Effective Altruism Foundation

Foundational Research Institute

Artificial Intelligence: Opportunities and Risks

Policy paper

Artificial intelligence (AI) and increasingly complex algorithms currently influence our lives and our civilization more than ever. The areas of AI application are diverse and the possibilities extensive: in particular, because of improvements in computer hardware, certain AI algorithms already surpass the capacities of human experts today. As AI capacity improves, its field of application will grow further. In concrete terms, it is likely that the relevant algorithms will start optimizing themselves to an ever greater degree—maybe even reaching superhuman levels of intelligence. This technological progress is likely to present us with historically unprecedented ethical challenges. Many experts believe that alongside global opportunities, AI poses global risks, which will be greater than, say, the risks of nuclear technology—which in any case have historically been underestimated. Furthermore, scientific risk analysis suggests that high potential damages should be taken very seriously even if the probability of their occurrence were low.

12 December 2015

Policy paper by the Effective Altruism Foundation.

Preferred citation: Mannino, A., Althaus, D., Erhardt, J., Gloor, L., Hutter, A. and Metzinger, T. (2015). Artificial Intelligence: Opportunities and Risks. Policy paper by the Effective Altruism Foundation (2): 1-16.

First published (in German): 12 December 2015.

www.foundational-research.org
www.ea-stiftung.org

Contents

Adriano Mannino, Philosopher & Co-President, Effective Altruism Foundation

David Althaus, Assistant Director, Foundational Research Institute

Dr. Jonathan Erhardt, Scientific consultant, Effective Altruism Foundation

Lukas Gloor, Researcher, Foundational Research Institute

Dr. Adrian Hutter, Physics Department, University of Basel

Prof. Thomas Metzinger, Professor of Philosophy, University of Mainz

Artificial Intelligence: Opportunities and Risks

Executive Summary

Artificial intelligence (AI) and increasingly complex algorithms currently influence our lives and our civilization more than ever before. The areas of AI application are diverse and the possibilities far-reaching, and thanks to recent improvements in computer hardware, certain AI algorithms already surpass the capacities of today's human experts. As AI capacity improves, its field of application will continue to grow. In concrete terms, it is likely that the relevant algorithms will start optimizing themselves to an ever greater degree and may one day attain superhuman levels of intelligence. This technological progress is likely to present us with historically unprecedented ethical challenges. Many experts believe that, alongside global opportunities, AI poses global risks surpassing those of e.g. nuclear technology (whose risks were severely underestimated prior to their development). Furthermore, scientific risk analyses suggest that high potential damages resulting from AI should be taken very seriously—even if the probability of their occurrence were low.

Current

In narrow, well-tested areas of application, such as driverless cars and certain areas of medical diagnostics, the superiority of AIs over humans is already established. An increased use of technology in these areas offers great potential, including fewer road traffic accidents, fewer mistakes in the medical treatment and diagnosing of patients, and the discovery of many new therapies and pharmaceuticals. In complex systems where several algorithms interact at high speed (such as in the financial market or in foreseeable military uses), there is a heightened risk that new AI technologies will be misused, or will experience unexpected systematic failures. There is also the threat of an arms race in which the safety of technological developments is sacrificed in favor of rapid progress. In any case, it is crucial to know which goals or ethical values ought to be programmed into AI algorithms and to have a technical guarantee that the goals remain stable and resistant to manipulation. With driverless cars, for instance, there is the well-known question of how the algorithm should act if a collision with several pedestrians can only be avoided by endangering the passenger(s), not to mention how it can be ensured that the algorithms of driverless cars are not at risk of hacking systematic failure.

Measure 1 The promotion of a factual, rational discourse is essential so that cultural prejudices can be dismantled and the most pressing questions of safety can be focused upon.

Measure 2 Legal frameworks must be adapted so as to include the risks and potential of new technologies. AI manufacturers should be required to invest more in the safety and reliability of technologies, and principles like predictability, transparency, and non-manipulability should be enforced, so that the risk of (and potential damage from) unexpected catastrophes can be minimized.

Mid-term

Progress in AI research makes it possible to replace increasing amounts of human jobs with machines. Many economists assume that this increasing automation could lead to a massive increase in unemployment within even the next 10-20 years. It should be noted that while similar predictions in the past have proved inaccurate, the developments discussed here are of a new kind, and it would be irresponsible to ignore the possibility that these predictions come true at some point. Through progressive automation, the global statistical average living standard will rise; however, there is no guarantee that all people—or even a majority of people—will benefit from this.

Measure 3 Can we as a society deal with the consequences of AI automation in a sensible way? Are our current social systems sufficiently prepared for a future wherein the human workforce increasingly gives way to machines? These questions must be clarified in detail. If need be, proactive measures should be taken to cushion negative developments or to render them more positive. Proposals like an unconditional basic income or a negative income tax are worth examining as possible ways to ensure a fair distribution of the profits from increased productivity.

Long-term

Many AI experts consider it plausible that this century will witness the creation of AIs whose intelligence surpasses that of humans in all respects. The goals of such AIs could in principle take on any possible form (of which human ethical goals represent only a tiny proportion) and would influence the future of our planet decisively in ways that could pose an existential risk to humanity. Our species only dominates Earth (and, for better or worse, all other species inhabiting it) because it currently has the highest level of intelligence. But it is plausible that by the end of the century, AIs will be developed whose intelligence compares to ours as ours currently compares to, say, chimpanzees. Moreover, the possibility cannot be excluded that AIs also develop phenomenal states—i.e. (self-)consciousness, and in particular subjective preferences and the capacity for suffering—in the future,which would confront us with new kinds of ethical challenges. In view of the immediate relevance of the problem and its longer-term implications, considerations of AI safety are currently highly underrepresented in politics as well as research.

Measure 4 It is worth developing institutional measures to promote safety, for example by granting research funding to projects which concentrate on the analysis and prevention of risks in AI development. Politicians must, in general, allocate more resources towards the ethical development of future-shaping technologies.

Measure 5 Efforts towards international research collaboration (analogous to CERN's role in particle physics) are to be encouraged. International coordination is particularly essential in the field of AI because it also minimizes the risk of a technological arms race. A ban on all risky AI research would not be practicable, as it would lead to a rapid and dangerous relocation of research to countries with lower safety standards.

Measure 6 Certain AI systems are likely to have the capacity to suffer, particularly neuromorphic ones as they are structured analogously to the human brain. Research projects that develop or test such AIs should be placed under the supervision of ethical commissions (analogous to animal research commissions).

Introduction

The pursuit of knowledge runs as a governing principle through human history. Whenever societies have undergone significant changes in their dynamics and structure, this has normally been the result of new technological inventions. Around two million years separate the first use of stone tools from the historic moment when Homo sapiens invented art and began to paint images on cave walls. Another thirty thousand years passed before the rise of arable farming and permanent settlement. The first symbols appeared a few thousand years after that, followed closely by the first written scripts. Then, around four hundred years ago, development began speeding up. The microscope was invented in the seventeenth century; industrialization in the nineteenth century enabled the first cities of a million people; and during the last century alone, the atom was split, humans set foot on the Moon, and the computer was invented. Since then, the processing capabilities and energy efficiency of computers have doubled at regular intervals [1]. But while technological progress often develops exponentially, the same is not true for human intellectual abilities.

