Science for good and science for bad[1]

Olle Häggström

1. Introduction

My aim in this text is to explain and defend my viewpoint concerning the role of science in society and research ethics which permeates the ethical arguments in my recent book *Here Be Dragons: Science, Technology and the Future of Humanity* (Häggström, 2016). To clarify my view, I will contrast it with two more widespread points of view which I will call the *academic-romantic* and the *economic-vulgar*. These will be sketched in Section 2. In Section 3 I explain what is missing in these approaches, namely, the insight that scientific progress may not only make the world better but may also make it worse, whence we need to act with considerably more foresight than is customary today. As a concrete illustration, I will in Section 4 discuss what this might mean for a specific area of research, namely artificial intelligence. In the concluding Section 5 I return to some general considerations about what I think ought to be done.

2. Two insufficient points of view

As a representative in this section for the first of the two viewpoints that I will criticize, the **academic-romantic**, I choose the Hungarian-Swedish cancer researcher and author Georg Klein, who passed away in 2016. I afford him this unrewarding role despite (or perhaps thanks to) the great influence he has had on my thinking during several decades, through his collections of essays from the 1980s and onward, in which he, in characteristically thoughtful style together with a wealth of personal recollections, discusses issues about the nature of science, ethics, creativity, the precariousness of the human species, and the meaning of life.

In his book *Korpens blick* (1988) Klein recapitulates an exchange of letters with Göran Rosenberg about the responsibility of the scientist for the possible consequences of the scientific result. Rosenberg argues that the scientist has a large such responsibility, while Klein disagrees maintaining that the creativity of the scientist is only fully functional when the pure and uncorrupted search for truth is in focus. The scientist should therefore isolate himself from issues of utility and considerations about societal consequences – considerations that disturb not only the scientist's ability to focus, but also the idea of objective search for truth, and thereby, in effect might damage the very credibility of science. This line of thinking is paralleled in journalism in what has been called consequence neutrality: for journalism to remain credible, journalists and editors must not allow decisions about what to publish and not to publish to be influenced by strategic considerations about the possible political or other consequences a publication may have.[2] Eager to liberate the scientist from responsibility for the consequences of the scientific results, Klein makes a comparison with other professions beyond that of journalism:

> Do we expect the butcher to ponder about the ethical justification of eating meat while exercising his profession? How often do flight attendants dwell on the noise damage of air traffic?

Klein goes on to give examples of scientists who have had great hopes for how their discoveries would contribute to a better world, only to discover that other forces beyond their control have used them otherwise. The consequences become unpredictable and unmanageable to the scientist. In

[1] Translation into English by Robert Callergård of the Swedish original *Vetenskap på gott och ont*.

[2] Fichtelius (2016). It will soon become clear that I do not fancy consequence neutrality taken too far.

other words: there is no point in even thinking about it. And in Klein´s own words: “Who will control the avalanches? Not the scientist, for sure.”

The academic-romantic conception of science, which holds that the search for truth (whatever it turns out to be) should be driven by curiosity but otherwise be impartial and uncorrupted by strategical considerations about consequences, dominates the academic environment where I come from – particularly among mathematicians. It may be contrasted with the **economic-vulgar** point of view, which is more common among politicians and to some extent in university management, as well as at some of the more applied departments at engineering schools such as my own alma mater and employer, Chalmers University of Technology.

The economic-vulgar view rejects the academic-romantic ideal about searching for knowledge for its own sake, and claims instead that the purpose of science is to generate innovations and patents for industry in order to fuel economic growth and competitiveness on the world market for the country. This may sound as a crude caricature. It comes very close, however, to the view of the present Swedish government, and it is easy to find unabashed statements of the economic-vulgar point of view. An interview with Thomas Nordström, at the time vice chancellor at Kristianstad University, in the journal *Universitetsläraren* in 2006, reaches almost parodic heights. Nordström asks rhetorically how it can be that “Scania, Volvo, Ericsson, Ikea and Sandvik, just to mention a few companies, are top ranked in the world while our universities are not” answering that what the universities need is a more purely application-oriented perspective and a firmer management, and continuing “Imagine if at Scania someone were to produce mopeds, someone else mowers, and another one makes toasters”.[3]

If we accept the economic-vulgar view of science, we consequently reject research about dinosaurs, about the Big Bang, about conceptions of genus in the works of Selma Lagerlöf, and everything else that does not quickly convert into industrial products. For me, as supposedly for most defenders of the academic-romantic view, this is by far evidence enough to show the untenability of the economic-vulgar view. I have, however, another line of criticism which hits equally hard against the academic-romantic view and the economic-vulgar, and which is the point of departure for the next section.

