible.[247] Weizenbaum explains this programming addiction with the satisfaction that the feeling of programming can give one: To be the creator of one's own universe on the computer, whose laws are determined solely by the programmer.[248]

This type of programmer is usually an excellent technician who knows every detail about his computer. For this reason, he is tolerated in the data centres. The data centres often draw on his expertise and use a number of his programmes themselves; after all, he programs effectively and at high speed. But since he does not document his work, the data centre becomes increasingly dependent on this type of programmer: "His position is rather like that of a bank employee who doesn't do much for the bank, but who is kept on because only he knows the combination to the safe."[249] The more complex his programmes become, the more unstable they become – because even the programmer inevitably loses track of his work due to a lack of documentation. According to Weizenbaum, a seemingly contradictory psychological situation arises here: while the programmer on the one hand exercises power over the computer by programming the computer accord-

[245] Weizenbaum considers the *compulsive programmer* to be an exclusively male phenomenon: "There are compulsive programmers all over the world [...]. But the funny thing is that they are exclusively men. There are no women who are compulsive programmers. [...] I have looked everywhere [...] for the last 30 years. In vain", Idem, Inseln der Vernunft im Cyberstrom, 119, translation: MW. He speculates that the power fantasies acted out in the compulsiveness of these programmers, which also manifested themselves in the desire of a researcher like Moravec for robot children, are ultimately an expression of envy of women's ability to have children, cf. ibid., 120f.

[246] Cf. Idem, Computer Power and Human Reason, 116.

[247] Cf. ibid., 116f.

[248] Cf. ibid., 115.

[249] Ibid., 117.

ing to his will, the computer on the other hand mercilessly shows the programmer his previous programming mistakes.[250] But instead of admitting to himself that he does not (any longer) understand his own programme, the programmer flees into his own world and continues programming because he sees his power challenged:

> *"[H]e will take enormous risks with his program, making substantial changes, one after another, in minutes or even seconds without so much as recording what he is doing, always pleading for just another minute. He can, under such circumstances, rapidly and virtually irretrievably destroy weeks and weeks of his own work. Should he, however, find a deeply embedded error, one that actually does account for much of the programme's misbehaviour, his joy is unbounded."*[251]

Weizenbaum compares the behaviour of the compulsive programmer to that of a compulsive gambler. Just like the programmer, the gambler imagines himself in control of a magical world whose rules are understood only by a select circle of people. What looks like superstition to outsiders is, for the gambler, a hypothetical construction of this world, which has been revealed to him by his luck: experience may have taught him, for example, that he will win more often with a rabbit's foot as a lucky charm. If this belief is falsified by reality, he flexibly adapts his hypothesis: Perhaps his lucky charm only helps him on Tuesdays and Thursdays. If he still loses on a Thursday, he finds other factors that could have influenced his luck: "Losing, therefore, doesn't mean that carrying a rabbit's foot, for example, is wrong or irrelevant, but only that some crucial ingredient for success has been overlooked so far."[252] Where the programmer flexibly modifies his programme in case of error, he acts no differently than the gambler who creates a highly complex concept of his gambling world, of which he is the sole expert.[253]

Weizenbaum believes that this description of compulsivity can be applied to large parts of science. Just like programmers and gamblers, scientists create their world from empirical observations. In doing so, they share the conviction: "[W]hat science has not done, it has not *yet* done; the questions science has not answered, it has not *yet* answered."[254] The compulsive gambler is convinced that all of life is a game of chance. The compulsive programmer believes that all of life is nothing more than a program on a giant computer and that every aspect of life can be explained in programming rules. Similarly, the obsessive scientist believes that every aspect of life and nature can be explained by

[250] Cf. ibid., 119.

[251] Ibid., 120.

[252] Ibid., 123.

[253] Cf. ibid., 123f.

[254] Ibid., 126 (emphasis as in the original).

scientific-empirical methods. The scientist's belief system is as unshakeable as that of the gambler, because every contradiction within the sciences is in turn resolved with the help of empirical-scientific methods.[255]

Weizenbaum warns against placing the world entirely in the hands of obsessive scientists and promotes a more comprehensive, not purely scientific, view of the world and human beings.[256]

1.4 The Illusion of Power over the Computer

As explained earlier, the computer gives its programmer a sense of power because it appears to do exactly what it has been programmed to do. Similarly, the user, whose input the computer programme reliably transforms into output, feels powerful. However, Weizenbaum suggests that this feeling of power over the computer is an illusion.

Not only because of the obsessive programmer who does not think of documenting his work, but also because of the sheer size and complexity of modern computer systems, they can no longer be fully understood by their own programmers.[257] As an example of the consequences of misunderstood systems, Weizenbaum cites the stock market crash in October 1987, also known as *Black Monday*.[258]

At this time, the first stockbrokers begin to carry out their transactions automatically via computer. The early broker computers analyse price gains and losses and make automated decisions about buying and selling shares. Because the computer can analyse stock market values much faster than a human being and every time advantage is worth money on the stock market, brokers promise themselves high profits through this kind of automation. The computers are not directly networked with each other, but are nevertheless indirectly connected via the market whose data they evaluate and on which they exert influence through transactions. The more of these computers are put into operation, the more they form an uncontrollable, autonomous system: computers react to the buying and selling of other computers by buying and selling. This system was not installed, let alone intended, by any human being – and yet it existed, and still exists today.[259] As an unstable system, it could and had to topple over at some point, and

[255] Cf. ibid.

[256] Cf. ibid., 127.

[257] Cf. Idem, Inseln der Vernunft im Cyberstrom, 116.

[258] Cf. also Fehr, Der vollautomatische Börsencrash.

[259] Cf. Weizenbaum, Islands of Reason in the Cyberstream, 112f. The problem described by Weizenbaum is still being discussed today. In the meantime, more than half of all stock exchange transactions are carried out by computers with the help of high-frequency trading systems, so-called *algo trading*.

according to Weizenbaum, that is exactly what happened in October 1987:[260] The Dow Jones fell by over 20 per cent in one day.[261] The fatal thing about such systems is that, because they were never intended to be systems, they have no off switch. The proposal of the stock exchange supervisory authority to simply not use such systems in the event of a crisis is naïve: especially in the event of a crisis, *every* broker will think that he is in a better position if he is the supposedly only one using the system right now.[262]

According to Weizenbaum, everyone who uses a computer today relies in some way on such a misunderstood and incomprehensible system. This already starts with the operating system: Due to the complexity of modern computer hardware, no individual can write his or her own operating system anymore; instead, he or she must rely on systems that already exist – without being able to know exactly how they function at their core: "Strictly speaking, then, someone else has told my computer what to do."[263] Today, the history of a computer system is rarely handed down: There are hardly any computer systems left that are developed by the same cohesive group of scientists. But if this history is lost, the system can no longer be understood.[264]

Weizenbaum describes two consequences: If computers are used as an aid to decision-making, the criteria for the computer's decision could no longer be questioned in sufficiently complex software systems because they are unknown. Further, the more complex a system becomes, the basic criteria would become immune to any form of modification – any substantial modification could destroy the misunderstood system. Such computer systems can therefore only grow after a certain point. Commonly, however, it is precisely the increasing complexity of such systems that is cited as legitimisation for the apparent correctness of their decisions[265] – a tendency that Weizenbaum considers "more than dangerous"[266] .

The price risks caused by algorithms cannot be estimated, see Welchering, Preisgestaltung im Millisekundentakt.

[260] Cf. Weizenbaum, Inseln der Vernunft im Cyberstrom, 114.

[261] Cf. Fehr, Der vollautomatische Börsencrash.

[262] Cf. Weizenbaum, Inseln der Vernunft im Cyberstrom, 114f.

[263] Cf. ibid., 116.

[264] Cf. ibid., 117.

[265] Cf. Idem, Computer Power and Human Reason, 236f.

[266] Idem, Inseln der Vernunft im Cyberstrom, 116, translation: MW.

According to Weizenbaum, we have handed over a large part of our responsibility to systems that we do not understand. Who bears the responsibility when, for example, a wrong decision is made because of a programming error that may have happened years ago, is no longer identifiable: "An important characteristic of our society is that it has developed the technique of distributing responsibility in such a way that no one has it."[267]

1.5 Artificial and Human Intelligence

Weizenbaum deplores an oversimplifying view of intelligence cultivated by the natural sciences. He sees an indication of this view in the invention of the intelligence quotient and the associated idea that intelligence can be measured quantitatively. However, an intelligence test can only measure a very specific part of human intelligence and can never provide a complete description of human intellectual abilities.[268] In the meantime, however, the human concept of intelligence has been influenced by the IQ test method to such an extent that a large proportion of people no longer think of intelligence as anything other than what an IQ test measures.[269] This simplistic view is partly responsible for the "perverse fantasy" of artificial intelligence and the idea that humans are nothing more than information-processing systems.[270]

While the proponents of strong AI claim that there is no area of human thought that cannot be replicated by machines, Weizenbaum argues that the intelligence of a computer will always be quite different from that of a human.[271] He justifies this, among other things, with the way the human brain works, or more precisely, the two hemispheres of the brain, which seem to function completely differently: While the left hemisphere thinks in a structured, sequentially ordered and commonly logical way, the right hemisphere thinks in a holistic way. For example, the left hemisphere is responsible for understanding language, while the right hemisphere is responsible for spatial orientation or creative processes such as making music.[272] The functioning of the left brain alone is perhaps best compared to the concept of a *general problem solver* when it transforms

[267] Ibid., 115, translation: MW.

[268] Cf. Idem, Computer Power and Human Reason, 203f. as well as Idem, Islands of Reason in the Cyberstream, 101f.: "[The] intelligence tests we are familiar with measure intellectual abilities that are regarded by influential representatives of the Western world as important (thinking) achievements. In other societies it will be quite different. [...] Intelligence is precisely not an objective quantity and not a linearly measurable phenomenon that exists independently of a particular frame of reference." (Translation: MW).

[269] Cf. ibid., 102.

[270] Cf. Idem, Computer Power and Human Reason, 203.

[271] Cf. ibid., 207.

[272] Cf. ibid., 214.

tasks such as "Tom has twice as many fish as Mary; if Mary has three guppies, how many fish does Tom have?" into functional representations such as *"$x = 2y$; $y = 3$; …"*. However, human problem solving, just like human communication, always combines the function of both hemispheres of the brain.[273] Functional representations alone are therefore not sufficient to describe human thinking, since people know much more than they can say or express in any symbol system – notes, mathematics or chemical formulae. Just because it is unspeakable, however, does not mean that it cannot be talked about, much less that it does not exist.[274]

Furthermore, human intelligence is always embedded in a social context. Even complete knowledge of all genetic and neurological structures of a living being is therefore not sufficient to understand the living being. Because it is in principle impossible to understand humans purely scientifically, the quest to produce robots in human form is absurd.[275]

Weizenbaum does not so much ask whether computers will eventually be *able to* make legal or psychiatric decisions. He asks instead whether computers, given their fundamentally different way of "thinking", *should* make such decisions:

> *"Computers can make judicial decisions, computers can make psychiatric judgments. They can flip coins in much more sophisticated ways than can the most patient human being. The point is that they* ought *not be given such tasks. They may even be able to arrive at 'correct' decisions in some cases – but always and necessarily on bases no human being should be willing to accept. […] What I conclude here is that the relevant issues are neither technological nor even mathematical; they are ethical. […] What emerges as the most elementary insight is that, since we do not now have any ways of making computers wise, we ought not now to give computers tasks that demand wisdom."*[276]

Weizenbaum sees application-specific expert systems as different from this. A computer system that lands an aeroplane, for example, perceives many factors at the same time much better than a human being – such an application-specific system, however, has nothing to do with intelligence. According to Weizenbaum, the vast majority of applications that are called [weak, MW] artificial intelligence today have precisely this applica-

[273] Cf. ibid., 219f.