In recent years, countless renowned scientists and entrepreneurs have warned of the urgent significance of AI, and how important it is that policy makers tackle the challenges raised by AI research [2]. Exponents of this movement for AI safety include Stuart Russell [3], Nick Bostrom [4], Stephen Hawking [5], Sam Harris [6], Max Tegmark [7], Elon Musk [8], Jann Tallinn [9] and Bill Gates [10].

In certain domain-specific areas, AIs have already reached or even overtaken human levels on several occasions. In 1997 the computer *Deep Blue* beat the reigning world champion Garry Kasparov at chess [11]; in 2011 *Watson* beat the two best human players on the language-based game show Jeopardy! [12]; and in 2015 the first variant of poker, *Fixed Limit Holdem heads-up*, was game theoretically fully solved by *Cepheus* [13]. Meanwhile, artificial neural networks can compete with human experts in the diagnosis of cancer cells [14] and are also more or less approaching human levels in the recognition of handwritten Chinese characters [15]. Back in 1994, a self-learning backgammon program reached the level of the world's best players by finding strategies that had never before been played by humans [16]. By now, there even exist algorithms that can independently learn many different games from scratch and thereby reach (or surpass) human levels [17, 18]. With these developments, we are slowly getting closer to a *general intelligence*, which at least in principle can solve problems of all sorts independently.

With great power comes great responsibility. Technology is in itself just a tool; what matters is how we use it. The use of existing AIs is already presenting us with considerable ethical challenges, which will be illuminated in the next section of this paper. The following chapter will outline developments in economic automation, and explain the mid-term prognosis that AI research will give rise to a significant restructuring of the labor market. Finally, the two last chapters will discuss the long-term and existential risks of AI research in relation to the possible creation of (super)human intelligence and artificial consciousness.

Advantages and risks of current AIs

Our individual lives and our civilization as a whole are governed to an ever-increasing extent by algorithms and domain-specific artificial intelligence (AIs) [19]. Well-known examples include such ubiquitous things as smartphones, air traffic control systems [20] and internet search engines [21]. Financial markets, too, are dependent on algorithms which are too large and complex for any single human being to fully understand [22, 23]. The operation of such algorithms, for the most part, proceed without incident, but there is always the possibility that an unlikely “black swan” event [24] might occur, threaten to plunge the whole system into chaos. We have already witnessed one such event: in 2010, an unexpected "flash crash" in a US stock market left the financial world dumbfounded. The crash occurred as a result of computer algorithms interacting with the financial market in an unforeseen manner [25, 26]. Within minutes, important shares lost more than 90% of their worth and then quickly returned to their high initial value. If such an event were to take place in a military context, a comparable “return to initial conditions” would be improbable [27]. To prevent devastating failures of this sort, it seems generally advisable to invest considerably more resources into the safety and reliability of AIs. Unfortunately, current economic incentives seem to favor increased AI capacity far more than safety.

Four criteria for the construction of AIs

Safety is essential to the construction of any sort of machine. However, new ethical challenges arise when constructing domain-specific AIs capable of taking over cognitive work in social dimensions—work that, until now, has been carried out by humans. For instance, an algorithm that judges the credit rating of bank customers might make decisions that discriminate against certain groups in the population (without this being explicitly programmed). Even technologies that simply replace existing actions could introduce interesting challenges for machine ethics [28]: driverless cars, for instance, raise the question of which criteria should be decisive in the case of an imminent accident. Should the vehicle ensure the survival of the passengers above all else or should it, in the case of an unavoidable accident, prioritize keeping the total number of casualties as low as possible [29]?

Because of this, both AI theorist Eliezer Yudkowsky and philosopher Nick Bostrom have suggested four principles which should guide the construction of new AIs [30]: 1) the functioning of an AI should be *comprehensible* and 2) its actions should be *basically predictable*. Both of these criteria must be met within a time frame that enables the responsible experts to react in time and veto control in case of a possible failure. In addition, 3) AIs should be *impervious to manipulation*, and in case an accident still occurs, 4) the responsibilities should be clearly determined.

Advantages of (domain specific) artificial intelligence

In principle, algorithms and domain-specific AIs bring many advantages. They have influenced our lives for the better and are expected to keep doing so at an ever-increasing rate in the future, provided that the necessary precautions are taken. Here we will discuss two instructive examples.

Driverless cars are no longer science fiction [31, 32]; they'll be commercially available in the foreseeable future. The *Google Driverless Car*, which is driven completely by autonomous AI algorithms, took its first test drive in the USA back in 2011 [33, 34]. Besides the time gained for work or relaxation, a second advantage to driverless cars consists in their higher safety. In 2010, 1.24 million people died worldwide in traffic accidents, nearly exclusively because of human error [35]. Countless human lives could therefore be saved every year, because driverless cars are already significantly safer than vehicles driven by humans [36, 37].

Naturally, a large number of people remain skeptical regarding driverless cars, mainly because they underestimate the safety benefits thereof whilst at the same time overestimating their own driving abilities. As an illustration of this latter point, one study came to the conclusion that 93% of all American drivers believe that their driving abilities are above the median [38]—which is statistically impossible. Unrealistic optimism [39] and the illusion of control [40] possibly also bias people towards underestimating the risks when they themselves are behind the wheel [41, 42].

Doctors, too, overestimate their abilities [43], which in the worst case can lead to deadly mishaps. In the USA alone, between an estimated 44,000 and 98,000 people die each year in hospitals because of treatment mistakes [44]. In this context, IBM's *Watson* [45] is a welcome development. This AI gained fame in 2011 when it beat the best human players on the quiz show *Jeopardy!* [12]. *Watson* isn't just better than humans in quiz shows, however. Hospitals have been able to hire Watson's computing power since 2014 for cancer diagnosis and other complex pattern-recognition tasks. Because "Doctor Watson" can rapidly collect and combine enormous quantities of information, it has partially overtaken the diagnostic skills of its human colleagues [46, 47].

The fact that a current AI can make more accurate medical diagnoses than human doctors may seem surprising at first, but it has long been recognized that *statistical inferences* are superior to clinical judgments by human experts in most cases [48, 49]. Seeing as AIs like *Watson* are ideal for making statistical inferences, it follows that using computers for certain types of diagnosis can save lives.

Cognitive biases: to err is human

One reason why human experts are less competent than AIs at statistical inferences is the aforementioned (and, unfortunately, all too human) tendency to overestimate one's own abilities. This tendency is known as *overconfidence bias* [50] and is just one of many documented cognitive biases that can lead to systematic errors in human thinking [51, 52]. AIs, on the other hand, can be built so as to avoid cognitive biases altogether. In principle, increasing confidence in the predictions of AIs could lead to a significantly more rational and efficient approach to many social and political challenges, provided they are made safely and according to comprehensible criteria. The problem here lies in using the strengths of AI without at the same time giving up human autonomy in the corresponding systems.

Conclusion and outlook

Irrational fears towards new and basically advantageous technologies are widespread, both now and in the past [53]. Such "technophobia" may also be one of the reasons that Watson or driverless cars are met with skepticism. However, being wary of kinds of technology is not always irrational. Most technologies can be used to the benefit of humanity, but can also be dangerous when they fall into the wrong hands, or when insufficient care is taken for safety and unforeseen side effects.

This also holds for artificial intelligence: driverless cars could make our lives easier and save human lives, but complex computer algorithms can also cause the stock market to crash unexpectedly. While the risks from domain-specific AIs appear limited in the near future, there are long-term developments to take into consideration: in the not-so-distant future, artificial intelligence could in principle pose an existential threat, similar in scope to the pandemic risks associated with biotechnology [54, 55, 4].

