Issue 1/2022 (Vol. 8)
ISSN 2190-6335

Intergenerational Justice Review

Issue topic:
Existential and unknown risks for future generations

Table of Contents

Issue topic:
Existential and unknown risks
for future generations

Editors of the IGJR

The peer-reviewed journal *Intergenerational Justice Review* (IGJR) aims to improve our understanding of intergenerational justice and sustainable development through pure and applied research. The IGJR (ISSN 2190-6335) is an open-access journal that is published on a professional level with an extensive international readership. The editorial board comprises over 50 international experts from ten countries, representing eight disciplines. Published contributions do not reflect the opinions of the Foundation for the Rights of Future Generations (FRFG) or the Intergenerational Foundation (IF).

Editorial

Climate breakdown, the loss of biodiversity, unaligned artificial intelligence, uncontrollable pandemics, and escalating armed conflicts: humanity faces cascading and overlapping risks that threaten its long-term survival. While each of these emerging crises alone has the potential to significantly degrade our species' future prospects, their interaction causes a danger even larger than the sum of its parts: a so-called polycrisis. At the extreme, a global polycrisis poses an existential challenge that could lead to civilisational collapse and ultimately human extinction. As a result of converging shocks, the World Economic Forum 2023 warned that the world may see such an event by the end of the decade.

How, then, do we tackle this new and burgeoning risk landscape? The traditional frameworks for managing risks are ill-equipped to deal with the complexity and magnitude of today's challenges. New ways of thinking and acting are urgently required for this task.

Against this background, a growing movement of researchers, policymakers, and activists is dedicated to the study and mitigation of existential risks. Toby Ord, a moral philosopher and leading figure in this field, offers the following definition: "An existential risk is a risk that threatens the destruction of humanity's long-term potential." On this account, humanity would not have to go literally extinct for an existential risk to be realised, since the destruction of its potential would already occur if humanity were no longer the master of its own fate. While this notion is thought-provoking, the idea of humanity's 'potential' is certainly open for different interpretations. 800 years ago, the European intellectuals of these times would have expressed views about the human potential that were entirely different from those of today. And academics in other parts of the world today might give quite different answers about the human potential than Oxford scholars (who sometimes delve in techno-utopian dreams).

An alternative definition of 'existential risks', offered here by the editors, has it that they are risks that lead to a breakdown of human-made systems to an extent that the survivors can barely fulfil their basic needs. While still being open to different interpretations as to which kind of ecological, social, technological or other catastrophes might cause this sort of breakdown, the idea of human needs provides a solid basis for assessing the standard of living of the residual mankind.

Existential threats can be divided into anthropogenic risks – those that stem from human actions – and natural risks – those that originate from conditions beyond human control, for instance major asteroids, massive volcanic eruptions, or gamma-ray bursts from stellar explosions. The odds of natural catastrophes have remained rather constant over the last millions of years, that is: constantly low. If they were different, we would likely not be here. From this, it seems reasonable to expect that their likelihood will remain low over the next thousands of years as well. On the other hand, anthropogenic risks, to the extent that we are aware of them, have massively increased and accelerated in the era of the Anthropocene.

The unfolding of such risks could involve massive immediate casualties, but also sustained and widespread decline in the quality of life of future generations. For this reason, protection from existential risks is an intergenerational public good on a global scale. To embrace this responsibility, today's societies urgently need to overcome their myopic biases and radically expand their timescales to encompass long-term futures. In this way, humanity could not only reduce existential risks but also imagine and unlock a pathway toward the flourishing of life in the long run – a pathway that could be called existential hope.

Within intergenerational justice research, the connection to the risk literature is rarely made. That is why the Foundation for the Rights of Future Generations devoted its biannually Intergenerational Justice Award to this topic. The best papers are published in a special double issue, IGJR 1-2022 and 2-2022.

In the first article, Johannes Kattan suggests that 'extinction risks' ought to be distinguished more sharply from other aspects of 'existential risks'. Human extinction is an outcome that can be ascertained rather precisely in biological terms. According to Kattan, however, it should be analysed separately from scenarios in which the subjective quality of human life is the concern. Nuclear war is taken as a primary example for illustrating an extinction risk and for discussing humanity's resilience to such threats. Kattan concludes that, despite the unprecedented damage it might cause, it is unlikely that a nuclear war would lead to the end of the human species.

The second article, written by Marina Moreno, covers a different aspect. To understand the background of her concerns, one should be aware that myopia is usually seen as something negative in the literature on intergenerational justice. Long-term thinking is key to human survival, write dozens of scholars, unisono. For Moreno, anti-presentism comes with its own problems, however. She takes issue with longtermism understood as a theory which holds that our moral focus should be on the long-term future, and that current and medium-term moral problems are comparatively insignificant. Moreno's paper explores the implications of rejecting the premise of moral aggregation of individuals. She concludes that non-aggregationism does not support longtermist conclusions.

Issue 1-2022 then concludes with two book reviews. Tolga Soydan reviews Toby Ord's influential: *The Precipice. Existential Risk and the Future of Humanity.* Finally, Grace Clover reviews William MacAskill's second monograph *What We Owe the Future*.

Jörg Tremmel, Editor
Felix Beer, Co-Editor
Markus Rutsche, Co-Editor

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

Does longtermism depend on questionable forms of aggregation?

Marina Moreno

We are constantly making choices about how to invest our time and resources. From a moral perspective, we must ask which moral concerns are most deserving of our attention. Longtermism, as e.g. defined by Greaves and MacAskill, holds that our moral focus should be on the long-term future, and that current and medium-term moral problems are comparatively insignificant. This theory is centrally based on the assumption that the moral importance of individuals can be aggregated. Since the number of individuals of future generations far exceeds the number of current individuals and those closer in time, future generations are to be morally prioritised, according to longtermism. This paper explores the implications of rejecting the premise of moral aggregation of individuals and adopting a strongly non-aggregationist position instead. It is argued that, according to strong non-aggregationism, the magnitude of the probability with which our intervention actually make a difference, as well as whether we look at the available interventions from an ex ante or ex post perspective, are relevant factors in their moral assessment. Ultimately, the conclusion is reached that strong non-aggregationism does likely not support strongly longtermist conclusions.

Keywords: *Longtermism; aggregation; extinction risks; unaligned AI*

Introduction

Ever since humanity has gained the technological potential to destroy itself, extinction risks have increasingly been the focus of political and philosophical debates which have in turn sparked efforts to minimise these risks. One important line of argument supporting this focus is suggested by recent literature, according to which it is one of the most important moral imperatives to let our actions be guided by a long-term perspective. The basic argument for this claim runs as follows: Humanity has the potential to go on existing for a very long time before going extinct. During this time, a very large number of morally considerable individuals (humans, animals, and potentially digital minds) could come into existence. Since these individuals are just as morally important as similar individuals living today, their aggregated moral importance outweighs the moral importance of present generations by far. Thus, ensuring the existence and influencing the welfare of the large number of generations in the far future is among the most important moral goals, if not the singular most important one.

Longtermism holds that future individuals are just as morally important as similar individuals living today, therefore their aggregated moral importance outweighs the moral importance of present generations by far. Thus, ensuring the existence and influencing the welfare of the large number of generations in the far future is among the most important moral goals, if not the singular most important one.

The above argument relies on a crucial moral assumption; namely that welfare can be aggregated across individuals to form better goods and worse bads. The thesis that moral importance is relative to the number of individuals affected lies at the heart of aggregationism. With his famous paper *Should the numbers count?* in 1979, John Taurek challenged this thesis and sparked a debate that is still on-going today (see Taurek 1977). For instance, in her recent monograph *Non-Aggregationismus: Grundlagen der Allokationsethik*, Weyma Lübbe defends a rigorous non-aggregationist position (see Lübbe 2015) – and various other authors have proposed alternative ways to morally count or not count the number of people affected by a decision as well.

In this paper, I aim to explore the challenge a strongly non-aggregationist position poses for longtermism and its implications, such as a strong moral focus on extinction risks, and point to the relevant research questions in this area. To this end, I will first locate and contrast the version of longtermism I am concerned with and show how it implies a certain anti-presentism in section 2. In section 3, I will then motivate non-aggregationism and defend it against an obvious objection. Subsequently, I will move on to section 4, where I consider what non-aggregationism implies regarding the longer-term future and large-scale decisions, such as where to invest resources and which career path to choose. At this point I will show that non-aggregationism seems to suggest a long-term focus as well. However, in section 5 I will raise and discuss complications regarding the probabilities with which individuals today and in the future can be affected. These complications, I will ultimately argue, suggest that non-aggregationism does not in fact recommend a long-term focus after all. Rather, it points to a near- to medium-term focus instead. Section 6 concludes with a summary and an outlook for further research into the issues discussed.

Aggregation, longtermism and (anti-)presentism

In *Reasons and Persons*, Derek Parfit asks us to consider and morally compare the following three scenarios: a) peace, b) a war that kills 99% of the human population and c) a war that kills 100% of the human population. While it stands to reason that a) is better than b) and b) is better than c), Parfit offers the striking thought that the difference in goodness between b) and c) is vastly larger than the difference between a) and b). This is due to the fact that c) implies the loss of all of humanity's possibly astronomically large future, while b) still allows for a full human recovery (Parfit 1984: 453-454). To the extent that the goodness of outcomes ought to crucially influence our moral decision-making, Parfit's reasoning supports a priority of the far off long-term over the short-term.

Based on such considerations, many longtermists have followed Parfit to further develop this argument. For instance, Nicholas Beckstead defends the claim that influencing the far future is morally *overwhelmingly* important, i.e., more important than most other moral goals, because the expected value of the far future is astronomically great (Beckstead 2013: 2, ch. 3). Similarly, Nick Bostrom suggests that existential risks reduction to avoid the above scenario c) ought to be our global priority and serve as a strongly action-guiding principle, more strongly so than other considerations exclusively affecting the well-being of present generations. That is, he argues that there is a moral case for existential risks reduction being more important than any other global public good, as well as that many existential risks will not arise right now, but rather in the foreseeable future, such that our global priority ought to be to build resilience against them. These two claims taken together imply that other policies which affect the well-being of present generations that might, for instance, starve within the next couple of years, are globally less important than existential risks reduction, whether this can be achieved in the short- or medium-term (Bostrom 2013: 1). A more comprehensive case for this kind of priority of the long-term over the near-term and medium-term has recently been developed by William MacAskill and Hillary Greaves (2021). They define the notion of the so-called "strong longtermism" which, in contrast to softer versions of longtermist thinking, holds that the long-term future is the *most important* moral feature of our present actions (Greaves and MacAskill 2021: 3-4, 26-27). Similarly, Toby Ord has recently argued that the prevention of extinction risks, whether presently relevant or likely to hit in the further future, ought to be one of, if not *the* most important of our global priorities (Ord 2020).