[274] Cf. Idem, Inseln der Vernunft im Cyberstrom, 161 as well as Idem, Wir gegen die Gier: "Unfortunately, I never got to know the poet Ionescu. From him comes the statement: 'Everything is sayable in words, except the living truth.' I would say to Ionescu: Very much is representable by the natural sciences, but not the living truth." (Translation: MW).

[275] Cf. ibid.

[276] Idem, Computer Power and Human Reason, 227 (emphasis as in the original).

tion-specific character. The designation as artificial intelligence, however, is part of the propaganda of AI researchers. They benefit from the fact that people outside universities and research institutes can no longer decide what is actually possible and what is not, due to the omnipresence of the AI term. In this way, the "fairy tale belief"[277] in artificial intelligence continues to spread.

1.6 Weizenbaum - an Enemy of Technology?

Weizenbaum defends himself against the accusation that his theses are irrational, anti-science and anti-technology. In this accusation, he sees an argumentation strategy of fundamentalist technologists and scientists who try to brand all objections to their megalomaniac visions as a rejection of reason, science and technological progress. The fundamentalist scientist considers intuition, emotion and religion irrational.[278]

Weizenbaum, on the other hand, sees himself precisely as an advocate of rationality. However, he opposes a conception of rationality that is completely decoupled from intuition and feeling. This decoupling ultimately leads to the mystification of science and technology. He pleads for a rational use of science that includes ethics as a fundamental element. His struggle is not with reason, but with the "imperialism of instrumentalised reason".[279] He warns against the addictive potential of a purely scientific worldview:

> *"It [...] used to be said that religion was the opiate of the people. [...] On the other hand, it may be that religion was not addictive at all. [...] But instrumental reason, triumphant technique, and unbridled science* are *addictive. They create a concrete reality, a self-fulfilling nightmare."* [280]

2. The Mind-Brain Debate

As previously indicated, a certain positioning within the mind-brain debate is required in order to claim, for example, a passed Turing test as proof of human-like thinking by machines. For a better philosophical location of the theses advocated by the strong artificial intelligence school, an overview of the current state of this debate is given here.

The mind-brain debate discusses the relationship between physical and mental processes of the brain. It distinguishes between the *first-person* and *third-person perspectives and* discusses a phenomenon that arises from our everyday experience: on the one hand, we

[277] Idem, Inseln der Vernunft im Cyberstrom, 128.

[278] Cf. Idem, Computer Power and Human Reason, 255f.

[279] Ibid., 256.

[280] Ibid., 256f (emphasis as in the original).

experience ourselves as beings whose body and mind are interdependent and interact with each other. On the other hand, there is already a linguistic separation between body and mind, so that an independence of the two entities, for example in the form of a disembodied world of consciousness, is at least theoretically conceivable.[281]

Brain research agrees that all mental and conscious processes occur either on the basis of or accompanied by physical neuronal processes. The discussion, however, begins with the question of whether these neuronal processes are only the necessary or already the sufficient condition for conscious processes.[282]

Fundamental to the mind-brain debate is a trilemma whose propositions are intuitively plausible but incompatible with each other:

1. "The physical world is causally closed without gaps." This sentence reflects the methodological principle of the natural sciences.
2. "From the causal closure of the physical world follows the causal ineffectiveness of mental events and properties." This sentence is nothing more than an explication of the first sentence. If the first sentence is completely true, the second must consequently also be true - even if this seems counterintuitive due to the fact that we can determine conscious experiential processes from the first-person perspective.
3. "Mental events are causally effective." This sentence reflects an inescapable basic intuition that corresponds to our consciousness of freedom. But it is incompatible with propositions 1 and 2.

The various positions of the mind-brain debate now try to "explain away" the incompatible proposition in favour of the other or others.[283]

A basic distinction can be made between *dualistic* and *monistic* positions. *Dualism* understands – as did Descartes with his distinction between the *res cogitans* and the *res extensa* – the physical and the mental as independent entities. However, this raises the question of how the two can be integrated. Descartes supposed that there was a so-called pineal gland in the brain that acted as a translating organ between mental and physical processes. However, this view is not biologically tenable. Modern *interactionists* like John Eccles refer to Heisenberg's uncertainty principle and see the physical quan-

[281] Cf. Pröpper, Theologische Anthropologie, 850.
[282] Cf. ibid., 858.
[283] Ibid., 851.

tum level as a gateway for mental processes. However, this does not solve the problem, but merely postpones it. *Parallelists* such as Arnold Geulincx and Nicolas Malebranche, on the other hand, argue in an occasionalist way when they assert a strict parallelism of all mental and physical processes brought about by God. However, this view ultimately cannot explain the meaning of created entities and clashes with the idea of freedom. For these reasons, dualistic views are now rarely held.[284]

Epiphenomenalism takes a middle position between dualism and monism, which interprets mental phenomena as merely accompanying physical processes in the brain. As such, mental phenomena do exist and are perceived from the first-person perspective, but they cannot have an effect on the physical – there is therefore neither intentionality nor freedom. Epiphenomenalism is extremely counterintuitive. It also does not answer the question of why mental phenomena have evolved evolutionarily at all if they are causally ineffective and thus irrelevant for survival.[285]

In contrast to dualism, *monism* denies a categorical difference between the mental and physical realms. Meanwhile, *mental monism,* which advocates a radical scepticism towards the external world and conceives the entire physical realm either as a pure construct of thought or as arising from mental processes, no longer plays a role. This idea also seems counterintuitive.[286]

In contrast, the *physical monism* that sets the tone in the current debate dissolves the difference between physical and mental processes in favour of the physical: As *reductive physicalism,* it identifies mental phenomena completely with their correlating neuronal processes. The proponents of the *identity thesis* thereby reduce internal, mental processes completely to behavioural dispositions in order to assert the complete identity of mental and physical states. However, current research shows that the same mental state can be realised by different physical states. *Functionalists* would agree with this: Mental states are functional states that can be realised in different ways by different carriers.[287] However, all reductive physicalism does not explain the subjective experiential quality of mental states, the so-called qualia.[288]

[284] Cf. ibid., 851f.

[285] Cf. ibid., 853.

[286] Cf. ibid.

[287] Cf. ibid., 853f.

[288] Cf. below, 62.

Non-reductive physicalism sees mental entities as fully realised by physical ones, but is concerned with a qualitative added value of the mental when it conceives mental properties as *emergent* properties. Emergent means that mental properties are determined by the physical system but cannot be traced back to it: The complex combination of physical entities in the brain could produce additional (mental) properties that the individual components do not possess by themselves. The question arises, however, whether the assertion of additionally produced properties does not represent a contradiction to the theory of physicalism.[289]

Eliminative materialism is the most extreme form of physical monism: It simply denies the existence of mental states. Such a notion is completely counterintuitive because it completely eliminates the first-person perspective. The question arises as to how such a theory is even defensible, let alone livable.[290]

So far unsolved, and possibly also unsolvable for methodological reasons, are four problems relating to the determination of the relationship between brain and mind:

1. The first problem is known as the *qualia problem.* It concerns the question of how subjectively perceptible experiences of consciousness, so-called qualia, can arise at all from neuronal processes. This is not only about the correlation of mental and physical processes, but rather about the question of *why* something feels *somehow at all.* This question arises in particular for the representatives of physical reductionism. So far, they have not been able to close this explanatory gap through a successful psychophysical reduction.[291]

2. Another problem is the question of *intentionality.* From a *naturalistic* perspective, intentionality of conscious processes is merely a linguistic construction, a shorthand we use for simplicity for complex physical processes that are in fact determined and can therefore, in principle, be replicated, as artificial intelligence attempts to do. Philosophers such as John R. Searle and Hilary Putnam, however, have shown that language,

[289] Cf. Pröpper, Theologische Anthropologie, 855f.

[290] Cf. ibid., 856f.

[291] Cf. ibid., 859-861.

logic and cognition seem to be too complex to be reduced to symbol processing systems. The propagation of a naturalistic thesis thus presupposes an unshakeable faith in science.[292]

3. The question of the *binding problem*, namely why a subjectively experienceable ego arises at all, is also unresolved. Since no neuronal correlate can be found for the ego, it can only be an illusion for the representatives of naturalism, provided they think their position through to its logical conclusion.[293] However, this idea is also counterintuitive in the end.

4. The last question concerns the *freedom of the will.* The Libet experiment[294] apparently showed that free will is only an illusion. Naturalists therefore occasionally refer to this experiment, while more recent brain research strongly questions Libet's results. Two positions can be identified in answering this question: The *incompatibilists* advocate a strong concept of freedom and see this as incompatible with a neuronal determination of the will. They do not deny that all freedom is conditioned by internal and external factors. Nevertheless, the individual has the possibility to act in agreement or disagreement with any situation. *Compatibilists*, on the other hand, claim that freedom and determination are compatible, but without providing a conclusive explanatory model for this – in this respect, the compatibilist concept of freedom ultimately amounts to a labeling fraud.[295]

Since there is no naturalistic concept that could completely explain the emergence of an "illusion of freedom", there is no reason from a scientific perspective to deny the intuitively experienced first-person perspective.[296]

The Heidelberg psychiatrist and philosopher Thomas Fuchs has made a current contribution to the debate in favour of human freedom. He attests that the entire standard scientific theory has a tendency towards dualism when it assumes a "bodiless and worldless subjectivity on the one hand and a physicalistically reduced, material world on the other"[297]. He sees the supposed alternative between a subjective ego in the sense

[292] Cf. ibid., 861-863.

[293] Cf. ibid., 863f.

[294] Cf. above, 32.

[295] Cf. Pröpper, Theologische Anthropologie, 865-870.

[296] Cf. ibid., 872f.

