Extinction risks and resilience: A perspective on existential risks research with nuclear war as an exemplary threat

by Johannes Kattan

A growing awareness of potential global catastrophes has recently given increased attention to the topic of existential risks. To date, there is still very limited consensus on the definition of existential risk, the likelihood of those risks, and their ethical implications. To achieve more clarity, it is proposed here that extinction risks should be discerned more clearly from other aspects of existential risks. Nuclear war is taken as a prime example to illustrate an extinction risk and to discuss humanity's resilience to such threats. It is concluded that it is unlikely that a nuclear war would lead to the end of the human species, despite the unprecedented damage it might cause. Further, some of the ethical aspects of longtermism and the communication of existential risks are discussed.

Keywords: *Extinction risks; existential risks; nuclear war; resilience factors; longtermism*

To achieve more clarity, it is proposed here that extinction risks should be discerned more clearly from other aspects of existential risks. Human extinction is an outcome that can be precisely defined in biological terms. It should be analysed separately from scenarios in which the subjective quality of human life is the concern.

Defining existential risk

Events in the last decade have led to an increased awareness of the dangers emanating from climate change, global pandemics, and the escalating tensions between nuclear superpowers. As a consequence, the study of existential risks has gained increasing attention, visibility, and funding (Cremer/Kemp 2021), and perhaps even run the risk of increasing harm. We highlight general challenges in ERS: accommodating value pluralism, crafting precise definitions, developing comprehensive tools for risk assessment, dealing with uncertainty, and accounting for the dangers associated with taking exceptional actions to mitigate or prevent catastrophes. The most influential framework for ERS, the "techno-utopian approach" (TUA). Its goal is to identify threats to humanity as well as their causes, implications, and respective countermeasures. An unsolved issue here is a missing consensus on what constitutes an existential risk (Steinmüller/Gerhold 2021). Toby Ord, currently among the most influential representatives of the field, has offered the following definition:

"An existential risk is a risk that threatens the destruction of humanity's long-term potential" (Ord 2020: 39).

This definition is very concise and intuitive, but the notion of human potential is vague and open to individual interpretation. However, this can be considered a necessary trade-off. What is set here as the potential of humanity is synonymous with what we deem desirable for our existence and future. A more determinate concept would amount to dictating a moral imperative for society. Without an authority or a collective agreement on the matter, what is desirable remains in the first instance a personal matter. In this sense, the term might act as a wildcard for the value embodied in humanity and its future as such; a value which might never reach a final shape. Nevertheless, this open-ended approach has been criticised (Friederich/Aebischer 2021; Cremer/Kemp 2021). First, it appears a difficult task to preserve something of which it is not clear what it is. Second, such definitions are too abstract to allow for robust analysis. Third, in the work of Ord, humanity's potential is not always expressed in a value-neutral way but along what Cremer and Kemp (2021) deem techno-utopian terms. These concepts are currently rather dominant in the discussion of existential risk, and we will consider some of their ethical implications later. Problematic here is that the subjects of global catastrophe and human extinction might be conflated with those specific moral ideas. Therefore, Cremer & Kemp (2021) suggested separating the study of existential risks into the areas of Extinction Ethics, Existential Ethics, atastrophic Risks, and Extinction Risks. I deem this a reasonable proposal. Human extinction is an outcome that can be precisely defined in biological terms. It should, if possible, be analysed separately from scenarios in which the subjective quality of human life is the concern. This would facilitate analysis and communication.

Existential threats

Existential threats can be divided into those that stem from the actions of humanity itself, called anthropogenic risks, and those that originate from conditions beyond the control of humanity, termed natural risks. Examples of natural threats are the impacts of major asteroids, massive volcanic eruptions, or gamma-ray bursts from stellar explosions (Ord 2020: 62-72; Steinmüller/Gerhold 2021). Luckily, the risk that any of these threats will trigger an extinction event in the near future can confidently be set as extremely low. The chances of natural catastrophes have remained rather constant over time. If they had a moderate likelihood, then the chances for *Homo sapiens* and its predecessors to have survived would be close to zero. Taking the age of humanity and the extinction rates of other mammals and hominid species into account, the upper bound for the annual probability of human extinction from natural causes was estimated to be lower than 1 in 870,000 (Snyder-Beattie et al. 2019). While natural risks have stayed almost constant over the span of human history, anthropogenic risks have not. Since the first detonation of an atomic bomb, several man-made risk scenarios have emerged, and even more may be revealed in the future. The most prominent anthropogenic risks and their estimated likelihood to threaten humanity's potential, according to Ord, are listed in Table 1. The fact that numbers are attached to the subject does not imply that any reliable statistical analysis of the risk has been

achieved and the reader should consciously correct for the human tendency to associate numbers with accuracy here. Partly due to such propensities, the presentation of concrete numbers regarding such risks has been criticised (Torres 2021; Cremer/Kemp 2021). Nevertheless, I would suggest that despite justified worries, it is still more useful to present these numbers with a warning than to rely solely on descriptive terms such as "very unlikely", which invite diverging interpretations and can in this context be close to meaningless. Moreover, such numbers allow for a more effective critique and discussion of estimated likelihoods.

Table 1: Estimates for the chances of an existential catastrophe curtailing humanity's potential

Existential catastrophe via	Chance within next 100 years
Asteroid or comet impact	~ 1 in 1,000,000
Supervolcanic eruption	~ 1 in 10,000
Stellar explosion	~ 1 in 1,000,000,000
Total natural risk	**~ 1 in 10,000**
Nuclear war	~ 1 in 1,000
Climate change	~ 1 in 1,000
Other environmental damage	~ 1 in 1,000
"Naturally" arising pandemics	~ 1 in 10,000
Engineered pandemics	~ 1 in 30
Unaligned artificial intelligence	~ 1 in 10
Unforeseen anthropogenic risks	~ 1 in 30
Other anthropogenic risks	~ 1 in 50
Total anthropogenic risk	**~ 1 in 6**
Total existential risk	**~ 1 in 6**

Adapted from Ord (2021: 140). The author noted that due to uncertainties some estimates might easily be off by three orders of magnitude.