3. Scientific progress may make the world better or worse

The academic-romantic point of view and the economic-vulgar concur in the implicit assumption that the worst thing that may happen with a scientific discovery is that it turns out to be irrelevant – that it gets zero citations, to use the language of bibliometry which regrettably has had an increasingly pervasive influence in the academic world over the last decade or two. This assumption is however dead wrong: far worse things might happen. A scientific discovery may create ripples that make the world worse. A research group that succeeds in sequencing the genome of the virus that caused the Spanish flue (that horrendous flu which killed more than 50 million people in the years 1918-1920) and then proceeds to make public the complete genome makes the world more insecure and worse.[4] On a scale which measures the impact of a scientific result, the total irrelevance (zero impact) should not be placed at the bottom end of the scale, but rather as a middle point on a scale that reaches from total catastrophe to blessed breakthrough.[5]

---

[3] Eliasson (2006). See also Häggström (2006b) for my (sarcastic) reaction to Nordström’s suggestion.

[4] This actually happened; see von Bubnoff (2005), van Aken (2007) and Häggström (2016).

[5] I do not wish to be understood to mean that it should be a *merit* to be at this middle point, which is what the great mathematician G.H. Hardy, in his famous essay *A Mathematicians Apology* from 1940, argues. Hardy claimed that the kind of mathematics he was working on did not have any applications (later this turned out to be wrong – though that is beside the point here), and that for this reason it is purer and more noble than the kind of chemistry which produces nerve gas and such. Many lesser mathematicians than Hardy have tried to follow suit in this line of argument, without realizing, however, that, if it can be seen at all as an argument for

To put this in the proper light, let me dwell a moment on the general situation for humanity now in the early 21st century. The condition of the world is better than ever in terms of several welfare measures (Pinker, 2018; Rosling, Rosling and Rosling Rönnlund, 2018). At the same time, however, humanity is confronted with huge challenges for the rest of this century, not only in terms of environmental problems, natural resources and the possibility of a nuclear war apocalypse, but also in terms of a collection of emerging technologies which, unless they are handled properly, may pose risks great enough to make the extinction of humanity a highly conceivable scenario. As physicist Max Tegmark recently said:

> It's now, for the first time in the 4.5 billion years history of this planet, that we are at this fork in the road. It's probably going to be within our lifetimes that we're either going to self-destruct or get our act together. [Harris, Goldstein and Tegmark, 2018, 1:16:40 into the recording.]

Attempts to investigate various scenarios of risks of extinction, as far as such investigations can be done systematically, tend towards two conclusions: (1) the probability that humanity perishes at some time during this century is far from insignificant, and (2) most of the risks involved are risks caused by us and our technologies, rather than from natural causes such as asteroid impacts (Bostrom and Circovic, 2008; Pamlin and Armstrong, 2015; Häggström, 2016). My stance on this is that we are facing a largely unknown territory of potential scientific and technological advancements, many of which may afford mankind prosperity and welfare, while others are deadly threats against our very existence.

It is therefore of utmost importance that we do our best to navigate this potentially fruitful but extraordinarily dangerous territory. Both the academic-romantic and the economic-vulgar viewpoints are tantamount to rushing blindly ahead in this minefield, and they both need to be rejected.

There is admittedly something to the arguments for the consequence neutrality of journalism and the corresponding principle of the academic-romantic about science. I can very well see that the efficiency and credibility of the search for truth fare best by not being disturbed by questions about what truths ought to or ought not to be made public. And surely, the credibility and efficiency of science is important. However, it is…

…not so important that it automatically trumps everything else! As an example of another value that I believe deserves to be considered and be weighed against the academic-romantic ideal, consider the survival of humanity and the prospect of a flourishing future which, if we play our cards properly, may stretch thousands or millions of years ahead, or more.