As I have outlined the arguments in the longtermism literature so far, longtermism seems to imply a certain anti-presentism, i.e., the view that the impact of our actions on present generations matters little in contrast to the impacts on the far future.[1] This anti-presentist implication could be resisted in at least three straightforward ways. Firstly, we might argue that even if their axiological claims regarding the importance of the far future, given its astronomical stakes, were true, this would not imply that our all-things-considered deontic duties must follow these axiological claims. That is, even if the consequences of our actions are actually best if we follow the maxim of seeking to optimise its long-term consequences, the goodness of those outcomes may not (exclusively) determine what we ought to do. Secondly, one might deny that there are in fact trade-offs between affecting the near-term and long-term future, i.e., that whatever benefits the present generations most will also benefit the far future most. While there would thus still be a conceptual priority of the far future and future generations, the priority would not matter much in practice. Lastly, one might reject the outlined arguments for longtermism altogether and champion a different version of longtermism, based upon considerations other than the astronomical aggregated stakes of humanity's (far) future. Neither one of these strategies for argueing against longtermism are the ones employed in this essay, but I will briefly address them now in turn in order to set up the context of my own debate.

Firstly, Greaves and MacAskill explicitly distinguish between axiological and deontic strong longtermism. Importantly, however, they defend both the axiological and deontic version of strong longtermism, i.e., they hold that the impact on the far future most importantly determines both the goodness of the respective outcomes as well as what we, all things considered, ought to do (Greaves and MacAskill 2021: 2).

Similarly, Bostrom explicitly asserts that the impacts existential risks have on the generations of the far future ought to be our central *action-guiding* principle, i.e., is deontically very relevant as well (Bostrom 2013: 1). Generally, central proponents of the version of strong longtermism that is based on the aggregative premise outlined above thus do indeed often make deontic claims. However, my arguments can be equally applied to both the axiological as well as the deontic versions of strong longtermism, since it will crucially concern itself with the underlying aggregative premise, which underlies both the axiological as well as the deontic version.

Furthermore, even if we grant that axiological strong longtermism does not directly translate into deontic strong longtermism, important questions on the deontic level are still likely to remain, to which my subsequent discussion will be relevant. Beckstead (2013: 9) and more recently MacAskill (2022) have left open the possibility that their aggregative axiology may not translate into deontic guidance directly. For instance, MacAskill allows for other important factors that should influence our decisions from a moral perspective, namely special relationships and reciprocity (MacAskill 2022: 8). However, consider the following possible scenario: A philanthropist has a certain sum of money that they would like to donate to a specific cause. They have already fully discharged both their duties regarding special relationships as well as reciprocity, such that these factors do not influence which cause the philanthropist is required to select to donate to. According to longtermists such as MacAskill, the deontic permissibility of their cause selection would then be entirely determined by how it influences the generations of the far future. Even this more limited deontic claim seems somewhat doubtful, however, given that many people (to whom the philanthropist has no special relation or reciprocal duties) alive today are in dire need of help as well.

This leads us to the second point above, namely whether the near- and long-term perspectives might not in fact simply converge on a more practically applied level. For instance, consider the example of nuclear disarmament. Ending the cold war was both important from a long-term perspective in order to mitigate the extinction risk posed by nuclear war, as well as beneficial for present generations on several levels; they no longer had to live in fear and a resulting peace dividend freed up resources which could in turn be spent on other humanitarian efforts.[2] To the extent that this kind of empirical convergence applies more broadly, the anti-presentism implied by strong longtermism would not actually make a relevant practical difference.

However, consider the following example inspired by Richard Pettigrew (n.d.: 2). Say, again, a philanthropist is choosing between two charities to which to donate.

On the one hand, they are considering donating to the cause of preventing malaria by net distributions. On the other hand, they are considering to fund researchers working on AI alignment.[3] From the perspective of future generations, it is mostly the latter that matters, since preventing malaria today is comparatively unimportant to them, given that it is unlikely to interfere with humanity's greater trajectory by which they will turn out to be influenced. AI alignment, however, is possibly very relevant to

future generations, insofar as it constitutes an extinction risk or a risk of very bad lock-in scenarios crucially shaping humanity's trajectory. Strong longtermists must therefore strongly advise the philanthropist to donate to AI alignment, given that the expected value of doing so will be larger by a significant amount in view of the vast number of people in the future influenced by AI but not by malaria. However, from the perspective of present generations, the case is not as clear at all. While it is nearly certain that malaria will affect people existing today within a short time period, risks resulting from misaligned AI may very well not affect people alive today at all, but certainly not the people who will die from malaria within, say, the next year if the necessary nets are not distributed. The priorities of present and future generations are thus in a trade-off here, simply as a matter of empirical reality. The same basic argument applies, mutatis mutandis, to time spent working on various causes as well as setting global priorities and developing respective policies. In general, it would seem surprising if the actions which have the best consequences for generations many thousands of years from now would converge exactly with the actions that have the best consequences for present generations. A certain anti-presentism thus does not seem to be avoidable within the framework of strong longtermism as outlined and defended by Greaves and MacAskill.

This leads to our third point, i.e., to considering other forms of longtermism which may be able to resist the anti-presentist implication. Note that none of these theories can be based as strongly on the aggregative premise outlined above, since this would lead us back to some form of anti-presentism. However, the general idea of introducing a long-term perspective has found its uptake in various frameworks, which I do not explicitly address in the present text. For instance, Roman Krznaric has suggested the notion of "long-term thinking", which refers to a type thinking which is aimed at mitigating the myopic short-term thinking that is pervasive in today's society (Krznaric 2020). It stands in sharp contrast to strong longtermism as outlined above insofar as it is not a fully comprehensive normative theory of evaluation but rather a guide to better incorporate the future perspective into our present thinking. The fact that Krznaric proposes a horizon of consideration of 100 years (Krznaric 2020: 14) further showcases the radical differences to strong longtermism in its normative underpinnings, where the relevant horizon includes up to several thousands if not billions of years (Greaves and MacAskill 2021: 6). In contrast to strong longtermism, Krznaric's long-term thinking can almost be viewed as presentist, i.e., biased towards the next 100 years.

Another broadly longtermist approach is concerned with the representation of future generations within the present political institutions. For instance, Jörg Tremmel (2021) suggests a four-branches model of government, where the three traditional branches are supplemented by an office for future generations. Tremmel states explicitly, however, that this office is not to have any veto power since that could significantly slow down the political process, but rather is intended to just propose policies benefitting future generations (Tremmel 2021: 18). Another political approach to the inclusion of future generations has been proposed by Dennis Thompson, who argues that democracies tend towards a harmful presentist bias which can be mitigated by the principle that "present generations should act to protect the democratic process itself over time" (Thompson 2010: 14). That is, present generations ought to act as trustees of the democratic process. Thompson likewise explicitly rejects theories which sacrifice the well-being of present generations in favour of some temporally distant collective good (Thompson 2010: 1).

Insofar as it can be expected of these political approaches that their practical implementation would still give a certain priority to the present generations, i.e., not weigh the interests of each person of the vast future generations exactly equally in an aggregative way, which would again lead to swamping out the interests of the present generations, they thus avoid the anti-presentist implications. I will, however, examine this latter strong longtermism for at least two pertinent reasons. Firstly, I am crucially considering how the implications of a non-aggregative normative theory differ from the explicitly aggregative theory which leads to the strong form of longtermism. Hence, the natural argumentative opponent of my arguments is the form of longtermism based on aggregative premises. Secondly, as Beckstead suggests, the strong form of longtermism would, if true, have the most radical and controversial implications. Greaves and MacAskill have recently argued that if longtermism were true, it would imply a strong focus on decreasing extinction risks and on influencing the path of high and long-term impact developments – such as the development of a superintelligent AI – (Greaves and MacAskill 2021: 13-14), since those are among the most promising interventions influencing the long-term future. By contrast, they argue that focusing e.g. on fighting malaria or factory farming today is of little moral importance, since the number of individuals affected by any of these is lower by several magnitudes. These implications are so counterintuitive that Beckstead devotes an entire section to make the claim plausible that most of what we believe about moral priorities might be wrong (Beckstead 2013: 25-53). Given the possibly radically revisionary nature of this theory, I believe it to be worthwhile to engage with it at length.

Greaves and MacAskill have recently argued that if longtermism were true, it would imply a strong focus on decreasing extinction risks and on influencing the path of high and long-term impact developments, such as the development of a superintelligent AI.

However, since such a rigorous anti-presentism is initially counterintuitive, it is frequently met with scepticism. Can a moral theory that implies anti-presentism, i.e., that we ought to essentially abandon efforts to help any currently existing individuals really be considered plausible? Is a very strong focus on long-term developments such as extinction risks really reasonable rather than fanatical, as longtermism is sometimes accused of being (Greaves and MacAskill 2021:24)? Longtermists may answer these questions by arguing that the mere fact that some conclusions are counterintuitive should not be sufficient to undermine a theory, at least not to the extent that we actually agree with its premises. If the argument sketched above proves sound upon reflection, philosophy might play a legitimate revisionary role here.

However, the fact that these conclusions do seem so counterintuitive ought to give us at least good reason to examine the premises upon which they rest. This paper critically discusses the aggregative premise and explores the implications its rejection has for the conclusions of strong longtermism. In what follows, I will refer to the position of strong longtermism based upon an aggregative

premise and as defended e.g. by Greaves and MacAskill as well as Beckstead and Bostrom, as "longtermism".