[297] Fuchs, Das Gehirn – ein Beziehungsorgan, 47, translation: MW.

of the Cartesian *res cogitans,* which rules over the entire body (*res extensa*), and the brain itself as the author of actions as too narrow.[298] The brain, as an organ, is not capable of making any decisions at all – concepts such as feeling, willing and deciding are not even applicable on a physiological level:

> *"The brain does not have mental states or consciousness, because the brain is* not alive *– it is only the* organ *of a living being, a living person. Not neuron assemblies, not brains, but only* persons *feel, think, perceive and act."* [299]

Of course, the capacity of a person is bound to his or her brain functions – the brain is thus of central importance for the possibility of personal existence. However, one should not confuse the *person* with a part of the body, as it is always a unity of body and soul, the living human being.[300] Fuchs sees the brain as an organ of freedom because its "increasing complexity in the course of evolution has loosened the rigid stimulus-response mechanism and thus enabled organisms up to the human being to have more and more degrees of freedom"[301]. It is the organ of possibilities: "It is not the mind that has to do what the neurons tell it to do, but the neurons make possible everything that unfolds in the mind".[302]

In summary, it can therefore be stated that the state of scientific research neither refutes nor contradicts the Christian-theological view of human freedom. Moreover, if one follows Fuchs' view, a mind independent of the body, as propagated by the school of strong artificial intelligence in the form of a software replication of the brain, can never be alive. If neuron associations within the brain are not sufficient to produce a person, this is all the more true for the simulation of neurons in the form of software.

As has been shown, the radical school of artificial intelligence believes that the human mind can be completely reproduced on a machine. To the extent that it considers the production of mental processes on a machine, a physical but non-biological carrier, to be possible, it argues in a functionalist manner. On the question of intentionality and free will, she takes a naturalistic position: only under this premise can the passing of the Turing test, for example, be regarded as proof of human-like thinking in machines.[303]

[298] Cf. ibid., 67.

[299] Ibid., 283 (translation: MW, emphasis as in the original).

[300] Cf. ibid.

[301] Ibid, 77.

[302] Ibid, 246.

[303] Cf. above, 21.

Ultimately, however, behind the idea of being able to isolate the human mind from the body and "upload" it onto a machine is a new form of dualism[304] – regardless of the fact that this position is hardly pursued in the current mind-brain debate.[305]

3. Artificial Intelligence in Philosophy and Theology

In this section, philosophers and theologians will have their say who critically engage with artificial intelligence and its theories in very different ways. Their different aspects and approaches are each in their own way fruitful for the debate on AI.

3.1 John R. Searle

The philosopher John R. Searle has already been mentioned in connection with the Turing Test.[306] With his Chinese Room argument, he shows that syntax and semantics are not the same thing and that syntax is not sufficient for semantics either: "A computer [...] could work through the steps of a program for some mental skill, such as understanding Chinese, without understanding a single word of Chinese."[307] From this he concludes that the human mind cannot be a computer programme. On the other hand, Searle answers the question of whether the processes of a brain can be simulated on a computer in the affirmative: According to the Church-Turing thesis[308], anything that can be described precisely enough as a sequence of steps can be simulated on a digital computer – just as the weather, the behaviour of the stock market or an airline schedule can be simulated on a computer.[309]

Contrary to what one might intuitively assume, the question of whether the brain is a computer is not yet settled that way. Even if the mind is more than a computer programme, it is still conceivable that mental processes correspond to computer processes. Searle calls the view that the mind is a computer programme strong AI, the view that the brain can be simulated on a computer weak AI, and the view that the brain is a digital computer cognitivism.[310]

[304] The dualism concept of strong artificial intelligence still goes far beyond that of the mind-brain dee bate, cf. below, 77.

[305] Cf. above, 59.

[306] Cf. above, 21.

[307] Searle, Ist das Gehirn ein Digitalcomputer? 212, translation: MW.

[308] The Church-Thuring thesis states that for every algorithm there is a Turing machine that can implee ment the algorithm, cf. ibid., 213.

[309] Cf. ibid., 212.

[310] Cf. ibid., 212f.

Cognitivism argues that the only alternative to understanding the brain as a digital computer is that of Cartesian dualism.[311] It sees the question of whether brain processes are computational processes as a question of empiricism that can be resolved in the same way as the question of whether the heart is a pump or whether green leaves perform photosynthesis. According to Searle, fundamental questions are often not clarified in the cognitivist literature, such as the question of what exactly a digital computer or a computational process actually is.[312]

A digital computer on the model of a Turing machine is a unit that can rewrite a zero on its tape into a one, a one on its tape into a zero and move its tape one field to the left or to the right, controlled by a programme with instructions. The problem with this definition is that in most computers we do not actually find zeros and ones, but representations of these digits, for example in the form of transistors and circuits. This means that, in theory, any system that is thought to have a representation of zeros and ones must be considered a digital computer, regardless of the material used. Therefore, if the brain is considered a digital computer, this also means that it can be made from any material[313] – a thesis that Kurzweil and Moravec would share. However, this multiple feasibility is accompanied by some problems:

1. There is a difference between functional multiple feasibility, when, for example, a carburettor can just as well be made of brass or of steel, and syntactic multiple feasibility, which would entail that all things representing zeros and ones are computers: "For this reason, no one says that carburettors, for example, can be made from pigeons. But the class of computers is syntactically defined in terms of the *attribution* of zeros and ones."[314] But then every object would be a digital computer because, according to Searle, everything could be described in terms of zeros and ones. Moreover, the attribution of properties always presupposes an external observer who considers certain phenomena to be syntactical. However, a position that brains are digital computers because all things are digital computers does not advance the debate.[315]

[311] Thomas Fuchs shows how narrow this view is, cf. above, 63f.

[312] Cf. Searle, Ist das Gehirn ein Digitalcomputer? 215f.

[313] Searle gives the example of an elaborate system of cats, mice and cheese that is interconnected in such a way that we might also understand, for example, the cat pulling a switch as a representation of zero or one, cf. ibid.

[314] Ibid., 218 (translation: MW, emphasis as in the original).

[315] Cf. ibid., 218f.

2. The second problem is that of the homunculus fallacy: every computer has a user, an external actor who attributes to it the properties of a computer and controls and interprets inputs and outputs. If the brain were indeed a computer, it would also need a user. In fact, cognitive scientists, often unconsciously, strive for a homunculus, an "internal user" that sits in the brain – and ultimately contradict themselves.[316]

3. The third problem is that an externally ascribed syntax of zeros and ones cannot have causal powers: According to the theory of cognitive science, a large pile of zeros and ones is manipulated in the digital computer brain, racing rapidly through the brain to produce cognition. However, as shown earlier, these zeros and ones exist only in the eye of the external observer. "The implemented program has no causal powers other than those of the implementing medium because the program has no real existence, no ontology independent of that of the implementing medium."[317] Without a homunculus, there are only patterns in the brain, as in the computer, which in themselves, without interpretation, cannot have causal powers. Cognitivism, then, cannot give a causal explanation for cognition.[318]

4. Cognitive science ultimately claims that the brain is an information-processing system. Searle shows that this is not the case: in a computer, an external actor makes inputs that are coded in such a way that the computer can process them. The computer then goes through a series of electrical states that can be interpreted syntactically or semantically by the external observer, but still have no intrinsic syntax or semantics of their own. Finally, the computer produces an output that is in turn interpreted syntactically or semantically by an external observer. In the brain, however, the relevant neurobiological phenomena take place independently of the observer. A computer can be used to produce an information-processing model of these phenomena. "But the phenomena themselves are not information-processing systems."[319]

[316] Cf. ibid., 221-223.

[317] Ibid., 224, translation: MW.

[318] Cf. ibid., 225.

[319] Ibid., 230f, translation: MW.

To regard the brain as a digital computer is thus a fallacy based on a whole series of flawed basic assumptions of cognitivism.

3.2 Margaret A. Boden

The British psychologist and philosopher Margaret Boden deals extensively with topics such as cognition and artificial intelligence. Among other things, she addresses the question of the fear that the advancement of artificial intelligence could call the image of man into question and thus push back or even negate human values.[320]

She admits that because natural science has no room for concepts such as intention, choice, action, creativity and above all subjectivity, it indirectly promotes an understanding of the world and of the human being in which these concepts also no longer take place. She describes the effects of such a worldview as inhumane: one does not do justice to human beings if one no longer grants them responsibility or treats them like machines.[321]

Against this background, Boden outlines two threats that society sees in artificial intelligence: firstly, that humans could become aware that they are not as intelligent as they thought, because machines seem to be more intelligent than they are; secondly, that humans could become convinced that they themselves are nothing more than a machine. This would put both human uniqueness and human dignity at stake.[322]

Boden, however, is of the opinion that the actual message of AI is exactly the opposite: thinking and intelligence performances of AI programmes are poor in comparison to human ones. This point is often overlooked because we are comparatively poor at the very things that machines are very good at: Mathematics and precisely specifiable scientific thinking skills. At the same time, however, computers fail almost completely at the things we are good at: language comprehension, reading, problem solving by means of *common sense.* Insofar as we become aware of this superiority, the threat to our self-image should also diminish:[323]

> *"Common sense and language comprehension are just two examples that show that the human mind – even the ordinary human mind – is much more powerful and much more intelligent than any AI programme. The space computer HAL (in* 2001 - A Space Odyssey*) is a fantasy and will remain so. To the extent that people realise*

[320] Cf. Boden, Künstliche Intelligenz und Menschenbilder, 235.

[321] Cf. ibid., 236.

[322] Cf. ibid., 237f.

[323] Cf. ibid., 238.

> *this, their self-esteem is not diminished by the advent of AI programmes. It may even become greater: as far as the human mind is concerned, the ordinary is indeed extraordinary."* [324]

Furthermore, computer simulations always constitute representations. Computer systems, just like humans, can only work with the information that has previously been made available to them. Thus, there can be no guarantee that a computer system can provide correct answers under conditions of incomplete information. To protect users from trusting too credulously in the answers provided by computer programs, Boden advocates that software remind its users from time to time that he is dealing with a *programme* that, by definition, cannot do all the reasoning that a human could do and was originally written by a human for human purposes. Programmers should also refrain from "plausibility tricks" in their software – such as addressing the user by name or greeting them with "hello". Computer education could also help users recognise the major limitations of software.[325]

The spectre of AI is therefore easy to debunk: artificial intelligence is much more similar to us in terms of fallibility than we think of it, while it is far inferior to us in intelligence and reason and thus highly dissimilar.[326]

3.3 Hans-Dieter Mutschler

Hans-Dieter Mutschler has already been discussed in the analysis of the singularity as a substitute religion.[327] Among other things, he deals with the question of whether man is a robot. He explains that, due to growing globalisation and economic constraints, there are already people today who are externally governed and only act according to optimisation criteria. These people would no longer act any differently than a machine – in relation to them, the question "Is man a robot?" could be answered in the affirmative. The real question should therefore be: "Does man want to be a robot?"[328]

Mutschler avoids using consciousness as a criterion for distinguishing between humans and robots because there is no consensus on the question "What is consciousness?"[329] Instead, he proposes the criterion of judgement.[330]

[324] Ibid., 241, translation: MW.

[325] Cf. ibid., 241-245.

[326] Cf. ibid., 246.

[327] Cf. above, 40.

[328] Mutschler, Ist der Mensch ein Roboter? 292.

[329] Cf. above, 59.

[330] Cf. Mutschler, Ist der Mensch ein Roboter? 293f.

