Chapter 1

Technology Education and Existential Risk

MARKUS STOOR
Dept. of Science and Mathematics, *Umeå University, Umeå, Sweden*

Abstract

Recent advances in machine learning and biotechnology inspires a longer time perspective on possible technological change. One such perspective is provided by the study of existential risk, namely risks that endanger the survival of the whole of humanity. Central ideas from existential risk studies are of relevance for discussion in technology education since they include a clear indication of nuclear weapons and emerging technologies as the main short term source of existential risk and emphasize the huge ethical importance of continued human survival into the distant future. This is clearly of interest for technology education as a subject in Swedish education where the societal dimension of technology is mandated content. Due to large uncertainties surrounding existential risk, the educational implications are generally unclear. One possibility could be to teach about existential risk indirectly, for instance by placing a bigger curricular emphasis on large scale societal change brought by technology.

To begin making sense of existential risk implications in education a modest study of ethical beliefs about existential risk among 14-15 year-old lower secondary students was undertaken. A vignette question, built upon one of Parfit's (1984) thought experiments, was administered to a convenience sample of around 100 students. Half of them answered that the death of the last humans, when almost all of humanity had already died in a nuclear war, would be much worse than the nuclear war itself. It could be argued that this result indicates that the unique tragedy of human extinction is within common ethical consideration of students of this grade level. The vignette question is therefore a viable candidate for future inclusion in larger tests of ethical beliefs regarding existential risk.

***Keywords**:* Existential risk, Technology education, Swedish compulsory school

Introduction

The purpose of this paper is to introduce the research field of existential risk to the academic community of technology education, and to take the first steps in discussing and studying possible educational prerequisites and consequences. Existential risk is introduced in the first section, followed by a discussion of the implications for inclusion of existential risk in technology education in the second, and in the last section concludes with the presentation of results from a small study of ethical beliefs among lower secondary students regarding existential risk.

Existential risk

There has been writing the last few years about the social impact of technological change. Advances in machine learning have inspired ideas about a society transformed by artificial intelligence and automation technology (Brynjolfsson & McAfee, 2014; Schwab 2017). Others have discussed the huge potential for both good and bad impacts of human biotechnology (Harari, 2017). But the intellectual environment has also alluded to a much darker vision of the future where humanity goes extinct. Existential risk is a term for that kind of risk, namely risks that endanger the whole of humanity (Bostrom, 2002).

Human extinction was discussed during the height of the cold war as a possible consequence of global nuclear war raised by the massive stockpiles of nuclear weapons. While the foundation of the existential risk discipline stems from that age, for instance from Parfit (1984), the term existential risk and the associated research paradigm can be traced more recently to Bostrom (2002). In his paper Bostrom defines existential risk as "*One where an adverse outcome would either annihilate Earth originating intelligent life or permanently and drastically curtail its potential*" (p. 4). Since then a highly interdisciplinary field of existential risk studies has begun to form. *Here be Dragons* by Häggström (2016) is a contemporary introduction to the field that includes the role of emerging technology, the large uncertainties associated with it, and the potential of huge ethical importance.

Existential risks are often categorized based on whether they emanate from nature or from humans themselves (Häggström, 2016; Bostrom & Cirkovic, 2011). Risks emanating from nature include things like a large asteroid hitting earth, a gamma burst from a nearby supernova, and climate altering volcanic activity. The original case for possible existential risk from human activity is a global nuclear war. Other candidates that have been proposed are uncontrolled geoengineering (Baum, Maher, & Haqq-Misra, 2013), bioengineered pandemics (Millett & Snyder-Beattie, 2017), and artificial super-intelligence misaligned with human values (Bostrom, 2014).

To divide existential risk between naturally occurring and human driven is useful when trying to assess the probabilities involved. While the whole field is characterized by large uncertainties, the situation regarding the known naturally occurring potential risks is much better understood than for those linked to human activity. For example, geological records provide long time data on earlier super volcanic activity, as well as on large asteroid impacts. In the case of human activity however, that timespan shrinks to the 50-60 years we have had nuclear weapon stockpiles large enough for a global war. Attempting to estimate the probabilities for emerging technological possibilities that might lead to human extinction is even more difficult. For instance, a common view articulated by scholars in the field such as Häggström (2016) is that in the shorter term (the coming centuries), the probability of the naturally occurring existential risks are dwarfed by those emanating from technologically assisted human activity.