> Recommendation 1 — **Responsible approach:** As with all other technologies, care should be taken to ensure that the (potential) advantages of AI research clearly outweigh the (potential) disadvantages. The promotion of a factual, rational discourse is essential so that irrational prejudices and fears can be broken down. Current legal frameworks have to be updated so as to accommodate the challenges posed by new technologies. The four principles described above should be followed for every extensive use of AIs [30]. ■

Automation and unemployment

In light of recent successes in the field of machine learning and robotics, it seems there is only a matter of time until even complicated jobs requiring high intelligence could be comprehensively taken over by machines [56].

If machines become quicker, more reliable and cheaper than human workers in many areas of work, this would likely cause the labour market to be uprooted on a scale not seen since the Industrial Revolution. According to economists like Cowen [57], McAfee and Brynjolfsson [58], technological progress will widen the income gap even further and may lead to falling incomes and rising unemployment in large segments of the population.

A 2013 analysis concluded that it will likely be possible to automate 47% of all jobs in the USA within 10-20 years [59]. The hardest jobs to automate are those which require high levels of social intelligence (e.g. PR consultation), creativity (e.g. fashion design) and/or sensitive and flexible object manipulation (e.g. surgery). In these domains, the state of AI research is still far below the level of human experts.

Advantages and disadvantages to automation by computers

Those who will benefit the most from technological progress are the people and nations that understand how to make use of new technological opportunities and the corresponding flood of "big data" [60]. In particular, countries with well-trained computer specialists are expected to prosper in the face of technological progress. Moreover, it is likely that a thorough understanding of the ways in which various computer algorithms compare to human decision-making and working abilities—as well as the (dis)advantages of each—will become increasingly important in the future, thus necessitating high standards of education [61].

Following the automation of the production and service industries, one might expect only the entertainment industry to remain; yet here, too, we are already witnessing extensive changes. With flawless computer graphics, novel entertainment technologies, and countless smartphone apps all becoming increasingly affordable, the addictive pull of videogames and internet usage is rising [62]. While we have not yet been able to research the long-term social and psychological consequences of this development, several factors currently indicate that these trends are profoundly changing our social behavior [63], attention spans, and childhood development [64]. These effects may be amplified by the increasing use of virtual reality technology, which is already available to consumers. As these become increasingly detailed and realistic, they may blur the user's boundaries between reality and simulation, thereby invading deeper into our everyday experience. The consequences of more regular immersion in virtual realities—including experiences like body-transfer illusions, in which subjective awareness is temporarily projected into a virtual avatar [65]—should receive greater attention.

While the entertainment industry does offer significant opportunities for better education through personalized AI teaching and the gamification of learning material [66], it also increases the risk that a growing proportion of young people will have trouble completing their education due to a pathological addiction to video games and/or the internet [67].

Utopias and dystopias

Technological progress increases societal productivity [68], in turn raising the average standard of living [69]. If more work is carried out by machines, this frees up time for leisure and self-development for humans—at least those in a position to profit from it. However, a drawback to increasing automation could be that the increases in productivity go along with increasing social inequality so that a rise in the *mean* standard of living doesn't coincide with a rise in the *median* quality of life. Experts like the MIT economics professor Erik Brynjolfsson even worry that technological progress threatens to make the lives of a majority of people worse [70].

In a competitive economy where AI technology has progressed to the point where many jobs are done by machines, the income for automatable human work will fall [58]. Without regulation, the incomes of many people could sink below subsistence level. Social inequality may rise sharply if economic output were to increase more rapidly than the wages needed to effect redistribution. To counteract this development, McAfee and Brynjolfsson suggest that limiting certain jobs to humans should be subsidized. Additional options for ensuring fair distribution of advantages from technological progress amongst the whole population include unconditional basic income, and a negative income tax [71, 72]

Some experts also warn of future scenarios in which the projected changes are even more drastic. For example, the economist Robin Hanson expects that it will be possible within this century to digitally run human brain simulations—so-called *whole brain emulations (WBEs)* [73]—in virtual reality. WBEs would be reproducible, and could (assuming that sufficient hardware is available) run many times faster than a biological brain, consequently implying a huge increase in labor efficiency [74]. Hanson predicts that in such a case, there would be a "population explosion" amongst WBEs, who could be used as enormously cost-efficient workers [75]. Hanson's speculations are contested [61], and it should not be assumed that they sketch out the most likely future scenario. Current research in this field, such as the Blue Brain Project at ETH Lausanne, is still very far from the first brain simulations—never mind supplying them in real time (or even faster) with inputs from a virtual reality. However, it is important to keep hardware developments in mind in relation to the possibility of WBEs. If the scenario sketched out by Hanson were to occur, this would be of great ethical relevance. For one thing, many humans replaced by complex simulations could become unemployed; for another, there is the question whether the WBEs deployed would have phenomenal consciousness and subjective preferences—in other words, whether they would experience suffering as a result of their (potentially forced) labor.

Recommendation 2 — Forward thinking: As in the case of climate change, incentives should be set for researchers and decision makers to deal with the consequences of AI research; only then can the foundations of precautionary measures be laid. In particular, specialist conferences should be held on AI safety and on assessing the consequences of AI, expert commissions should be formed, and research projects funded. ■

Recommendation 3 — Education: The subsidization of human work, an unconditional basic income, and a negative income tax have all been proposed as measures to cushion the negative social impacts of increased automation. Research should be conducted toward finding additional options, as well as identifying which set of measures has the maximum effect. Moreover, advantages and disadvantages must be systematically analyzed and discussed at a political level, and research grants should be established in order to answer any empirical questions that will inevitably arise as a result of this discussion. ■

Recommendation 4 — Transparency over new measures: The subsidisation of human work, an unconditional basic income or a negative income tax have been proposed as measures to cushion the negative social impacts of increasing automation. It is worth clarifying which further options exist and which set of measures has the maximum effect. In addition, advantages and disadvantages must be systematically analysed and discussed at a political level. Research grants should be established to answer the empirical questions thrown up by this discussion. ■

General intelligence and superintelligence

General intelligence measures an agent's ability to achieve goals in a wide range of environments [76, 77]. This kind of intelligence can pose a (catastrophic) risk if the goals of the agent do not align with our own. If a general intelligence reaches a superhuman level, it becomes a *superintelligence*; that is, an algorithm superior to human intelligence in every way, including scientific creativity, "common sense", and social competence. Note that this definition leaves open the question of whether or not a superintelligence would have consciousness [78, 79].

Comparative advantages of general artificial intelligence over humans

Humans are intelligent, two-legged "bio-robots" possessing a conscious self-awareness, and were developed over billions of years of evolution. These facts have been used argue that the creation of artificial intelligence may not be so difficult, [80, 81, 82] seeing as AI research can be conducted in a faster, more goal-oriented way than evolution (which only progresses through the slow accumulation of successive generations). Alongside the fact that evolution is a precondition for the feasibility of AIs, it naturally also permits directed human research to borrow from biological design and thereby proceed considerably faster.

