Effective Altruism Foundation

Foundational Research INSTITUTE

Artificial Intelligence: Opportunities and Risks

Policy paper

Artificial intelligence (AI) and increasingly complex algorithms influence our lives and our civilisation more than ever. The areas of AI application are diverse and the possibilities far reaching: 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 optimising *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, Director of Strategy, 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 influence our lives and our civilisation more than ever. The areas of AI application are diverse and the possibilities far reaching: 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 optimising 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.

Current

In narrow, well-tested areas of application (for example driverless cars or some areas of medical diagnostics) the superiority of AIs over humans is already established. An increased use of technology in these areas offers great potential (fewer road traffic accidents for example, fewer mistakes in the medical treatment of patients, or the discovery of many new kinds of therapy). In complex systems, where several algorithms interact at high speed (for example 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 favour of speed. In any case, it is crucially important which goals or ethical values should be programmed into AI algorithms, and that it can be technically guaranteed that the goals remain stable and resistant to manipulation. With driverless cars for instance, the classical question arises of how the algorithm should act if a collision with several pedestrians can only be avoided by endangering the passenger — and how it can be ensured that the algorithms of driverless cars do not fail systematically.

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

Measure 2 The legal frameworks should be adapted with regards to 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 observed, so that the risk of unexpected catastrophes can be minimised.

Mid-term

Progress in AI research makes it possible to replace more and more human work with that of machines. Many economists assume that this increasing automation could lead to a massive increase in unemployment within even the next 10-20 years. (Economists do this in the knowledge that similar predictions in the past have not proved accurate; the current developments are of a new kind, and it would be irresponsible to close our eyes to the possibility that these predictions come true at some point.) Through progressive automation, the (statistically) 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 sensibly with the consequences of AI automation? Are current social systems suitable for this? 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 to ensure a fair distribution of the profits from increased productivity.

Long-term

Many AI experts think it plausible that within this century, AIs will be created whose intelligence is far superior to human intelligence in all respects. The goals of such AIs, which in principle could take all possible forms (human ethical goals represent a tiny proportion of all possible goals), would influence the future of our planet decisively — which 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 probable 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 in future, that is (self-)consciousness and in particular subjective preferences and the capacity for suffering, 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, both in politics and also in 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 supply more resources to ethically guide the development of future-shaping technologies.

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

Measure 6 Certain AI architectures 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 (in analogy to animal research commissions).

Introduction

The pursuit of knowledge runs as a governing principle through human history. When societies have changed significantly in structure and their dynamics, this has normally been based upon 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 in caves. Several tens of thousands of years lie between arable farming and permanent settlement. The first symbols appeared a few thousand years after that, followed later by the first written scripts. In the seventeenth century the microscope was invented. Industrialisation in the nineteenth century enabled the first cities of a million people. Only a century later the atom was split and people landed on the Moon. The computer was invented, and since then the processing capabilities and energy efficiency of computers have doubled at regular intervals [1]. Technological progress often develops exponentially. For human intellectual abilities on the other hand, this is not the case.

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 specific areas (that is, domain-specifically) AIs have already repeatedly reached or even overtaken human levels. 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 simply a means; what matters is how we use it. Already the use of existing AIs presents us with considerable ethical challenges, which will be illuminated in the next part of this position paper. The following chapter will describe developments in economic automation and the prognosis that in the mid-term, AI research will give rise to a significant restructuring of the labour market. Finally, the two last chapters are concerned with 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 life and our civilisation are governed to an ever increasing extent by algorithms and domain-specific artificial intelligences (AIs) [19]: just think of smartphones, air traffic [20] or internet search engines [21]. Finance markets too are dependent on increasingly complex algorithms, which we understand less and less [22, 23]. Mostly the operation of such algorithms proceeds without incident, but there is always the possibility that an unlikely "black swan" event [24] occurs, threatening to throw the whole system into chaos. So for instance in 2010 in the USA, where an unexpected stock market crash shocked the financial world. The crash happened because computer algorithms interacted in an unforeseen manner with the financial market [25, 26]. Within minutes, important shares lost more than 90% of their worth and then quickly returned to their high initial value. In military contexts, such a "return to the initial situation" would be improbable[27]. To prevent devastating failures of this sort, it seems generally advisable to invest considerably more into the safety and reliability of AIs. Unfortunately the current economic incentives favour increasing AI capacity over safety.

Four criteria for the construction of AIs

Safety is essential with any kind of machine, but the construction of domain-specific AIs comes with new kinds of ethical challenges as soon as they take over cognitive work with social dimensions, work that was formerly carried out by humans. For instance, an algorithm that judges the

creditworthiness 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 example raise the question of which criteria should be decisive in the case of an imminent accident. Should the vehicle prioritise the survival of the passengers most highly, or, in the case of an unavoidable accident, should it be trying to keep the total number of casualties as low as possible [29]?

