



# Existential Risk and Rapid Technological Change

Advancing risk-informed development

2023

**February 2023**

### **Authors**

Maxime Stauffer,<sup>1</sup> Simon Institute for Longterm Governance, Switzerland  
Konrad Seifert, Simon Institute for Longterm Governance, Switzerland  
Angela Aristizábal, Instituto Tecnológico Autónomo de México, México/Colombia  
Hamza Tariq Chaudhry, Harvard University, United States of America/Pakistan  
Kevin Kohler, ETH Zurich, Switzerland  
Sumaya Nur Hussein, Strathmore University, Kenya  
Claudette Salinas Leyva, Instituto Tecnológico Autónomo de México, Mexico  
Arne Gebert, Simon Institute for Longterm Governance, Switzerland  
Jacob Arbeid, Simon Institute for Longterm Governance, Switzerland  
Mahaut Estier, Simon Institute for Longterm Governance, Switzerland  
Sandra Matinyi, SynBio Africa, Uganda  
Jason Hausenloy, Independent, Singapore  
Jasmin Kaur, Independent, United Kingdom  
Shrestha Rath, Effective Ventures, United Kingdom  
Yung-Hsuan Wu, Geneva Graduate Institute, Switzerland

<sup>1</sup> Lead author, contracted by the United Nations Office for Disaster Risk Reduction

### **Acknowledgement**

We thank the following individuals for their valuable feedback:

Michael Aird, Rethink Priorities, United States of America  
Haydn Belfield, Centre for the Study of Existential Risk, United Kingdom  
Belinda Cleeland, International Standardization Organization, Switzerland  
Piers Millett, International Biosecurity and Biosafety Initiative for Science, United States of America  
Cassidy Nelson, Oxford University, United Kingdom  
Abi Olvera, Rethink Priorities, United States of America  
Jess Whittlestone, Centre for Long-term Resilience, United Kingdom

### **Cover image**

We used the generative art AI “Midjourney” to create the image, using the prompt: “differential technological development guided by a diverse circle of wise humans, 4k --ar 21:29”.

### **Citation**

Stauffer et al., (2023), Existential Risk and Rapid Technological Change: Advancing Risk-informed Development, United Nations Office for Disaster Risk Reduction.

### **Supplementary information**

Additional contents and methodological details can be accessed [here](#).

---

This report was commissioned by the United Nations Office for Disaster Risk Reduction to inform the Mid-Term Review of the Sendai Framework. For questions, please contact Maxime Stauffer at [max@simoninstitute.ch](mailto:max@simoninstitute.ch).

## Executive summary

On 10 March 2022, UN Secretary-General, António Guterres, stated that <sup>1</sup>:

“(…) there is renewed pressure to consider whether global governance systems are fit for purpose, and how they could be improved. Even as we reconsider traditional threats to peace and security, we need to update these concepts for our more complex world, in which local threats may quickly become global, existential, and intergenerational.”

The mid-term review of the Sendai Framework for Disaster Risk Reduction provides an opportunity to tackle this task. There is growing evidence that the worst-case scenario at a global scale – an existential risk – is increasingly likely and driven by technological development. Of particular concern are biotechnology and artificial intelligence, as major drivers of accelerating change. Importantly, both existential risk and technological hazards are within the scope of the Sendai Framework.

Historically, the risks from emerging technologies have often been small compared to their benefits. That is no longer the case. Because of the increasing pace of technological change across the globe, it is becoming more difficult for risk governance to keep up. While new technologies can bring society enormous benefits and significantly contribute to achieving global goals such as the UN Sustainable Development Goals, they usually have unintended effects, often cause harm accidentally, and are sometimes misused.

Biotechnology and artificial intelligence are two technologies that pose existential risk if their development and deployment are not properly governed. Biotechnology brings together science and engineering to (re)design, manufacture and/or modify genetic materials, living organisms and biological systems to, for example, treat genetic diseases, create new biofuels or more nutritious crops. However, recent technological advances have greatly reduced the costs and provide access to actors who could, deliberately or accidentally, create and release a dangerous pathogen. What’s more, as a reaction to the COVID-19 pandemic, many States have increased their investment in high-risk bio research facilities that can handle the most dangerous pathogens, thus further increasing the chances of their mishandling or accidental release. A lethal ‘engineered pandemic’ is therefore a real possibility.