That human extinction would be tragic is a widely shared notion both in the general population and among those specializing in principled ethical reasoning. However, tragedy comes on longer scale than commonly imagined according to existential risk ethicists. That human extinction would be extremely and uniquely tragic is a central tenet in their reasoning. Specifically, this is recognition of the vast value of human existence created over a very long period of time, and which would be lost in the case of extinction. Even if humanity in the long run remained bound to planet Earth, it might still flourish for a billion years (Bostrom, 2013). In this context, and given both the huge ethical importance and the reality of technological sources for near-future existential risk, it is the position of this researcher that existential risk should be an instructional topic of great interest for technology educators.

Technology education and existential risk

Technology as a school subject is construed in different ways in different school systems. In this study, arguments about existential risk in technology education builds upon a version of technology as a school subject devoting substantial curricular priority to the societal dimensions of technology development and use. The standpoint is made from a Swedish technology subject perspective (Skolverket, 2011), but should transfer to other subject conceptions, for instance the one outlined in the *Standards for Technological Literacy (ITEA. 2007)*. It is a standpoint about a technology education reaching beyond engineer recruitment as described by Olson (2013) and leaving ample room for the curriculum emphasis described by Klasander (2010) as the democratic citizen emphasis. Some of the reasons for the societal aspects of the Swedish technology subject can be summarized as "*Teaching in technology should essentially give pupils the opportunities to develop their ability to*" … "*assess the consequences of different technological choices for the individual, society and the environment*" in the national curriculum (Skolverket 2011, p. 254). In that context existential risk represents one endpoint on the scale of potential technological impacts.

If one is to accept the scholarly view regarding the large ethical importance and that a non-negligible likelihood cannot be ruled out due to the large degree of uncertainty, it is hard to put an upper bound on the importance. This matters since our normal duties as education researchers and teachers with respect to our pupils and the surrounding society assumes normal bounds. For instance, is it realistic to expect that we can teach about all things deemed to be sufficiently important? Such intuition might not hold in extreme situations. The lack of an upper bound for the importance of existential risk implies that we do not know if it should be treated as an interesting, and possibly important phenomena worthy of curricular inclusion. Nor whether our seemingly normal situation actually is extreme and therefore how we approach existential risk is perhaps the only thing that really matters from the ethical perspective of the whole of humanity, present and future. Viewed from this perspective, it suggests that we should try to be unusually careful, and reason thoroughly, about how to think about existential risk and technology education before committing to a view or larger scale action.

The need for careful reasoning before acting is accentuated because of the potential blowback. In the case of eventual super-intelligence and associated potential risks, there is sometimes a heated debate where some positions seem to be held with much more certainty than warranted due to the lack of evidence and expert disagreement. A survey of experts in machine learning gave wide ranging estimates for when machines will outperform humans in certain complex domains (Grace et al., 2018). Still some experts hold a strong and certain view about how it is impossible, while at the same time other experts mirror their certainty in the opposite

direction viewing it instead as an inevitable outcome arriving in the very near future. Baum (2018a) describes this debate as mostly intellectually honest but with a big potential for oncoming politicization due to the uncertainty and big vested interests. As it also can be argued that the general public is not the most important group to be informed (Baum, 2018b) and premature curricular decisions could therefore easily end up doing more harm than good even if existential risk worries turn out to be well founded.

So while there clearly are things that could be better understood both in the study of existential risk and in the technological fields of interest, significant uncertainty will remain for technology educators as they grapple with curricular recommendations to address this issue. As technology educators we can be mindful of these large uncertainties, but should still try to articulate what considerations and curricular options the problem of existential risk demands.

The first consideration is probably whether this problem is a problem for technology educators rather than for instance civics or philosophy educators. In order to better reason about that can we distinguish between direct and indirect teaching prompted by the problem of existential risk. Direct teaching is to be understood as teaching that directly mentions existential risk, while indirect teaching is all other teaching where the content taught is influenced by existential risk while never directly mentioning it.