Today's robots have no power of judgement. Mutschler believes that this will not be the case in the future either, but refrains from making a definitive prognosis:

> *"It is [...] futile to speculate about future technologies anyway, since we are not even able to predict the technological development of the next 20 years. If Hans Moravec thinks he can predict this development for 100 years, then he is doing science fiction, not science."* [331]

He agrees with the statement "Computers may have intelligence, but they certainly don't have the power of judgement": expert systems in medical diagnosis have failed, for example, where the doctor has recognised a certain disease at first go. Computers do not have the power of judgement, but react schematically. Even at the highest political level, the final decision lies with people, not computers: During the Cold War, computers repeatedly diagnosed alleged Russian attacks that were actually based on a calculation error. For these reasons, one would be well advised not to let a computer take away one's power of judgement. The faculty of judgement is therefore something specifically human that distinguishes us from robots. Applied to robots, anthropomorphic terms such as "learning", "deciding", "making experiences" would never describe what we call "learning" in humans.[332]

If it were nevertheless possible at some point to develop autonomous robots with the power of judgement, the question would still arise as to why we would want to build such machines in the first place: An autonomous robot would also have to be able to go on strike and would thus no longer be a means to the ends we have set for it.[333]

Nevertheless, it is conceivable that modern man, due to the constraints of increasing efficiency, feels his power of judgement to be a burden and prefers to marginalise it: "It could be shown relatively easily that man is not a robot. But if he *wants to be* a robot, then any argumentation is powerless."[334]

3.4 Dirk Evers

The Protestant systematic theologian Dirk Evers has already been discussed in the context of the Turing Test.[335] From a theological point of view, he considers it unproblematic to attribute intelligence to machines and to speak of "artificial intelligence" as long as

[331] Ibid., 294, translation: MW.

[332] Cf. ibid., 298f.

[333] Cf. ibid., 300f.

[334] Ibid., 306 (emphasis as in the original).

[335] Cf. above, 21.

one understands *intelligence* in this context as a purely rule-governed process. From a theological point of view, it only becomes problematic when this form of intelligence – like naturalism – is regarded as sufficient.[336] Evers explains that the Turing test only has meaning at all when viewed from this naturalistic perspective and concludes that there is a difference between artificial and natural intelligence: the former is rule-governed, the latter understands meanings.[337]

Mere functionality, as desirable in robots, Evers considers a poor basis for the image of man: "It ignores the ultimate unavailability in which we face each other as acting subjects, and it ignores that we owe ourselves to relationships that cannot be formally controlled without being destroyed."[338] However, like Mutschler, he sees a danger in the fact that predetermined criteria of efficiency lead to an ever more effective instrumentalisation of modern man:

> *"The problem in dealing with robotics and artificial intelligence therefore seems to me not so much that robots and computers are becoming more and more like us, so that we would be offended in our uniqueness, but that we feel compelled or even tempted to become more and more like* them.*"* [339]

[336] Cf. Evers, Der Mensch als Turing-Maschine? 103.

[337] Cf. ibid., 105-107.

[338] Ibid., 111 (translation: MW).

[339] Ibid., 112 (translation: MW, emphasis as in the original).

IV. Artificial Intelligence as a Challenge

"The old man robot is wonderful. He doesn't caress you because he hopes for your inheritance or simply can't find another job. He is simply there because he is yours. Sure, he bathes you and pushes you out into the fresh air if that's what you want. But, the real best thing about him is that he can listen..."[340]

Edward Feigenbaum

"I would like to hear one sensible reason, just one sensible reason, for building a human-like robot. And the answer is always the same: it is for elderly people who are lonely and who need help [...]. A colleague has said that a great advantage of such robots would be that this robot would never say when the old man tells it a story: Oh, listen, you've already told me this story ten times. That's better than a human being even."[341]

Joseph Weizenbaum

The further development of Artificial Intelligence will pose some challenges to us in the future. The school of strong artificial intelligence poses a challenge to Christian anthropology. According to Christian understanding, the human being, as the image of God, must be more than a machine. As has been shown, personhood, free will and intentionality are concepts that cannot be described in scientific categories. In this chapter, I will therefore look at the agenda of strong artificial intelligence in the mirror of the Christian image of man.

Regardless of the unlikely realisation of strong artificial intelligence, ethical and moral questions arise in the practical application of weak artificial intelligence that is already available in our everyday lives. In this chapter, I will take as examples the topics of robots in the care of the elderly, autonomous drones in military operations and computer-controlled, driverless cars in road traffic.

[340] Feigenbaum, quoted after: Evers, Der Mensch als Turing-Maschine? 115, translation: MW.

[341] Weizenbaum, in: Schanze, Plug & Pray, minute 58f, transcription: MW.

1. Artificial Intelligence as a Challenge for Anthropology

1.1 Artificial Intelligence in the Mirror of the Christian Image of Man

If strong artificial intelligence is to lead to a *Human 2.0* (Kurzweil) or to our *mind children* (Moravec), the question of the interrelationship between Christian anthropology and artificial intelligence arises. This section therefore attempts to illuminate the agenda of strong artificial intelligence from the perspective of the Christian image of man. The Pastoral Constitution *Gaudium et Spes* of the Second Vatican Council will serve as a first orientation.

1.1.1 Man in the Image of God

According to the Christian view, the human dignity to which every human being is entitled is based on the fact that God created man in his own image (cf. Gen 1:26). As God's image, man is "capable of knowing and loving his Creator".[342] Because God is a person, man is also a person. However, being made in God's image does not only mean closeness, but also distance: as an image, man is ontologically distinct from the archetype God.[343]

Man, partaking in "the light of the divine mind",[344] is endowed with reason. In the application of his spiritual endowments, he is creatively active and has developed science, technology and art. Reason strives for wisdom, which leads man to the "quest and a love for what is true and good"[345]. Creation is entrusted to him for responsible stewardship (cf. Gen 1:28).

If man is God's image, and this image of man now creates an image of himself in the form of a robot, does the image of God and the human dignity associated with it then also apply to the robot? Foerst would certainly answer this question in the affirmative if she herself ascribes personhood to robots[346] and would even be prepared to baptise (!) them, should they first ask for it.[347] On the other hand, it seems more appropriate to speak of a robot being made in the image of human beings. If there is an ontological difference between God and man, this difference also concerns the divine and human act of creation. Thus, in relation to God, man is the image, in relation to the robot, the

[342] Vat. II, *GS, No. 12.*

[343] Cf. Zsifkovits, Das Menschenbild der christlichen Theologie, 14.

[344] Vat. II, *GS, No. 15.*

[345] Ibid.

[346] Cf. Foerst, Von Robotern, Mensch und Gott, 194.

[347] Cf. Dworschak, Mitarbeiterin der Woche.

archetype. The robot-image of man must therefore be separated ontologically from man once again.[348] Because the robot image remains different from the human archetype, it should not be attributed human dignity. The philosopher Robert Spaemann emphasises that simulations (images) can indeed make the functioning of our own life processes comprehensible, but only if they are interpreted from the point of view of bringing about this purpose, i.e. teleologically. But this teleological interpretation "is only possible for beings who know from themselves, namely because they live, what it means to aim at something."[349] Simulation thus always presupposes the original in order to be understood as a system structure at all.[350]

With the necessary humility and the awareness of not being able to reproduce humans, the construction of human-like robots can therefore actually be an expression of the creative activity of humans.[351] Finally, the attempt to simulate human bodily functions can contribute to a better understanding of the human being and thus of divine creation. On the other hand, wanting to create an artificial consciousness in the sense of strong artificial intelligence and thus take the place of God himself is certainly an expression of hubris.

1.1.2 The Human Being as a Bodily Being

The Church does not regard the human being as a duality, but as a unity of body and soul. Through his bodily composition, man unites "the elements of the material world in himself".[352] Through him they reach the height of their destiny: the praise of the Creator. A disregard for the human body would therefore violate the honour of creation.[353] The bodily composition of the human being also means community and historicity: corporeality implies descent from one another, human beings "live in a very real and at the same time in a very complex sense one from the other".[354]

For powerful artificial intelligence, corporeality is no longer a category. Kurzweil, for example, explicitly speaks of corporeality becoming arbitrary or obsolete in cyberspace.[355] In doing so, he denies an essential dimension of human identity.

[348] Cf. Herms, Künstliche Intelligenz, 291.

[349] Spaemann, Schritte über uns hinaus, 132, translation: MW.

[350] Cf. ibid., 131f.

[351] Cf. Foerst, Von Robotern, Mensch und Gott, 47.

[352] Vat. II, *GS, No. 14.*

[353] Cf. ibid.

[354] Ratzinger, Einführung in das Christentum, 231, translation: MW.

[355] Cf. below, 77.

The New Testament scholar Klaus Berger sees bodily composition as a prerequisite for being a name bearer in the biblical sense:

> *"Whoever can be called by his name is individual and unmistakable. This is not a biological-physical concept, but strictly social. The name is only attributed to me in togetherness. Of course, you can also assign a name to a computer or a car. But then you do so in a very inauthentic sense. Against all dualistic attempts to divide man into a material and an immaterial side, the Bible asserts the indivisibility of man and would therefore not declare machines to be bearers of names."* [356]

The aspect of historicity is also lost in AIs: robots or virtual AIs have no childhood to shape their identity. In cyberspace, temporality becomes obsolete to the highest degree.[357] However, bodily composition and temporality are prerequisites for personhood, which takes place in dialogical form. Spaemann also sees the temporality of the human being as a basic prerequisite for subjectivity because subjects always already identify with themselves through time, i.e. they have a history.[358] The historical identity of a person seems to me to no longer apply at the latest when a hypothetical brain upload is performed.

1.1.3 Human Freedom and Sin

Man is created as a free being, for "only in freedom can man direct himself toward goodness."[359] The gift of freedom is necessary so that man can seek his Creator by his own decision and thus reach perfection in union with God.[360]

Freedom, however, also includes the possibility of its abuse. "[M]an finds that he has inclinations toward evil too, and is engulfed by manifold ills which cannot come from his good Creator."[361] Because of freedom, man is capable of sin; the effort to choose the good represents a lifelong struggle for him. Through sin, man is alienated from himself because it prevents him from reaching his fulfilment. Through Christ, however, man is freed from the bonds of his sin.[362]

[356] Berger, Brauchen wir eine Theologie der Roboter? Translation: MW.

[357] Cf. above, 26.

[358] Cf. Spaemann, Personen, 116. Weizenbaum also sees history-less consciousness as one of the weak points of strong AI: "I'm thinking of a famous researcher who said: 'In 50 years, we'll have human-identical robots and even humans who marry these robots'." But what this robot lakcs, Weizenbaum says, is history. He was never a child, he had experienced nothing that "makes a human out of any material", cf. Schanze, Plug & Pray, minute 17f, translation: MW.

[359] Vat. II, *GS, No. 17.*

[360] Cf. ibid.

[361] Ibid., No. 13.

[362] Cf. ibid.

As a reductionist worldview, strong AI tends to deny real freedom.[363] Kurzweil, for example, sees the Libet experiment as evidence against free will.[364] Without freedom, the concept of "sin" is also obsolete. However, a worldview that denies freedom must inevitably remain incompatible with the Christian view of man.