Strong non-aggregationism: the rationale

In what follows, I will first outline the rationale and arguments motivating strong non-aggregationism. The literature on these issues has grown over the past years and I will only attempt to outline the specific view that I will be working with in this paper, which is inspired by Taurek, Lübbe and Mannino.

To illustrate his view, Taurek asks us to consider the following interpersonal rescue conflict: Six people are in need of medication to survive and you have a certain supply of said medication. Five of those six only need one fifth of the amount available to survive, while the sixth, call him David, needs all of it, if he is to survive. Is there a moral duty to give the medication to the five people and letting David thereby die? And relatedly, is it better in any relevant sense to save the five and let David die? To answer questions of the latter sort, Taurek argues one ought to ask: Better for whom? (Taurek 1977: 299) If the five people receive the medication, that is better for each of the five people and worse for David. If David receives the medication, it is better for David and worse for each of the five other people. However, Taurek argues that there is no one for whom it is better if five people are being saved instead of only one. There is no one perspective from which the fact that the greater number is saved is better.

> Taurek's thought experiment: Six people are in need of medication to survive and you have a certain supply of said medication. Five of those six only need one fifth of the amount available to survive, while the sixth, call him David, needs all of it, if he is to survive. Is it better in any relevant sense to save the five and let David die?

Taurek goes on to ask what it would even mean to evaluate the death of five people as worse than David's death (Taurek 1977: 299-300). What do evaluative judgements such as better or worse really mean if there is no person for whom they are better or worse? It is precisely the fact that something can be bad from someone's own perspective that makes them ethically relevant individuals in the first place. However, there is no meaningful level of evaluation above and beyond this personal perspective. According to Taurek, this point extends to all numbers cases, i.e., that the number of affected people never morally counts. For instance, even if, say, 50 family members were to be suffering if the five people die (10 each), and only 10 in case David dies, the same argument applies again: Since there is no meaningful aggregated perspective from which it is worse that more people suffer (i.e., 50 instead of 10), there is no corresponding reason to prioritise the greater number of separate people regarding the family.[4] Note that this example also shows that the lack of an impersonal perspective Taurek alludes to here is not an implication of the fact that the people *die* if they are not saved, i.e., the problem is not that their perspectives vanish once they die. The example would still hold if we consider the grieving family or if we modified the example such that each of the six people get a severe (but not deadly) illness without the medication. While receiving the medication is better for each of them from their own personal perspective, there is no one perspective from which five prevented illnesses are better than one prevented illness.

Contrast the above case with a rescue situation in which there is a unique pareto-optimal solution. Suppose you could either save the five people or save the very same five people and David. Here it is clear that the best option is saving everyone, because that is better from everyone's perspective. However, in a genuine conflict situation it seems unclear what it would even mean that one outcome is worse, period, than the other. According to defenders of this "no worse claim", there is thus no reason to save the greater number based on any notion of realising better world states (Taurek 1977: 300). In this sense, Taurek's non-aggregationism undermines both axiological as well as deontic versions of how numbers may morally count.

This argument bears a similarity to Rawls' critique of utilitarianism, according to which utilitarianism does not take the separateness of persons sufficiently seriously (Rawls 1971: 23, 163). One way to interpret this idea is exactly what motivates non-aggregationism: Since people's perspectives are separate, and do not in this sense aggregate into one superbeing whose well-being consists of all the separate individuals, there cannot be an equivalent aggregate moral value. Call this view regarding the strict non-aggregation across persons strong non-aggregationism.

However, strong non-aggregationism ought to offer a different solution to interpersonal rescue conflicts in order to constitute a genuine alternative to aggregationism. How exactly this solution is spelled out differs for various theories. What most of these suggestions have in common is the following general thought: In interpersonal rescue conflicts, all of the people affected equally hold a very strong claim to being helped. As such, all claims involved ought to be respected. Since the situation is such that we cannot actually fully satisfy all of the claims jointly, we have to instead find another way to express our equal concern for all the people involved. Now if we simply aggregated and opted to save the greater number, the people in the majority group would receive all of our concern and the people in the minority group would receive none. Instead, many strong non-aggregationists embrace a lottery solution: Since the good of survival cannot be given to everyone who has a claim on it, one ought to instead at least distribute the chances of survival equally among everyone.

There again exist different versions of how the chances are distributed exactly. Due to the scope of this paper, I cannot go into detail about all the advantages and disadvantages each of those versions offer. Rather, I will present and further discuss one particular version defended by Lübbe (2015). The reason I chose this version is twofold: Firstly, I personally believe it to best capture the spirit of equal and maximal concern, and secondly, because I believe it poses the most interesting challenge to longtermism, which is what I aim to explore in the present paper.

Consider again Taurek's case of five people versus one person. Lübbe holds that each of the people involved in this rescue conflict ought to be respected equally and maximally. Hence, she argues, the maximal chance of survival to be allocated that is consistent with equality is a 50-50 lottery between the five people and David. This way, each individual person involved receives a chance of 50% of survival, which is as good as it can get for everyone without decreasing anyone else's chance for survival.

Even though not all of the claims involved can thus be respected to the degree that they will eventually be satisfied, as this is unfortunately simply the empirical reality we find ourselves in, every claim is still equally respected to the maximal possible degree.

Such proposals indeed seem to be in line with many societal practices in place today. In many jurisdictions, for instance, it is illegal to deprioritise patients with an increased need of a certain scarce medical resource on the grounds that doing so would likely save a greater number of people. For instance, the Organ Procurement and Transplantation Network in the United States functionally assigns a higher chance to be saved to people who need two organs than to people who only need one organ (OPTN 2020: 137). Similarly, the policies from Eurotransplant do not opt to save the greater number either, but rather specifically assign an equivalent chance to multi-organ patients as to single-organ patients (Eurotransplant 2019: 10, 23, 36). That is, everyone's claims are being respected by granting everyone access to medical resources by virtue of an organ waiting list, i.e., a fair queue. The fair queue arguably implements a type of natural lottery, where the random process of admission to the queue distributes the respective chances equally from an ex ante perspective (see John and Millum 2020). The triage law recently developed and passed in Germany likewise does not ensure that the greater number is saved by e.g. deprioritising patients who will need scarce medical resources such as ventilators for a longer period of time than others. If e.g. two patients each need the last ventilator for a week to survive and a third patient needs it for two weeks, the triage law will not make the former two patients a priority, but will give an equal chance to all three of them by equally admitting to the fair queue (see Deutscher Bundestag 2022).

Consider again Taurek's case of five people versus one person. Lübbe holds that each of the people involved in this rescue conflict ought to be respected equally and maximally. Hence, she argues, the maximal chance of survival to be allocated that is consistent with equality is a 50–50 lottery between the five people and David. This way, each individual person involved receives a chance of 50% of survival, which is as good as it can get for everyone without decreasing anyone else's chance for survival.

Quantitative Catastrophes and Unequal Stakes

In their paper, Greaves and MacAskill (2021) briefly consider a non-aggregationist view according to which an aggregate good or aggregate bad may not be considered sensible notions, conceding that such a view might challenge their longtermist conclusions.

They reject this line of thought by the following argument: Any minimally plausible theory must be able to explain the fact that special norms seem to apply to situations where there is a huge amount at stake. For instance, if millions of people are threatened by a natural catastrophe, other comparatively minor moral constraints get justifiedly overriden. Greaves and MacAskill (2021:27-28) hold that any attempt at explaining this fact from a non-aggregationist perspective will lead back to supporting their own conclusions.

Their argument can be reformulated to pose a challenge to non-aggregationism more straightforwardly: Is it not simply entirely implausible to not believe that a large-scale natural catastrophe affecting and possibly killing billions of people is not any worse than a small-scale natural event killing one person would be? And relatedly: Is it not obvious that in catastrophic circumstances, saving as many people as possible is the right thing to do?

Indeed, emergencies are often thought to be a context in which a thoroughly consequentialist aggregative ethics is most adequate, given the huge number of people affected (e.g. Rakić 2018). While one may be inclined to follow Taurek in cases where the number differences are small, such as one versus five people, the bigger the number differences become, the harder it is to swallow Taurek's claim. While our intuitions in the case of catastrophic circumstances may still be able to be accounted for by the particular badness catastrophic circumstances may come with (which I will come back to below), even just in comparing, say, 10,000 car accidents killing one person each versus one other car accident killing one other person, it seems very counterintuitive not to claim that the former is clearly worse. If non-aggregationism is unable to account for these intuitions in any way, it may simply be implausible at the outset.

Is it not simply entirely implausible not to believe that a large-scale natural catastrophe affecting and possibly killing billions of people is not any worse than a small-scale natural event killing one person? And relatedly: Is it not obvious that in catastrophic circumstances, saving as many people as possible is the right thing to do?

While these intuitions do indeed pose a great challenge to non-aggregationism, they do not completely undermine its plausibility. This is due to at least three features a plausible version of non-aggregationism ought to include.

Firstly, while non-aggregationists do not consider the numbers intrinsically morally relevant in cases of interpersonal conflicts, there are ways in which they do let the numbers count after all. In the case in which we could either save the five people or the five people and David, additionally saving David is the right thing to do even for a non-aggregationist, as it is the unique Pareto-optimum. A non-aggregationist can thus make sense of the intuition that additionally saving more people from a catastrophe is better than saving fewer – so long as saving these additional people is not in direct conflict with saving other people with equal claims.

Secondly, people who are in a sufficiently "veiled" ex ante position may reasonably agree on an aggregative *policy* without necessarily referring back to aggregative *moral premises*. (The notion of "veiling" employed here is Rawls's (1971), but the ex ante situation is supposed to be empirically real, not a hypothetical "original position".) Consider a government which ought to decide whether, as a general policy, they send help to a place where there are more people to be saved versus to a place where there are less people to be saved, if both of them are affected by the same devastating catastrophe. From an ex ante perspective, each person living under the government is generally more likely to end up in the place where there are more people, assuming that where you end up is sufficiently chancy, and uniformly so. Non-aggregationists may thus jointly commit to an aggregative emergency policy to the extent that it is beneficial to everyone from an ex ante perspective because this policy gives everyone the greatest chance to be among the ones who will be saved. Thus they adopt an aggregative policy not because saving more people is better in an impersonal sense, but because the policy benefits everyone separately.