According to Ord, the threat to human potential emanating from human progress is estimated to be far higher than the one originating from natural risks. Moreover, even within anthropogenic risks, almost all the risk stems from a few risk factors, with unaligned artificial intelligence alone being responsible for more than half of the total existential risks. It should be noted that if only human extinction would be used as a criterion, some of the chances of these risks might be lower, as the criteria applied by Ord include other outcomes as well. In such a case, a separate analysis for extinction risks could offer more clarity.

Some of these estimates have been deemed as much too low on climate change and nuclear war (Sears 2021), or too high in the case of unaligned AI (Sand 2021). However, in the former case, the rebuttal does not offer any specific counterargument for why the estimates are too low. Instead, it is only stated that they "seem" too low. Strikingly, it appears that in many discussions on the topic there is a tendency to ignore or underestimate resilience factors and mechanisms which would protect modern humanity from extinction.

Here we will discuss some of these resilience factors, using nuclear war as a detailed example. Nuclear war has been the first existential threat humanity has become aware of and it lately achieved a comeback in public awareness. It also shares some characteristics with risks such as massive volcanic eruptions, making it possible to generalise at least some conclusions. Following an analysis of nuclear war and the resilience factors of humanity regarding extinction, we will discuss some of the ethical aspects in the current discussion of these threats.

The nuclear threat

The nuclear attacks on Japan did not immediately change the nature of warfare, as the casualties were not higher than those suffered in one of the raids on Tokyo by conventional bombing (Searle 2002; Harwell/Grover 1985). What changed was the ease with which casualties in the hundreds of thousands could be inflicted. However, the invention of fusion weapons and increase in number of warheads since then has amplified the potential for destruction by many magnitudes. The three atomic bombs the US possessed in 1945 had a combined explosive yield of 55 kt (55.000 tons of TNT equivalent). In 2018, the armament of the US consisted of about 4,000 active warheads, with a total yield that can be estimated at roughly 700.000 kt (Kristensen/Norris 2018). Not included therein are warheads awaiting dismantlement as well as those of other nations, adding up to a total inventory of almost 13.000 warheads worldwide today (Kristensen/Korda 2022). Notably, this is but a fraction of the cold war arsenal, which has been reduced by 82% since 1986, thus demonstrating that disarmament is feasible. On the contrary, international tensions have caused a shift to modernisation and rearmament of national arsenals (de León 2019).

NATO and Russia together field over 90% of the current global nuclear arsenal. The conflict in Ukraine has without a doubt immensely increased the risk of an escalation between these power blocks and with it the deliberate or accidental usage of their nuclear weaponry.

NATO and Russia together field over 90% of the current global nuclear arsenal. The conflict in Ukraine has without a doubt immensely increased the risk of an escalation between these power blocks and with it the deliberate or accidental usage of their nuclear weaponry. No robust statistics are available on the chances of a nuclear war breaking out, as there never has been any historical precedent of a nuclear exchange. We can only analyse events that posed the threat of an escalation, such as the Cuban Missile Crisis or the numerous accidents of nuclear arsenals, to estimate how close we might have come in the past. Recently, the president of the Nuclear Threat Initiative has stated a personal estimate of a 0.5% chance of a nuclear war for each year (Rohlfing 2022). What is sure is that it is a substantial threat that is currently increasing in its urgency to be addressed.

Direct effects of the detonations

The two bombs dropped on imperial Japan have given the world a horrifying preview of what the consequences of a global nuclear war might look like. If the nuclear power blocs of NATO and Russia were to slide into a full exchange of their arsenal, millions of people would die within hours. Some of them would succumb to fatal burns from thermal radiation, others would be killed by collapsing buildings and other effects of the blast wave, and some would be trapped in the spreading fires. Within the next weeks, more would die from fatal radiation exposure.

It is already difficult to predict the extent of these direct casualties. One cannot simply scale up the effects of the bombs dropped on Japan. First, the increase of explosive yield in nuclear weapons

does not translate into an equal increase in destruction. With an increasing yield of the bomb, the fraction of energy released that is travelling over a two-dimensional landscape is becoming smaller compared to the total energy that is released in three-dimensional space. Thus, the shockwave of a typical Russian warhead destroys an area 9 times larger than the bomb dropped on Nagasaki, even though it possesses 27 times the explosive yield (Bell/Dallas 2007). Second, these Japanese cities were densely populated centres. In a realistic nuclear war scenario, the combatants would not distribute their arsenal on cities worldwide equally. Instead, industrial and military facilities would be targeted as well and might even be preferred over civilian targets (McKinzie et al. 2001). Especially ICBM (intercontinental ballistic missile) sites would be of high priority, given that their destruction would be the best chance to reduce one's own casualties. Nuclear strikes could also be expected to be focused on the participants of the conflict, with overkills of employed warheads in some areas. At the same time, the continents of South America, Australia, and Africa might be spared direct attacks completely. Estimates for direct casualties in the US alone through a Russian strike range from 30 million to over 100 million deaths (Helfand et al. 2002; Rodriguez 2019). Including other NATO countries and Russia, the total amount of casualties can likely be tripled or be estimated even higher if more countries are considered involved.