According to my (perhaps somewhat primitive) view of ethics, every human is obliged to consider the consequences of their actions. This applies to scientists just as it applies to journalists, butchers and flight attendants. For scientists, specifically, I would like to inculcate the following ethical rule:

It is never acceptable to pursue research where the risk that it plunges humanity into misery and extinction exceeds its potential of creating human flourishing and welfare. Neither is it permissible to start a research project without having carefully and honestly considered this issue.

When I talk to other scientists about this, I tend to receive a mixed response. Some think that my ethical rule is obvious, while others are more reluctant. Some starts negotiating: "Such a rule may make sense in applied science, but for basic research, isn´t this too much to ask?"[6] On this issue I am

---

publicly financed mathematical research, its logical conclusion would be that mathematicians are so useless that one must be satisfied if they can be employed with something completely harmless. I do not believe in this conclusion, and let me emphasize: mathematical research without obvious application may very well be worthy of support, but then, not *thanks to* the want of applications but *despite* it.

[6] I will not reveal who said this, but he is highly positioned in Swedish academia.

adamant. I cannot see any good reasons why scientists doing basic research (however these terms are defined) should be exempted from the general human ethical requirement of reflecting beforehand on one´s actions and to try to avoid actions that cause more harm than good. I realize of course that the consequence analysis that I am advocating is difficult, and that we seldom or never can expect answers that are certain. I cannot accept, however, this as an excuse for ignoring the question and for not trying at all.

Please do not misunderstand. A year ago, when I talked to a reporter about these things and the reporter let me have a look at the text prior to publication, I was ascribed the position that "all dangerous research should be forbidden". Good fortune that I was given the opportunity to correct that before printing! The question of what research should perhaps be forbidden and what research can be handled in other ways is difficult, and I am not prepared to reject all research that involve risks, in case the risks are balanced by the potential good.

The most famous historical case where scientists faced such a dilemma is the Manhattan project in the 1940s in which an extraordinarily talented group of physicists developed the A-bomb. The horrible consequences of this scientific discovery – firstly, the bombs over Hiroshima and Nagasaki, secondly, the billions of people during the following decades being held hostage by the death grip of a number of political and military leaders – were not predictable in any detail by the participating scientists, though they were fully aware that they were about to develop a horrible weapon of vast consequences for humanity.[7] The consequences of not pursuing the project appeared, given what was known at the time, horrible as well: The belief was that Nazi-Germany were pursuing a similar project that might give them world dominance if they finished first. Whether it was right or wrong to pursue and participate in the project is a difficult question that I do not need to settle here. I claim however that it would have been wrong to enter the project without even reflecting over the question. Several of the participating physicists have generously shared their ethical considerations. See for instance Dyson (1979), Feynman (1985), Bethe (1991), and Ottaviani and Myrick (2011). I find Richard Feynman´s observations particularly interesting from a psychological point of view. Both before the project and afterwards he wrestled with difficult ethical doubts, while during the project he was so absorbed by it that the ethical issues shuffled into the shadow completely, to the extent that he did not even notice that when Germany capitulated in May 1945 his initial reason to be in the project had disappeared.

What is the Manhattan-project of our time? What are the research areas that we need to look extra carefully at in terms of the future consequences for mankind and the ethical issues involved? I have already clarified that I am reluctant to set up borders that save parts of the territory of research from the imperative of making critical consequence analyses. Still, it is possible to point out areas of research in which such analyses are more urgently needed than, say, in comparative studies of early modern literature. These include some rapidly developing technological areas like bio- and nanotechnology and artificial intelligence (AI), the last of which is the subject of the next section. See also Häggström (2016) for a broader overview of technologies that can be expected to have far-reaching consequences, for better or for worse.

4. The case of AI

No reader can have missed the reports about recent advances in AI technology: new and more powerful apps for our cell phones, a wave of automatization that is about to revolutionize one sector

[7] In addition, they faced the question of whether the first A-bomb blast in the desert of New Mexico, July 16th, 1945 (only a few weeks before Hiroshima) would ignite the whole atmosphere and thereby wipe out humanity and all living creatures on earth. There were theoretical calculations that indicated that that would not happen. There was, however, some remaining disagreement among the participating scientists on whether this issue had been conclusively resolved. See Ellsberg (2017).

after the other, and so on. These bring hopes (justified in my opinion) that innovations in AI and robotics may contribute a large part of the expected economic growth in next couple of decades. In the longer perspective the possibilities are practically without limits, except for those set by the laws of physics. Still, besides these wonderful possibilities there are also great risks. Allow me to highlight some of these risks in this section (which is partly based on Häggström, 2018c).