Because of this, the AI theorist Eliezer Yudkowsky and the 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 within a time frame that gives the responsible experts sufficient room for reaction and veto control in case of a possible failure. In addition, 3) AIs should be *robust 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 with them many advantages. They have influenced our lives for the better and will do so even more in future, provided that the necessary precautions are taken. In the following, two instructive examples are discussed.

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].

Of course, a large number of people remain sceptical of driverless cars, because they overestimate both the risks of such cars and their own driving abilities. For instance, a study came to the conclusion that 93% of all American drivers believe that their driving abilities are above the median [38] — something 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 can lead to deadly mistakes. 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, the AI *Watson* [45], developed by IBM, is to be welcomed. *Watson* became famous in 2011, beating the best human players of the quiz show *Jeopardy!* [12]. *Watson* isn't just better than humans in quiz shows: since 2014 hospitals could hire *Watson*, for instance to make cancer diagnoses. Because "Doctor Watson" can pick up and combine enormous quantities of information within the shortest space of time, it has partially overtaken its human colleagues diagnostically [46, 47].

The fact that a current AI can make more accurate medical diagnoses than human doctors may be surprising. But it has long been recognised that *statistical inferences* are mostly superior to clinical judgements by human experts [48, 49]. And of course, AIs like *Watson* are downright ideal for making statistical inferences. Using computers for the right kinds of diagnoses can therefore save lives.

Cognitive biases: to err is human

One reason that human experts are less competent than AIs at statistical inferences is the above mentioned, all too human tendency to overestimate one's own abilities. This tendency is called *overconfidence bias* [50]. The overconfidence bias is just one of a number of cognitive biases that can lead to systematic errors in human thinking [51, 52]. AIs on the other hand can be built so that they display no cognitive biases. In principle, increasing confidence in the predictions of AIs, so long as these are made safely and according to comprehensible criteria, could lead to a significantly more rational and efficient approach to many social and political challenges. The problem here lies in using the strengths of AI without 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 regarded with scepticism. However, having doubts with regards to new 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. 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 to that of biotechnology (for instance through the possible creation of new kinds of viruses) [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. Outmoded legal frameworks have to be updated with regards to 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 the light of the successes in the field of machine learning and robotics in recent years, it seems 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, the consequences on the labour market would be far reaching. According to economists like Cowen [57], McAfee and Brynjolfsson [58], technological progress will widen the income gap even further and could lead to falls in income and massively increased unemployment in large segments of the population.

A 2013 analysis concluded that it 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) or sensitivity and flexibility with movements (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 people and countries that understand how to make use of new technological opportunities and the corresponding flood of big data will benefit the most from technological progress [60]. In particular, countries with well-trained computer specialists are expected to prosper. Moreover, in future it will become ever more important that people have a good understanding of the advantages and disadvantages of different computer algorithms in comparison to purely human decision making and work capacity — something to which good education is central [61].

In the entertainment industry too, there will be far reaching innovations: with improved graphics, new entertainment technologies and new functions for mobile devices, which are all becoming increasingly cheaper, the addictive pull of videogames and internet usage is rising [62]. The social and psychological consequences of this development have not yet been fully researched, but there are several indications that these trends are profoundly changing our social behaviour [63], our attention spans and the way in which children develop [64]. In the foreseeable future, if detailed virtual realities become available to non-scientists, invading deeper into our everyday experience, then these effects could come into play much more forcefully. The consequences of more regular immersion in virtual realities, or of experiences like body-transfer illusions, in which subjective awareness is temporarily projected into a virtual avatar [65], should receive greater attention.

Altogether, the entertainment industry offers big opportunities for education through the gamification of learning content [66]; at the same time there is the risk that an increasing proportion of young people will have trouble completing their education because of pathological video game or internet consumption [67].

Utopias and dystopias

Technological progress increases societal productivity [68], raising the average standard of living [69]. If more work is carried out by machines, this creates time for leisure and self-development for humans — at least for those humans 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 *average* 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 in which AI technology has progressed so far that many jobs could be carried out by machines, the income for automatable human work will fall [58]. Without regulation, the incomes of many could sink below subsistence level. Social inequality could rise starkly if economic output rises without the income payments necessary to effect redistribution. To counteract this development, McAfee and Brynjolfsson suggest that having certain jobs be carried out by humans could be subsidised. Further possibilities for sharing the advantages of technological progress amongst the whole population are an 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 thinks it plausible 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 — which would consequently imply a huge increase in labour 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 — for example 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 occurred, 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) labour.