Nuclear winter

Yet, most casualties will likely be caused by the onset of a nuclear winter. The intensity of the fires is expected to carry soot and small particles up into the higher stratosphere, blocking a significant percentage of sunlight. This in turn is predicted to lead to a drastic cooling of global temperature of about 8°C, with some continental regions suffering decreases of temperatures by 20 to 30°C during the first year (Coupe et al. 2019; Robock et al. 2007). Combined with a decrease in precipitation, this means that many regions will suffer an almost complete loss of crop yields during the first years (Harwell/Grover 1985; Robock 2010). Soot aerosols have a long residence time in the atmosphere, so it might take more than a decade for surface climate to recover (Robock et al. 2007; Wagman et al. 2020). Even a local nuclear war scenario between India and Pakistan could put more than two billion people at risk of starvation – and over five billion people are estimated to starve after a potential nuclear war between the United States and Russia (Xia et al. 2022).

An important factor to consider here is that food supplies will vary highly between different localities. The change in climate will not be the same globally, with the southern hemisphere and maritime regions being much less affected. Some places are predicted to experience a comparatively mild cooling of 2 to 3°C (Coupe et al. 2019). Thus, in some regions at least parts of the harvest could probably still be brought in. Next, there are differences in the size of food stockpiles that countries have available. Some countries will quickly run out of reserves, while others will tend to have storages large enough to help them through the first months (Robock 2010). Certain sources of food like fishing and animal herding will be less impacted by the drop in temperature, giving the populations with access to them at least some sources of calories until the climate starts to recover (Robock 2010). Additionally, greenhouses could be used to mitigate the impact of fallout and lower temperatures. In short, the factors influencing the chances of communities to survive a nuclear winter vary considerably depending on location, available resources, and sheer luck.

The factors influencing the chances of communities to survive a nuclear winter vary considerably depending on location, available resources, and sheer luck.

Fallout

About 500.000 kt worth of nuclear weapons tests were conducted above ground until 1971. In 1961/1962 alone, a total of 340.000 kt was detonated, corresponding to about half of the currently active nuclear arsenal of the United States. The fallout created by these tests did inflict serious harm locally, but globally the exposure remained far below the natural background radiation. In a nuclear war scenario, the fallout might be considerably higher and increase cancer and birth defect rates globally, but it would not be high enough to threaten general survival. Regarding agriculture, there are large differences, up to four orders of magnitude, when it comes to the proclivity in which plants take up radioactive isotopes (Rantavaara 1987). Thus, preferencing certain vegetables and fruits might help to substantially reduce exposure through food intake. At least until recently, one hundred residents who had resisted evacuation were still living in the heavily contaminated exclusion zone of Chernobyl, despite its radiation (Global Resilience Institute 2019). Ironically, the ecosystem around the power plant has recovered to such an extent that it is now richer and more stable than it was before the incident (Hopkin 2005). The radioactive contamination has proven to be less hazardous for wildlife than the previous human settlements.

Threat of extinction through nuclear war

Taking into account everything we know about nuclear war and nuclear winter, it is unlikely that it would directly lead to the eradication of all of humanity (Ord 2020: 87; Oman 2012; Robock 2010). The creators of current nuclear winter climate models themselves make no such claim, and some of them outrightly deny that a nuclear winter is expected to lead to human extinction (Robock 2010). A recent study has estimated that a nuclear war between NATO and Russia would cause about five billion casualties from direct effects and starvation, meaning that 67% of the world population would die within two years (Xia et al. 2022). These numbers are of course enormous in their implications, but they are relatively far away from an extinction scenario. Nonetheless, beyond the extensive nuclear testing in the past, there has never been a precedent for such an event, so the threat of extinction cannot be excluded either. Different considerations apply when it comes to the collapse of nation-states or civilisations on a global scale as possible consequences. Heavy destruction of infrastructure in the combatant states, breakdown of trade, and desperate competition for food and other resources might very well cause a breakdown of social order and supply chains worldwide.

Taking into account everything we know about nuclear war and nuclear winter, it is unlikely that it would directly lead to the eradication of all of humanity.

Prevention and deterrence

The only certainty about a nuclear war is that it would be a disaster of hellish dimensions. The obligation to prevent it can be derived rather directly from that fact. It is more complicated to decide which policies should be enacted to do so. For example, if a state tries to gain advantages by intimidation through the threat of nuclear escalation, then our intuition might suggest that compromising to such demands is the best strategy to avoid a nuclear disaster. This can be backed up by several historical examples in which confrontational doctrines have almost caused catastrophic escalations. However, a successful intimidation might encourage further aggressive actions, thus potentially increasing the threat of escalation in the long run. Most questions of such concrete policies are too complex and situation-specific to be given sufficient justice here.

To name just a single concrete suggestion, a comparatively simple and attainable change of doctrine would be the cutback of land-based ICBMs. A severe problem of current nuclear deterrence strategy is that the reaction time to any assumed attack is very short. If a nation does not launch its ICBMs in time, they will likely be incapacitated by an incoming nuclear strike. This limited reaction time increases the chances that an attack is carried out by a false alarm. The enemy side would then be compelled by their deterrence doctrine to conduct the strike they had been wrongly suspected of, causing a full exchange. A solution proposed by former Secretary of Defence W. Perry is to give up land-based ICBMs entirely in favour of weapons carried by aircraft and submarines, as these do not have to be fired immediately for effective deterrence (Perry/Harris 2020).

An interesting unknown is the likelihood and extent to which leaders and military personnel would follow through with a retaliation strike. Mikhail Gorbachev is reported to have refused to give the order for a nuclear strike as part of a war simulation, creating the impression on soviet generals that he would neither do so under a real nuclear attack (Sebestyen 2010). The cold-war paradigm of "mutually assured destruction" might dictate nuclear retaliation as the vital part of deterrence, but very limited reason for it remains once deterrence has already failed to protect a nation from a full first strike. To assure deterrence, some might consider installing an AI with control over the arsenal, which would be programmed to retaliate once certain parameters are met. If it is kept protected from cyber-attacks, which is a critical assumption, the program should be incorruptible and, being devoid of any emotions like doubt, guilt, or mercy, retaliate faster and more reliably than human personnel. It can also still retaliate if the entire military leadership is already taken out. Thus, the enemy should be even more discouraged from launching a first strike and the reaction time to an enemy strike might be increased. In fact, an assumingly semi-autonomous system for this very purpose, named Dead Hand, has already existed in Russia since the Cold War and is considered to be still operational. Whether such a programme increases or decreases in total the likelihood of a nuclear war remains up for debate. Something to be considered here is the substantial number of technical errors that have already led to false alarms during the Cold War (Forden et al. 2000).