1.1.4 The Human Being as a Social Being

Man is a social being: He was created "male and female" (cf. Gen 1:27). By his innermost nature he is a social being: "unless he relates himself to others he can neither live nor develop his potential."[365] Through Christ this community is made perfect: "There is no longer Jew or Greek, there is no longer slave or free, there is no longer male and female; for all of you are one in Christ Jesus." (cf. Gal 3:28).

The social idea also comes up short in the theories of strong artificial intelligence. Today's artificial intelligences are always single "beings", be it the chess computer *Deep Blue* or the computer programme *Eugene*, which is credited with passing the first Turing test. Kurzweil's and Moravec's theories remain weak when it comes to social interaction of machines or a description of the social fabric in cyberspace.

1.1.5 The Human Being: Called to Eternal Communion

Man fears death as "perpetual extinction".[366] In the face of death, he is confronted with the riddle of human existence. However, he concludes correctly when he "he abhors and repudiates the utter ruin and total disappearance of his own person":[367]

> *"He rebels against death because he bears in himself an eternal seed which cannot be reduced to sheer matter. All the endeavors of technology, though useful in the extreme, cannot calm his anxiety; for prolongation of biological life is unable to satisfy that desire for higher life which is inescapably lodged in his breast."* [368]

But man is called by God to the "endless sharing of a divine life beyond all corruption"[369]. Through his resurrection, Christ won eternal life for mankind. Therefore, faith is an answer to man's fear of his future and at the same time the hope of a reunion with those who have already died.[370]

[363] Cf. above, 59.

[364] Cf. above, 32.

[365] Vat. II, *GS, No. 12.*

[366] Ibid, no. 18.

[367] Ibid.

[368] Ibid.

[369] Ibid.

[370] Cf. ibid.

Because man is known and loved by God, he cannot perish. In contrast to dualistic concepts of immortality, the Church believes in a holistic and dialogical form of immortality: "The essence of man, the person, remains".[371]

Contrary to church doctrine, the strong artificial intelligence school believes it can eliminate death with the help of technological progress. To this end, it propagates the detachment of the spirit from the body: in cyberspace, "resurrection" takes place bodilessly and purely spiritually.[372] Such a conception of eternal life is thus diametrically opposed to the Christian one. The materialistic worldview of strong AI also cannot satisfactorily answer the question about those who have already died – the deceased have simply lived too early to be able to participate in the "fruits" of the technological revolution.

1.2 Strong AI - a New Form of Dualism

While the concept of dualism within the mind-brain debate merely describes the view that mental and physical phenomena exist separately from each other, the concept of dualism in the history of religion goes beyond this: it describes the idea that the cosmos is determined by two opposing principles. Dualistic systems call these principles spirit and matter, light and darkness or heaven and earth.[373] In all dualistic schools of thought, the sphere of the thinkable and the divine is valued more highly than the material world of the senses.[374] It is striking how clearly dualistic topoi appear in the explanations of representatives of strong AI.

Both Kurzweil and Moravec describe cyberspace as an immaterial sphere that will transcend the material world in the course of the Singularity. Associated with the mundane sphere are scourges such as disease and death, from which humanity will be freed by progressive technological development. Here the old dualistic opposition of spirit and matter, this world and the hereafter, mortality and immortality becomes very clear. It is used to exalt technology as the saviour of humanity.[375]

Kurzweil's dualistic view of the world seems unshakeable: He counters the reproach of neuroscientist Anthony Bell that the brain – unlike the computer – cannot be understood as a dualistic entity with a steadfast adherence to his view: "This argument

[371] Cf. Ratzinger, Einführung in das Christentum, 335, translation: MW.

[372] Cf. above, 25.

[373] Cf. Hutter, Dualism, 387.

[374] Cf. Wetz, Dualism, 389.

[375] Cf. Böhme, Die technische Form Gottes.

is easily dispensed with. The ability to separate in a computer the program from the physical instantiation that performs the computation is an advantage, not a limitation."[376] Without seriously addressing Bell's criticism, Kurzweil presents the ability to arbitrarily replace hardware independently of software as an advantage. He also desires this advantage for the human mind, which he wants to free from the "immutable" and "severely limited" structure of the brain.[377]

In 1996, the science journalist Margaret Wertheim described the promise of cyberspace as follows:

> *"Today's 'angels' are [...] to be found on the Internet: Millions of cybernauts 'surf' here, stripped of their bodies, in an idealised immaterial realm. As beings of the ether, the cybernauts, like the angels, are stripped of all physical limitation. They are free from deformity, disease and ugliness. All frailty of the body is left behind when they enter 'net-space'. Obesity, acne, short stature, short-sightedness or rotten joints are simply thrown overboard. In cyberspace, say the freaks, you can just 'be' – a pure soul that transcends physical and national boundaries."* [378]

This description, which certainly applies to contemporary users of cyberspace worlds such as online gamers, also includes the dualistic idea of transcending bodily composition in favour of an idealised appearance detached from the material substrate.

The Church has always condemned dualism. Already in the Syriac Didascalia, an early Christian church order from the late 3rd century, there is a warning against association with such heretics, who imagine the resurrection as a bodiless and purely spiritual reality – an idea which is surprisingly close to that of an "eternal" life in cyberspace.[379] Further early condemnations of dualism take place at the Synods of Toledo (400)[380] and Braga (561-574)[381] , where the Church's doctrine of creation, Christ and man is emphasised as anti-dualistic. In the Middle Ages, too, the Church takes action against dualistic tendencies, such as those of the Albigensians and Cathars.[382]

A strong artificial intelligence that – as in the case of Moravec and Kurzweil – represents a world view with such clear dualistic tendencies is therefore incompatible with Catholic teaching.

[376] Kurzweil, The Singularity Is Near, 444.

[377] Cf. ibid., 445.

[378] Wertheim, Ehre sei Gott im Cyberspace, translation: MW.

[379] Cf. Syriac Didascalia, no. 26.

[380] Cf. DH, nos. 188-208.

[381] Cf. ibid., nos. 451-464.

[382] Cf. Ganoczy, Dualism, 392.

2. Artificial Intelligence as a Challenge for Practice

2.1 Robots in Elderly Care

In view of an ageing society and the high utilisation of nursing staff, care for the elderly is one of the most pressing issues of our future. By 2050, the need for full-time care workers in elderly care is expected to double to 1.3 million in Germany alone.[383] So it is not surprising that artificial intelligence and robotics are also being discussed as part of the solution to the nursing shortage problem.

The developments in this field are therefore manifold. With *Paro,* the Japanese *National Institute of Advanced Industrial Science and Technology* (AIST) has developed a therapy robot in the form of an artificial seal. It is to be used wherever natural animals are currently used in animal therapy. Equipped with electric motors, touch sensors, loudspeakers, a microphone and artificial fur, it reacts to touch with movements and sounds. According to its manufacturers, *Paro* does the same as real animals in animal therapy, but without their disadvantages: It does not trigger allergies, has an unlimited fitness and does not require any additional care for its keeping. When used with dementia patients, he can have a motivating effect, stimulate communication and reduce stress.[384]

In fact, dementia patients seem to react positively to *Paro.* Especially in cases of agitation and aggression, it helps as a "non-medicinal therapy option".[385] It is certainly the better alternative compared to administering sedating drugs. In view of its rather high price of almost 3,000 €[386], however, questions arise as to how distributive justice is to be achieved and whether this money could not be better invested in natural forms of occupational therapy.

After all, *Paro's* inventor Takanori Shibata makes a point of stating that his robot cannot replace a human being and that its use must be accompanied by qualified personnel.[387] He chose the shape of the seal because humans know seals in principle, but not exactly, and in a direct comparison between the original and the imitation, the robot always loses[388] – its success is therefore based on a deliberate deception. This is another reason

[383] Cf. Reintjes, Drei Finger reichen den Saft.

[384] Cf. ibid.

[385] Cf. AA.VV., Japanische Kuschelroboter für Demenzkranke.

[386] Cf. Schulz / Barth, Plüsch-Tech für Senioren.

[387] Cf. Reintjes, Drei Finger reichen den Saft.

[388] Cf. Schulz / Barth, Plüsch-Tech für Senioren.

why the Catholic theologian Jürgen Manemann remains critical of *Paro*. In view of the care crisis, he sees it at best as a short-term support. Dementia patients, however, are first and foremost entitled to human attention and interaction with nature:

> *"Instead of using Paro, animals should be involved. Their therapeutic effect has been proven. Animals, such as cats and dogs, are capable of real communication because of their ability to empathise. It shows our disturbed relationship with nature when we tell ourselves that the seal Paro can replace contact with humans and animals. In doing so, we are not only cheating the dementia patients, but also ourselves. Dementia patients need a lot of comfort because they are constantly confronted with the loss of experiences and competences. Comfort can only come from people, animals and the rest of nature."* [389]

Autonomous service robots such as the Care-O-Bot from the Stuttgart Fraunhofer Institute go into a different direction. The 1.45-metre tall, 180-kilogram household robot finds its way around rooms on rubber wheels, can identify people and objects and interact with them using a three-finger hand, a screen and voice output. In the future, it will be used to help in the household; its creator Birgit Graf sees it as a logical continuation of household appliances such as washing machines and microwaves. It has already been successfully tested in a retirement home where it was used to distribute water from a water dispenser to residents, whom it could address specifically by name.[390] The nursing scientist Heiner Friesacher welcomes such technical support systems, but criticises that nursing science has not been involved enough in the development of such devices. A service robot that brings water or keeps the kitchen in order may seem sensible at first glance. However, especially with regard to elderly people, the question arises: "Do people want this kind of support? Or do they want more human attention, more opportunities to live in community, more human support system?"[391]

Protestant ethicist Arne Manzeschke also takes an ambivalent view of such systems. Technical assistance systems could help older people to live longer in their own homes. However, technology becomes a problem when it promotes the deterioration of intellectual and motor skills. An intelligent medication dispenser, for example, can be a great help to elderly people, but may take away their mental training.[392]

[389] Manemann, Paro ist ein Ausdruck für den Verlust von Kommunikation, translation: MW.

[390] Cf. Reintjes, Drei Finger reichen den Saft.

[391] Friesacher, quoted from: Reintjes, Drei Finger reichen den Saft, translation: MW.

[392] Cf. Manzeschke, Stumme Roboter.

At the University of Electro-Communications in Tokyo, *Digoro,* a robot that can play with children and senior citizens to relieve their boredom, is currently being designed. It currently masters games such as rock-paper-scissors and memory, albeit at an early stage.[393] However, real human interaction, an essential element especially when playing with children and seniors, does no longer take place here. If such robots are used to relieve caregivers, there is a danger that both children and seniors will be denied human contact.