The time frames for the different risks varies. One risk that is already at our door step involves AI-based image processing, which is very useful in the movie industry. It has, however, a darker side which became evident at the end of 2017 when a collection of pornographic videos was published on the Internet; these gave the deceitful but very realistic appearance of showing some of the world's most famous actresses. They were made using so called face swap – an AI technology by which a person´s face can be switched with someone else's. Soon after that, an app was released which enables anyone to engage in such image processing (Jerräng, 2018). Whether the consequence of this will be a wave of revenge porn and other harmful applications remains to be seen. Thinking optimistically, the problem might solve itself since the access to face swapping enables the victim to claim that such pictures are falsifications. But then again, what will then happen to video evidence presented in court? And how will our ability to distinguish fake news from correct news reports be affected? What will efficient face recognition software do to our personal integrity? Problems are accumulating.

Another problem which is already here concerns the development of military drones and related AI-technology for so called autonomous weapons – or, with a less polite term, killer robots. In the summer of 2015 I joined thousands of scientists in signing an open letter with the title *Autonomous Weapons: An Open Letter from AI and Robotics Researchers,* in which we point out the risks involved in this development and insist that a ban on offensive autonomous weapons beyond meaningful human control. The seriousness of the situation is evident from the following passage in the petition:

> If any major military power pushes ahead with AI weapon development, a global arms race is virtually inevitable, and the endpoint of this technological trajectory is obvious: autonomous weapons will become the Kalashnikovs of tomorrow. Unlike nuclear weapons, they require no costly or hard-to-obtain raw materials, so they will become ubiquitous and cheap for all significant military powers to mass-produce. It will only be a matter of time until they appear on the black market and in the hands of terrorists, dictators wishing to better control their populace, warlords wishing to perpetrate ethnic cleansing, etc. Autonomous weapons are ideal for tasks such as assassinations, destabilizing nations, subduing populations and selectively killing a particular ethnic group. We therefore believe that a military AI arms race would not be beneficial for humanity.

The last sentence is obviously a stark understatement. I usually try to be nuanced and to avoid dead certain conclusions about whether some research is ethically justified or not. In this case, however, I dare to be adamant: Anyone who contributes to the AI arms race of autonomous weapons is making the world worse.

Another consequence of the development of AI that needs to be addressed is what it does to the labour market. Will it lead to so called technological unemployment, that is, unemployment caused by rationalization due to technical advancements? That human work tasks are transferred to machines is not a new phenomenon. We may use Swedish agriculture as an example: While employing 75% of the workforce in the middle of the 19th century, rationalization of the agricultural sector has diminished it to employ only 3% today. But the remaining 72% have largely found

employments in other sectors of the labour market (Schermer, 2017). Whether this phenomenon – that the work force finds new employments in other sectors at about the same pace as they are being rationalized away in an old sector – will last is not something that we can take for granted. There are new circumstances in play, such as not only physical work being automatized today, but also intellectual work. It is not self-evident that we will for all eternity be able to find new tasks in which humans outcompete machines.

The assumption that technological unemployment is inherently bad seems to rest on the further assumption that wage labour is necessary for the good life, which certainly can be questioned. In his song *Maskinerna är våra vänner* (The machines are our friends), Swedish singer-songwriter Kjell Höglund celebrates their taking over our work to give us more time for art, culture, love and amusements. A society with 100% unemployment is a society of equality in a certain sense, but how such a society could be organized is hardly obvious. And once we have specified what we want we will need a plan of implementation, which would presumably have to take us through all intermediate levels of say 20%, 50% and 90% unemployment. How can we pass through these stages without drastically increasing economic inequality, along with the risk for further social instability? These are difficult questions that we do not have clear answers today.

The present unemployment figures should probably not be interpreted as the beginning of an escalating technological unemployment (Alexander, 2018), but the situation may change quickly. The development of autonomous vehicles may come to eradicate a whole work sector in just a couple decades, and similar developments might very well take place in other sectors (Frey and Osborne, 2013; Brynjolfsson and McAfee, 2014).