Compared to the biological brain of a person, computer hardware offers several advantages[4, p. 60]: the basic computational elements (modern microprocessors) "fire" millions of times faster than neurons; signals are transmitted millions of times faster; and a computer can store considerably more basic computational elements in total (a single supercomputer can easily take up an entire factory floor). A future digital intelligence would also have big advantages over the human brain in relation to software components [4, pp. 60–61]: for instance, it is easy to both modify and multiply, meaning that potentially relevant information can be called upon at any time. In a few important areas such as energy efficiency, resilience to purely physical damage, and *graceful degradation* [83], artificial hardware still lags behind the human brain. In particular, there is still no direct relation between thermodynamic efficiency and complexity reduction at the level of information processing [84, 85], but this may change as computer hardware improves in coming decades.

In view of these comparative advantages and the predicted rapid improvement of hardware [86] and software, it seems probable that human intelligence will someday be overtaken by that of machines. It is important to assess more precisely how and when this could take place, and where the implications of such a scenario lie.

Timeframes

Different experts in the area of AI have considered the question of when the first machines will reach the level of human intelligence. A survey of the hundred most successful AI experts, measured according to a citation index, revealed that a majority consider it likely that human-level AI will be developed within the first half of this century [4, p. 19]. The belief that humans will create a superintelligence by the end of this century, as long as technological progress experiences no large setbacks (as a result of global catastrophes), was also held by the majority of experts [4, p. 20]. The variance among these estimates is high: some experts are confident that there will be machines with at least human levels of intelligence no later than 2040; (fewer) other experts think that this level will never be reached. Even if one makes a somewhat conservative assumption, accounting for the tendency of human experts to be overconfident in their estimates [87, 88], it would still be inappropriate to describe superintelligence as mere "science fiction" in the light of such widespread confidence among relevant experts.

Goals of a general intelligence

As a rational agent, an artificial intelligence strives towards just what its goals/goal function describes [89]. Whether an artificial intelligence will act *ethically*, that is, whether it will have goals which are not in conflict with the interests of humans and other sentient beings, is completely open: an artificial intelligence can in principle follow all possible goals [90]. It would be a mistaken anthropomorphisation to think that every kind of superintelligence would be interested in ethical questions like (typical) humans. When we build an artificial intelligence, we also establish its goals, explicitly or implicitly.

These claims are sometimes criticized on the grounds that any attempt to direct the goal of an artificial intelligence according to human values would amount to "enslavement," because our values would be *forced* upon the AI [91]. However, this criticism rests on a misunderstanding, as the expression "forced" suggests that a particular, "true" goal already exists, one the AI has *before* it is created. This idea is logically absurd, because there is no pre-existing agent "receiving" the goal function in the first place, and thus no goal independent of the processes that have created an agent. The process that creates an intelligence determines inevitably its functioning and goals. *If* we intend to build a superintelligence, then we, and nothing and nobody else, are responsible for its goals. Furthermore, it is also not the case that an AI must experience any kind of harm through the goals that we inevitably give it. The possibility of being harmed in an ethically relevant sense requires consciousness, which we must ensure is not achieved by a superintelligence. Parents inevitably form the values and goals of their children's "biological intelligence" in a very similar way, yet this does obviously not imply that children are thereby "enslaved" in an unethical manner. Quite the opposite: we have the greatest ethical duty to impart fundamental ethical values to our children. The same is true for the AIs that we create.

The computer science professor Stuart Russell warns that the programming of ethical goals poses a great challenge [3], both on a technical level (how would complex goals in a programming language be written so that no unforeseen consequences resulted?) and on an ethical level (which goals anyhow?). The first problem is called the *value-loading problem* in the literature [92].

Although the scope of possible goals of a superintelligence is huge, we can make some reliable statements about the actions they would take. There is a range of instrumentally rational subgoals that are useful for agents with highly varied terminal goals. These include goal- and self-preservation, increasing one's intelligence, and resource accumulation [93]. If the goal of an AI were altered, this could be as negative (or even more so) to the achievement of its original goal as the destruction of the AI itself. Increased intelligence is essentially just an ability to reach goals in a wider range of environments, and this opens up the possibility of a so-called *intelligence explosion*, in which an AI rapidly undergoes an enormous increase in its intelligence through recursive self-improvement [94, 95] (a concept first described by I.J. Good [96] which has since been formalized in concrete algorithms [97].) Resource accumulation and the discovery of new technologies give the AI more power, which in turn serves better goal achievement. If the goal function of a newly developed superintelligence ascribed no value to the welfare of sentient beings, it would cause reckless death and suffering wherever this was useful for its (interim) goal achievement.

One could tend towards the assumption that a superintelligence poses no danger because it is only a computer, which one could literally unplug. By definition, however, a superintelligence would not be stupid; if there were any probability that it would be unplugged, a superintelligence could initially behave itself as the makers wished it to, until it had found out how to minimize the risk of an involuntary shutdown [4, p. 117]. It could also be possible for a superintelligence to circumvent the security systems of big banks and nuclear weapon arsenals using hitherto unknown gaps in security (so-called *zero day exploits*), and in this way to blackmail the global population and force it to cooperate. As mentioned earlier, in such a scenario a "return to the initial situation" would be highly improbable.

What is at stake

In the best-case scenario, a superintelligence could solve countless problems for humanity, helping us overcome the greatest scientific, ethical, ecological and economic challenges of the future. If, however, the goals of a superintelligence were incompatible with the preferences of human beings or any other sentient beings, it would amount to an unprecedented existential threat, potentially causing more suffering than any preceding event in the known universe [98].

Rational risk management

In decision situations where the stakes are very high, the following principles are of crucial importance:

1. Expensive precautions can be worth the cost even for low-probability risks, provided there is enough to win/lose thereby [89].
2. When there is little consensus in an area amongst experts, epistemic modesty is advisable. That is, one should not have too much confidence in the accuracy of one's own opinion either way.

The risks of AI research are of a global nature. If AI researchers fail to transfer ethical goals to a superintelligence in the first attempt, there quite possibly won't be a second chance. It is absolutely tenable to estimate the long-term risks of AI research as even greater than those of climate change. In comparison to climate change, however, AI research is receiving very little attention. With this paper, we want to emphasize that it is therefore even more valuable to invest considerable resources into AI safety research.

If the scenarios discussed here have a non-infinitesimal chance of actually happening, then artificial intelligence and the opportunities and risks associated with it should be a global priority. The probability of a good outcome of AI research can be maximized through a number of measures, including the following: If the scenarios discussed here have (a perhaps small, but) more than an infinitesimal chance of actually happening, then artificial intelligence and the opportunities and risks associated with it should be a global priority. The probability of a good outcome of AI research can be maximised through the following measures, amongst others:

Recommendation 5 — **Information:** An effective improvement in the safety of artificial intelligence research begins with awareness on the part of experts working on AI, investors, and decision-makers. Information on the risks associated with AI progress must, therefore, be made accessible and understandable to a wide audience. Organizations supporting these concerns include the Future of Humanity Institute (FHI) at the University of Oxford, the Machine Intelligence Research Institute (MIRI) in Berkeley, the Future of Life Institute (FLI) in Boston, as well as the Foundational Research Institute (FRI). ■

Recommendation 6 — **AI safety:** Recent years have witnessed an impressive rise in investment into AI research [86], but research into AI safety has been comparatively slow. The only organization currently dedicated the theoretical and technical problems of AI safety as its top priority is the aforementioned MIRI. Grantors should encourage research projects to document the relevance of their work to AI safety, as well as the precautions taken within the research itself. At the same time, high-risk AI research should not be banned, as this would likely result in a rapid and extremely risky relocation of research to countries with lower safety standards. ■