For instance, the six people in Taurek's medication case might have precommitted to an aggregative policy in advance, i.e., be-

fore they knew how much of the medication they would need. With such a policy, all of them have a higher chance to end up being saved, since it is, *ceteris paribus*, likelier for everyone to end up in the bigger group of five people who only need a fifth of the medication. As such, this precommitment benefits each of the people separately, by virtue of giving each of them a higher ex ante survival chance, and it is thus not necessary to hold that saving five people is simply better from some objective, impersonal perspective. While this reply is not able to completely answer the objection raised by large moral catastrophes, it is able to weaken its strength, as it is able to account of some aggregative policies without referring back to aggregative premises, i.e., without giving up the central idea of Taurekianism.

Note, however, that the same argument is not analogously available in the context of future generations for at least two reasons. Firstly, present generations cannot agree to aggregative policies that include all future generations, since the situation is not sufficiently veiled in this case: People living today would know in advance that they will be among the deprioritised ones and at no point is such an aggregative policy in their interest. Secondly, it is hard to make sense of an ex ante consensus regarding people who have not been born yet. For an analogous argument to work, one would have to argue that for each person that has not been born yet, any time in the future is roughly equally likely to be their actual birth time. If this is even a sensible proposition, it is certainly at least metaphysically doubtful. An argument for aggregation based on a veiled ex ante situation does not seem very promising if the veiled situation vanishes as soon as the relevant individuals *come into existence*.

Alternatively, one may interpret the veiled ex ante situation as a "mere" operationalisation of impartiality (see Rawls 1971), and thus argue that it must apply to future generations as well. But this would beg the question against non-aggregationists, who hold that impartiality implies that equal rescue chance must be given to the individuals actually involved in a rescue conflict, even if saving the greater number of people would be preferable behind a hypothetical veil.

Finally, non-aggregationism also has resources to account for the particular badness such catastrophes often come with. While non-aggregationists do not consider the interpersonal quantitative dimension of a catastrophe, they should consider the *interpersonal comparability* of the quantitative dimension of a catastrophe. That is, for non-aggregationism to be plausible, it ought to consider the stakes of the affected individuals as a relevant criterion to determine the strength of a claim these individuals may hold. For instance, if the five people in the above example only need the medication to cure their broken hand while David needs all of it to survive, a 50-50 lottery seems hardly adequate. Instead, non-aggregationists would have reason to prioritise David. This prioritisation may take the form of a lottery weighted in his favour or choosing him directly. I will consider both of these types of prioritisation below. These cases with unequal stakes are rarely comprehensively addressed by non-aggregationists, but e.g. Mannino (2021) argues that a plausible theory of non-aggregationism allows for interpersonal comparability of this sort, and prioritises people according to their claim strengths, which are dependent on their stakes. One reason why non-aggregationists can include this type of interpersonal comparability is the fact that this type of comparison does not necessarily leave the person-relative framework: When comparing a headache of person A with torture-level suffering of person B, one can plausibly hold that B's suffering is worse *for* B than the headache is *for* A.

It is in this sense that non-aggregationists can make sense of the particular badness of many intuitively catastrophic states of affairs. Many intuitively catastrophic events come with devastating consequences for the affected people, such that their claim to be helped is particularly strong. Even though it is not the number of people affected that constitutes the catastrophe, non-aggregationists can still account for the fact that these catastrophes are very bad for the separate individuals and create a strong reason to help them. Vice versa, if the number of people affected by an event is very high but the individual stakes are sufficiently low, the intuition of it being a catastrophic event is usually way less strong. This seems to suggest that our understanding of catastrophes is not as far away from the non-aggregationist picture as may be thought initially.

While interpersonally quantitatively large catastrophes may thus still pose a challenge to non-aggregationism, I do not believe this objection to undermine it sufficiently such that it ought not to be considered a serious theoretical rival to aggregationism.

In this context, it is also worth noting that Heikkinen has recently argued that non-aggregationism poses a challenge for longtermism *even if* one accepts that interpersonal quantitative catastrophes make for a *worse* outcome in an axiological sense than e.g. smaller scale accidents. This is because accepting an aggregative axiology does not force us to adopt an aggregationist deontic stance according to which this axiological fact gives us reason to prioritise avoiding the axiologically worse outcome (Heikkinen 2022: 13-14). In fact, Greaves and MacAskill also note that many non-aggregative theories are mostly concerned with deontic claims, rather than axiological claims (Greaves and MacAskill 2021: 26-27). While they provide a separate deontic argument along the lines mentioned above, according to which very high stakes always ought to override deontic side constraints, Heikkinen shows convincingly that their argument is not sufficient to reject non-aggregationism in the deontic sense, even when we accept the corresponding axiological stance. While the non-aggregationist theory I am defending here includes the rejection of aggregation both in terms of value axiology *and* deontic obligations, Heikkinen's reasoning showcases even further how pervasive the challenge for longtermism posed by non-aggregationism seems to be. The arguments in the following sections can, mutatis mutandis, likewise be applied to non-aggregationist theories which are concerned with the deontic rather than the axiological questions.

Before moving on to considering the implications non-aggregationism likely has for longtermism, I will briefly summarise the particular non-aggregationist picture I have outlined so far and will be working with in the following sections: Non-aggregationism holds that evaluative statements involving notions such as "better" or "worse" can only be made from personal perspectives. For this reason, saving the greater number of people cannot be said to be better from anyone's perspective, unless it is indeed better for everyone involved, and is thus a Pareto-optimal solution. In the case of an interpersonal conflict, however, everyone involved ought to be respected equally by giving everyone a maximal chance of having their claim satisfied that is consistent with everyone else receiving the same chance. This way, each and every person is given maximal and equal concern. However,

even though non-aggregationists do not prioritise people on the grounds that they are part of a majority group, they do consider the individual stakes at hand. That is, they prioritise people according to the strength of their claims, rather than according to their number. Hence, a non-aggregationist ought not to be on the lookout for the largest groups to save, but rather for the person with the strongest claim to satisfy. In what follows, I will further elaborate on what factors determine the individual strength of a claim.

Non-aggregationism holds that evaluative statements involving notions such as "better" or "worse" can only be made from personal perspectives. For this reason, saving the greater number of people cannot be said to be better from anyone's perspective, unless it is indeed better for *everyone* involved. In the case of an interpersonal conflict, however, everyone involved ought to be respected equally by giving everyone a maximal chance of having their claim satisfied that is consistent with everyone else receiving the same chance.

Locating the Strongest Claims

The world importantly consists of an enormously large and complex interpersonal rescue conflict. An astronomical number of beings across the space-time continuum have held, currently hold or will hold claims to be helped. As individuals and collectives, we have limited resources to affect those beings. For instance, a philanthropist can spend her money to support certain causes and not others. As individuals, we also have a limited amount of time we can invest to contribute to providing help. As such, we constantly face an interpersonal rescue conflict. The conflict examined in the present context pertains to the trade-off between interventions benefitting people of the present, near- or medium-term future versus people of the long-term future. As outlined above, Greaves and MacAskill argue that it is a moral imperative to invest one's resources into affecting the long-term future, since the most beings who are morally relevant will be affected by that. However, how does a non-aggregationist resolve this enormous conflict?

As argued above, the non-aggregationist moral imperative plausibly holds that one ought to locate the strongest individual claims and prioritise accordingly. How could one go about identifying the strongest claims? I argue that at least two criteria ought to be taken into account: magnitude of the benefit and priority to the worst-off. Let us consider their implications in turn.

Magnitude of the Benefit and Priority to the Worst-Off

Firstly, as explained above, the stakes at hand for each individual matter. The first question to be answered is thus the following: To which individuals across time and space can we offer the largest benefits? To answer this question, we first need to ask whether increasing someone's happiness is equally as important as saving someone from harm. That is, should we consider the stakes for someone for whom we can increase their level of well-being from, say, 0 to +10 as equally as high as the stakes for someone for whom we can increase their level of well-being from -10 to 0? This same debate is being held among aggregationists as well, with some suggesting that there is a certain asymmetry between suffering and happiness which warrants the prioritisation of suffering. Many of the arguments brought forward by aggregationists in this debate will likely also apply to the analogous debate for non-aggregationists. However, there may be an additional argument pointing to a suffering-focused view for non-aggregationists: Even if it is indeed true that we tend to disvalue suffering more than we tend to value happiness, aggregationists may still be able to argue that happiness ought to be prioritised in certain cases wherever there is a sufficient amount of people whose happiness can be affected, such that their number outweighs the suffering on the other side. For non-aggregationists, however, this line of argument is not available since it is not the aggregate happiness that matters, but the individual claims. Thus, non-aggregationists may more often have reason to focus on suffering than aggregationists do, at least to the extent that we assume a certain asymmetry between happiness and suffering to hold.

Is increasing someone's happiness equally as important as saving someone from harm? Should we consider the stakes for someone for whom we can increase their level of well-being from 0 to +10 as equally as high as the stakes for someone for whom we can increase their level of well-being from -10 to 0?

Be that as it may – who among those individuals whose fate we may be able to influence are the ones with the strongest claims regarding the magnitude of the benefit they might receive by us? And where are they located across all of time and space?

Even though non-aggregationists do not consider the number of people as directly morally relevant, the enormous amount of people that may come into existence in the future does play a relevant role in answering this question. Since the vast majority of people holding a claim will live in the future, it is overwhelmingly likely that the people with the strongest claims in terms of the magnitude of the benefit will also live in the future. Considering the magnitude of the benefit we may be able to provide for individuals thus seems to point to a long-term focus. Note, however, that this argument does not take into account the probability by which these benefits will actually be conferred, but rather just the net benefit that would be conferred in case our help is successful. I will consider complications regarding the probabilities of success in section 4.

Let us turn to the second criterion. Many theories of distributive justice share a certain basic normative feature according to which – *ceteris paribus* – the worse-off ought to receive a certain priority. For instance, if a certain minority group is particularly bad off in a society, this constitutes a reason to redistribute and allocate resources to them disproportionately, i.e., to give them more than their arithmetically equal share.