Riba is a robot designed in Nagoya that is to be used explicitly for care tasks for bedridden patients. With 128 touch-sensitive sensors on the upper arms and 94 on the forearms, it perceives body contacts very sensitively. In principle, it should be able to be used for all physical care activities, which is why the safety of the system is one of the main aspects in its development. The legal responsibility in case of an accident has not yet been clarified.[394]

Here, too, Manzeschke is sceptical: because most nursing activities require a holistic view, he thinks it is wrong to leave them to machines: "When a good caregivers reposition a patient, they talks to him, pay attention to his body tension or skin moisture. That way they can learn a lot about the patient's condition."[395] Machines, on the other hand, only perform predefined tasks in a dull manner, without taking care of the patient in the true sense of the word.

In principle, intelligent machines in elderly care are to be welcomed wherever they can help to maintain an elderly person's autonomy for as long as possible. Intelligent voice-controlled systems, for example, could mean a great deal of quality of life for visually impaired people and make them less dependent on external help. Similarly, service robots could help physically impaired patients maintain their independence longer than before. For caregivers, too, such inventions could one day mean a great deal of help and relief: If the introduction of such technologies leads to more time for personal contact with the patient, this would be a welcome gain in humanity in elderly care.

Artificial intelligence in elderly care becomes dangerous, however, when it is seen as a substitute for real human interaction – a danger that cannot be dismissed in view of the economic pressure, especially in personnel-intensive elderly care. Already, it often seems that old people and those in need of care are "shunted off" to where they are least

[393] Cf. Reintjes, Drei Finger reichen den Saft.

[394] Cf. ibid.

[395] Manzeschke, Stumme Roboter.

inconvenienced. This poses a serious problem, not only for those affected: A society that no longer maintains contact with its elderly inevitably becomes poorer. Not only does it forego valuable experiences of the elderly; it increasingly loses sight of an important stage of life, namely the end of life.

In view of economic constraints, it is conceivable that elderly people will be deprived of more and more human interaction opportunities and have to settle for robots like *Care-O-Bot* and *Digoro* instead. However, as has been shown before, a robot cannot replace real human interaction due to a lack of consciousness, emotionality, intentionality and free will. Only human beings are made in the image of God, and as social beings, each of these images has a right to contact with other images.

Technological progress could further accelerate such a development of alienation. However, it would be a mistake to conclude that artificial intelligence technology has no place in the care of the elderly: what matters is that it is used wisely and humanely. Technology can help to relieve the burden on carers and relatives and guarantee the independence of elderly people for a longer period of time. However, it cannot replace social ties:

> *"We are richer than ever before, but we have achieved this partly by increasingly reducing social ties. Everyone can and should work. Those who can no longer work should be well looked after. But we now lack the social ties for that. Instead, technology is supposed to solve the social problems. That will hardly work."*[396]

2.2 Drones in Warfare

Nobel Peace Prize winner Barack Obama has given war a new face. Due to the war-weariness of the American population, he is gradually withdrawing human soldiers from crisis areas such as Afghanistan and Iraq and is relying instead on remote-controlled drones. For him as president, this "war by joystick" has many advantages: Fewer US soldiers return in coffins, more interventions can take place in secret.[397] The operation is cost-effective because the needs of pilots and drivers no longer have to be taken into account. The drone pilot usually sits at a control console thousands of kilometres away from the operation site. He can go home in the evening and complete "war in shifts". Little is known so far about the psychological effects of such a war mission.[398]

[396] Manzeschke, Stumme Roboter, translation: MW.

[397] Cf. Kolb, Wie Obama lernte, die Drohnen zu lieben.

[398] Cf. Rieger, Das Gesicht unserer Gegner von morgen, 31.

However, the widespread use of drones, equipped with cameras, sensors, radio surveillance instruments and missiles, also means the entry of algorithms into the military. No one responsible for controlling the drone is on site; its navigation is usually carried out by autopilot according to a predefined route. Due to the satellite connection, the control console can only intervene with a time delay in the range of seconds. In view of the mass use of drones, the flood of data from their sensors is enormous and already overwhelms their controllers: one reason for the use of artificial intelligence. Algorithms search for typical movement patterns that correspond to suspected movements of insurgents, for example, and make recommendations for action. Only the final push of a button is still left to humans for legal and moral reasons.[399]

From a purely technical point of view, drones are already capable of independently recognising and attacking targets with the help of algorithms. In view of the barely manageable flood of data, human command is already only abstract:

> *"At some point in the foreseeable future the final push of a button will become just a ritual, an action that is actually superfluous because it no longer represents a conscious decision, but only the time-honoured tradition, in the face of technical possibilities an increasingly inefficient and antiquated moral obligation."* [400]

However, the software underlying drones is so extensive and complex that it inevitably makes mistakes. As Weizenbaum and Boden have shown,[401] expert systems often lull us into the illusion that they are infallible. In this respect, the calculation of recommendations for action that amount to decisions about life and death is morally highly questionable. Without human control, such systems would become completely irresponsible.

The Catechism of the Catholic Church does oblige every citizen and ruler "to work for the avoidance of war."[402] Nevertheless, it states that the absence of war does not necessarily mean to peace:

> *"Peace is not merely the absence of war, and it is not limited to maintaining a balance of powers between adversaries. Peace cannot be attained on earth without safeguarding the goods of persons, free communication among men, respect for the dignity of persons and peoples, and the assiduous practice of fraternity."* [403]

[399] Cf. ibid.

[400] Ibid, translation: MW.

[401] Cf. above, 55 as well as 68.

[402] CCC, No. 2308.

[403] CCC, No. 2304.

If war cannot be avoided, the Catechism makes the "doctrine of just war" the basis of warfare:

> *"The strict conditions for legitimate defense by military force require rigorous consideration. The gravity of such a decision makes it subject to rigorous conditions of moral legitimacy. At one and the same time:*
>
> - *the damage inflicted by the aggressor on the nation or community of nations must be lasting, grave, and certain;*
> - *all other means of putting an end to it must have been shown to be impractical or ineffective;*
> - *there must be serious prospects of success;*
> - *the use of arms must not produce evils and disorders graver than the evil to be eliminated. The power of modem means of destruction weighs very heavily in evaluating this condition."*[404]

Regardless of the legitimacy of current wars, which cannot and should not be discussed here, the use of drones seems to blatantly violate the fourth point of this just war doctrine. Those who know neither the programme basis nor the flaws of targeting algorithms have no way whatsoever to "pay careful attention to the awesome destructive power of modern weapons". The killing of innocents, when taken on the basis of a faulty software recommendation, quickly results in damage greater than the evil to be eliminated.

The use of drones should be critically questioned for several reasons anyway. First of all, the question arises whether the supposed distance of a drone pilot from the actual theatre of war does not lower an inhibition threshold that would still exist if the pilot were actually deployed on the ground. Furthermore, data from the air, which can never be fully analysed by a human being, must not be used as a basis for decisions about life and death. Even analysing algorithms do not change this, because no software is error-free and, moreover, no morality can be attributed to it.

Therefore, machines must not be allowed to kill on their own under any circumstances. Software-based recommendations for action are also morally unjustifiable if they become the basis for decisions on life and death. For these reasons, the outlawing of war drones, analogous to the outlawing of chemical, biological or nuclear weapons, should be urgently discussed:

[404] CCC, No. 2309.

> *“War takes place between people, even if machines carry it out. And only those who defeat other humans, force them to surrender or even subjugate them, will be able to win the war. The dangers of robotising war are so great that automated killing machines must be outlawed. Time is pressing. For modern armies are already planning with machine warriors. Weapons can be banned if a consensus can be reached that they are too dangerous. There are encouraging examples: Chemical weapons, landmines, cluster munitions – robot warriors should be next.”* [405]

2.3 The Driverless Car

Since 2008, the search engine manufacturer Google has been working on the driverless car. While the AI driver was limited to motorway journeys at the beginning of the project, the Google cars have covered over 100,000 miles in city traffic and on country roads in the last year and a half – without any incidents worth mentioning. With eyes closed, there is said to be no discernible difference between the car and a human driver: The car accelerates and brakes smoothly, changes lanes regularly, stops for pedestrians, avoids cyclists and follows changed traffic layouts at road works. During a 45-minute test drive in May 2014, the Google technician, who is still behind the wheel for safety reasons, did not intervene once – so the car is already driving completely autonomously.[406]

The head of the *Google Self-Driving Car Project,* Chris Urmson, cites traffic fatalities as one of the main reasons for the development of autonomous driving systems. Currently, 32,000 people die in traffic accidents in the USA alone each year, and the figure is 1.2 million worldwide. More than 90 percent of accidents are due to human error. Artificial intelligence, on the other hand, is attentive at all times and never distracted – for this reason alone, a software driver would cause far fewer errors and accidents than a human driver.[407]

Nevertheless, even with highly developed software, some accidents cannot be avoided because also software cannot suspend the laws of physics. A pedestrian suddenly running onto the road, the sudden braking of another road user or sudden deer crossing are just a few examples of how even the minimal reaction time of control software will not be able to completely prevent accidents. Legally, this raises the question of liability: if

[405] Ladurner, Wenn Roboter töten, translation: MW.

[406] Cf. Schulz, Testfahrt in Google Self-Driving Car.

[407] Cf. Google Self-Driving Car Project, Behind the Google Self-Driving Car Project.

an AI driver causes an accident, who is at fault in the legal sense – the software or the programmer of the software? According to current law, a human being must be able to control a moving car at all times – here, technology is years ahead of jurisprudence.[408]

An ethical problem is posed by accident situations in which a collision with one of two objects is unavoidable. A human driver would usually decide instinctively and rather randomly in which direction to turn the steering wheel. Software, however, would be able to make a hazard assessment within fractions of a second as to which collision would cause the least damage. If the software had to choose between two cars, for example a heavy SUV with front and side airbags and a light small car without safety equipment, the software could decide to collide with the SUV because it absorbs more energy in the collision than the small car and its driver would also be better protected by the airbag than the driver without an airbag. The situation would be similar if the software were allowed to decide between two cyclists, only one of whom was wearing a helmet: Statistically speaking, the cyclist with a helmet would have a greater chance of survival than the one without, which is why a collision with the helmet-wearer would make more sense from the point of view of "accident optimization".[409]

The problem of such a consideration is obvious: while a human driver cannot consciously make such a consideration in the short time available, the driver AI could make possible decisions about life and death in an accident situation based on statistical criteria. However, this would penalise precisely those road users who behave responsibly, for example by driving a safe car or wearing a helmet.[410]

This problem is a modern variant of the *trolley dilemma*, that describes the dilemma of a tram driver who notices that his brake is defective. There are five people on the track in front of him who cannot get to safety in time. A switch before the critical section of track leads to the right, but there is another person on the deviating track. Should the tram driver now decide to actively throw the track switch and thus kill one person, or not do anything and thus kill five people?[411]

In the sense of utilitarian ethics, which does not evaluate actions but only consequences, the tram driver would have to decide to change the switch and thus accept the killing of one person in order to save the lives of five people. Deontological ethics, on the other

[408] Cf. Markoff, Google Cars Drive Themselves, in Traffic.

[409] Cf. Lin, The Robot Car of Tomorrow May Just Be Programmed to Hit You.

[410] Cf. ibid.