Recommendation 7 — **Global cooperation and coordination:** Economic and military incentives create a competitive environment in which a dangerous AI arms race will almost certainly arise. In the process, the safety of AI research will be reduced in favor of more rapid progress and reduced cost. Stronger international cooperation can counter this dynamic. If international coordination succeeds, then a “race to bottom” in safety standards (through the relocation of scientific and industrial AI research) would also be avoided. ■

Artificial consciousness

Humans and many non-human animals have what is known as phenomenal consciousness—that is, they experience themselves to be a human or a non-human animal with a subjective, first-person point of view [99]. They have sensory impressions, a (rudimentary or pronounced) sense of self, experiences of pain upon bodily damage, and the capacity to feel psychological suffering or joy (see for example the studies of depression in mice [100]). In short, they are *sentient* beings. Consequently, they can be *harmed* in a sense that is relevant to their own interests and perspective. In the context of AI, this leads to the following question: Is it possible for the functional system of a machine to also experience a potentially painful “inner life”? The philosopher and cognitive scientist Thomas Metzinger offers four criteria for the concept of suffering, all of which would apply to machines as well as animals:

1. Consciousness.
2. A phenomenal self-model.
3. The ability to register negative value (that is, violated subjective preferences) within the self-model.
4. Transparency (that is, perceptions feel irrevocably “real”, thus forcing the system to self-identify with the content of its conscious self-model) [101, 102].

Two related questions have to be distinguished actually: firstly, whether machines could ever develop consciousness and the capacity for suffering at all; and secondly, if the answer to the first question is yes, which *types* of machines (will) have consciousness.

In addition to the above, two related questions have to be distinguished: Firstly, whether machines could technically develop consciousness and the capacity for suffering at all; Secondly, if the answer to the first question is yes, which types of machines (will) have consciousness. These two questions are being researched by philosophers and AI experts alike. A glance at the state of research reveals that the first question is easier to answer than the second. There is currently substantial, but not total, consensus amongst experts that machines could in principle have consciousness, and that it is at least possible in *neuromorphic* computers [103, 104, 105, 106, 107, 108, 109]. Such computers have hardware with the same functional organization as a biological brain [110]. The question of identifying which types of machines (besides neuromorphic computers) could have consciousness, however, is far more difficult to answer. The scientific consensus in this area is less clear [111]. For instance, it is disputed whether pure simulations (such as the simulated brain of the *Blue Brain Project*) could have consciousness. While some experts are confident that this is the case [109, 105], others disagree [111, 112].

In view of this uncertainty among experts, it seems reasonable to take a *cautious* position: According to current knowledge, it is at least conceivable that many sufficiently complex computers, including non-neuromorphic ones, could be sentient.

These considerations have far-reaching ethical consequences. If machines could have consciousness, then it would be ethically unconscionable to exploit them as a workforce, and to use them for risky jobs such as defusing mines or handling dangerous substances [4, p. 167]. If sufficiently complex AIs will have consciousness and subjective preferences with some probability, then similar ethical and legal safety precautions to those used for humans and non-human animals will have to be met [113]. If, say, the virtual brain of the Blue Brain Project was to gain consciousness, then it would be highly ethically problematic to use it (and any potential copies or "clones") for systematic research of e.g. depression by placing it in depressive circumstances. Metzinger warns that conscious machines could be misused for research purposes. Moreover, as "second class citizens", they may lack legal rights and be exploited as dispensable experimental tools, all of which could be negatively reflected at the level of the machines' inner experience [106]. This prospect is particularly worrying because it is conceivable that AIs will be made in such huge numbers [4, 75] that in a worst-case scenario, there could be an astronomical number of victims, outnumbering any known catastrophe in the past.

These dystopian scenarios point toward an important implication of technological progress: Even if we make only "minor" ethical mistakes (e.g. by erroneously classifying certain computers as unconscious or morally insignificant), then by virtue of historically unprecedented technological power, this could result in equally unprecedented catastrophes. If the total number of sentient beings rises drastically, we must ensure that our ethical values and empirical estimates improve proportionally; a mere marginal improvement in either parameter will be insufficient to meet the greatly increased responsibility. Only by acknowledging the uncertain nature of possible machine consciousness can we begin to take appropriate cautionary measures in AI research, and thus hope to avoid any of the potential catastrophes described above.

Recommendation 8 — **Research:** In order to make ethical decisions, it is important to have an understanding of which natural and artificial systems have the capacity for producing consciousness, and in particular for experiencing suffering. Given the apparent level of uncertainty and disagreement within the field of machine consciousness, there is a pressing need to promote, fund, and coordinate relevant interdisciplinary research projects (comprising philosophy, neuroscience, and computer science). ■

Recommendation 9 — **Regulation:** It is already standard practice for ethics commissions to regulate experiments on living test subjects [114, 115]. In light of the possibility that neuromorphic computers and simulated beings could also develop consciousness, it is vital that research on these, too, is carried out under the strict supervision of ethics commissions. Furthermore, the (unexpected) creation of sentient artificial life should be avoided or delayed wherever possible, as the AIs in question could—once created—be rapidly duplicated on a vast scale. In the absence of pre-existing legal representation and political interest in artificial sentience, this proliferation would likely continue unchecked. ■

Conclusion

Already today, we are witnessing the spread of novel AI technologies with surprising potential. The AI technology currently behind driverless cars, *Watson*-assisted medical diagnosing, and US military drones will gradually become available for general use in the foreseeable future. It is crucial that carefully constructed legal frameworks are in place before this happens, so as to realize the potential of these technologies in ways that safely minimize any risks of a negative overall development.

The more progress is made in the field of AI technology, the more pressing a rational, far-sighted approach to the associated challenges becomes. Because political and legal progress tends to lag behind technological development, there is an especially large amount of responsibility resting on the individual researchers and developers who directly take part in any progress being made.

Unfortunately, however, there are strong economic incentives for the development of new technologies to take place as fast as possible without "wasting" time on expensive risk analyses. These unfavorable conditions increase the risk that we gradually lose our grip on the control of AI technology and its use. This should be prevented on all possible levels, including politics, the research itself, and in general by anyone whose work is relevant to the issue. A fundamental prerequisite to directing AI development along the most advantageous tracks possible will be to broaden the field of AI safety. This way, it can be recognized not only among a few experts but in widespread public discourse as a great (perhaps the greatest) challenge of our age.

As a final addition to the concrete recommendations given above, we would like to conclude by pleading that AI risks and opportunities be recognized as a global priority—akin to climate change, or the prevention of military conflicts—as soon as possible.