Similarly, there may be a reason to prioritise someone whose well-being we can affect such that they will go from -20 to -10, rather than helping someone for whom our help would make a difference in terms of going from -10 to 0. Even though the magnitude of the benefit is the same, and we are concerned with avoiding suffering in both cases, there may still be a reason to prioritise those in particularly bad situations. If this criterion to determine the strength of a claim is correct, what does it suggest for the identification of the strongest claims across time and space? The same argument from above seems to apply here too: Given

the presumably enormously large number of people living in the future, it is overwhelmingly likely that the being living the worst life ever lived, i.e., the worst-off, will live in the future. Thus, in isolation, this criterion again seems to suggest a long-term focus for non-aggregationists.

How to Prioritise

Considering the two criteria just discussed, it may thus initially seem as though non-aggregationism also suggests a moral long-term focus. Since the strongest claims of people – in terms of the magnitude of the stakes and of how bad off they are – are likely located in the far future, non-aggregationists have moral reasons to prioritise the far future.

Before I turn to complications regarding the probabilities by which our actions can actually affect different people, I would like to address the question of how exactly non-aggregationists would prioritise the individuals in the far future, if they were to do so. Firstly, it is important to distinguish between conflict cases regarding divisible goods and conflict cases regarding indivisible goods. For instance, if A and B are both very poor and need monetary support, but A needs more than B, A might have to be prioritised. If there is a sufficient amount of money available, however, a sensible prioritisation of A does not necessarily consist of giving all the money to A, but rather of splitting the money according to their needs, i.e., the strength of their claims. If it is an indivisible good, however, and A's claim is stronger than B's, then one might either only support A, or perform a lottery that is weighted in A's favour in order to respect each of the claims according to their strength. Which of these two options is sensible may depend on the exact claim strengths: It is plausible that e.g. the claims resulting from one broken arm and two broken arms should both be considered in a conflict, and thus in a case of indivisible resources a lottery weighted in favour of the two broken arms should be performed. On the other hand, in a conflict between a broken arm and a threat of death, the claim resulting from the potential death should perhaps be prioritised outright.

How exactly non-aggregationists should prioritise the far future thus depends on at least two factors: Firstly, it depends on whether the respective resources invested are divisible; and secondly, it depends on whether the claims of individuals in the far future are likely to be sufficiently strong such that present and nearer-term claims should be ignored or whether present- or nearer-term claims are also sufficiently strong, such that they should merely be given a lower weight. There are certainly cases both in which the respective resources are divisible, such as when money is to be donated to charities, as well as cases in which they are not divisible, such as when choosing certain career paths that limit one's ability to contribute to causes other than the chosen ones.

Depending on the exact details of the non-aggregationist position and the empirical realities of different claim strengths and the resources available, the exact nature of the prioritisation may look different; an area of research certainly worth exploring further. While some positions might suggest an outright prioritisation of the strongest claims, others might suggest that we perform a complex (weighted) lottery between all relevant claims, possibly including the weaker near-term claims and the stronger long-term claims.

In this context, it is also interesting to note that an allegation often raised against consequentialist moral theories, such as various forms of utilitarianism, seems to apply just as much to this inherently non-consequentialist theory, namely the problem of cluelessness. It is often claimed that e.g. utilitarian theories are implausible because it is too difficult to determine all the empirically relevant factors influencing which actions will have the best overall outcomes. However, determining which claim strengths across time and space all deserve our consideration and respect, and potentially even performing an extremely complex lottery in order to determine which of those individuals eventually receive our help, seems to be at least as difficult, if not more so, as making a comprehensive consequentialist impact assessment.

Relatedly, both the aggregationist and the non-aggregationist must rely on complex empirical analyses regarding the quantification of individual (and collective) stakes. That is, it presupposes some form of ethical calculability of benefits and needs which is interpersonally comparable. To the extent that such analyses are too empirically difficult, they pose a problem for both longtermism and the non-aggregationist alternative I am sketching here. However, this is a problem any possible theory of large-scale moral prioritisation must ultimately solve, and a whole different canon of literature is already dedicated to it, particularly in health economics where concepts such as such as the QALY (quality-adjusted life years) or the DALY (disability adjusted life years) have been developed.

Probability Discounting

As hinted at above, an important factor has not been addressed so far. Even though the individuals with the strongest claims in terms of the two criteria outlined above may indeed live in the far future, the probability with which our actions will actually make a difference for these individuals may be vanishingly small. Consider, for instance, a rescue conflict between A and B where you can either decrease the probability of death for A by 0.00001 or certainly save B from becoming paraplegic. Even if A has the stronger claim in terms of all two aforementioned criteria, the fact that the probability of successfully helping A is so small must have an influence on A's claim strength. Indeed, it seems that in this case, B ought to prioritised, since, discounted by the respective probability, B has a lot more at stake regarding our help. Since there is a lot of uncertainty regarding the potential help we can provide to people in the far future, such cases may be analogous to the rescue conflict between individuals of the present- and the near-term future and individuals of the far future: We can be fairly sure that the right donation will in fact save someone from contracting malaria or from starvation, while an attempt to influence people in the far future may very likely not have an effect at all. Thus, if we discount the assistance by its probability of success, the actual benefit we can offer people in the far future may be much lower than the benefit we can offer people in the present or nearer-term future. For aggregationists such as Greaves and MacAskill, this fact does not play as much of a role, particularly given their endorsement of expected value theory. Since even very small probabilities of affecting a vast future population by e.g. reducing the risk of extinction results in a large expected value given aggregationist premises, discounting every individual single claim with a very small probability does not undermine their longtermist conclusions. In what follows, however, I will discuss the factors which determine whether the argument does undermine longtermist conclusions for non-aggregationists.

Consider a rescue conflict between A and B in which you can either decrease the probability of death for A by 0.00001 or certainly save B from becoming paraplegic. Even if A has the stronger claim in terms of all two aforementioned criteria, the fact that the probability of successfully helping A is so small must have an influence on A's claim strength.

Stochastic (In)dependence

Expected value theory, by itself, does not distinguish between the following two cases.

Case 1: There are 1000 people on a ship which is in danger of sinking, thereby killing all the people on it. You have the possibility of negatively affecting the probability of it sinking by 0.001. In expectation, you save one person with this action. You give thus each of the people on the ship an ex ante benefit of 0.001 of survival.

Case 2: There are 1000 people threatened to be killed by a vicious disease. You can give each of those 1000 people a medication which lowers each of their respective probabilities of dying by 0.001. In expectation, you save one person with this action. You give thus each of the sick people an ex ante benefit of 0.001 of survival.

In Case 1, the probabilities of providing help for the 1000 people are stochastically dependent: you either save all of them or you save none. In Case 2, on the other hand, the probabilities are stochastically independent, i.e., some may die and some may survive. For aggregationist expected value theorists such as Greaves and MacAskill, there is little reason to distinguish between these two cases. However, there does seem to be a morally important difference between them. There are two important perspectives to be taken into account here: ex ante and ex post. Ex ante, i.e., before the action has taken place, it is indeed the case that the actions seem to benefit each of the people in Case 1 and Case 2 equally. They each get an ex ante benefit of 0.001 of survival and the ex ante expected value is the same, namely one person saved.
However, ex post, i.e., after the action has played out, the probabilities of which outcomes have actually occurred are radically different: While in Case 1, the probability that at least one of the people has actually survived is 0.001, while the probability that at least one person survived in Case 2 is higher than 0.6, since the outcomes for the stochastically independent chances are more spread out. From an ex post perspective, we can be much more sure that we will indeed have satisfied at least one claim in Case 2, which arguably constitutes a reason to prioritise the group in Case 2 over the group in Case 1. Even though from the ex ante perspective, each person in Case 1 and Case 2 gets the same expected benefit, having satisfied at least one claim ex post must be relevant for non-aggregationists as well. If we hold both the ex ante and post perspective to be important, then both of these perspectives have to be taken into account when evaluating the relevance of probability discounting.

Longtermism Ex Ante and Ex Post

Consider a non-aggregationist who attempts to compare the claim strengths between individuals of the far future and individuals in the present- or near-term future. As outlined in the previous section, notwithstanding success probabilities, it may look as though individuals in the far future hold the stronger claims. However, what does it look like if we add the discounts due to the probabilities of success? This depends crucially on whether we take an ex ante or an ex post perspective.
As explained above, it seems that the claims of future individuals would have to be discounted by a lot, when considered from an ex ante perspective. The benefit we can offer to an individual of the present or near-term future when discounted by the probability of success is a lot higher than the benefit we can offer an individual of the far future, given that it is extremely uncertain whether our help will affect them at all. Individuals of the present and nearer-term future thus clearly seem to hold a stronger claim from an ex ante perspective, once the discounts given the probabilities of success are added.
However, it may look differently from an ex post perspective. This depends crucially on whether affecting the longer-term future is more analogous to Case 1 or analogous to Case 2. While the numbers do not count in and of themselves for a non-aggregationist, the number of people influenced does make a difference even for a non-aggregationist if the benefits offered to them are stochastically independent. To see why, consider two further cases:

Case 3: You can either relieve individual A of medium pain with a probability of 1, or provide a probability of 1/100 of a relief of intense pain for individual B.

Case 4: You can either relieve individual A of medium pain with a probability of 1, or provide a probability of 1/100 of a relief of intense pain for each person of a group of 100,000 individuals.