I have saved the ultimate vision of AI research for last – the creation of an artificial general intelligence (AGI), that is, a machine whose intelligence matches or exceeds that of humanity over the whole spectrum of relevant abilities, including creativity and the ability to think outside the box. Experts disagree heavily about when – if at all – such a breakthrough will come, spreading their estimates all over our present century and beyond (Muller and Bostrom, 2016; Häggström, 2016, 2018b; Tegmark, 2017). In this situation it is wise to be prepared for all possibilities. There is also great disagreement about the consequences of such a breakthrough, though it is a common view among AI futurologists that an AGI breakthrough may ignite an accelerating self-improvement process which very quickly leads to an AGI with so-called *superintelligence*, meaning an intelligence which *vastly* exceeds that of humanity over the whole spectrum of relevant cognitive abilities. This rapid (though so far only hypothetical) dynamics has sometimes been called the *Singularity* and sometimes the *intelligence explosion* (Yudkowsky, 2013; Bostrom, 2014).

Already in 1951, Alan Turing, the father of computer science, seriously discussed the possibility of a future superintelligent machine (Turing, 1951). It was not until 2005, when the author, inventor and futurologist Ray Kurzweil´s book *The Singularity is Near* appeared, that this issue gained broader notice (except in science fiction literature). Kurzweil describes the breakthrough of superintelligence as the final step which will help us humans to liberate ourselves from our frail bodies and to give us everything we could wish for, including the conquest of outer space. Since then the discussion has significantly shifted character, away from Kurzweil´s evangelic tone to a more balanced consideration and the insight that an AI breakthrough may come with great risks such as the risk of total extinction of mankind (Yudkowsky, 2008; Bostrom, 2014).

Turing realized that once a superintelligent AGI is in place and we humans no longer are the most intelligent beings on the planet we will probably not remain in control, and our destiny will be in the

hands of the machines.[8] The crucial issue then is what drives and goals the machines have. The main suggestion for handling this is to make sure, in some way, that the first superintelligent AGI has values that prioritize human wellbeing, and which is generally aligned with human values (whatever that might mean). This project was named *Friendly AI* by Yudkowsky (2008) but is nowadays called *AI Alignment*. For several reasons it is believed to be very difficult. One reason for this is something that all programmers know: when we code we tend to make mistakes, and when there is a discrepancy between our code and our intention the former overrules the latter. Another reason is the basic instability of the outcome as a function of the goal, where seemingly small deviations from the desired goal may have catastrophic consequences (Bostrom, 2014). We also need to take extreme care in deciding what goals we want: a seemingly attractive goal like "maximize the amount of wellbeing and minimize the amount of suffering in the world" would probably be generally good for the universe as a whole but also lead to the extinction of mankind. Our bodies and brains are far from optimal in terms of the amount of hedonistic wellbeing per kilogram matter.

On top of this, we have the potential problem of managing a kind of arms race scenario. If two or more companies or nations compete to become the first to create a superintelligence, the one that tries to find solutions to both the superintelligence problem and the AI Alignment problem will have a considerably harder task than the one that chooses to focus only on creating superintelligence. The more ambitious competitor will risk falling behind in the race. This sort of economic-vulgar logic may result in the AI Alignment problem getting less attention than it needs (Miller, 2012).

I have barely scratched the surface of the emerging and important research area that studies the consequences of an AGI and superintelligence breakthrough and strategies for managing it. A common first reaction of those not familiar with the area is to immediately make up some counterargument to the point that nothing dangerous could happen and decide that the argument is conclusive. I would like to ask any reader who happen to feel an urge in that direction to calm down and instead employ a fair balance between intellectual openness and critical thinking when digesting the exciting ongoing debate. I discuss some of the most tempting counterarguments in Häggström (2018b), and for anyone who wishes to dig deeper I recommend the books by Bostrom (2014) and Tegmark (2017).

5. Conclusions

In this essay I have argued that more precaution is called for in the choice of what research breakthroughs to strive for, and I have claimed that lack of precaution may have fatal consequences. Who is responsible for this precaution and that it is implemented in practice? In Sections 2 and 3 I emphasized the responsibility of the individual scientist. I claim, however, that it would be foolish of society to rely solely on that. Among scientists, whether they lean towards the academic-romantic viewpoint or the economic-vulgar, unawareness of these problems is large, as is their creativity in finding ways to evade moral responsibility.