Resilience factors against extinction

The proliferation of nuclear weaponry and international tensions are undoubtedly risks to humanity's existence. In the following section, we will on the other hand look at elements and mechanisms that are protecting humanity from extinction. These will be considered here as resilience factors. One factor that has already been mentioned is that modern humanity is stratified in a vast variety of habitats, each differing in their susceptibility to specific catastrophic scenarios. In case of dramatic temperature changes, there will likely still be zones that remain or would become habitable. This can limit threats of extinction posed by a nuclear winter, super volcano, or climate change scenarios. Diversity of cultures and lifestyles are further factors that reduce the likelihood of one threat causing extinction. For example, there are still tribes with limited connection to the outside world – and some actively avoid any contact (Sasikumar 2018). This reduces the likelihood that those communities would be affected by a global pandemic spreading between otherwise interconnected societies. Technology, though being the main source of current existential risks, it is at the same time an extremely valuable protective factor. It can directly mitigate risks, for example through vaccine development respective to pandemics or carbon capture respective to climate change. Even if mitigation is impossible, it might still help humanity to survive. In a nuclear winter, gardening lamps could be used to grow food, while some of the renewable energy sources could be utilised at least for a limited time independent from fossil fuel supply chains.

Nuclear war on its own might possess only a low likelihood to wipe out humanity, but it has been argued that several such catastrophic events combined could be sufficient to cause extinction. These events might arise either in parallel or cause each other sequentially in a cascade effect (Marques 2020; Steinmüller/Gerhold 2021). A catastrophe could exacerbate certain other risks, for example by increasing international tensions. However, anthropogenic existential risks can also limit each other in negative feedback loops. Anthropogenic threats stem from the growing power potential and impact of humanity. A catastrophe which severely diminishes humanity will in many cases also decrease the prevalence of anthropogenic threats. It might be our intuition that, like a boxer, humanity will be even more vulnerable once it took a hit. The COVID-19 pandemic has at least partly been an example counter to this. With the beginning of lockdowns, worldwide CO2 emissions have decreased substantially in 2020 compared to previous years (Liu et al. 2020; Sikarwar et al. 2021). It stands to reason that a pandemic, or any other event which disrupts transportation and industry, will cause a decrease in emissions. Another example would be the mentioned recovery of the ecosystem around Chernobyl. A catastrophe can also be self-limiting. For instance, a lethal and contagious pathogen will destroy its own means of replication by decimating the host population. As a consequence,

Nuclear war on its own might possess only a low likelihood to wipe out humanity, but it has been argued that several such catastrophic events combined could be sufficient to cause extinction. However, anthropogenic existential risks can also limit each other in negative feedback loops. The COVID-19 pandemic has at least partly been an example to this. With the beginning of lockdowns, worldwide CO2 emissions have decreased substantially in 2020 compared to previous years.

it would in most cases die out before it could infect and kill all of humanity (Adalja 2016). Obviously, falling victim to a global catastrophe that cuts humanity short is not an acceptable solution. Nonetheless, at least a degree of reassurance lies in the thought that if humanity fails to prevent one global catastrophe, the chance that another one sets in right afterwards might be in some cases lower, not higher than before.

These are just a few examples of factors and mechanisms that protect humanity from extinction. The list is far from exhaustive, and each factor offers protection against some threats and not against others. One possible threat that ignores most of these protective factors is the emergence of an artificial general intelligence (AGI) that acts against human interests. In such a scenario it would for example probably matter little what technologies humanity possesses, as an AGI could likely utilise them more effectively. Several scenarios of how an AGI might become dangerous have been proposed and despite little consensus on how likely these are, there are by now several experts who believe AGI to be one of the biggest threats to human existence (Bostrom 2014; Ord 2020: 124-126; Vold/Harris 2021).

General resilience against extinction

The bubonic plague, also known as Black Death, is estimated to have killed about a full third of the European population in the 14th century (Glatter/Finkelman 2021). The event was traumatic in nature, caused people to expect the advent of the apocalypse, and affected the power balance in Europe. However, it did not lead to the full collapse of any major society. In fact, during the decades following the plague, the life of the common people improved in many regions of Europe. The number of workers had decreased, while the infrastructure and farmland remained largely untouched. This resulted in cheaper land and a rise in the price of labour, thus favouring the poor. Employers were forced to pay workers better wages, offer food of higher quality, and grant more freedoms (Scheidel 2018: 291-313). Despite such catastrophes, the existence of humanity was not seriously endangered until the modern age. When tribes and cultures vanished, the reasons were – in most cases, at least – societal changes and not the extinction of the whole community (Middleton 2012; Hunt/Lipo 2012). While being apocalyptic for the people it directly affected, the plague and similar catastrophes have become on a historical scale mere steps of human progress. This perspective should not relativise the human suffering involved, but it may help to preserve confidence in the future of our species.