From an ex ante perspective, individual A probably ought to be prioritised in both cases according to non-aggregationism, because their claim is stronger when probability discounting is added. However, from an ex post perspective, it is incredibly likely that at least one of the 100,000 individuals in Case 4 will have been spared from intense pain, and thus the large number of the group in Case 4 provides a reason to prioritise them from an ex post perspective. Ex post, it is simply very likely that at least one person in the large group will have benefited more from your helping them than A would have, if you had helped them.
Returning to the question of longtermism, non-aggregationists would have to ask the following question: Are there actions that grant the vast populations of the future probabilistically small but stochastically independent benefits? Or is influencing the future more analogous to Case 1, such that the probability of success for having helped any of the people in the future is very small ex post?
If our situation is analogous to Case 1, the ex post perspectives would converge with the ex ante perspective and suggest a nearer-term focus for non-aggregationism. However, if our situation is analogous to Case 2, the ex post perspective likely supports a long-term focus, since there would be a very high likelihood that a strong claim will have been satisfied ex post.
So which of the two cases are we actually faced with? Unfortunately for longtermists, I believe that influencing far off future populations will turn out to be mostly analogous to Case 1. In particular, their focus on interventions such as reducing extinction risks or influencing high-impact developments such as work-

ing on AI alignment, as Greaves and MacAskill suggest, seems to influence future populations in a stochastically dependent way. Extinction influences the future (non-)existence of all future populations simultaneously and different AI alignment scenarios influence the shape of the world and universe for all future populations at the same time. For if humanity goes extinct, this extinction event will have been the cause for the non-existence of all future populations simultaneously. Similarly, the world created by a misaligned superintelligent AI will influence future populations simultaneously. To the extent that this reasoning is correct, it also does not matter much which exact extinction risk we are considering, for *if* humanity does go extinct, it will have influenced the nonexistence of future populations simultaneously, no matter which exact extinction risk was ultimately realised.

One may object that some extinction risks influence existing people differently, rather than simultaneously, before extinction actually occurs. For instance, climate change will have adverse effects on different parts of the global population in different ways, and various interventions will influence people differently. However, this is not an argument the longtermist view has at its disposal, since its focus on extinction risk is primarily justified by the vast number of people who will not come into existence in case of extinction, rather than the people who are influenced by these risks before extinction occurs.

If this reasoning is correct, neither extinction risks nor AI alignment influence people in a stochastically independent way. Thus, if the probabilities of success regarding respective interventions is sufficiently small, the vast number of people who are influenced by them is still not relevant for a non-aggregationist from an ex post perspective. For this reason, I believe that non-aggregationism does ultimately not support longtermist conclusions. Consequently, a strong focus on reducing extinction risks must be justified in a different manner to non-aggregationists, if at all.

There may be a certain sweet spot between the present and the far future, such that the probabilities of success are still sufficiently high and the number of people sufficiently large such that some of the people affected are likely going to have very strong claims in terms of the two criteria outlined above. For instance, the number of people who will be living in the next few centuries is likely going to be a lot larger than the number of people living today. Hence it is also likely that some of these people will hold some of the strongest claims. At the same time, the probability of success regarding possible help for these future people may be sufficiently high for it to still be worth it, even if the respective help is stochastically dependent. In particular, non-aggregationists may get behind efforts to reduce extinction risks or other high-impact developments to the extent that they are sufficiently tractable, such that the probability of success is sufficiently high to be worth it. Non-aggregationism may thus support a much weaker form of anti-presentism, according to which the very strong focus on present issues is viewed critically, but not replaced with a focus on the very long-term future but a near- to medium-term focus instead.

Conclusions and Outlook

In this paper, I have argued that strong non-aggregationism is a relevant alternative to aggregationism. It has a compelling rationale and can be defended against obvious objections, such that it should be taken seriously when considering which large-scale moral imperatives we are faced with. For this reason, it is important to consider what exact moral recommendations follow from non-aggregationism. In this paper in particular, I explored its implications for longtermism and a focus on extinction risks and other high impact long-term developments. To this end, I outlined one version of non-aggregationism and argued that such non-aggregationists ought not to be on the lookout for the largest groups to save, but rather for the individuals with the strongest claims to satisfy. I have furthermore argued that at least two criteria are relevant for determining claim strengths: magnitude of the benefit and priority to the worst-off.

Considering these two criteria in isolation, non-aggregationism seems to suggest a long-term focus as well, and thus influencing the existence and welfare of generations in the far future would be a central imperative of non-aggregationism, too.

However, when taking the probabilities of success into account, this result cannot be maintained. From an ex ante perspective, the discounts resulting from the uncertainty of being able to affect the far future suggest that non-aggregationists ought not to prioritise the far future. From an ex post perspective, it depends on whether the probabilities with which people in the far future can be influenced are stochastically dependent or independent. In the former case, the ex post perspective likewise suggest more of a near- to medium-term focus, rather than a long-term focus. In the latter case, longtermist conclusions look more attractive. However, since the empirical reality seems to be more similar to the former case, I have tentatively concluded that non-aggregationism likely does not recommend a long-term focus. In particular, reducing extinction risks is an example of stochastically dependent probabilities with which future generations can be influenced. However, to the extent that the extinction risks are sufficiently tractable, non-aggregationists can support a respective focus as well, even if less strongly so.

There is ample space for further research in this area. All the different versions of non-aggregationism could be examined regarding their implications for longtermism. Furthermore, more research into the criteria which determine claim strengths might turn out to be very important in answering the questions I have outlined. In this context, investigating both different versions of prioritisation and the strength of the cluelessness objection for non-aggregationist theories may likewise be very interesting. Finally, further work on the correct analysis of the implications of stochastically dependent and independent probabilities of success, and the correct analysis and combination of the ex ante and ex post perspectives will also likely significantly influence the answers to the questions I have outlined.

Notes

1 Note, however, that the term "anti-presentism" in this context is not meant to imply that there is normative discounting of present interests. Rather, insofar as present and future generations all count equally, the future generations just vastly outweigh the present generations such that the latter end up mattering much more. Thus, to the extent that there are trade-offs in benefitting the present generations and future generations, one ought to opt for benefitting the vastly greater future generations. "Anti-presentism" in this context is not meant to refer to anything other than this basic implication of the classic longtermist argument.

2 I thank an anonymous reviewer for this helpful example.

3 The former is often referred to as one of the most cost-effective

ways to save lives in the present (Greaves and MacAskill 2021: 2), while the latter is one of the explicit priorities of many longtermist organisations, such as e.g. the career advice centre "80000 hours".

4 I thank an anonymous reviewer for the suggestion to clarify this point.

References

Beckstead, Nicholas (2013): On the Overwhelming Importance of Shaping the Far Future.
Ph.D. thesis, Rutgers University, New Jersey. https://rucore.libraries.rutgers.edu/rutgers-lib/40469/PDF/1/play/. Viewed 9 December 2022.

Bostrom, Nick (2013): Existential Risk Prevention as Global Priority. In: Global Policy, 4 (1), 15-31. https://existential-risk.org/concept.pdf. Viewed 9 December 2022.

Deutscher Bundestag (2022): Entwurf eines Zweiten Gesetzes zur Änderung des
Infektionsschutzgesetzes. https://dserver.bundestag.de/btd/20/038/2003877.pdf. Viewed 9 December 2022.

Eurotransplant (2019): Eurotransplant Manual. https://www.eurotransplant.org/allocation/eurotransplant-manual/. Viewed 9 December 2022.

Greaves, Hillary, and William MacAskill (2021): The case for strong longtermism.
GPI Working Paper no. 5-2021. Global Priorities Institute. https://globalprioritiesinstitute.org/wp-content/uploads/The-Case-for-Strong-Longtermism-GPI-Working-Paper-June-2021-2-2.pdf. Viewed 9 December 2022.

Heikkinen, Karri (2022): Strong longtermism and the challenge from anti-aggregative moral views. GPI Working Paper no. 5-2022. Global Priorities Institute. https://globalprioritiesinstitute.org/wp-content/uploads/Karri-Heikkiken-GPI-working-paper-1.pdf. Viewed 9 December 2022.

John, Tyler M. / Millum, Joseph (2020): First Come, First Served? In: Ethics, 2 (130), 179–207.

Krznaric, Roman (2020): The good ancestor: How to think long term in a short-term world. London: Ebury Publishing..

Lübbe, Weyma (2015): Nonaggregationismus: Grundlagen der Allokationsethik. Münster: Mentis.

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

Mannino, Adriano (2021): Wen rette ich, und wenn ja wie viele? Stuttgart: Reclam.

OPTN Organ Procurement and Transplantation Network (2019). Ethical Implications of Multi-Organ Transplants. Richmond. https://optn.transplant.hrsa.gov/media/eavh5bf3/optn_policies.pdf. Viewed 9 December 2022.

Ord, Toby (2020): The Precipice. Existential Risk and the Future of Humanity. New York: Hachette Books.

Parfit, Derek (1984): Reasons and Persons. Oxford: Oxford University Press.

Pettigrew, Richard (2022): Should Longtermists Recommend Hastening Extinction Rather Than Delaying It? Unpublished manuscript.

Rakić, Vojin (2018): Disaster Consequentialism. In: O'Mathúna, Dónal P. / Dranseika, Vilius / Gordijn, Bert. (eds.): Disasters: Core Concepts and Ethical Theories. Cham: Springer International Publishing, 145–156.

Rawls, John (1971): A Theory of Justice. Cambridge, MA: The Belknap Press of Harvard University Press.

Taurek, John (1977): Should the Numbers Count? In: Philosophy and Public Affairs, 6 (4), 293–316.

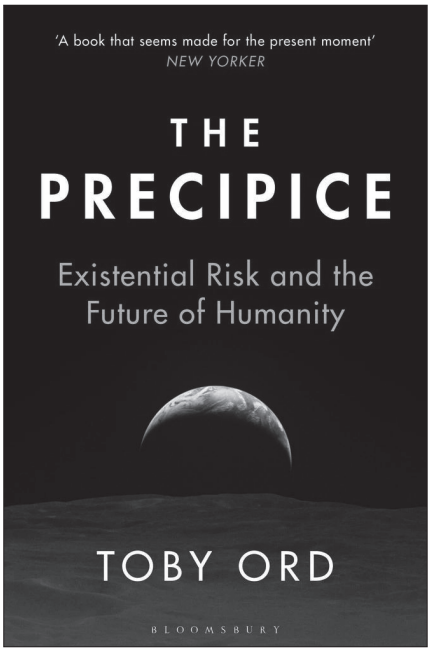
Thompson, Dennis F. (2010): Representing Future Generations: Political Presentism and Democratic Trusteeship. In: Critical Review of International Social and Political Philosophy, 13 (1), 17–37.