[411] Cf. Thomson, Killing, Letting Die, and the Trolley Problem, 206.

hand, assesses actions in themselves, regardless of their consequences. Depending on the deontological interpretation, therefore, throwing the switch could be seen as an intrinsically bad action because it actively causes the death of a human being. In general, deontological ethics assumes that the negative duty not to kill outweighs the positive duty to save lives. In the case of the trolley dilemma, opinions about the right action therefore diverge depending on whether one sees the sparing of the five as a negative or positive duty. In the first case, the negative duty to avoid killing five would outweigh the negative duty to avoid killing one – so the tram driver would have to change the switch. In the second case, the positive duty to spare five lives would not outweigh the negative duty to avoid one killing – so the tram driver would not be allowed to throw the switch.[412]

Software-controlled driver systems could – as described above – be equipped with a kind of utilitarian ethics for minimising harm. The realisation of a deontological ethic is unlikely to be technically feasible because software cannot take on any qualitative evaluation of its action.

The extent to which such a utilitarian ethic makes sense if it leads to a penalisation of responsible road users and could ultimately contribute to road users foregoing protective measures must be discussed. It becomes clear that even a seemingly indisputable application of artificial intelligence such as the autonomous control of a car can have ethical implications that are difficult to resolve.

[412] Cf. ibid., 206f.

V. Conclusion

Artificial intelligence poses a challenge to both theological systematics and theological practice. In view of the rapid technological progress, the school of strong artificial intelligence gains apparent plausibility. Theology has to address the strong AI conception of the world and of man, which is based on naturalistic empiricism, has strong dualistic tendencies and reduces man to a machine – designed rather badly than well by evolution – and present an alternative conception of man.

As shown with Hans-Dieter Mutschler, technological progress has been accompanied by religious promises since the industrial revolution. However, in times of erosion of traditional religious convictions, the idea of the singularity represents an extensive substitute for religion that realises many aspects of traditional religions with the promise of eternal life as well as the overcoming of diseases and physical weaknesses, the image of God as a cosmos-spanning artificial intelligence and the belief in exponential progress as a sacred element.

Protagonists of strong AI such as Kurzweil and Moravec see their theses as scientifically sound, although much of their work has more the ring of science fiction than science. Similarly, they do not realise how often they postulate religious-style beliefs without being able to support them scientifically: Kurzweil, for example, when he presupposes that a perfectly simulated consciousness is indeed a consciousness without being able to explain how the leap from simulation to consciousness is supposed to be made; Moravec, when he asserts the equivalence of simulation and reality without being able to make this plausible. Although both claim to be scientific, they become extremely unscientific in many places, for example when they try to infer the overall performance of the brain by merely extrapolating a small part of the human neural network, the retina.

The Judeo-Christian tradition explicitly warns against human hubris – already the first book of the Old Testament bears witness to this in the story of the Tower of Babel. One task of the theology of our day must be to expose as such the hubris of artificial intelligence, which demands a new creation of man and thus puts itself in the place of God. A science that becomes a substitute religion dehumanises man and no longer serves him.

Joseph Weizenbaum showed that in view of the omnipresence of modern technology, people are inclined to believe its promises uncritically. Human beings willingly allow themselves to be deceived by systems like ELIZA and trust in the computer's power of judgement, whose actual susceptibility to error can no longer be verified in view of its growing complexity. The technologisation of our world also contributes to the fact that man compares himself more and more with technology. However, a comparison is only meaningful if both comparison variables are known. Humans cannot be described in purely scientific, empirical definitions. Therefore, such a comparison inevitably leads to a dangerous reduction of the image of man.

The mind-brain debate shows that serious science cannot simply "explain away" consciousness phenomena and human freedom. Neither epiphenomenalism nor naturalistic reductionism can convincingly explain the existence of qualia. Their attempts to present human freedom as an illusion remain inconclusive. In this respect, it is not wrong to assume, in accordance with one's own intuition and the Christian image of man, that human freedom is real.

The strong AI image of man contradicts the Christian image of man on all levels. Lacking a theistic foundation, it denies that human beings are made in the image of God, contradicts the unity of body and soul, postulates a new dualism and does not take human freedom and sin into account.

However, this work would be misunderstood if it were read as a mere warning against technological progress. Progress, also in artificial intelligence, is to be supported as long as it puts people at the centre and serves them. Artificial intelligence already represents a great benefit in many areas of our everyday lives – research into driverless cars or labour-saving household robots, to name just two examples, will continue this development.

What is dangerous is a conception of artificial intelligence that sees robots as a substitute for humans or even as "better humans". Artificial intelligence will not be able to replace humans in their wholeness, their freedom and their social integration in the foreseeable future – and in all probability not at all.

List of Abbreviations

AA.VV.	Various authors
AI	Artificial Intelligence
AAS	Acta Apostolicae Sedis
CCC	Catechism of the Catholic Church
cps	Calculations per Second
DH	Denzinger, Heinrich / Hünermann, Peter (Hg.), Enchiridion symbolorum definitionum et declarationum de rebus fidei et morum, Freiburg i. Br. ^{43}2010.
FAZ	Frankfurter Allgemeine Zeitung
GS	Vatican II Pastoral Constitution „Gaudium et Spes“
LThK	Lexikon für Theologie und Kirche
MIPS	Million Instructions per Second
MIT	Massachusetts Institute of Technolgy
NZZ	Neue Zürcher Zeitung
RGG	Religion in Geschichte und Gegenwart
Vat. II	Second Vatical Council

List of Sources and Literature

Sources

Catechism of the Catholic Church, New York 1995, quoted as: CCC.

Denzinger, Heinrich / Hünermann, Peter (ed.), Enchiridion symbolorum definitionum et declarationum de rebus fidei et morum, Freiburg i. Br. 43 2010. Quoted als: DH.

Google Self-Driving Car Project, Behind the Google Self-Driving Car Project. Video, 27 May 2014, URL: https://www.youtube.com/watch?v=cdeXlrq-tNw (visited 20 August 2023).

Holy Bible. New Revised Standard Version. Catholic Edition, Nashville 2008.

Kurzweil, Ray, Curriculum Vitae, URL: https://web.archive.org/web/20140701122347/http://www.kurzweilai.net/ray-kurzweil-curriculum-vitae (visited 19 August 2023).

Idem, Press release from 14 December 2012, URL: https://web.archive.org/web/20230307092257/https://www.kurzweilai.net/kurzweil-joins-google-to-work-on-new-projects-involving-machine-learning-and-language-processing (visited 19 August 2023).

News article „EU wählt zwei Projekte der Spitzenforschung aus", in: Frankfurter Allgemeine Zeitung 65 (2013) No. 24, 29 January 2013, 11.

News article „Computerprogramm ‚Eugene' besteht Turing-Test", URL: https://www.heise.de/newsticker/meldung/Computerprogramm-Eugene-besteht-Turing-Test-2217857.html (visited 19 August 2023).

News article „Google kauft zum Jahresende Militärroboter-Hersteller", URL: https://www.golem.de/news/boston-dynamics-google-kauft-zum-jahresende-militaerroboter-hersteller-1312-103387.html (visited 19 August 2023).

News article „Googles autonome Autos unterwegs in der Stadt", URL: http://www.heise.de/ix/meldung/Googles-autonome-Autos-unterwegs-in-der-Stadt-2178981.html (visited 19 August 2023).

Moravec, Hans P., Curriculum Vitae, URL: https://web.archive.org/web/20140724023818/https://frc.ri.cmu.edu/~hpm/hpm.cv.html (visited 19 August 2023).

Pope Benedict XVI, Address on 12 September 2006, in: Acta Apostolicae Sedis 98 (2006), 728-739.

Second Vatican Council, Pastoral Constitution „Gaudium et Spes“, in: AAS 58 (1966), 1025-1120 (English version URL: https://www.vatican.va/archive/hist_councils/ii_vatican_council/documents/vat-ii_cons_19651207_gaudium-et-spes_en.html, visited 19 August 2023). Quoted as: Vat. II, GS.

Syriac Didascalia, in: Vööbus, Arthur (ed.), The Didascalia Apostolorum in Syriac. Vol. 2 (= Corpus Scriptorum Christianorum Orientalium, 408), Leuven 1979.

Literature

AA.VV., Japanische Kuschelroboter für Demenzkranke, in: Die Welt (online only), URL: http://www.welt.de/gesundheit/article3071134/Japanische-Kuschelroboter-fuer-Demenzkranke.html (visited 20 August 2023).

Adams, Douglas, The Hitch Hiker’s Guide to the Galaxy. A Trilogy in Four Parts, Basingstoke 1986.

Berger, Klaus, Brauchen wir eine Theologie der Roboter? Interview on 21 June 2000, in: FAZ 52 (2000) No. 142, 21 June 2000, 55.

Boden, Margaret A., Künstliche Intelligenz und Menschenbilder, in: Schefe, Peter u.a. (ed.), Informatik und Philosophie, Mannheim 1993, 235-244.

Böhme, Hartmut, Die technische Form Gottes. Über die theologischen Implikationen von Cyberspace, in: Neue Zürcher Zeitung 217 (1996) No. 86, 13/14 April 1996, 53.

Borchers, Detlef / Ziegler, Peter-Michael, Der letzte Service: zum Tode von Joseph Weizenbaum, URL: https://www.heise.de/newsticker/meldung/Der-letzte-Service-zum-Tode-von-Joseph-Weizenbaum-188114.html (visited 20 August 2023).

Brandt-Herrmann, Gila, Typische Biographien untypischer Informatiker. Bildungsprozesse in Berufsbiographien von Informatikern (= Internationale Hochschulschriften, 500), Münster 2008.

Bürkle, Horst, Religion. III. Religionswissenschaftlich, in: Kasper, Walter u.a. (ed.), Lexikon für Theologie und Kirche^3, Vol. 8 Pearson–Samuel, Freiburg i. Br. 1999, 1039-1041.

Capurro, Rafael, Leben im Informationszeitalter, Berlin 1995.

Dworschak, Manfred, Mitarbeiterin der Woche, in: Die Zeit 54 (1999) No. 39, 23 September 1999, BL1.

Evers, Dirk, Der Mensch als Turing-Maschine? Die Frage nach der künstlichen Intelligenz in philosophischer und theologischer Perspektive, in: Neue Zeitschrift für Systematische Theologie und Religionsphilosophie 47 (2005) 101-118.

Fehr, Benedikt, Der vollautomatische Börsencrash, in: FAZ 60 (2008) No. 48, 26 February 2008, 23.

Feil, Ernst, Religion I, in: Betz, Hans Dieter et al (ed.), Religion in Geschichte und Gegenwart^4. Vol. 7 R–S, Tübingen 2004, 263-267.

Fisch, Florian, Der Griff nach dem Bewusstsein. Im Human Brain Project wird eine Software entwickelt, die psychische Krankheiten simulieren soll, in: NZZ 232 (2011) No. 109, 11 May 2011, 60.

Foerst, Anne, Von Robotern, Mensch und Gott. Künstliche Intelligenz und die existenzielle Dimension des Lebens, Göttingen 2008.

Freeman, Walter J., Intentionality, in: Scholarpedia 2(2):1337, Revision 123821.