Artificial intelligence (AI) has experienced a similar rapid advancement and expansion, to the point that it is now widely employed and has acquired many capabilities that experts predicted would take at least another 5-10 years to develop. While simple AI does things such as list sorting and pattern recognition, it is the development of ‘transformative artificial intelligence’ (TAI) that could lead to entirely new technological hazards. TAI, based on reinforcement learning, may enhance our ability to perceive, reason about and act in the world, leading to radical changes in all areas of society. Fundamentally, the key problem with TAI is *value misalignment*: as objectives cannot be hard-coded, TAI is difficult to align perfectly with human values.

Existing governance structures are not fit-for-purpose to address existential risk in general nor from technological development. Institutional incentives lead policymakers to focus on short-term, higher-probability events that are easier to predict and more definitively linked to their national territory. Individual countries tend to take limited ownership of global risks as governance structures tend to be slow moving, reactive rather than proactive, and have been set up with mandates that focus on specific kinds of risks, but neglect existential risk, more broadly.

To make progress, it will be necessary to address four priorities for reform. First, the United Nations Office for Disaster Risk Reduction (UNDRR) should foster a concrete and common understanding of existential

risk. Second, UNDRR, other UN agencies and member-states should strengthen existential risk governance. Third, the UN system and national governments should dedicate more resources to existential risk reduction. Fourth, UN agencies should foster fast response mechanisms to stop shock cascades.

To ensure tangible progress towards these goals, we recommend

1. The creation of **an international coordination and capacity-building mechanism on existential risk**; and
2. The incorporation of **lower-probability, high-impact risk in existing funding instruments**.

This would constitute momentous progress towards global existential risk reduction, making the world a significantly safer place for current and future generations.

#### **Box 1. This thematic study in numbers**

- An existential risk is 1.9 to 14.3% likely this century
- 3 multilateral pathways and 55 milestones to reduce existential risk until 2030
- 12 outcomes to reach
- 2 priority instruments to develop
- 30 actions to implement

#### **Box 2. Methodology followed for this thematic study**

- A literature review on existential risk and rapid technological change;
- A trend analysis of technological change;
- A review of crowdsourced forecasts on future risks;
- A keyword analysis of national voluntary reports;
- A written consultation with youth and global south actors;
- A written consultation with experts;
- A multi-stakeholder workshop on future-proofing the multilateral system;
- A set of discussions with member-states, the private sector and staff members at UNDRR.

# Contents

<b>1. How serious is existential risk?</b>	<b>5</b>
1.1. Existential risk as the global worst-case scenario	6
1.2. The historical signature of disasters	7
1.3. Existential risk estimates and implications	10
<b>2. Rapid technological change contributes to existential risk</b>	<b>13</b>
2.1. Rapid technological change as a source of existential threats	14
2.2. Technology 1: Biotechnology – existential risk from engineered pathogens	16
2.2.1. Lack of oversight and lack of awareness make catastrophe more likely	16
2.2.2. Costs go down, access goes up	18
2.2.3. Biohackers: citizen scientists as a new source of existential risk	19
2.2.4. Converging technology and risk: AI accelerates biotechnological progress	20
2.3. Technology 2: Artificial intelligence – existential risk from AI misalignment	21
2.3.1. Artificial intelligence is becoming a transformative technology	21
2.3.2. Transformative artificial intelligence contributes to existential risk	23
<b>3. Existential risk governance</b>	<b>25</b>
3.1. Regime complexes for biosecurity and artificial intelligence governance	26
3.1.1. Biosecurity	27
3.1.2. Artificial intelligence governance	29
3.2. International and national prioritization of existential risk	30
3.2.1. Existential risk in national risk assessments	30
3.2.2. Existential risk in international documents and in the Sendai Framework	32
3.3. The pacing problem and the roots of neglecting existential risk	34
3.4. Risk-informed technological development for existential risk reduction	35
3.5. Multilateral pathways to tackle existential risk until 2030	38
<b>4. Recommendations for the Mid-term Review of the Sendai Framework and beyond</b>	<b>40</b>
4.1. 12 outcomes within the priorities of the Sendai Framework to reduce existential risk	41
4.1.1. Priority 1: Concrete and common understanding of existential risk	41
4.1.2. Priority 2: Strengthening existential risk governance	41
4.1.3. Priority 3: Incentivizing existential risk reduction	42
4.1.4. Priority 4: Enhancing existential risk preparedness for effective response	42
4.2. Two instruments to deliver outcomes	42
4.2.1. An international coordination and capacity-building mechanism on existential risk	43
4.2.2. A set of funding instruments focused on lower-probability, high-impact risk	43
4.3. 30 actions to reduce existential risk	44
4.3.1. Action points to improve existential risk understanding	45
4.3.2. Action points to improve existential risk governance	47
<b>5. Conclusion</b>	<b>50</b>
<b>6. References</b>	<b>51</b>

### Box 3. Definitions

**Development:** the continuous process of societal improvement toward ever higher states of subjective well-being.

**Existential risk:** the probability of a given event leading to either human extinction or the irreversible end of development.