Besides extinction, another catastrophic outcome that is often considered existential is the collapse of civilisation on a global scale. It has been proposed that in such a case humanity might find itself in a world so ravaged that it would never fully recover again, thus remaining in a "primitive" state (Steinmüller/Gerhold 2021; MacAskill 2022). While possible, I would argue that such a fate is at least not a likely one. There are only few catastrophes from which Earth would not recover eventually. Ash clouds precipitate, radioactivity declines, and ecosystems adjust. With recovery of the environment, humanity should be able to recover as well. Especially since it will be surrounded by artefacts of former civilisations, pointing the way to what it has already achieved in the past. Even if a catastrophe is significant enough to cause the total collapse of society, not all knowledge would be lost, as there would still be written records and the memory of the survivors. There would also be many resources available by scavenging destroyed cities, the tombs of the former civilisation. Precious metals that had to be dug up and purified with great effort in our early history would be scattered on the surface and thus be easily available. A major hurdle might be to attain energy sources, as there will be much fewer fossil fuel sources available than during the industrial revolution. In a case of a second industrialisation, other sources of energy might be utilised in addition. Plastic, left over from the previous civilisation, for example has a relatively high energy density and could be collected as a fuel. Even without facilities to create modern machines, the survivors could likely still use some of the remaining machinery for years, decades, or centuries. Those relics and the remaining records should speed up the technological recovery by serving as direct blueprints. Some of the modern crops, fruits, and farm animals, for which it took millennia of breeding, would likely survive as well (MacAskill 2022), allowing for more efficient farming than in early agricultural societies. The millions of ruins of abandoned houses would give valuable shelter, for which most caveman would have probably traded their favourite flint stones. As long as no other catastrophe sets in to finish what the first started, humanity would probably recover. If conditions after a catastrophe were too harsh for recovery, it is unlikely that humanity would survive for long at all. In the end, even a catastrophe killing 99% of humanity and making many areas of the planet temporarily uninhabitable would not necessarily destroy the capacity for humanity to recreate societies as advanced as our own in the long-term.

Techno-utopian ideas in longtermism

As laid out before, it would be unlikely that a nuclear war would directly lead to the extinction of humanity. Yet, I do not wish to suggest that this estimation reduces our moral obligations to prevent such a hellish event in any real sense. It would be an even worse fate if a nuclear war would not only cause the death of billions of people but would also lead to the extinction of humanity. However, from a practical point of view, the death and immense suffering of billions is already such an extreme scenario that the additional threat of extinction, no matter how significant in its implications, can barely increase the urgency of the matter, because its importance is already close to the absolute. The situation is similar with threats such as synthetic pathogens or climate change.

Compared to such threats, the possibility of an unaligned AGI is more hypothetical and appears of little urgency considering our immediate future. However, in case of its emergence, it might pose a significant chance to cause extinction or other long-term catastrophic consequences. Yet, the level of resources and research spent on AGI safety is currently minimal (Ord 2020: 53). Therefore, some argue that such threats should receive additional, if not our utmost attention. "Longtermism" is the idea that positively influencing the long-term future is a key moral priority of our time (MacAskill 2022). The main argument of longtermism is quite straightforward. The life of a human being in the future should be fundamentally considered just as valuable as one in the present. However, there are further implications and arguments made by some longtermists which go beyond this simple acceptance of the value of the future.

Several longtermists are influenced by the mentioned techno-utopian ideas. These are mostly predicated on utilitarianism, transhu-

manism, and a belief that technological progress will radically improve the well-being of humanity. Utilitarianism prescribes that the best action is the one that brings the most well-being to the most people. Transhumanism invokes the idea that the human race should evolve beyond its current physical and mental limitations, primarily by means of technology (Bostrom 2005). Lastly, humanity's potential is considered dependent, if not in some cases synonymous, with progress in science, technology, and exploration of space (Bostrom 2013). Therefore, supporters of these ideas consider events which will close off such progress to be existential risks as well.

> Longtermists like Nick Bostrom are influenced by techno-utopian ideas. These are mostly predicated on utilitarianism, transhumanism, and a belief that technological progress will radically improve the well-being of humanity.

Moreover, there exists a strong version of longtermism, which proposes that positively influencing the long-term future is not only important but fundamentally ought to take priority over other concerns (Greaves/MacAskill 2021). According to it, the value of future generations is almost infinitely higher than the one of current generations (Bostrom 2013; Greaves/MacAskill 2021; Torres 2017). This derives from the premise that the future might contain an almost countless number of human individuals. Further, those yet to be born are assumed to have better lives than we currently do, mainly due to technological progress. Consequentially, the moral value of all these future generations would be far higher by quantity and quality than that of currently living humans. While this argument may be internally coherent, it is based on assumptions which are not necessarily shared by a majority of people (Cremer/Kemp 2021). Even more importantly, some of the proponents of strong longtermism have pushed this line of argument to the point that it appears to effectively undermine the worth and rights of human beings by statements such as:

"One might consequently argue that even the tiniest reduction of existential risk has an expected value greater than that of the definite provision of any 'ordinary' good, such as the direct benefit of saving 1 billion lives." (Bostrom 2013).

Another controversial assertion is that it should be open to considerations to introduce surveillance systems that would fully monitor every person on the planet in real-time (Bostrom 2019). Such controversial argumentation at least begs serious questions about its underlying motivation, worldview, and assumptions.

Considering the latter, it is for example questionable to which degree the moral aspects of human existence can be reduced in any manner to calculations. It is also debatable to which degree continued rapid progress in technology will be more likely to make humanity's existence better and safer. After all, the largest fraction of current existential risk comes from technological advances. Moreover, there is pragmatic wisdom to applying a certain degree of temporal discounting to ethical decisions. Considering the far future as less predictable than the near future, interventions oriented toward the near future might be overall more effective (Fawcett et al. 2012). This issue is intensified by our ignorance about the degree to which non-existential problems can exacerbate existential risks (Liu et al. 2018). Even the effectiveness of planning based primarily on predictions can be put into doubt by the Black Swan theory, which assumes that the most influential events are the ones that are most difficult to predict (Taleb 2016). Besides extinction, longtermists are also worried about the possibility of a lock-in of negative values, meaning that certain undesired values might become so entrenched in the culture of the future that they will persist over an extremely long time (MacAskill 2022). Therefore, the formation and guarding of good moral values are considered as an essential step towards a better future. However, strictly acting out some of the more fanatic suggestions of longtermism, such as sacrificing millions or more if this is perceived to be a necessary step to protect a desired future, would likely foster totalitarian values. An extreme version of longtermism might itself create one of the catastrophic outcomes it is setting out to prevent.