Fuchs, Thomas, Das Gehirn – ein Beziehungsorgan. Eine phänomenologisch-ökologische Konzeption, Stuttgart ^{2}2009.

Ganoczy, Alexandre, Dualismus. IV. Systematisch-theologisch, in: LThK^3, Vol. 3 Dämon–Fragmentenstreit, Freiburg i. Br. 1995, 391f.

Grolle, Johann, Aufruf zur Verschwendung, in: Der Spiegel 67 (2013) No. 6, 4 February 2013, 104.

Herms, Eilert, Künstliche Intelligenz. Wesen und sozialethische Probleme aus Sicht des christlichen Menschebildes, in: Idem (ed.), Gesellschaft gestalten. Beiträge zur evangelischen Sozialethik, Tübingen 1999, 285-295.

Hutter, Manfred, Dualismus. I. Religionsgeschichtlich, in: LThK^3, Vol. 3 Dämon–Fragmentenstreit, Freiburg i. Br. 1995, 387f.

Janssen, Jan-Keno, Warum Glass (noch) nicht funktioniert. Ernüchternde Langzeiterfahrungen mit Google Glass, in: c't 31 (2013) No. 15, 76f.

Knapp, Alex, Ray Kurzweil's Predictions For 2009 Were Mostly Inaccurate, URL: http://www.forbes.com/sites/alexknapp/2012/03/20/ray-kurzweils-predictions-for-2009-were-mostly-inaccurate/ (visited 20 August 2023).

Kolb, Matthias, Wie Obama lernte, die Drohnen zu lieben, in: Süddeutsche Zeitung (online online), URL: http://www.sueddeutsche.de/politik/amerikas-kriege-wie-obama-lernte-die-drohnen-zu-lieben-1.1795640 (visited 20 August 2023).

Kornwachs, Klaus, Prothese, Diener, Ebenbild. Warum sollen wir denkende Maschinen bauen? In: Herder Korrespondenz 56 (2002) 402-407.

Kurzweil, Ray, Interview from December 2012 („Wie baut man ein Gehirn, Herr Kurzweil?"), in: Technology Review dt. 10 (2012) No. 12, 56f.

Idem, The Age of Spiritual Machines. When Computers Exceed Human Intelligence, New York 1999.

Idem, How to Create a Mind. The Secret of Human Thought Revealed, New York 2013.

Idem, The Singularity Is Near. When Humans Transcend Biology, New York 2005.

Ladurner, Ulrich, Wenn Roboter töten, in: Die Zeit 68 (2013) No. 3, 10 January 2013, 10.

Lem, Stanslaw, Golem XIV, in: Idem, Imaginary Magnitude, San Diego 1984, 97-248.

Lin, Patrick, The Robot Car of Tomorrow May Just Be Programmed to Hit You, URL: https://www.wired.com/2014/05/the-robot-car-of-tomorrow-might-just-be-programmed-to-hit-you/ (visited 21 August 2023).

Loos, Andreas, Pionier und Pazifist. Zum Tode des Informatikers und Computerkritikers Joseph Weizenbaum, in: Jüdische Allgemeine 63 (2008) No. 11, 13 March 2008, 13.

Manemann, Jürgen, Paro ist ein Ausdruck für den Verlust von Kommunikation, URL: https://www.wissenschaftsjahr.de/2013/die-themen/themen-dossiers/besser-leben-mit-technik/contra-position.html (visited 20 August 2023).

MacNeil, Ted, Don't Be Misled by MIPS, in: TechChannel, URL: https://techchannel.com/11/2004/dont-be-misled-by-mips (visited 19 August 2023).

Manzeschke, Arne, Stumme Roboter. Interview on 3 January 2013, in: Die Zeit 68 (2013) No. 2, 3 January 2013, 28.

Markoff, John, Google Cars Drive Themselves, in Traffic, in: New York Times 160 (2010) n. N., 10 October 2010, A1.

Markram, Henry et al, Introducing the Human Brain Project, in: Procedia Computer Science 7 (2011), 39-42.

McCorduck, Pamela, Denkmaschinen. Die Geschichte der künstlichen Intelligenz, Haar 1987.

McGinn, Colin, Hello, HAL. Three books examine the future of artificial intelligence and find the human brain is in trouble, in: New York Times 149 (1999) n. N., 3 Januar 1999, Section 7, 11f.

Meijas, Jordan, Unsterblichkeit für alle, in: FAZ 63 (2011) No. 181, 6 August 2011, 31.

Monya, Hannah et al., Das Manifest. Elf führende Neurowissenschaftler über Gegenwart und Zukunft der Hirnforschung, in: Gehirn & Geist 3 (2004) No. 6, 30-37.

Moravec, Hans P., Robot. Mere Machine to Transcendent Mind, New York 1999.

Mutschler, Hans-Dieter, Die Gottmaschine. Das Schicksal Gottes im Zeitalter der Technik, Augsburg 1998.

Idem, Ist der Mensch ein Roboter? In: Koßler, Matthias / Zecher, Reinhard (ed.), Von der Perspektive der Philosophie. Beiträge zur Bestimmung eines philosophischen Standpunkts in einer von den Naturwissenschaften geprägten Zeit (= Schriftenreihe Boethiana, 56), Hamburg 2002, 291-308.

Nadella, Satya / Pall, Gurdeep Singh, Presentation on 27 May 2014 at the Inaugural Code Conference, URL: https://www.vox.com/2014/5/27/11627276/microsofts-skype-star-trek-language-translator-takes-on-tower-of-babel (visited 19 August 2023).

Nagel, Thomas, What Is It Like to Be a Bat? In: The Philosophical Review 83 (1974), 435-450.

Oppy, Graham / Dowe, David, The Turing Test, in: Zalta, Edward N. (ed.), The Stanford Encyclopedia of Philosophy. Spring 2011 Edition, URL: http://plato.stanford.edu/archives/spr2011/entries/turing-test/ (visited 19 August 2023).

Pröpper, Thomas, Theologische Anthropologie. Zweiter Teilband, Freiburg i. Br. 2011.

Ratzinger, Joseph, Einführung in das Christentum. Vorlesungen über das Apostolische Glaubensbekenntnis, München 2000.

Reintjes, Thomas, Drei Finger reichen den Saft. Therapie-, Pflege- und Service-Roboter sollen Pflegebedürftige unterstützen – sinnvolle Hilfe oder ein weiterer Schritt zur Vereinsamung? In: FAZ 63 (2011) No. 165, 19 July 2011, T1.

Rieger, Frank, Das Gesicht unserer Gegner von morgen. Von der Zukunft des Krieges im Zeitalter der Maschinen, in: FAZ 64 (2012) No. 221, 21 September 2012, 31, 33.

Rojas, Raúl, Analoge versus Digitale Seele. Ray Kurzweil, der Tod und die Singularität, in: Rötzer, Florian (ed.), Können Roboter lügen? Essays zur Robotik und Künstlichen Intelligenz (E-Book), Hannover 2013.

Idem, Die Angst des Roboters beim Elfmeter, in: Ibid.

Idem, IBM vs. Blue Brain: Wettlauf der Himmelstürmer, in: Ibid.

Idem, Warum „Watson“ ein Durchbruch ist, in: Ibid.

Schanze, Jens, Film „Plug & Pray. Von Computern und anderen Menschen“, farbfilm verleih 2010.

Schult, Thomas J., Crazy Hans: Ein Portrait des vielgehaßten Roboterforschers Hans Moravec, in: Die Zeit 51 (1996) No. 27, 28 June 1996, 62.

Schulz, Sandra / Barth, Theodor, Plüsch-Tech für Senioren. Paro, der Glücklichmach-Roboter, in: Spiegel Online, URL: http://www.spiegel.de/panorama/gesellschaft/pluesch-tech-fuer-senioren-paro-der-gluecklichmach-roboter-a-443593.html (visited 21 August 2023).

Schulz, Thomas, Testfahrt in Google Self-Driving Car. Dieses Auto kommt ohne Sie aus, in: Spiegel Online, URL: https://www.spiegel.de/auto/aktuell/google-auto-unterwegs-im-selbstfahrenden-auto-a-969532.html (abgerufen am 15. Juni 2014).

Searle, John R., Chinese room argument, in: Scholarpedia 4(8):3100, Revision 66188.

Idem, Ist das Gehirn ein Digitalcomputer? In: Schefe, Peter u.a. (ed.), Informatik und Philosophie, Mannheim 1993, 211-232.

Spaemann, Robert, Personen. Versuche über den Unterschied zwischen „etwas“ und „jemand“, Stuttgart 1996.

Idem, Schritte über uns hinaus. Gesammelte Reden und Aufsätze II, Stuttgart 2011.

Teilhard de Chardin, Pierre, Der Mensch im Kosmos, München ^{7}1964.

Thomson, Judith Jarvis, Killing, Letting Die, and the Trolley Problem, in: The Monist 59 (1976) 204-217.

Weizenbaum, Joseph, Albtraum Computer. Ist das menschliche Gehirn nur eine Maschine aus Fleisch? In: Die Zeit 27 (1972) No. 3, 21 January 1972, 43.

Idem, Computer Power and Human Reason. From Judgment to Calculation, San Francisco 1976.

Idem, Wir gegen die Gier. Die Erde könnte ein Paradies sein – wenn wir sie nur richtig deuten würden, in: Süddeutsche Zeitung 64 (2008) No. 6, 8 January 2008, 13.

Idem, Wo sind sie, die Inseln der Vernunft im Cyberstrom? Auswege aus der programmierten Gesellschaft (mit Gunna Wendt), Freiburg i. Br. 2006.

Welchering, Peter, Preisgestaltung im Millisekundentakt, in: FAZ 63 (2011) No. 201, 30 August 2011, T2.

Wertheim, Margaret, Ehre sei Gott im Cyberspace. Virtuelle Welten im Mittelalter und im Internet, in: Die Zeit 51 (1996) No. 22, 24 May 1996, 31.

Wetz, Franz Josef, Dualismus. II. Philosophisch, in: LThK³, Vol. 3 Dämon–Fragmentenstreit, Freiburg i. Br. 1995, 388f.

Wright, Robert, Did The Universe Just Happen? In: Atlantic Monthly 261 (1988) No. 4 (April), 29-44.

Zilles, Karl, Hirnforschung widerlegt nicht Freiheit. Libet-Experiment mißt keine Willensentscheidung. Presentation on 17 November 2004 at a conference of Wissenschaftszentrum NRW in Düsseldorf, URL: http://web.archive.org/web/20050330162252/information-philosophie.de/philosophie/neurophilosophie5.html (accessed 19 August 2023).

Zirker, Hans, Religion. I. Begriff, in: LThK³, Vol. 8 Pearson–Samuel, Freiburg i. Br. 1999, 1034-1036.

Zsifkovits, Valentin, Das Menschenbild der christlichen Theologie, in: Jahrbuch für Christliche Sozialwissenschaften 22 (1981), 13-22.

Zuse, Konrad, Rechnender Raum, in: Elektronische Datenverarbeitung 8 (1967), 336-344.