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. Then the bases for precautionary measures could 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: Targeted adjustments to educational content could help people to prepare better for the new kinds of challenges. For example, IT and programming knowledge are quickly gaining in relevance, while knowledge learned by heart is losing value. Gamification offers great potential which should be promoted. The social and psychological consequences of the internet should be further researched and the pathological consumption of videogames and online media prevented. ■

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, then it becomes a *superintelligence*: a superintelligence is superior to human intelligence in every way, including scientific creativity, "common sense", and social competence. This definition of superintelligence leaves open 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", who possess a conscious self-model and were developed over billions of years of evolution. These facts have been used in support of arguments that the creation of artificial intelligence may not be so difficult [80, 81, 82], since AI research can be conducted in a faster, more goal-orientated way than evolution, which only progresses through slow, meandering generational steps. 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 to proceed considerably faster.

In comparison 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 — supercomputers could e.g. be the size of a 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, software is easy to edit or to multiply, so that potentially relevant information can be called upon at any time. In a few important areas, for instance in 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]. In the coming decades, however, computer hardware will be continually further developed.

In view of these comparative advantages and the predicted meteoric improvement of hardware [86] and software, it seems probable that human intelligence will be overtaken by that of machines. It is worth finding out and assessing more precisely how and when this could be the case and where the implications of such a scenario lie.

Timeframes

Different experts in the area of artificial intelligence 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 of these experts think it likely that this will happen in the first half of this century [4, p. 19]. A majority of the experts further hold 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) [4, p. 20]. The variance amongst these time estimates is high: some experts are very sure that there will be machines with at least human levels of intelligence by 2040 at the latest, while (fewer) other experts think that this level will never be reached. Even if one makes a somewhat conservative assumption, to factor in the tendency of human experts to be overconfident in their estimates [87, 88], it would still be completely inappropriate to describe superintelligence as "science fiction": even conservative assumptions imply that there is a non-negligible probability that an AI with human levels of intelligence will be developed within this century.

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.

Sometimes these claims are criticised on the grounds that any attempt to direct the goal of an artificial intelligence according to human values would amount to an "enslavement," because our values would be *forced* upon the AI [91]. This criticism rests on a misunderstanding though. The expression "forced" suggests that a particular, "true" goal already exists, one the AI has *before* it is created. This idea is absurd: there is no "ghost in the machine," 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 – a requirement that must not be fulfilled by a superintelligence.) We inevitably form the values and goals of our biological children – "biological intelligences" – in a very similar way. Of course this does 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 artificial intelligences 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 breadth of possible goals of a superintelligence is huge, we can make some reliable statements about the actions they would take. There are a range of instrumentally rational subgoals that are useful for agents with the most various 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) for the achievement of its original goal as the destruction of the AI itself. Increasing intelligence is important because it means nothing other than increasing the ability of reaching goals in a wide range of environments — this opens up the possibility of a so-called *intelligence explosion*, in which an AI increases hugely in intelligence through recursive self-improvement in a short amount of time [94, 95]. (The basic idea of recursive self-improvement was first conceptualised by I.J. Good [96]; since then concrete algorithms for this have been made [97].) Resource accumulation and the discovery of new technologies give the AI more power, which also 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 a danger that it would be unplugged, then it would initially behave itself as the makers wished it to, until it had found out how to minimise 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 at the beginning, in such a case 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, and help us to overcome the great scientific, ethical, ecological and economic challenges of the future. If however the goals of a superintelligence did not coincide with our preferences or the preferences of all sentient beings, then it would become an existential threat and could potentially cause more suffering than there has ever been [98].

Rational risk management

In decision situations where the stakes are very high, the following principles are important:

1. Expensive precautions can be worth it even for low-probability risks, if 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 emphasise that it is therefore even more valuable to invest considerable resources into AI safety research.

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 be made accessible in an easily understandable fashion. Organisations which support these concerns are 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: In the past years there has been an impressive rise in investment into AI research [86]. In comparison, research into AI safety has lagged behind. The only organisation that researches the theoretical and technical problems in AI safety as its highest priority is the Machine Intelligence Research Institute (MIRI). Grantmakers should encourage projects to document the relevance of their work to AI safety as well as the precautions that are being taken. A ban on all high-risk AI research on the other hand would not be practicable and would lead to a fast and dangerous 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 arms race will occur with a likelihood bordering on certainty. In the process, the safety of AI research will be reduced in favour of faster progress and cost reduction. 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 phenomenal consciousness — 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, experience pain upon bodily damage, and can feel psychological suffering or joy (see for example the studies of depression amongst mice [100]). In short, they are *sentient* beings. Consequently, they can be *harmed* in sense relevant to their own interests and perspective. In the context of AI, the question arises: can there also be machines whose material-functional structure experiences a painful "inner life"? The philosopher and cognitive scientist Thomas Metzinger offers four criteria for the concept of suffering, that would also need to be fulfilled by relevant machines:

1. Consciousness.
2. A phenomenal self-model.
3. The ability to register negative valences (that is, violated subjective preferences) within the self-model.
4. Transparency (that is, perceptions feel irrevocably "real" – so the system is forced to 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.