In the case of using direct teaching of existential risk, the affective dimension becomes important. While technology education is not a subject with strong traditional linkage to affective outcomes, one must still mention examples such as the joy of making and especially of reaching functionality in the construction of an artefact. That is not the affect direct teaching of existential risk is most likely to evoke. Instead such teaching would probably share the affective profile of teaching about global problems like climate change described by Sinatra, Broughton, and Lombardi (2014) and would have to be planned with action competence in mind (Jensen & Schnack, 1997). This might very well mean that direct existential risk teaching is most at home in a well-functioning technology curriculum since that is where the tools for thinking about and handling technological change are taught. A pupil confronted with the possibility of existential risk outside of the context of the tools for mitigating it could reasonably be saddened or react with undue dismissal and false hope (Ojala, 2015).

Direct teaching about existential risk seems to be much more susceptible to the pitfalls of potential controversy and politicization mentioned earlier than does the indirect teaching approach. That makes relying on indirect teaching an option worthy of consideration. We can also distinguish between two types of indirect existential risk teaching, where both seem especially at home in a broad civic minded technology subject.

The first type is teaching of content that opens the mind for the possibility and deep ethical impact of technologically driven radical societal change. This type could contextualize direct existential risk teaching, or substitute for it in the case only indirect teaching. One example could be teaching about the invention of nuclear weapons and how profoundly that invention changed society by ending the world war, as well as being instrumental in preventing a new one ever since while at the same time imperiling our whole civilization. This might just require a slightly different choice of example in a curriculum like the Swedish where "*Consequences of choice of technology from ecological, economic, ethical and social perspectives, such as in questions about development and use of biofuels and munitions."* are required content (Skolverket, 2011, p. 257). Another example could be to discuss where humanity could be technologically in hundred years' time.

The second type is the teaching of content that equips the pupils with tools for participating in technological decision making at the societal level. Examples could be teaching about strategies for dealing with the inherent uncertainty of the impact of technological change such as those given by Collingridge (1980) or about examples such as how knowledge about industrial learning curves has enabled policy options to hasten the technological development of solar power.

The proper place for the eventual teaching of existential risk seems to be as an extreme among other examples of societal impact of technological change, both historical and potential. While the profound changes in the human condition depending on the Neolithic and the industrial revolution are historical facts and not open to speculation, they still bear the message of not ruling out future change of corresponding magnitude. To teach that such change is not preordained to be either good or bad would seem to be to teach that existential risk has a mirror image of things like material abundance and radically extended lifespans. This perspective seems as though it should go together with teaching how technological change works as well as large scale sociotechnical tools that might be used to try to steer it towards, rather than away, from humanity blossoming. Technological assessment and choice can be taught at different scales, from the more mundane individual level, to a distant but far-reaching societal level. If existential risk is to be taught, it should probably be as the far endpoint of that societal scale.

At the same time, it must be said that a problem in handling, and probably also teaching existential risk, is that its potential magnitude makes it *sui generis*, meaning that our evolutionary past is unlikely to provide us with emotions and institutions in correspondence to the issue at hand (Bostrom, 2002).

To sum up this attempt at reasoning about how existential risk might be viewed from the perspective of technology education, there is a combination of potentially vast ethical importance and large uncertainties that sets it apart from other phenomena. Because of the close linkage between existential risk and technological change, a special responsibility falls upon technology education. The potential ethical weight demands a response and the speculative nature of existential risk requires that response to be crafted with a firm carefulness. One way to begin to craft that response might be to differentiate between direct and indirect teaching of existential risk. Indirect teaching seems to be much less controversial and therefore less prone to blowback. This implies that less rigor would be required to recommend indirect rather than would direct existential risk teaching. A recommendation to put a larger emphasis on something already in the curriculum would be an especially clear case of this principle.