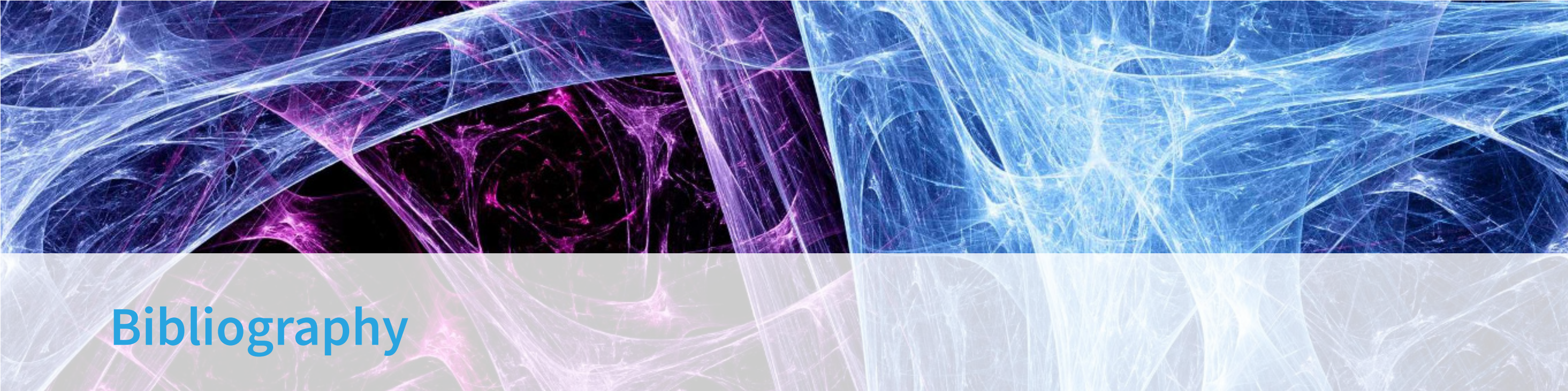
Acknowledgements

We thank all those who helped us in the research or writing of this position paper. Worthy of particular mention are: Kaspar Etter and Massimo Mannino, for their suggestions on the structure of the paper; professor Oliver Bendel, for suggestions to the chapter "Advantages and risks of current AIs"; and professor Jürgen Schmidhuber, both for his inputs to the chapters "General intelligence and superintelligence" and "Artificial consciousness" and for his valuable contributions to the current state of knowledge in various fields of AI research.

Supporters

The central points of this position paper are supported by:

- **Prof. Dr. Fred Hamker**, Professor of Artificial Intelligence, Technical University of Chemnitz
- **Prof. Dr. Dirk Helbing**, Professor of Computational Social Science, ETH Zürich
- **Prof. Dr. Malte Helmert**, Professor of Artificial Intelligence, University of Basel
- **Prof. Dr. Manfred Hild**, Professor of Digital Systems, Beuth Technical College, Berlin
- **Prof. Dr. Dr. Eric Hilgendorf**, Director of Research in Robotic Law, University of Würzburg
- **Prof. Dr. Marius Kloft**, Professor of Machine Learning, Humboldt University, Berlin
- **Prof. Dr. Jana Koehler**, Professor of Information Science, Luzern College
- **Prof. Dr. Stefan Kopp**, Professor of Social Cognitive Systems, University of Bielefeld
- **Prof. Dr. Dr. Franz Josef Radermacher**, Professor of Databases and Artificial Intelligence, University of Ulm



Bibliography

[1] Koomey, J. G., Berard, S., Sanchez, M., & Wong, H. (2011). Implications of Historical Trends in the Electrical Efficiency of Computing. *IEEE Annals of the History of Computing*, *33*(3), 46–54.

[2] Brockman, J. (2015). *What to Think About Machines That Think: Today's Leading Thinkers on the Age of Machine Intelligence*. Harper Perennial.

[3] Russell, S. (2015). Will They Make Us Better People? (http://edge.org/response-detail/26157)

[4] Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

[5] BBC. (2015a). Stephen Hawking Warns Artificial Intelligence Could End Mankind. (http://www.bbc.com/news/technology-30290540)

[6] Harris, S. (2015). Can We Avoid a Digital Apocalypse? (https://edge.org/response-detail/26177)

[7] The Independent. (2014). Stephen Hawking: 'Transcendence Looks at the Implications of Artificial Intelligence — But Are We Taking AI Seriously Enough?' (http://www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-at-the-implications-of-artificial-intelligence--but-are-we-taking-ai-seriously-enough-9313474.html)

[8] The Guardian. (2014). Elon Musk Donates $10m to Keep Artificial Intelligence Good for Humanity. (http://www.theguardian.com/technology/2015/jan/16/elon-musk-donates-10m-to-artificial-intelligence-research)

[9] SBS. (2013). Artificial Irrelevance: The Robots Are Coming. (http://www.sbs.com.au/news/article/2012/07/18/artificial-irrelevance-robots-are-coming)

[10] BBC. (2015b). Microsoft's Bill Gates Insists AI Is a Threat. (http://www.bbc.com/news/31047780)

[11] Silver, N. (2012). *The Signal and the Noise: Why So Many Predictions Fail – But Some Don't*. Penguin.

[12] PCWorld. (2011). IBM Watson Vanquishes Human Jeopardy Foes. (http://www.pcworld.com/article/219893/ibm_watson_vanquishes_human_jeopardy_foes.html)

[13] Bowling, M., Burch, N., Johanson, M., & Tammelin, O. (2015). Heads-up Limit Hold'em Poker Is Solved. *Science*, *347*(6218), 145–149.

[14] Ciresan, D. C., Giusti, A., Gambardella, L. M., & Schmidhuber, J. (2013). Mitosis Detection in Breast Cancer Histology Images Using Deep Neural Networks. MICCAI 2013. (http://people.idsia.ch/~juergen/deeplearningwinsMICCAIgrandchallenge.html)

[15] Ciresan, D., Meier, U., & Schmidhuber, J. (2012). Multi-Column Deep Neural Networks for Image Classification. *Computer Vision and Pattern Recognition 2012*, 3642–3649.

[16] Tesauro, G. (1994). TD-Gammon, a Self-Teaching Backgammon Program, Achieves Master-Level Play. *Neural Computation*, *6*(2), 215–219.

[17] Koutník, J., Cuccu, G., Schmidhuber, J., & Gomez, F. (2013). Evolving Large-Scale Neural Networks for Vision-Based Reinforcement Learning. In *Proceedings of the 15th Annual Conference on Genetic and Evolutionary Computation* (pp. 1061–1068). ACM.

[18] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., … Ostrovski, G. et al. (2015). Human-Level Control Through Deep Reinforcement Learning. *Nature*, *518*(7540), 529–533.

[19] Slavin, K. (2012). How Algorithms Shape Our World. (http://ed.ted.com/lessons/kevin-slavin-how-algorithms-shape-our-world)

[20] Tagesanzeiger. (2008). Computer-Panne legt US-Flugverkehr lahm. (http://www.tagesanzeiger.ch/ausland/amerika/ComputerPanne-legt-USFlugverkehr-lahm/story/13800972)

[21] Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank Citation Ranking: Bringing Order to the Web. (http://ilpubs.stanford.edu:8090/422/)

[22] Wired. (2010). Algorithms Take Control of Wall Street. (http://www.wired.com/2010/12/ff_ai_flashtrading/all/)

[23] Lin, T. C. (2012). The New Investor. *UCLA L. Rev. 60*, 678–735.

[24] Taleb, N. N. (2010). *The Black Swan: The Impact of the Highly Improbable Fragility*. Random House.

[25] Lauricella, T. & McKay, P. (2010). Dow Takes a Harrowing 1,010.14-point Trip. *Wall Street Journal (May 7, 2010)*.

[26] Securities, U., Commission, E., & the Commodity Futures Trading Commission. (2010). Findings Regarding the Market Events of May 6, 2010. *Report of the Staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues*.

[27] Spiegel. (2015). Denkende Waffen: Künstliche-Intelligenz-Forscher Warnen vor Künstlicher Intelligenz. (http://www.spiegel.de/netzwelt/netzpolitik/elon-musk-und-stephen-hawking-warnen-vor-autonomen-waffen-a-1045615.html)

[28] Bendel, O. (2013). Towards Machine Ethics. In *Technology Assessment and Policy Areas of Great Transitions* (pp. 343–347). Proceedings from the PACITA 2013 Conference in Prague.