Tremmel, Jörg (2021): The Four-Branches Model of Government: Representing Future Generations. In: Cordonier Segger, Marie-Claire / Szabó, Marcel / Harrington, Alexandra R. (eds.): Intergenerational Justice in Sustainable Development Treaty Implementation: Advancing Future Generations Rights through National Institutions. Cambridge University Press, 754–780.

Marina Moreno is a PhD student at the Munich Center for Mathematical Philosophy at LMU Munich and holds an MA in logic and philosophy of science.
Email: marinaestrellamoreno@gmail.com

Toby Ord:
The Precipice: Existential Risk and the Future of Humanity

Reviewed by Tolga Soydan



Imagine yourself rolling a dice, but instead of winning at a game, you find yourself rolling the dice on the fate of humanity, having a chance of 1 in 6 of destroying it over the next hundred years. Would you do it? Probably not, unless you are ridiculously confident or careless. In Toby Ord's book *The Precipice: Existential Risk and the Future of Humanity*, the senior researcher at the Future of Humanity Institute in Oxford argues that unless humanity does not take the possibility of existential risks more seriously, it stands the same chance of getting itself or its potential destroyed in the next century. So, the question arises: Why do we all roll the dice with such stakes?

Ord's ambition is clear: Showing humanity the risks it faces, warn us and even more, showing us the heights we could theoretically achieve in the long term, if we play our cards right.

To accomplish this, Ord divides *The Precipice* into three parts.

The first part, *The Stakes*, takes the reader to the humble beginnings of mankind, how we tamed nature and worked together and eventually reached the top of the food chain. But with great power also came great potential for destruction, reaching its practical pinnacle through the use of nuclear weapons in the Second World War. 1945 therefore marks the very beginning for Ord, where we reached the Precipice: the state in which humanity eventually possesses the means to destroy itself. Ord fears that there may be too much of a difference between our power and our wisdom to wield such power responsibly at the moment, putting us in a situation of higher existential risk.

Existential risks are defined as all the risks that could destroy humanity, leading to its extinction or permanently destroy its future potential, for example by getting locked into a dystopian scenario or suffering a permanent social collapse. Existential catastrophes are impossible to be undone and can never be allowed to happen. The importance of the matter is founded in the possibility of the trillions of people who could be born in the future, if we manage to avoid existential risks, as well as in all the lives before us that made the present possible. Ord calls the protection of existential risks an „intergenerational global public good" (59), as it especially benefits future humans. This good is insufficiently funded, comparing the billions of dollars that are spent on the work on AI to the millions of dollars that are spent on making sure AI is aligned with human values. One further glaring deficit in avoiding existential risks is the lack of a centrally coordinated institution.

In the second part, *The Risks*, Ord divides the existential risks into natural, anthropogenic and future risks. The natural risks are cases such as an asteroid or comet impact, supervulcanic eruptions or stellar explosions. He argues that we are well equipped in the case of a potential asteroid impact, because we have identified over 95% of the dangerous objects. As for stellar explosions and supervulcanic eruptions, the fossil record gives reasons to be fairly optimistic that those risks will stay minimal in the foreseeable future. Still, Ord pleads for more research on the field. Compared to the anthropogenic risks, he estimates the danger of natural existential risks a thousand times smaller (87).

Anthropogenic risks are risks such as nuclear weapons, climate change and general environmental damage. Even though each of those risks presents more of an existential risk by itself than all the three natural risks combined, Ord suggests it would be speculative to assume these anthropogenic risks to be sufficient to destroy humanity as a whole or its long-term potential. Nevertheless, Ord is in favour of more research on the effects of anthropogenic risks as well.

Ord finally locates the greatest danger for humanity in future risks connected to technology. He closely inspects the dangers of pandemics and biotechnology, unaligned artificial intelligence, dystopian scenarios and a few other risks, such as nanotechnology. Talking about pandemics, Ord highlights the dangers of biotechnology and information hazards, as unfiltered public information could lead bad actors to try and capitalize on the available technology and release deadly viruses. To date, the hypothesis that SARS-CoV-2 escaped from a laboratory in Wuhan, China, has not been completely dismissed. But Ord's main concern seems to be unaligned artificial intelligence, where he estimates the risk over the next hundred years to be on a 1 out of 10. If humanity were to successfully create a general AI smarter than human beings, our own fate would not necessarily be in our hands anymore. We do not know how to implement our values into AI, and yet we steadily upgrade the capabilities of AI making it more and more likely to put ourselves at risk.

In the third part, *The Path Forward*, Ord maps out in detail how he calculated the risks we could potentially face, how those risks could combine and how specific risk factors such as climate or economic failures could raise the danger of existential risks, and how specific safety measures could in turn lower it, such as achieving peace between the powerful nations. In addition, he urges us to re-evaluate the way we deal with risks on a theoretical and

practical level, strongly advising a more centrally organised policy making and binding powers to protect humanity and suggests representatives who stand in for future generations.

Looking to the future, Ord proposes three phases in which humanity could fulfil its potential. First, we have to reach Existential Security. For him this means to preserve and protect our potential by taking the risks seriously and managing them from their onset or avoiding them. The second phase, the Long Reflection, should be the time humanity literally spends time reflecting on the road it wants to take, choosing its best options. The third and last phase should see us achieving our potential. He keeps this section quite vague, explaining that humanity should first focus on reaching security.

As the state of knowledge on this field is quite young, he advises researchers to be more specific on possible risks and to be cautious about what not do, for example regulating prematurely and ignoring the positives for the sake of exaggeration. He advises everyone interested in the field to make a change through their professional careers or by donating money. He finishes the last part of his book by drawing upon the imagination of a humanity colonising the universe and maybe even changing its nature to reach the next stage in evolution, if needed. The humans of tomorrow need a chance to fulfil all the things we today can only dream of.

Ord presents an exciting and very good introduction for all those interested in the field of existential risks. He writes eloquently and yet very understandable, avoiding technical terminology wherever possible while explaining it well whenever he can't, making it an altogether interesting read even for a non-academic audience. The structure of the book is inherently sound and his overall tone of voice sounds calm and rational. And yet, this very interesting book is not without its flaws.

First, Ord leaves out a major part of philosophical debates revolving around population ethics, dedicating only a few pages in the appendices to it. The book could have benefited immensely from this if it dived deeper into the debates of human nature, ethics, population and potential. Especially the debate around s-risks (risks of astronomical suffering) that explain how a future does not only have to include happiness but also an huge amount of potential suffering could have been helpful. S-risks put into question whether extinction would be the worst scenario if the alternative would be to cause unprecedented amounts of suffering. Thus in some scenarios, we could not find ourselves in an existential risk, but a s-risk. Lowering the existential risk could therefore raise the s-risks. How then do we avoid existential and suffering risks and still find the best future? Ord argues that we constantly made progress, fighting poverty, strengthening women's rights, and making education possible for more humans than ever before, but that this does not guarantee our steady progress in the future. We could still evolve back on issues or never find a consensus on important subjects. The Long Reflection part of the book is made out to be the time when humanity finally gets its act together and decides its path in unison – but we should already be talking about all these important issues now, because they determine the way we will walk. Hence we should not worry about bringing people into existence first, but worry about whether those people can live a life worth living. Ord could have given his opinion on the procreation asymmetry and how this influences longtermism and dealing with existential risks.

Second, Ord spends a lot of time on the danger of unaligned AI for humanity, but he neglects the dangers of such a powerful existence for the universe. An unaligned AI could theoretically not only control our planet but decide to colonize space and extend its influence into the galaxy causing irreparable damage and suffering not only for us, but also for other sentient beings, if they exist.

Third, Ord talks about the potential of humanity as if it were an individual, but it is a collective. There is not „one humanity" with its intentions and hopes, but instead people hold many different views and values. He imagines the potential of humanity to be one of high art and science, but one inevitably wonders about the negative potential mankind has also shown to possess, its aggressiveness. Every year we kill billions of animals as a food resource, we wage war against each other and still allow people to starve to death in some parts of the world. What potential for inflicting pain might we possess in the future?

Fourth, Ords view on longtermism, deciding what to do depending on the long-term effects, may be logical from the viewpoint of existential risks, but it could come with catastrophic consequences for present people. For example, if you had to let millions of people suffer now so that in the long-term humanity could benefit from it, you would be inclined to let it happen. But are we not morally obliged to stop suffering whenever we encounter it? Does the suffering of now really pale compared to the happiness of tomorrow? And what kind of quality does the happiness of the future hold, if it was at least partially founded on the sorrow of the past? Talking about the trillions of potential humans in the future suggests that a few million who suffer now don't matter as much, but they do. They are real, they exist and they suffer in contrast to the non-existent humans of the future. There is a real danger of trivialising human lives for the sake of the big picture. Climate change will likely not be the end of humanity, but it will still bring immeasurable pain and suffering to many people, if not stopped – but still this does not make it an existential risk for Ord. But I argue it is an existential risk for all those who will die because of it, will lose land and family and lose hope for the future because of it.

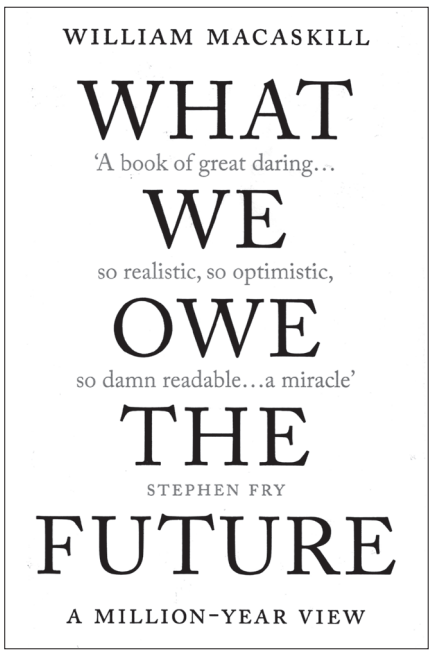
Fifth and finally, the chapter on the risk landscapes seems at times a bit problematic. Ord believes, all things considered, that our odds of facing an existential risk in the next century stand at 1 in 6. Yet, we are talking about risks that have never occurred and that can often only be estimated in rough ways, or that could potentially be much bigger or lower than we might dare think. Ord admits that all of his estimates are just his best guesses and should not be taken as precise mathematics, but those evolutions need a stronger ground on which to base our actions on if we were to take existential risks more seriously. We will need more work on the field of risk theory to better understand existential risks.