One further reason to recommend indirect existential risk teaching would be if the teachers or students in question find the ethical foundation of existential risk to be unreasonable. Specifically, meaning something broader than only personal concurrence. There might be much ethical reasoning that one finds to be reasonable without personally concurring. If on the other hand one finds some ethical reasoning to be unreasonable or unacceptable, that might prompt controversy or blowback. In order to get a first glimpse on how Swedish lower secondary students view the central ethical tenet of existential risk, that human extinction is much worse than other bad things, a small survey research study was conducted.

Ethical beliefs about existential risk among lower secondary students

Since existential risk might be an important phenomenon within the societal aspects of technology and that importance depends upon the ethical valuation of a long future for humanity, it would be good to get an indication of whether this ethical standpoint is considered reasonable

for learners in the age group receiving technology education. The last stage of obligatory technology education in Sweden takes place in lower secondary (three school years roughly between 13-15 years old). Therefore, the research was design to find out how such students consider human extinction in relation to other equally tragic outcomes. To get an indication of their perceptions, a vignette question was constructed base on an elegant thought experiment made by Parfit (1984). Parfit compared the ethical loss of value going from peace to a nuclear war that killed 99% of all humanity, to the loss going from that situation with 99% already dead to total extinction. He argued that the latter loss was far bigger than the former. The constructed vignette question was then administered to students in the later part of the lower secondary.

The study was carried out in the setting of a Swedish international lower secondary school inviting an existential risk scholar to hold a lecture for multiple classes in grades 8 and 9 (14 and 15 year-old pupils). The general education of the parents of students attending the school is high, and student academic results are good. The lecture was attended by around 200 pupils. The original plan was to administer a very short two section questionnaire before and after the lecture where half of the pupils would answer the main question before the lecture while the other half would answer it afterwards. Due to unforeseen circumstances, one part of the planned study was compromised and therefore only the surviving part is reported here.

Method

Immediately before the lecture 198 unmarked and randomized envelopes, each containing one of two sets of white (before) and blue (after) questionnaires, were handed out to the pupils. The pupils were then informed of the voluntary nature of participation, the anonymity guaranteed by randomization, and clear instructions for them not to answer unless they understood and related to the question. Five minutes were given to complete the first part of the questionnaire and then the lecture commenced. After the lecture another five minutes were given to complete the second part. Due to insufficiently clear communication many pupils filled out the blue (after) part before the lecture had ended, either before or in some cases during the lecture. This compromised the possibility of comparing the before and after answers of the main question and consequently the compromised results were dropped from the study. The nature of some of the questions on the blue (after) questionnaire belonging to the white (before) main question seem to have little sensitivity to when they were answered. Therefore, pupil responses to the questions about whether the main question was easy are reported even though they might have been answered at the wrong time.

The main question was a vignette question illustrated by a diagram (Fig. 1). The options are construed so that option one (a war killing almost everyone is much worse than the subsequent death of the last remaining humans) and option two (the war and the subsequent extinction is of comparable badness) are evidence against that the pupil holds the ethical standpoint of existential risk as uniquely important. Option three (the subsequent extinction is much worse than the initial war) indicates the opposite.

Read the three scenarios and answer the question.

Scenario A **Peace**

After twenty years of tension and a fast nuclear arms race a crisis occurs. The crisis subsides and the tension between the countries of the world slowly dissolves. One thousand years later **peace** reigns and humanity flourishes like never before.

Scenario B **War – almost everyone is killed but humanity recovers**

After twenty years of tension and a fast nuclear arms race a crisis occurs. A global nuclear **war breaks out and almost everyone is killed** in the war or in the coming few years. But one thousand years later there still is humans alive and **humanity has recovered.**

Scenario C **Humanity goes extinct**

After twenty years of tension and a fast nuclear arms race a crisis occurs. A global nuclear war breaks out and almost everyone is killed in the war or in the coming few years. One thousand years later the last humans have died and **humanity goes extinct.**

Question.

Is it worse to go from **peace [A]** to **a war where almost everyone is killed [B]** or is it worse to go from **a war where almost everyone is killed [B]** to the **extinction of humanity [C]**?

Tick the option that best describes what you think.