[29] Goodall, N. J. (2014). Machine Ethics and Automated Vehicles. In *Road Vehicle Automation: Lecture Notes in Mobility* (pp. 93–102). Springer International Publishing.

[30] Bostrom, N. & Yudkowsky, E. (2013). The Ethics of Artificial Intelligence. In *Cambridge Handbook of Artificial Intelligence*. Cambridge University Press.

[31] Dickmanns, E. D., Behringer, R., Dickmanns, D., Hildebrandt, T., Maurer, M., Thomanek, F., & Schiehlen, J. (1994). The Seeing Passenger Car 'VaMoRs-P'. In *International Symposium on Intelligent Vehicles 94* (pp. 68–73).

[32] Dickmanns, E. (2011). Evening Keynote: Dynamic Vision as Key Element for AGI. 4th Conference on Artificial General Intelligence, Mountain View, CA. (https://www.youtube.com/watch?v=YZ6nPhUG2i0)

[33] Thrun, S. (2011). Google's Driverless Car. (http://www.ted.com/talks/sebastian_thrun_google_s_driverless_car)

[34] Forbes. (2012). Nevada Passes Regulations for Driverless Cars. (http://www.forbes.com/sites/alexknapp/2012/02/17/nevada-passes-regulations-for-driverless-cars/)

[35] Organization, W. H. et al. (2013). *WHO Global Status Report on Road Safety 2013: Supporting a Decade of Action*. World Health Organization.

[36] Simonite, T. (2013). Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks. *MIT Technology Review, Oct*, *25*.

[37] CNBC. (2014). Self-Driving Cars Safer Than Those Driven by Humans: Bob Lutz. (http://www.cnbc.com/id/101981455)

[38] Svenson, O. (1981). Are We All Less Risky and More Skillful Than Our Fellow Drivers? *Acta Psychologica*, *9*(6), 143–148.

[39] Weinstein, N. D. (1980). Unrealistic Optimism about Future Life Events. *Journal of Personality and Social Psychology*, *39*(5), 806.

[40] Langer, E. J. (1975). The Illusion of Control. *Journal of Personality and Social Psychology*, *32*(2), 311.

[41] Von Hippel, W. & Trivers, R. (2011). The Evolution and Psychology of Self-Deception. *Behavioral and Brain Sciences*, *34*(1), 1–56.

[42] Trivers, R. (2011). *The Folly of Fools: The Logic of Deceit and Self-Deception in Human Life*. Basic Books.

[43] Berner, E. S. & Graber, M. L. (2008). Overconfidence as a Cause of Diagnostic Error in Medicine. *The American Journal of Medicine*, *121*(5), S2–S23.

[44] Kohn, L. T., Corrigan, J. M., Donaldson, M. S. et al. (2000). *To Err Is Human: Building a Safer Health System*. National Academies Press.

[45] The New York Times. (2010). What Is IBM's Watson? (http://www.nytimes.com/2010/06/20/magazine/20Computer-t.html)

[46] Wired. (2013). IBM's Watson Is Better at Diagnosing Cancer Than Human Doctors. (http://www.wired.co.uk/news/archive/2013-02/11/ibm-watson-medical-doctor)

[47] Forbes. (2013). IBM's Watson Gets Its First Piece Of Business In Healthcare. (http://www.forbes.com/sites/bruceupbin/2013/02/08/ibms-watson-gets-its-first-piece-of-business-in-healthcare/)

[48] Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical Versus Actuarial Judgment. *Science*, *243*(4899), 1668–1674.

[49] Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical Versus Mechanical Prediction: A Meta-Analysis. *Psychological Assessment*, *12*(1), 19.

[50] West, R. F. & Stanovich, K. E. (1997). The Domain Specificity and Generality of Overconfidence: Individual Differences in Performance Estimation Bias. *Psychonomic Bulletin & Review*, *4*(3), 387–392.

[51] Tversky, A. & Kahneman, D. (1974). Judgment Under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131.

[52] Pohl, R. (Ed.). (2004). *Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory*. Psychology Press.

[53] Brosnan, M. J. (2002). *Technophobia: The Psychological Impact of Information Technology*. Routledge.

[54] Yudkowsky, E. (2008). Artificial Intelligence as a Positive and Negative Factor in Global Risk. *Global Catastrophic Risks*, *1*, 303.

[55] Bostrom, N. (2002). Existential Risks. *Journal of Evolution and Technology*, *9*(1).

[56] Smith, A. & Anderson, J. (2014). AI, Robotics, and the Future of Jobs. Pew Research Center.

[57] Cowen, T. (2013a). *Average Is Over: Powering America Beyond the Age of the Great Stagnation*. Penguin.

[58] Brynjolfsson, E. & McAfee, A. (2014). *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. WW Norton & Company.

[59] Frey, C. B. & Osborne, M. A. (2013). The Future of Employment: How Susceptible Are Jobs to Computerisation? *Oxford Martin Programme on Technology and Employment*. (https://web.archive.org/web/20150109185039/http://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf)

[60] Helbing, D. (2015). *Thinking Ahead — Essays on Big Data, Digital Revolution, and Participatory Market Society*. Springer.

[61] Cowen, T. (2013b). EconTalk Episode with Tyler Cowen: Tyler Cowen on Inequality, the Future, and Average is Over. (http://www.econtalk.org/archives/2013/09/tyler_cowen_on.html)

[62] Griffiths, M., Kuss, D., & King, D. (2012). Video Game Addiction: Past, Present and Future. *Current Psychiatry Reviews*, *8*(4), 308–318.

[63] Srivastava, L. (2010). Mobile Phones and the Evolution of Social Behaviour. *Behavior & Information Technology*, *24*(2), 111–129.

[64] Prensky, M. (2001). Do They Really Think Differently? *On the Horizon*, *47*(2).

[65] Metzinger, T. (2015a). Virtuelle Verkörperung in Robotern. *SPEKTRUM*, *2*, 48–55.

[66] Kapp, K. M. (2012). *The Gamification of Learning and Instruction: Game-Based Methods and Strategies for Training and Education*. Pfeiffer.

[67] Bavelier, D., Green, S., Hyun Han, D., Renshaw, P., Merzenich, M., & Gentile, D. (2011). Viewpoint: Brains on Video Games. *Nature Reviews Neuroscience*, *12*, 763–768.

[68] Fagerberg, J. (2000). Technological Progress, Structural Change and Productivity Growth: A Comparative Study. *Structural Change and Economic Dynamics*, *11*(4), 393–411.

[69] Galor, O. & Weil, D. N. (1999). From Malthusian Stagnation to Modern Growth. *American Economic Review*, 150–154.

[70] Brynjolfsson, E. (2014). EconTalk Episode with Erik Brynjolfsson: Brynjolfsson on the Second Machine Age. (http://www.econtalk.org/archives/2014/02/brynjolfsson_on.html)

[71] Hughes, J. J. (2014). Are Technological Unemployment and a Basic Income Guarantee Inevitable or Desirable? *Journal of Evolution and Technology*, *24*(1), 1–4.

[72] Krugman, P. (2013). Sympathy for the Luddites. *New York Times*, *13*. (http://www.nytimes.com/2013/06/14/opinion/krugman-sympathy-for-the-luddites.html)

[73] Bostrom, N. & Sandberg, A. (2008). Whole Brain Emulation: A Roadmap. Oxford: Future of Humanity Institute.