These two questions are being researched both by philosophers and AI experts. A glance at the state of research shows that the first question if easier to answer than the second. There exists a relatively solid (but not total) consensus amongst experts that machines could in principle have consciousness and that machine consciousness is possible at least in *neuromorphic* computers [103, 104, 105, 106, 107, 108, 109]. Such computers have hardware with the same functional organisation as a biological brain [110]. The second question is harder to answer: which types of machines, besides neuromorphic computers, could have consciousness? In this area the scientific consensus is less clear [111]. It is for example disputed whether pure simulations — like the simulated brain of the *Blue Brain Project* — could have consciousness. The question is indeed answered positively by various experts [109, 105], but is also rejected by others [111, 112].

In view of this uncertainty amongst experts, it seems reasonable to take a *cautious* position: it is at least conceivable according to current knowledge that many sufficiently complex computers, including non-neuromorphic ones, will 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 like defusing mines or handling dangerous substances [4, p. 167]. If sufficiently complex artificial intelligences 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 need to be met [113]. If, say, the virtual brain of the Blue Brain Project had consciousness, then it would be highly ethically problematic to place it (and with it countless copies or "clones") in depressive circumstances in order to systematically research e.g. depression. Metzinger warns that conscious machines could be misused for research purposes and, as "second class citizens", might not only have no rights and be used as dispensable experimental tools, but that these facts could also be negatively reflected at the level of the machines' inner experience [106]. Furthermore, this prospect is particularly worrying because it is conceivable that artificial intelligences will be made in huge numbers [4, 75]. So, in a worst-case scenario, there could be an astronomical, historically unprecedented number of victims.

These dystopian scenarios point towards an important implication of technological progress: even if we make only "small" ethical mistakes, like falsely classifying certain computers as unconscious or morally insignificant, then by dint of historically unprecedented technological power, this could lead to historically unprecedented catastrophes. If the total number of sentient beings rises steeply, then a marginal improvement in our ethical values and empirical estimates will not be enough — both must improve *massively*, to meet the greatly increased responsibility. Therefore we should exercise particularly great caution in the field of AI in view of our uncertainty with regards to machine consciousness. Only in this way can we hope to avoid potential catastrophes of the manner described.

> Recommendation 8 — **Research:** In order to make ethical decisions, it is important to have an understanding of which natural and artificial systems have consciousness and especially the capacity for suffering. In the field of machine consciousness there remains great uncertainty over this. It therefore seems sensible to promote relevant interdisciplinary research (in philosophy, neuroscience, computer science). ■

> Recommendation 9 — **Regulation:** It is already standard practice for ethics commissions to regulate experiments on living test subjects [114, 115]. Because of the possibility that neuromorphic computers and simulated beings too could develop consciousness, research on them should also be carried out under the strict supervision of ethics commissions. The (unexpected) creation of sentient artificial life should be avoided or delayed, especially because this could happen in very great numbers and then — in the absence of the representation of legal and political interests — continue completely unchecked.
>
> ■

Conclusion

Already today, there are initial versions of new AI technologies with surprising potential, be it driverless cars, *Watson* as an assistant in medical diagnoses, or the newest drones sanctioned by the US military. In the foreseeable future, these applications will be available on the market for general use. Then, at the very latest, we will need well thought-through legal frameworks to realise the potential of these technological possibilities in such a way that the risks of a negative overall development remain as small as possible.

The more progress there is in the central field of AI technology, the more important and urgent becomes the rational, forward looking approach to the associated challenges. The researchers and developers of new technologies also carry responsibility for how their contributions will impact the world. In contrast to the realm of politics and law, which usually lag behind the newest developments, AI researchers and developers participate directly in the events; they are the ones who know the material best.

Unfortunately, there are strong economic incentives to undertake the development of new technologies as fast as possible, without "losing" time for expensive risk analyses. These unfavourable conditions heighten the risk that control of AI technology and its use will slip further and further from our grasp. This should be countered on as many levels as possible: in politics; in the research itself; and in general by all individuals whose work is relevant to the issue. A fundamental prerequisite to directing AI development along the most advantageous tracks possible will be

broadening the field of AI safety, so that it is recognised not only amongst a few experts but in widespread public discourse as a great (perhaps the greatest) challenge of our age.

Besides the concrete recommendations given above, we'd like to conclude this with the a plea that the topic "risks and opportunities of AI", like climate change or the prevention of military conflicts, is to be recognised as soon as possible as a global priority.

Acknowledgements

We thank all those who helped us in the research or writing of this position paper. Particularly noteworthy 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 for inputs to the chapters "General intelligence and superintelligence" and "Artificial consciousness", as well as for his contributions to the current state of research in various areas of AI.

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