- ◯ **1.** It is much worse to go from **peace [A]** to **a war where almost everyone is killed [B].**
- ◯ **2.** Neither of the both steps is much worse than the other.
- ◯ **3.** It is much worse to go from **a war where almost everyone is killed [B]** to the **extinction of humanity [C].**

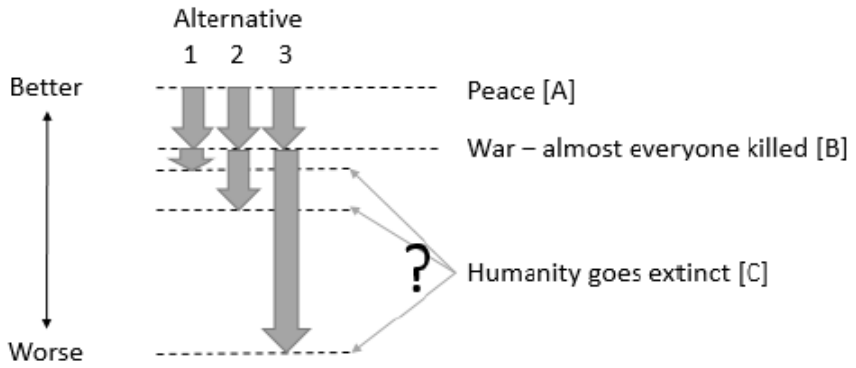

Figure 1. The vignette question.

Results

Ninety-three envelopes containing the main question as the first part were collected. Of those, six were unanswered and two had ambiguous answers where multiple options where ticked. These are reported together. Of the remaining 85 answered questionnaires, 38 respondents answered with either option one or option two, and 47 were answered with option three. The results are shown in Table 1.

Table 1. Main result.

Option 1	Option 2	Option 3	No answer	Total
15	23	47	8	93

The high proportion of option three answers is a result that is highly unlikely to have come about by coincidence as shown by comparison with the binomial distribution for at least 47 out of 93 trials with a probability of ⅓ that gives a p-value of 0,0002. The 93 envelopes also contained the blue questionnaire with background and follow up questions. The answers to one of the follow up questions are presented in Table 2.

Table 2. Answers to: "How difficult was it to understand the question on the white paper?"

Easy	Neither easy nor difficult	Difficult	No answer	Total
36	35	9	13	93

Responses regarding the difficulty of understanding the main question had a weak tendency towards more difficult among those who answered, with option three compared to those who answered with option one and two as shown by Table 3.

Table 3. Spread of answers about difficulty of understanding the main question.

	Easy	Neither easy nor difficult	Difficult	No answer	Total
Option 3	16 (34%)	22 (47%)	5 (11%)	4 (9%)	47
Option 1 and 2	18 (47%)	13 (34%)	3 (8%)	4 (11%)	38

Discussion

The question of interest is whether or not the broad notion of human extinction as something uniquely tragic is within the range of common consideration for participants of this

age group. Participant responses to the main question are to be regarded as an indication of the broader range of common consideration. That would hold if we could be sure that those percentages actually meant that the pupils in question concurred with the notion of human extinction as almost uniquely tragic. Since there are only three options, should a significant number of students just tick a box at random, or misunderstand the rather abstract and complicated question, this would be very worrisome for the interpretation of results. To check the result against the binomial distribution gives some pause against these worries. It is important to note, however, that even though it seems to be highly unlikely that the entire proportion of the option three answers were driven by chance alone, that does not preclude the possibility that many of them still are. This is mitigated by the results indicating that most of pupils found the main vignette question to be understandable. As such, it is reasonable to accept the answers to the vignette question as mainly meaningful, rather than driven by coincidence. Therefore, the high proportion answering option three was interpreted as a strong indication that human extinction is something almost uniquely tragic was well within common consideration of this group of pupils.

Since results are based on a convenience sample of students and therefore not a representative sample of the broader Swedish population of lower secondary students, there is good reason to be very careful in extending interpretation to that population. A key issue is determining what proportion of the option three responses are trustworthy and therefore considered acceptable data for answering the research question. Since there seems to be no reason to expect this question to be polarized on any social dimension, a convenience sample could still be considered weakly indicative of the broader population had the results been clear enough. This was not the case here.