Moreover, we should not forget that we are not uninvolved decision-makers when it comes to ethical problems. Ignoring the plight of humans close to us for the hypothetical benefit of future generations might not only be ethically questionable (Torres 2021) but might also have an impact on our psyche. After all, it has been shown that we subconsciously utilise our past behaviour for our decision-making, self-informing ourselves by our former actions (Albarracín/Wyer Jr. 2000). If we ignore the suffering of others, because we believe that doing so will bring a better outcome, then this might generalise such an unempathetic response. Moreover, prioritising existential threats over other problems might create an incentive for people involved in these discussions to paint the issue they are lobbying for as an existential risk. While it is important to bring attention to a problem, there can be downsides to presenting a problem as a matter of general human survival.

On the other hand, longtermists such as Toby Ord or William MacAskill have offered inspiring visions of a successful path into the future and some well-founded arguments for taking responsibility for ensuring the prosperity of forthcoming generations. This can motivate us to be even more engaged in preventing outcomes that would not only harm future but also present generations. Not to mention the many other individuals, organisations, and schools of thought that emphasise the need for long-term thinking in our societies along their own specific ideas and ideals. In total, many of the arguments made by longtermist have worth and validity, but I agree with their critics that these should still be challenged by other ethical and philosophical perspectives before being handed to policymakers.

Public communication of risks

Nuclear war is often depicted as an event that would annihilate humanity's existence. One possible reason for that portrayal is that it offers a potent picture to warn the public of its dangers. It is very salient, easy to comprehend, and emotionally charged. If the framing of a risk as an extinction risk is a superior strategy for gaining support and facilitating the prevention of such catastrophic risks, then it might be considered justifiable to do so. However, there are likewise costs attached to overstating a risk which should be considered.

A direct consequence of hyperbolical messaging might be the deterioration of the reputation of the corresponding activists and agencies. Therefore, some people will become sceptical of any valid information given by them as well. Further, if there is a multitude of threats that are discussed in such an intense man-

ner, then the anxiety-provoking input might become so intense that it causes counter-productive coping mechanisms such as withdrawal, paralysis, fatalism, or nihilism. Climate change has for example manifested in too many minds as a comparatively quick transition from denial to despair. None of those mental states generally allow for effective action. In several countries over half of the young population now believes that humanity is doomed (Marks et al. 2021). Some activist statements even caused climate researchers to warn against needlessly frightening children (Courtney-Guy 2019).

With nuclear war, one would hope it to be sufficient to communicate that a large percentage of people in the West would likely die from the consequences of a full nuclear exchange between NATO and Russia. A vivid imagery of what that would mean is given by the eyewitness accounts from the bombing of Hiroshima and Nagasaki (Nuclear Weapon Archive 1995). Similarly, concrete scenarios can be drawn for the consequences of climate change, without falling back to claims of imminent extinction.

And last, assuming that we are not one of the last generations might motivate us even more to avoid global catastrophes. After all, if people assume that humanity will vanish very likely anyway due to all the threats looming ahead, then they might comfortably fall back to a state of nihilism. This way they may reject having any responsibility for the future at all. However, if we assume we will be judged by future generations for our actions and inactions, then we will have to face being remembered for how we have handled ourselves in the face of the coming challenges.

Conclusion

The good news is that humanity seems in most metrics currently quite resistant against being fully wiped out. At the same time, events that would not terminate humanity but vanquish modern civilisation or cause the death of millions are much more likely.

A simple explanation why people might overestimate the likelihood of human extinction is that with a scenario such as nuclear war, it is indeed likely that we and the world we know would be annihilated. Such a mental image can understandably be mistaken for the end of humanity. However, it might do us well to remember that humanity does not vanish with us, our community, or our nation. It might be at least a little bit of solace that the future of humanity does not solely rest on us and that others will likely carry on if we do not make it.

As expressed, a threat should not need the label of an extinction risk to be taken seriously enough. Even without the biological survival of our species on the line, we should have plenty of incentives to avoid pandemics, ecological catastrophes, or nuclear exchanges. Longtermists are fully right in their diagnosis that our societies suffer from a pathological case of short-termism. For sure more must be done to safeguard our future. Nevertheless, how the well-being of current generations should be balanced against that of future generations remains a difficult problem. What can be said firmly is that any approach which seriously neglects one of the two sides will fall short morally and practically.

Therefore, it only makes sense to have an extra place on our agenda for threats which currently pose little immediate danger, but which have a realistic chance of cancelling humanity forever. In this regard, AGI stands out as a black box regarding its risks, which should be a reason to be cautious and to invest more resources than currently in preventive measures. While no precise prediction can be made of all the beneficial and harmful consequences of an AGI, I would agree to put it as the currently most dangerous long-term risk, partly because of its potential ability to nullify almost all resilience factors of humanity.

In this paper, risks not related to extinction were largely left aside. That is not to say that they are of less importance. The threat of humanity being trapped in a totalitarian or otherwise dystopian state might very well be greater than the one of extinction. Further, only few of the possible interactions between different risks were considered. These might play important roles and are currently insufficiently investigated. Possible interactions might make it even harder to find clear policies – especially in cases in which certain interventions against one risk might increase other risks. Regarding nuclear war, any careless escalation must be avoided. At the same time, appeasement towards authoritarian governments might increase the chance of other existential risks manifesting. In this sense, it might be useful to imagine humanity walking not only along a precipice, as described by Ord, but on a mountain ridge, with precipices falling off to both sides. No single doctrine can be safely trusted. Instead, a wise balance will be needed to reach the other summit.