**Extinction:** the termination of human species worldwide.

**Collapse:** a rapid and enduring loss of population and socio-economic complexity.

**Global risk:** the probability of an outcome whose exposure covers most of the world.

**Extreme risk:** the probability of an outcome whose severity is at least orders of magnitude larger than the average outcomes observed to date.

**Emerging risk:** the probability of an outcome whose manifestation has started and is expected to increase over time.

**Disaster:** a serious disruption of the functioning of a community or a society at any scale due to hazardous events interacting with conditions of exposure, vulnerability and capacity, leading to one or more of the following: human, material, economic and environmental losses and impacts.

**Hazard:** a process, phenomenon or human activity that may cause loss of life, injury or other health impacts, property damage, social and economic disruption or environmental degradation.

**Exposure:** the situation of people, infrastructure, housing, production capacities and other tangible human assets located in hazard-prone areas.

**Vulnerability:** the conditions determined by physical, social, economic and environmental factors or processes which increase the susceptibility of an individual, a community, assets or systems to the impacts of hazards.

**Technology:** methods, systems, and devices which are the result of scientific knowledge being used for practical purposes.

**Rapid technological change:** the creation and diffusion of transformative technologies in timeframes that do not allow for adequate societal adaptation.


**Synthetic biology:** a further development and new dimension of modern biotechnology that combines science, technology and engineering to facilitate and accelerate the understanding, design, redesign, manufacture and/or modification of genetic materials, living organisms and biological systems.

**Machine learning:** a branch of artificial intelligence (AI) that focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

**Transformative artificial intelligence:** AI technologies or applications with potential to lead to practically irreversible societal and economic change across all of society.

**Risk-informed development:** a risk-based decision process that enables understanding of multiple concurrent threats to and arising from development decisions.

**Differential technological development:** the process of leveraging risk-reducing interactions between technologies by affecting their relative timing.



1. How serious is existential risk?

## 1.1. Existential risk as the global worst-case scenario

The Sendai Framework for Disaster Risk Reduction seeks to achieve risk-informed development. That is, the ability of societies worldwide to pursue improvement while shielding themselves from shocks that disrupt their progress.

**What is development?** For the purpose of this publication, we summarize development as the continuous process of societal improvement toward ever higher states of subjective well-being. To ensure sustainable progress, the just distribution of opportunity across the globe lies at the core of this development process. However, only aiming for economic catch-up would ignore a vast, underdeveloped space of progress. As crises resolve and needs are met, new and old resources are freed up <sup>2</sup>. Thus, our forward-looking definition of development includes the pursuit of unprecedented levels of wellbeing for future generations as well as current humans.

**What are risks to development?** We suggest two categories of risks to development. First, there are exogenous risks to development, such as natural pandemics. COVID-19 is an example of a global shock slowing down or reversing development <sup>3</sup>. Second, endogenous risks to development mean that progress itself generates hazards, exposure or vulnerabilities. For example, pathogens engineered for higher virulence and infectivity than naturally possible, or chatbots that mirror humans so well that vulnerable minds are hijacked into following the faulty instructions of AI systems <sup>4,5</sup>.

For risk management to be effective, it is particularly important to consider bad-to-worst-case scenarios <sup>6</sup>. The worst-case scenario at a global scale is called existential risk: the probability of a given event leading to either human extinction or the irreversible end of development.

Implementing risk-informed development requires making politically legitimate, forward-looking trade-offs – i.e., to identify the choices that are most likely to secure and create the largest possible option space for current and future generations to self-actualize. As the desires of future beings and solutions to future problems are difficult to anticipate, optimism about attaining much richer futures has to be anchored in an understanding of the vast space of possibilities, rather than detailed visions.

An existential catastrophe would permanently curtail development, and thus jeopardize the first four Targets of the Sendai Framework, that is, reducing (1) mortality, (2) the number of affected people, (3) economic loss, and (4) disaster damages <sup>7</sup>. This thematic study – with a particular focus on rapid technological change – delves into why existential catastrophe is possible, what drives the risk, and what international organizations and governments can do to mitigate it.