In August 2015, not long after the publishing of the open letter about autonomous weapons that I quoted in Section 4, I attended a talk by computer scientist Patrick Doherty about his fascinating research on AI technology for drones. The intended application was non-military. Since the letter was so recent and had achieved considerable attention Doherty felt that it was necessary to comment on

[8] Some thinkers have tried to avoid this conclusion. Most of the arguments are variations on "We can always pull the plug" and tend to be utterly naïve (Häggström, 2014). Somewhat more promising is the so called AI-in-a-box approach, in which the machines are to be held isolated from the world outside except for a narrow and carefully controlled communications channel. The tentative conclusion so far is that this sort of solution can only function over at most a short transitory period (Armstrong, Sandberg and Bostrom 2012; Häggström, 2018a).

it, saying that he had "not signed the letter, as no technologies are good or evil; only the uses of a technology can be good or evil" (quoted from memory). To abandon all responsibility with this kind of sweeping statement, and to withdraw from any concern about whether one's research might lead to a situation in which terrorists have access to a devastatingly forceful technology, is just not acceptable in my view. I would not even call it a thought or a point of view; it is merely a simplistic slogan devised to shield the scientist from having to think at all.

As an even more unabashed example – frightening but impressively honest – we may consider a statement by AI researcher Geoffrey Hinton. In an interview in *The New Yorker* (Khachaturian, 2015) Hinton is strongly pessimistic about the societal consequences that the technology that his research aims to develop will have, saying that it will probably be used by the state to suppress the people. To the question why then he is conducting this research, Hinton answers that he could have given the "the usual arguments" but that the truth is that "the prospect of discovery is so *sweet*".

Doherty and Hinton are no rare outliers in the scientific community. Feynman´s psychological experience mentioned in Section 3 is not uncommon either. Scientists are only human, and they are therefore simply not to be trusted when it comes to the consequence ethics I find imperative. While scientists should not, of course, be relieved of responsibility, other actors need to step in and take their share of responsibility: universities, technology companies, research foundations, investors, politicians, media and ordinary citizens. This responsibility is not taken sufficiently seriously today, except very sporadically.

I do not have a readymade opinion about whether a new authority should be established to preside over what research should be allowed in view of the possible consequences for the future of humanity. I realize of course that to give a single governing body such power is problematic. But even if we postpone such institutional issues, there is a lot to be gained if all the actors mentioned in the last paragraph (each of which, in one way or another, has some influence over research) took seriously the question about the long-term societal consequences and risks. Of course, it is not an easy thing for these actors to make well-informed consequence analyses and risk estimations. However, if a public authority were to be established with the task of producing well-balanced reports of the state-of-the-art concerning the societal consequences of new and future technologies – for instance, following the example of the UN panel IPCC (Intergovernmental Panel on Climate Change) – it might be a great help. That is at least what I think. What we on all accounts should not do to continue the present course of ignoring the problem.

Bibliography

van Aken, J. (2007) Ethics of reconstructing Spanish Flu: Is it wise to resurrect a deadly virus? *Heredity* **98**, 1-2.

Alexander, S. (2018) Technological unemployment: much more than you wanted to know, *Slate Star Codex*, February 19.

Armstrong, S., Sandberg, A. and Bostrom, N. (2012) Thinking inside the box: controlling and using an oracle AI, *Minds and Machines* **22**, 299-324.

Bethe, H. (1991) *The Road from Los Alamos*, American Institute of Physics, New York.

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

Bostrom, N. and Cirkovic, M. (2008) *Global Catastrophic Risks*, Oxford University Press, Oxford.

Brynjolfsson, E. and McAfee, A. (2014) *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*, W.W. Norton, New York.

von Bubnoff, A. (2005) The 1918 flu virus is resurrected, *Nature* **437**, 794-795.

Dyson, F. (1979) *Disturbing the Universe*, Harper and Row, New York.

Eliasson, P.-O. (2006) Ny strategi behövs för nyttoperspektiv, *Universitetsläraren*, no. 9.

Ellsberg, D. (2017) *The Doomsday Machine: Confessions of a Nuclear War Planner*, Bloomsbury, New York.