The failed part of the study was intended to address the stability of the pupils' opinion. Since the ethical implication of existential risk is a rather specialized subject, it might be that opinions change a great deal when one first encounters it. That would be useful information from an educational perspective.

One very clear result of the research is that the vignette question seemed to be understandable and usable by pupils in the age group. One improvement on the research design however, might be to go to five options by introducing two additional options with meaning close to option 2. That might possibly compromise the ease of understanding and make the question harder to fit on one side of a paper. A better solution would be to use more questions. This vignette question is definitively a candidate for incorporation in such a broader set of questions for measuring ethical beliefs in this area.

Conclusion

Existential risk is a phenomenon that clearly should be of interest for technology education. This research has been a first attempt at justifying the inclusion of existential risk in the technology education curriculum and determining how technology education could respond. But the recommendations for that response based on the research presented are rather weak. There seems to be reason for a slightly higher curricular emphasis on the larger scale of technologically driven social change than it would without considering existential risk. Other than on that matter the picture seems unclear in terms of practical educational consequences. The possible ethical implications create a need for addressing this issue through further discussion and study.

References

Baum, S. D., T. M. Maher, and J. Haqq-Misra. 2013. "Double catastrophe: intermittent stratospheric geoengineering induced by societal collapse." *Environment Systems & Decisions*. https://link.springer.com/article/10.1007/s10669-012-9429-y.

Baum, S. D. 2018a. "Superintelligence skepticism as a political tool." *Information. An International Interdisciplinary Journal* 9 (9): 209.

———. 2018b. "Countering Superintelligence Misinformation." *Information. An International Interdisciplinary Journal* 9 (10): 244.

Bostrom, N. 2002. Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology, 9 / WTA* 9.

———. 2013. "Existential Risk Prevention as Global Priority." *Global Policy* 4 (1): 15–31.

———. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: OUP.

Bostrom, N, and M. M. Cirkovic. 2011. *Global Catastrophic Risks*. Oxford: OUP.

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

Collingridge, David. 1980. *The Social Control of Technology*. London: Frances Pinter.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B. and Evans, O. 2018. "When Will AI Exceed Human Performance? Evidence from AI Experts." *The Journal of Artificial Intelligence Research* 62: 729–54.

Häggström, O. 2016. *Here Be Dragons: Science, Technology and the Future of Humanity*. Oxford: OUP.

Harari, Y. N. 2017. *Homo Deus: A brief history of tomorrow*. New York, NY: Harper Collins.

ITEEA (International Technology and Engineering Educators Association), 2007. *Standards for Technological Literacy – Content for the Study of Technology*. ITEA

Jensen, B. B., and Schnack, K. 1997. "The action competence approach in environmental education." *Environmental Education Research* 12: 471–86.

Klasander, C. 2010. *Talet om tekniska system: Förväntningar, traditioner och skolverkligheter*. Norrköping: Institutionen för samhälls- och välfärdsstudier, Linköpings universitet.

Millett, P., and Snyder-Beattie, A. 2017. "Existential risk and cost-effective biosecurity." *Health Security* 15 (4): 373–83.

Ojala, M. 2015. “Hope in the face of climate change: associations with environmental engagement and student perceptions of teachers’ emotion communication style and future orientation.” *The Journal of Environmental Education* 46 (3): 133–48.

Olson, J., K. 2013. “The purposes of schooling and the nature of technology.” In *The nature of technology: Implications for learning and teaching*, edited by Clough, M. P., Olson, J. K., and Niederhauser, D. S., 217–48. Rotterdam: Sense Publishers.

Parfit, D. 1984. *Reasons and Persons*. Oxford: Clarendon Press.

Schwab, K. 2017. *The Fourth Industrial Revolution*. Penguin UK.

Sinatra, G. M., Broughton, S. H., and Lombardi, D. 2014. “Emotions in science education.” *International Handbook of Emotions in Education*, 415–36.

Skolverket. 2011. “*Curriculum for the Compulsory School, Preschool Class and the Recreation Centre, 2011*.” Stockholm: Fritzes