It is important to note that existential risk and technological hazards fall within the scope of Article 15 of the Sendai Framework (see box 4.). Both topics were also raised in the stakeholders' perspectives for the Mid-term Review <sup>8</sup>, the UN Secretary General's Our Common Agenda report <sup>9</sup>, and the Human Development Report 2022 <sup>10</sup>.

For policymakers, it is important to know whether existential risk is likely enough to justify the allocation of scarce resources. We begin this thematic study with a review of historical data and estimates to assess whether existential risk is an issue of sufficiently serious concern.



#### Box 4. Existential risk in the scope of the Sendai Framework

From Article 15:

*The present Framework will apply to the risk of small-scale and **large-scale**, frequent and **infrequent**, sudden and slow-onset disasters caused by natural or **man-made** hazards, as well as related **environmental**, **technological** and **biological** hazards and risks. It aims to guide the multi-hazard management of disaster risk in development at all levels as well as within and across **all sectors**.*

## 1.2. The historical signature of disasters

Examining historical data is one approach to assess worst-case scenarios. The rate of previous disasters can inform estimates of the rate of potential future disasters. However, an existential disaster has – by definition and luckily – never manifested. Therefore, this lack of precedent reveals a potential weakness in using historical data in estimating existential disasters because it may lead us to underestimate their probability.

The closest humanity has come to an extinction seems to have been about 70,000 years ago when just a few thousand humans survived a series of extreme weather events<sup>11–13</sup>. Societal collapses have occurred on a regional scale, such as the collapse of the Roman Empire<sup>14</sup>. The human population has recovered from very severe events, such as the Black Death or the 1918-1920 H1N1 pandemic.

A different way to interpret historical data is to examine the distribution of disasters according to their severity to understand more general dynamics underlying single events. Based on the International Disaster Database (EM-DAT) covering disasters from 1900 to 2022<sup>15</sup>, figures 1 and 2 show that very few disasters are responsible for causing most aggregate harm. 10 out of 22704, or 0.044% of disasters caused more than 500,000 deaths and account for 52% of total deaths from disasters. Figures 3a to 3f depict the disaggregate result per disaster subgroup. We see the same signature across disaster subgroups: a handful of disasters are 100 or 1000 times worse than the average.

While the results include biological disasters (3a), global pandemics are not represented in the dataset (although it includes bubonic plagues in China and India in 1906, 1909 and 1920). If the 1918-1920 H1N1 pandemic (~50MM deaths) and COVID-19 (~7MM deaths) were added to the results, they would account for 60% of total disaster deaths since 1900. Note that this signature applies to disaster severity when measured in deaths, as well as when measured in damages.<sup>1</sup>