Feynman, R.P. (1985) *Surely You're Joking, Mr. Feynman? Adventures of a Curious Character*, W.W. Norton, New York.

Fichtelius, E. (2016) Journalister ska inte ta hänsyn till konsekvenserna, *Svenska Dagbladet*, January 16.

Frey, C.B. and Osborne, M. (2013) The future of employment: how susceptible are jobs to computerisation?, preprint, http://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf

Häggström, O. (2006a) Angående attityder inom vetenskapen, *Svenska Matematikersamfundets Medlemsutskick*, 1 februari.

Häggström, O. (2006b) Ett anspråkslöst förslag rörande svensk forskning, *Tentakel,* no.7.

Häggström, O. (2014) Om Singulariteten i DN, *Häggström hävdar*, June 22.

Häggström, O. (2016) *Here Be Dragons: Science, Technology and the Future of Humanity*, Oxford University Press, Oxford.

Häggström, O. (2018a) Strategies for an unfriendly oracle AI with reset button, to appear in *Artificial Intelligence Safety and Security* (ed. R. Yampolskiy), CRC Press, Boca Raton, FL.

Häggström, O. (2018b) Remarks on artificial intelligence and rational optimism, *Should we Fear Artificial Intelligence?*, The EU Parliament's STOA Committee, Brussels, pp 19-26.

Häggström, O. (2018c) AI-utvecklingen och dess yttersta risker, to appear in *Livet med AI*, Stiftelsen för Strategisk Forskning, Stockholm.

Hardy, G.H. (1940) *A Mathematician's Apology*, Cambridge University Press, Cambridge, UK.

Harris, S., Goldstein, R. and Tegmark, M. (2018) What is and what matters, *Waking Up Podcast*, March 19.

Jerräng, M. (2018) Deepfakes är det läskigaste på nätet just nu – och ett tydligt exempel på riskerna med AI, *ComputerSweden*, January 31.

Khatchadourian, R. (2015) The doomsday invention, *The New Yorker*, November 23.

Klein, G. (1998) *Korpens blick: Essäer om vetenskap och moral*, Bonniers, Stockholm.

Kurzweil, R. (2005) *The Singularity Is Near: When Humans Transcend Biology*, Viking, New York.

Marr, B. (2017) Another example of how artificial intelligence will transform news and journalism, *Forbes*, July 18.

Miller, J. (2012) *Singularity Rising: Surviving and Thriving in a Smarter, Richer, and More Dangerous World*, Benbella, Dallas, TX.

Müller, V. and Bostrom, N. (2016) Future progress in artificial intelligence: A survey of expert opinion. *Fundamental Issues of Artificial Intelligence,* Springer, Berlin*,* pp 553-571.

Ottaviani J. and Myrick, L. (2011) *Feynman*, First and Second, New York.

Pamlin, D. and Armstrong, S. (2015) *12 Risks That Threaten Human Civilization*, Global Challenges Foundation, Stockholm.

Pinker, S. (2018) *Enlightenment Now: The Case for Reason, Science, Humanism, and Progress*, Viking, New York.

Rosling, H., Rosling, O. and Rosling Rönnlund, A. (2018) *Factfulness: Ten Reasons We're Wrong About the World – and Why Things Are Better Than You Think*, Flatiron Books, New York.

Russell, S. et al (2015) *Autonomous Weapons: An Open Letter from AI and Robotics Researchers*, Future of Life Institute.

Schermer, I.G. (2017) Strukturförändringar i sysselsättningen, *EkonomiFakta*, https://www.ekonomifakta.se/Fakta/Arbetsmarknad/Sysselsattning/Strukturforandringar-i-sysselsattningen/

Tegmark, M. (2017) *Life 3.0: Being Human in the Age of Artificial Intelligence*, Brockman Inc, New York.

Turing, A. (1951) Intelligent machinery: a heretical theory, BBC, http://philmat.oxfordjournals.org/content/4/3/256

Yudkowsky, E. (2008) Artificial intelligence as a positive and negative factor in global risk, pp Bostrom and Cirkovic (2008), s 308-345.

Yudkowsky, E. (2013) *Intelligence Explosion Microeconomics*, Machine Intelligence Research Institute, Berkeley, CA.