In the end, Toby Ord has delivered a very compelling book on one of the most interesting and maybe underrepresented subjects in the public discourse. He manages to give a well written introduction into existential risks, even though it ignores a large spectrum of philosophical debate, but leaves the reader wanting to learn more about our potential and the risks we could face. Its maybe biggest accomplishment is to give the reader a sense of hope, even in the face of our potential doom. One can only agree with Ord, that things are always largely in our hands.

Ord, Toby (2020): The Precipice. Existential Risk and the Future of Humanity. London: Bloomsbury Publishing. 480 pages. ISBN 9781526600219 (hardback), ISBN 9781526600233 (paperback), ISBN 9781526600196 (e-book). Price: hardback $34.00/£25.00; paperback $14.95/£10.99; e-book $11.96/£8.79

William MacAskill: What We Owe the Future: A Million-Year View

Reviewed by Grace Clover



As a teenager living in Glasgow, the philosopher William MacAskill enjoyed urban climbing, on one occasion putting his foot through a skylight and narrowly escaping puncturing his internal organs on broken glass. At the time, he saw the likelihood of falling and dying as insignificant and thus untroubling. But now aged 35, MacAskill admits that his youthful insouciance was foolish, not because his death was *likely*, but because it 'wasn't *sufficiently unlikely*' to warrant risking such severe consequences (39). This is how MacAskill – a founding member of the Effective Altruism movement, now a researcher at the Global Priorities Institute at the University of Oxford – represents current generations in this book: as a short-sighted teenager, obliviously making decisions which will impact its long-term future. While we cannot exactly predict the likelihood or value of existential risks, he argues that they are now far too likely to remain overlooked.

MacAskill's latest work, *What We Owe the Future,* is indicative of a wider trend within the Effective Altruism movement in the last ten years, which has seen its priorities shifting away from utilitarian charitable spending on global poverty towards a greater concern with existential risks and the entrenchment of global values. His book offers a moral justification for longtermism and a framework for dealing with uncertain expected value. As in his previous book *Doing Good Better* (2015), MacAskill calls upon the reader to take a rational and disimpassioned approach to improving the world, challenging the assumptions which guide our actions, and leading us to seemingly counterintuitive but logically argued conclusions.

MacAskill calls 'longtermism', understood as an ethical theory, a "key moral priority of our time." (3) His justification for longtermism is as follows: People in the future could exist, and there could be a lot of them (9). These people should matter no less, morally, than people alive today. He writes: "I am not claiming that the interests of present and future people should always and everywhere be given equal weight. I am just claiming that future people matter significantly." (11) So long as these people live sufficiently happy lives (he does note ethical and practical problems in measuring this), it is of moral value that they are able to live. Even if the human race only exists for a fraction of the evolutionary lifespan of the average mammal (one million years), billions of people could still live in the future. This foundational thought underpins the rest of the book.

MacAskill begins by considering how we can improve the value of life in the future, theorising about how we can ensure that the future is a 'morally exploratory world', which prioritises improving wellbeing (99). He suggests that historical value changes – such as the abolition of slavery – were the contingent outcomes of one value system becoming culturally 'fitter' over time and outcompeting others, partly due to the work of activists. However, he warns that in the future such moral progress may become increasingly difficult due to a 'value lock-in' caused by the premature convergence of a global culture or by the creation of an artificial general intelligence able to implement its own values or those of a specific group. To demonstrate this, MacAskill employs a metaphor of history as molten glass, with periods during which our values are malleable, before the glass sets and they become enduringly entrenched.

MacAskill then moves onto existential risks, assigning one chapter respectively to extinction risks, civilisational collapse, and technological stagnation. MacAskill emphasises the risk posed by developments in nuclear warfare and engineered pathogens, even if they are never intentionally deployed: after all, lab leaks and nuclear false alarms occur with alarming frequency. Such catastrophes, if not causing extinction, could also drastically reduce our ability to collaborate internationally on other risks, such as climate change. Here civilisational collapse is defined a non-extinction threat through which we lose 'the ability to create most industrial and post-industrial technology'(124). He does remain optimistic, however, and notes that in the past, mankind as a whole has been remarkably resistant to catastrophes such as epidemics and global warfare. He suggests that an existential risk scenario, such as a nuclear winter, would not affect the entire globe equally, likely leaving areas such as Australasia relatively unharmed. He poses this as positive, as it would allow our species to survive, re-industrialise, and re-develop. Finally, MacAskill emphasises the risk posed by technological stagnation, arguing that as global birth rates slow, so must the rate of technological development. He sees this as detrimental to our ability to respond to existential risks. To this end, he cautions against the familiar environmentalist narrative that having children is unsustainable and instead promotes having children as a way to personally ward against civilisational collapse.

Following this, MacAskill introduces a theory of population ethics influenced by the moral philosopher Derek Parfit (1942–2017). He argues that the biological extinction of the human race would be, morally speaking, significantly worse than a non-extinction risk that killed 99.9% of the world's population, as it would prevent the

lives of millions of people who might otherwise have lived in the future. He critiques the logical asymmetry of the 'intuition of neutrality', a philosophical viewpoint which sees bringing an unhappy life into the world as morally bad but bringing a happy life into the world as morally neutral (171). Instead, MacAskill argues – with some caution – that only 10% of the world's population today have below-neutral wellbeing, and thus on balance, the future will more likely be good than bad for the people living in it. He predicts that global wellbeing will increase overtime, drawing a causal relationship between increased wealth, happiness, and moral progression (assuming that we avoid value lock-in and stagnation). As such, he suggests that we have a moral obligation to ensure that future populations are able to live, and potentially grow indefinitely.

Finally, MacAskill offers practical advice about what individuals can do to implement longtermism. The arguments here are mostly familiar from his earlier writings: he suggests that the focus on personal consumption in the environmentalist movement is often misplaced and instead emphasises the good individuals can do by donating to effective charities, having children, and making well-considered career choices.

Structurally the text might have benefited if the discussion of population ethics presented in *Part IV: Assessing the End of the World* had immediately followed the moral argument for longtermism in part I, but otherwise the book's argument proceeds logically. MacAskill could have also focused more on the impact of longtermism on ecosystems and non-human animals, which remain largely overlooked. That said, the book is expansive in scope and very coherently written. As a philosopher, MacAskill is no stranger to the use of thought experiments to justify extrapolating moral positions. It is perhaps more impressive that his case studies from fields as diverse as history and zoology are so effective and evocative. The moments of personal reflection about his own life as well as his friends and colleagues also offer particularly engaging touches of warmth.

Implicit however in every part of the book – from his metaphor of humanity as a singular teenager, to his use of aggregate moral value and quality-adjusted life years – is MacAskill's treatment of humanity as an individual, rather than a collective made up of many parts with independent needs. Such a premise is foundational to his utilitarian emphasis on doing the maximum amount of good for humanity as a whole, whilst avoiding emotional assessments of individual need. This dehumanising tendency could easily alienate many readers from his conclusions. For example, most people who are concerned about climate change would agree that decarbonisation is key to our path to a sustainable future, improving the health of current people and the safety of future generations. This view is entirely coherent with the model of longtermism which MacAskill proposes here. He even describes decarbonisation as the yardstick for judgeing all longtermist action. However, many would find it absurd, or at least too abstract, that he justifies decarbonisation partly on the basis that we must leave easily accessible fossil fuels available for re-industrialisation following civilisational collapse. Under this logic, the deaths of billions in such a collapse are brushed aside, so as to emphasise the moral benefit of future population growth.

This is indicative of a more integral problem with MacAskill's work: his unwillingness to engage with the practical and emotional implications of death, or the social systems which underpin global suffering. MacAskill does note that it is a 'colossal injustice' that developing countries who contributed least to the climate crisis are likely to be most impacted by it (36), but he fails to engage with what this injustice means in practice: the intense suffering caused by drought, flooding, famine, and natural disasters, and the lack of financial resources to recover from it. Nor does he indicate any global structural changes which could even out this injustice, such as the Loss and Damage Fund agreed upon on the world climate conference in Sharm el-Sheikh in 2022. As his earlier work has shown us, MacAskill is certainly not ignorant of global suffering. But in emphasising the moral obligations we have for the future, suffering in the present appears to have lost some of its emotional weight. Regardless of what one thinks about longtermism, in an ideology framed around improving wellbeing, this seems like a contradiction.

Despite this, MacAskill offers an urgent but upbeat call to action to deal with existential risks, written in an accessible and engaging style. Though MacAskill remains deliberately cautious when drawing conclusions about the future and warns against complacency, the overriding impression left by *What We Owe the Future* is an optimism about our ability to positively impact the longterm and about the expected value of the future itself.

MacAskill, William (2022): What we owe the future: A million-year view. London: Oneworld Publications. 352 Pages. ISBN: 9780861542505 (hardback), ISBN: 9780861542512 (paperback). Price: hardback £20; paperback £10.99.

Imprint

Publisher: The Foundation for the Rights of Future Generations (Stiftung für die Rechte zukünftiger Generationen) and The Intergenerational Foundation
Permanent Editor: Jörg Tremmel
Co-editor for IGJR 1/2022:
Markus Rutsche, Felix Beer
Additional Editor: Grace Clover
Layout: Angela Schmidt, Obla Design
Print: Kuhn Copyshop & Mediacenter, Nauklerstraße 37a, 72074 Tübingen
Website: www.igjr.org
Email: editors@igjr.org

Editorial offices:
Foundation for the Rights of Future Generations (Stiftung für die Rechte zukünftiger Generationen)
Mannspergerstraße 29
70619 Stuttgart, Germany
Tel.: +49(0)711 – 28052777
Email: kontakt@srzg.de
Website: intergenerationaljustice.org

The Intergenerational Foundation
19 Half Moon Lane
Herne Hill
London SE24 9JU
United Kingdom
Email: info@if.org.uk
Website: if.org.uk