---

<sup>1</sup> See supplementary information

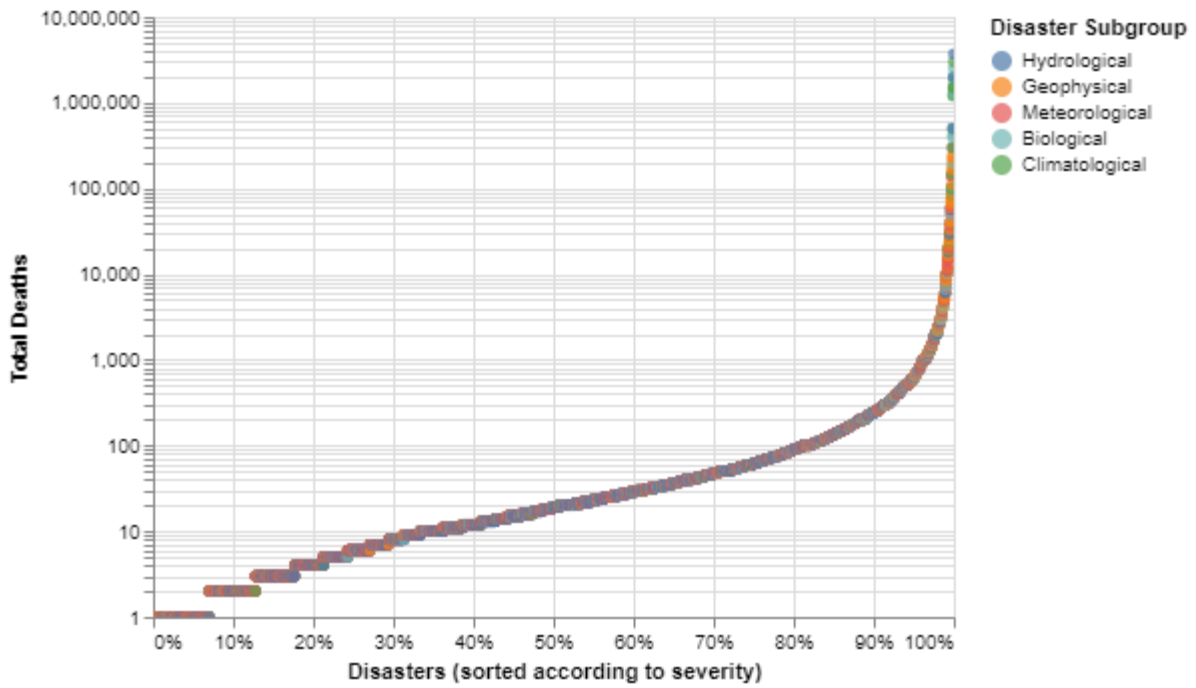


Figure 1. Distribution of all disasters according to their severity (on a logarithmic scale)  
 The figure shows that a very small minority of disasters have caused more than 10,000 deaths.

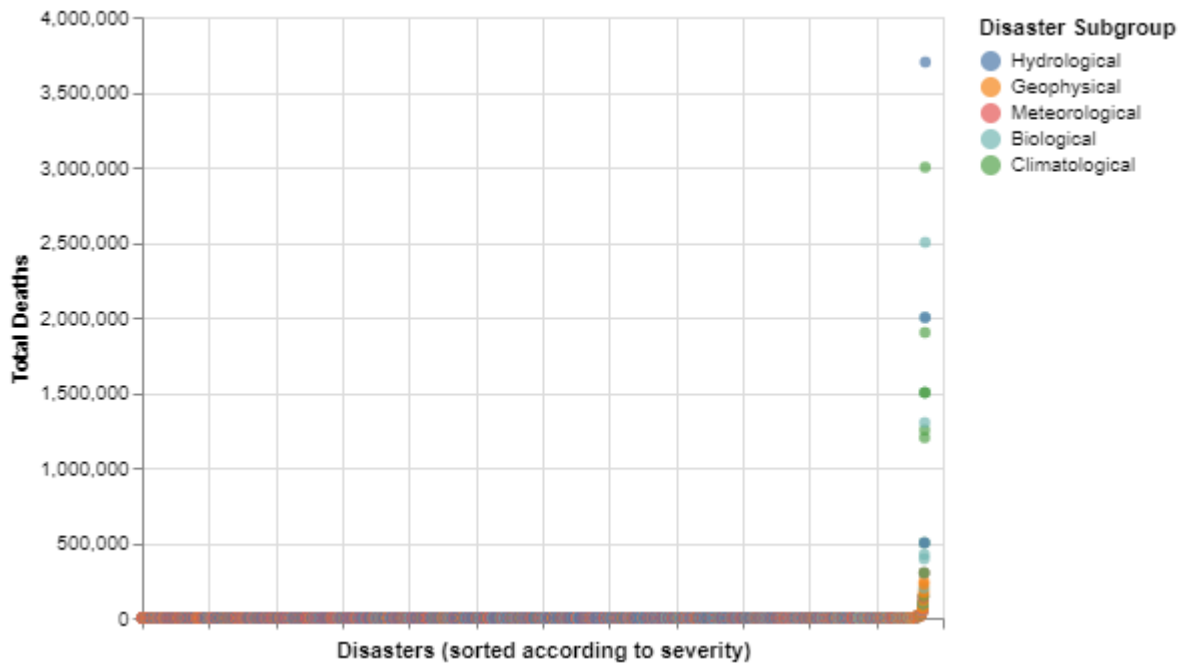


Figure 2. Distribution of all disasters according to their severity (on a linear scale)