References

Adalja, Amesh (2016): Why hasn't disease wiped out the human race? In: The Atlantic. https://www.theatlantic.com/health/archive/2016/06/infectious-diseases-extinction/487514/. Viewed 31 May 2022.

Albarracín, Dolores / Wyer, Robert (2000): The cognitive impact of past behavior: Influences on beliefs, attitudes, and future behavioral decisions. In: Journal of Personality and Social Psychology, 79 (1), 5-22.

Bell, William / Dallas, Cham (2007): Vulnerability of populations and the urban health care systems to nuclear weapon attack: Examples from four American cities. In: International Journal of Health Geographics, 6 (1), 5.

Bostrom, Nick (2019): The Vulnerable World Hypothesis. In: Global Policy, 10 (4), 455-476.

Bostrom, Nick (2014): Superintelligence: Paths, dangers, strategies. Reprint edition. Oxford University Press: Oxford.

Bostrom, Nick (2013): Existential risk prevention as global priority. In: Global Policy, 4 (1), 15-31.

Bostrom, Nick (2005): Transhumanist Values. In: Journal of Philosophical Research, 30 (Supplement): 3–14. https://philpapers.org/rec/BOSTV. Viewed 2 December 2022.

Coupe, Joshua / Bardeen, Charles / Robock, Alan / Toon, Owen B. (2019): Nuclear winter responses to nuclear war between the United States and Russia in the Whole Atmosphere Community Climate Model Version 4 and the Goddard Institute for Space Studies ModelE. In: Journal of Geophysical Research: Atmospheres, 124 (15), 8522-8543.

Courtney-Guy, Sam (2019): Scientists blast Extinction Rebellion speaker who told kids they may not grow up. In: Metro. https://metro.co.uk/2019/10/29/climate-scientists-blast-extinction-rebellion-speaker-told-kids-may-not-grow-11006887. Viewed 28 May 2022.

Cremer, Carla Zoe / Kemp, Luke (2021): Democratising Risk: In Search of a Methodology to Study Existential Risk. arXiv:2201.11214. https://arxiv.org/abs/2201.11214. Viewed 2 December 2022.

Fawcett, Tim / McNamara, John / Houston, Alasdair (2012): When is it adaptive to be patient? A general framework for evaluating delayed rewards. In: Behavioural Processes, 89 (2), 128-136.

Forden, Geoffrey / Podvig, Pavel / Postol, Theodore (2000): False alarm, nuclear danger. In: IEEE Spectrum, 37 (3), 31-39.

Friederich, Simon / Aebischer, Emilie (2021): At the precipice now, in eternal safety thereafter? In: Metascience, 30 (1), 135-139.

Glatter, Kathryn / Finkelman, Paul (2021): History of the plague: An ancient pandemic for the age of COVID-19. In: The American Journal of Medicine, 134 (2), 176-181.

Global Resilience Institute (2019): Forty-three years after meltdown in Chernobyl, social and economic resilience help drive recovery. https://globalresilience.northeastern.edu/fourty-three-years-after-meltdown-in-chernobyl-social-and-economic-resilience-help-drive-recovery. Viewed 15 May 2022.

Greaves, Hilary / MacAskill, William (2021): The Case for Strong Longtermism. GPI Working Paper No. 5-2021.

Harwell, Mark / Grover, Herbert (1985): Biological effects of nuclear war I: Impact on Humans. In: BioScience, 35 (9), 570-575.

Helfand, Ira / Forrow, Lachlan / McCally, Michael / Musil, Robert (2002): Projected U.S. casualties and destruction of U.S. medical services from attacks by Russian nuclear forces. In: Medicine & Global Survival, 7, 68-76.

Hopkin, Michael (2005): Chernobyl ecosystems „remarkably healthy“. In: Nature. https://doi.org/10.1038/news050808-4. Viewed 30 October 2022.

Hunt, Terry / Lipo, Carl (2012): Ecological catastrophe and collapse: The Myth of „Ecocide“ on Rapa Nui (Easter Island). SSRN Scholarly Paper 2042672. Rochester, NY: Social Science Research Network.

Kristensen, Hans / Korda, Matt (2022): Status of World Nuclear Forces. https://fas.org/issues/nuclear-weapons/status-world-nuclear-forces/. Viewed 30 October 2022.

Kristensen, Hans / Norris, Robert (2018): United States nuclear forces, 2018. In: Bulletin of the Atomic Scientists, 74 (2), 120-131.

León de, Ernesto (2019): New era of nuclear rearmament. In: YaleGlobal Online. https://archive-yaleglobal.yale.edu/content/new-era-nuclear-rearmament. Viewed 31 May 2022.

Liu, Hin-Yan / Lauta, Kristian /Maas, Matthijs (2018): Governing boring apocalypses: A new typology of existential vulnerabilities and exposures for existential risk research. Futures, 102: 6–19. https://www.sciencedirect.com/science/article/abs/pii/S0016328717301623. Viewed 2 December 2022.

Liu, Zhu / Ciais, Philippe / Deng, Zhu et al. (2020): Near-real-time monitoring of global CO2 emissions reveals the effects of the COVID-19 pandemic. In: Nature Communications, 11 (1): 5172.

MacAskill, William (2022): What we owe the future: A million-year view. London: Oneworld Publications.

Marks, Elizabeth / Hickman, Caroline / Pihkala, Panu / Clayton, Susan / Lewandowski, Eric / Mayall, Elouise / Wray, Britt / Mellor, Catriona / Susteren van, Lise (2021): Young people's voices on climate anxiety, government betrayal and moral injury: A global phenomenon. SSRN Scholarly Paper 3918955. Rochester, NY: Social Science Research Network.