[74] Hanson, R. (2012). Extraordinary Society of Emulated Minds. (http://library.fora.tv/2012/10/14/Robin_Hanson_Extraordinary_Society_of_Emulated_Minds)

[75] Hanson, R. (1994). If Uploads Come First. *Extropy*, *6*(2), 10–15.

[76] Legg, S. & Hutter, M. (2005). A Universal Measure of Intelligence for Artificial Agents. In *International Joint Conference on Artificial Intelligence* (Vol. 19, p. 1509). Lawrence Erlbaum Associates ltd.

[77] Hutter, M. (2007). Universal Algorithmic Intelligence: A Mathematical Top-Down Approach. In *Artificial General Intelligence* (Vol. 6, *2*, pp. 227–290). Springer.

[78] Bostrom, N. (1998). How Long Before Superintelligence? *International Journal of Future Studies*, *2*.

[79] Schmidhuber, J. (2012). Philosophers & Futurists, Catch Up! Response to The Singularity. *Journal of Consciousness Studies*, *19*(1-2), 173–182.

[80] Moravec, H. (1998). When Will Computer Hardware Match the Human Brain. *Journal of Evolution and Technology*, *1*(1), 10.

[81] Moravec, H. (2000). *Robot: Mere Machine to Transcendent Mind*. Oxford University Press.

[82] Shulman, C. & Bostrom, N. (2012). How Hard Is Artificial Intelligence? Evolutionary Arguments and Selection Effects. *Journal of Consciousness Studies*, *19*(7-8), 103–130.

[83] Sengupta, B. & Stemmler, M. (2014). Power Consumption During Neuronal Computation. *Proceedings of the IEEE*, *102*(5), 738–750.

[84] Friston, K. (2010). The Free-Energy Principle: A Unified Brain Theory? *Nature Reviews Neuroscience*, *11*, 127–138.

[85] Sengupta, B., Stemmler, M., & Friston, K. (2013). Information and Efficiency in the Nervous System — A Synthesis. *PLoS Comput Biol*, *9*(7).

[86] Eliasmith, C. (2015). On the Eve of Artificial Minds. In T. Metzinger & J. M. Windt (Eds.), *Open mind*. MIND Group. (http://open-mind.net/papers/@@chapters?nr=12)

[87] Armstrong, S., Sotala, K., & ÓhÉigeartaigh, S. S. (2014). The Errors, Insights and Lessons of Famous AI Predictions — And What They Mean for the Future. *Journal of Experimental & Theoretical Artificial Intelligence*, *26*(3), 317–342.

[88] Brenner, L. A., Koehler, D. J., Liberman, V., & Tversky, A. (1996). Overconfidence in Probability and Frequency Judgments: A Critical Examination. *Organizational Behavior and Human Decision Processes*, *65*(3), 212–219.

[89] Peterson, M. (2009). *An Introduction to Decision Theory*. Cambridge University Press.

[90] Armstrong, S. (2013). General Purpose Intelligence: Arguing the Orthogonality Thesis. *Analysis and Metaphysics*, (12), 68–84.

[91] Noë, A. (2015). The Ethics Of The ‘Singularity’. (http://www.npr.org/sections/13.7/2015/01/23/379322864/the-ethics-of-the-singularity)

[92] Bostrom, N. (2012). The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents. *Minds and Machines*, *22*(2), 71–85.

[93] Omohundro, S. M. (2008). The Basic AI Drives. In *Proceedings of the First AGI Conference, 171, Frontiers in Artificial Intelligence and Applications* (Vol. 171, pp. 483–492).

[94] Solomonoff, R. (1985). The Time Scale of Artificial Intelligence: Reflections on Social Effects. *Human Systems Management*, *5*, 149–153.

[95] Chalmers, D. (2010). The Singularity: A Philosophical Analysis. *Journal of Consciousness Studies*, *17*(9-10), 7–65.

[96] Good, I. J. (1965). Speculations Concerning the First Ultraintelligent Machine. In *Advances in Computers* (pp. 31–88). Academic Press.

[97] Schmidhuber, J. (2006). Gödel Machines: Fully Self-Referential Optimal Universal Self-Improvers. In *Artificial General Intelligence* (pp. 119–226).

[98] Tomasik, B. (2011). Risks of Astronomical Future Suffering. Foundational Research Institute. (http://foundational-research.org/publications/risks-of-astronomical-future-suffering/)

[99] Nagel, T. (1974). What Is it Like to Be a Bat? *The Philosophical Review*, 435–450.

[100] Durgam, R. (2001). Rodent Models of Depression: Learned Helplessness Using a Triadic Design in Tats. *Curr Protoc Neurosci*, (8).

[101] Metzinger, T. (2012). Two Principles for Robot Ethics. In H. E & G. J-P (Eds.), *Robotik und Gesetzgebung* (pp. 263–302). NOMOS. (http://www.blogs.uni-mainz.de/fb05philosophie/files/2013/04/Metzinger_RG_2013_penultimate.pdf)

[102] Metzinger, T. (2015b). *Empirische Perspektiven aus Sicht der Selbstmodell-Theorie der Subjektivität: Eine Kurzdarstellung mit Beispielen*. Selbstverlag. (http://www.amazon.de/Empirische-Perspektiven-Sicht-Selbstmodell-Theorie-Subjektivitat-ebook/dp/B01674W53W)

[103] Moravec, H. P. (1988). *Mind Children: The Future of Robot and Human Intelligence*. Harvard University Press.

[104] Chalmers, D. J. (1995). Absent Qualia, Fading Qualia, Dancing Qualia. *Conscious Experience*, 309–328.

[105] Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

[106] Metzinger, T. (2010). *The Ego Tunnel: The Science of the Mind and the Myth of the Self* (First Trade Paper Edition). New York: Basic Books.

[107] Metzinger, T. (2015c). What If They Need to Suffer? (https://edge.org/response-detail/26091)

[108] Dennett, D. C. (1993). *Consciousness Explained*. Penguin UK.

[109] Bostrom, N. (2003). Are We Living in a Computer Simulation? *The Philosophical Quarterly*, *53*(211), 243–255.

[110] Hasler, J. & Marr, B. (2013). Finding a Roadmap to Achieve Large Neuromorphic Hardware Systems. *Frontiers in Neuroscience*, *7*(118).

[111] Koch, C. (2014). What it Will Take for Computers to Be Conscious, MIT Technology Review. (http://www.technologyreview.com/news/531146/what-it-will-take-for-computers-to-be-conscious/)

[112] Tononi, G. (2015). Integrated Information Theory. *Scholarpedia*, *10*(1), 4164. (http://www.scholarpedia.org/article/Integrated_Information_Theory)

[113] Singer, P. (1988). Comment on Frey's 'Moral Standing, the Value of Lives, and Speciesism'. *Between the Species: A Journal of Ethics*, *4*, 202–203.

[114] Swissethics, Verein anerkannter Ethikkommissionen der Schweiz. (n.d.). (http://www.swissethics.ch/)

[115] Senatskommission für Tierexperimentelle Forschung. (2004). Tierversuche in der Forschung. (http://www.dfg.de/download/pdf/dfg_im_profil/geschaeftsstelle/publikationen/dfg_tierversuche_0300304.pdf, publisher=Deutsche Forschungsgemeinschaft)

www.foundational-research.org

www.ea-stiftung.org