Marques, Luiz (2020): Pandemics, existential and non-existential risks to humanity. In: Ambiente & Sociedade, 23. https://www.scielo.br/j/asoc/a/M6BMn4gtwyTZHnkWTDJDt8C/?lang=en. Viewed 2 December 2022.

McKinzie, Matthew / Cochran, Thomas / Norris, Robert / Arkin, William (2001): The U.S. nuclear war plan: A time for change. Natural Resources Defense Council. https://www.nrdc.org/sites/default/files/us-nuclear-war-plan-report.pdf. Viewed 20 October 2022.

Middleton, Guy (2012): Nothing lasts forever: Environmental discourses on the collapse of past societies. In: Journal of Archaeological Research, 20 (3), 257-307.

Nuclear Weapon Archive (1995): Eyewitness accounts. https://nuclearweaponarchive.org/Japan/Eyewit.html. Viewed 30 May 2022.

Oman, Luke (2012): Overcoming bias: Nuclear winter and human extinction: Q&A with Luke Oman. https://www.overcomingbias.com/2012/11/nuclear-winter-and-human-extinction-qa-with-luke-oman.html. Viewed 23 May 2022.

Ord, Toby (2020): The precipice: existential risk and the future of humanity. Bloomsbury Publishing.

Perry, William / Harris, Sam (2020): #210 – The logic of Doomsday. https://www.samharris.org/podcasts/making-sense-episodes/210-logic-doomsday. Viewed 16 May 2022.

Rantavaara, Aino (1987): Radioactivity of vegetables and mushrooms in Finland after the Chernobyl accident in 1986. Finnish Centre for Radiation and Nuclear Safety. Report STUK-A--59.

Robock, Alan (2010): Nuclear winter. In: WIREs Climate Change, 1 (3), 418-427.

Robock, Alan / Oman, Luke / Stenchikov, Georgiy (2007): Nuclear winter revisited with a modern climate model and current nuclear arsenals: Still catastrophic consequences. In: Journal of Geophysical Research: Atmospheres, 112 (D13).

Rodriguez, Luisa (2019): How many people would be killed as a direct result of a US-Russia nuclear exchange? https://forum.effectivealtruism.org/posts/FfxrwBdBDCg9YTh69/howmany-people-would-be-killed-as-a-direct-result-of-a-us. Viewed 24 May 2022.

Rohlfing, Joan (2022): Joan Rohlfing on how to avoid catastrophic nuclear blunders. https://80000hours.org/podcast/episodes/joan-rohlfing-avoiding-catastrophic-nuclear-blunders. Viewed 24 October 2022.

Sand, Martin (2021): The Precipice – Existential Risk and the Future of Humanity. In: Journal of Applied Philosophy, 38 (4), 722-724.

Sasikumar, Mundayat (2018): The Sentinelese of the North Sentinel Island: Concerns and perceptions. In: Journal of the Anthropological Survey of India, 67 (1), 37-44.

Scheidel, Walter (2018): The Great Leveler: Violence and the history of inequality from the stone age to the twenty-first century. Princeton University Press.

Searle, Thomas (2002): „It made a lot of sense to kill skilled workers“: The firebombing of Tokyo in March 1945. In: The Journal of Military History, 66 (1), 103-134.

Sears, Nathan (2021): The precipice: Existential risk and the future of humanity. In: Governance, 34 (3), 937-941.

Sebestyen, Victor (2010): Revolution 1989: The fall of the Soviet empire. New York: Vintage.

Sikarwar, Vineet / Reichert, Annika / Jeremias, Michal / Manovic, Vasilije (2021): COVID-19 pandemic and global carbon dioxide emissions: A first assessment. In: Science of The Total Environment, 794, 148770. https://pubmed.ncbi.nlm.nih.gov/34225159/. Viewed 2 December 2022.

Snyder-Beattie, Andrew / Ord, Toby /Bonsall, Michael (2019): An upper bound for the background rate of human extinction. In: Scientific Reports, 9 (1), 11054. https://www.nature.com/articles/s41598-019-47540-7. Viewed 2 December 2022.

Steinmüller, Karheinz / Gerhold, Lars (2021): Existentielle Gefahren für die Menschheit als Gegenstand für die Zukunftsforschung. In: Zeitschrift für Zukunftsforschung, (2021), 30-80. https://www.zeitschrift-zukunftsforschung.de/ausgaben/2021/1/5370. Viewed 2 December 2022.

Taleb, Nassim Nicholas (2016): The black swan: The impact of the highly improbable. New York: Random House.

Torres, Phil (2021): Why longtermism is the world's most dangerous secular credo. In: Aeon Essays. https://aeon.co/essays/why-longtermism-is-the-worlds-most-dangerous-secular-credo. Viewed 29 May 2022.

Torres, Phil (2017): Morality, foresight, and human flourishing: An introduction to existential risks. Durham, North Carolina: Pitchstone Publishing.

Vold, Karina / Harris, Daniel (2021): How Does Artificial Intelligence Pose an Existential Risk? In: Véliz, Carissa: Oxford Handbook of Digital Ethics (forthcoming). https://philarchive.org/rec/VOLHDA. Viewed 2 December 2022.

Wagman, Benjamin / Lundquist, Katherine / Tang, Qi / Glascoe, Lee / Bader, David (2020): Examining the climate effects of a regional nuclear weapons exchange using a multiscale atmospheric modeling approach. In: Journal of Geophysical Research: Atmospheres, 125 (24), e2020JD033056.

Xia, Lili / Robock, Alan / Scherrer, Kim et al. (2022): Global food insecurity and famine from reduced crop, marine fishery and livestock production due to climate disruption from nuclear war soot injection. In: Nature Food, 3 (8), 586-596.

Johannes Kattan is a student of psychology at the University of Würzburg (Germany). He holds a PhD in bionanoscience from the Technical University of Delft in the Netherlands and a MA in biology from the University of Würzburg.
Email: j.kattan@protonmail.com