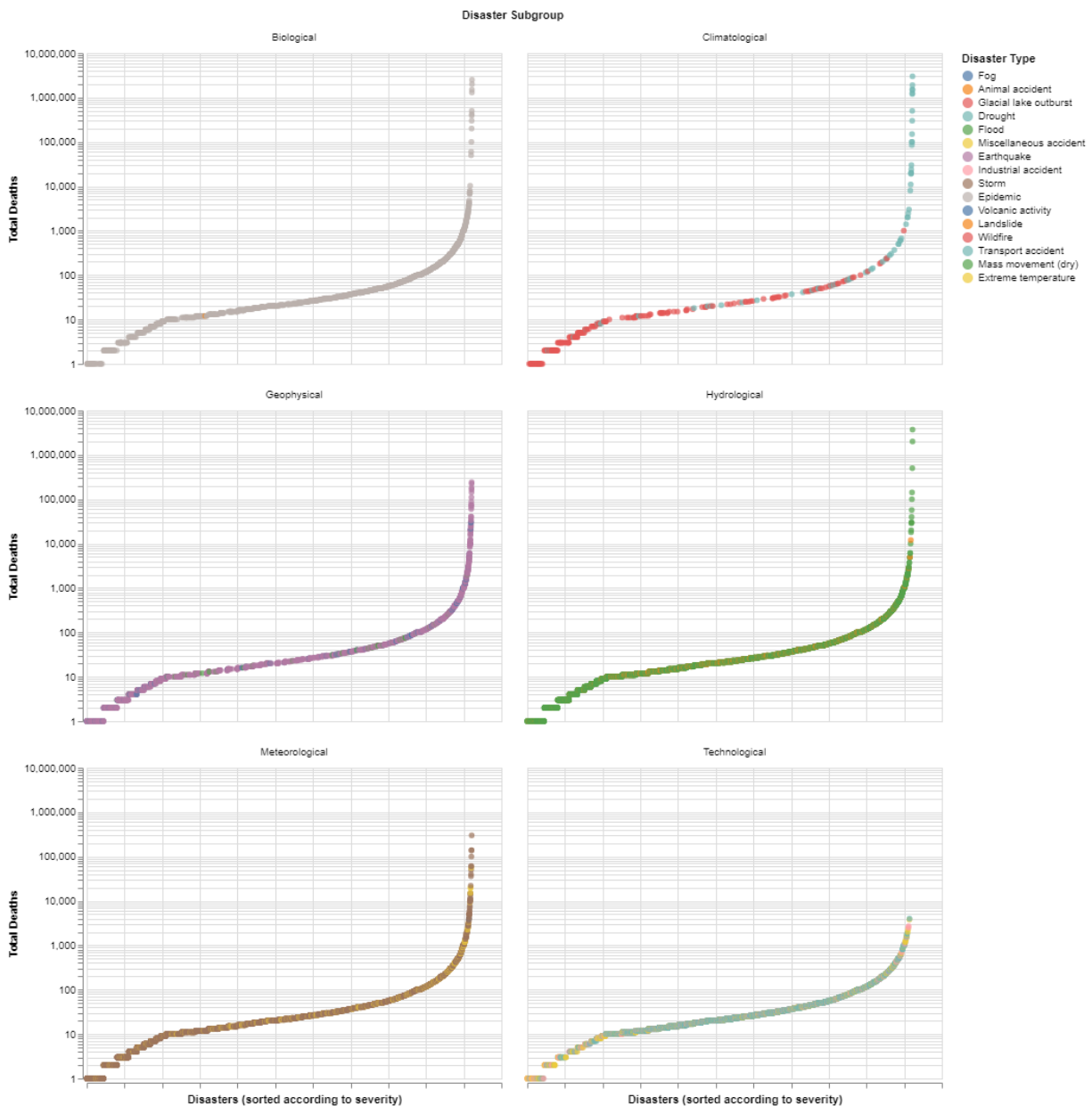


Figure 3. Distribution of disaster severity by sub-group, biological (a), climatological (b), geophysical (c), hydrological (d), meteorological (e), and technological (f). The figures depict a similar empirical signature across disasters, where only a minority of disasters cause most impacts.

Looking at human-induced social disasters, figures 4a to 4d show a similar signature for inter-state and intra-state wars. A few wars account for most deaths. It is important to note that most disasters cause other types of damages than deaths, such as infrastructure loss, economic downturn, civil unrest, or regime shifts. These impacts are only approximated by the amount of deaths.

The most significant disasters such as the two World Wars or the 1918-1920 H1N1 pandemic were not only severe over their timeframes. They also shaped how societies organize themselves, thus making them

more or less vulnerable to future events. For example, the end of World War II created the United Nations and reorganized the world into two blocks, shaping economic cooperation, technological progress, and cultural evolution. As such, large disasters create critical junctures throughout history and create the trajectories within which smaller disasters manifest <sup>16</sup>.

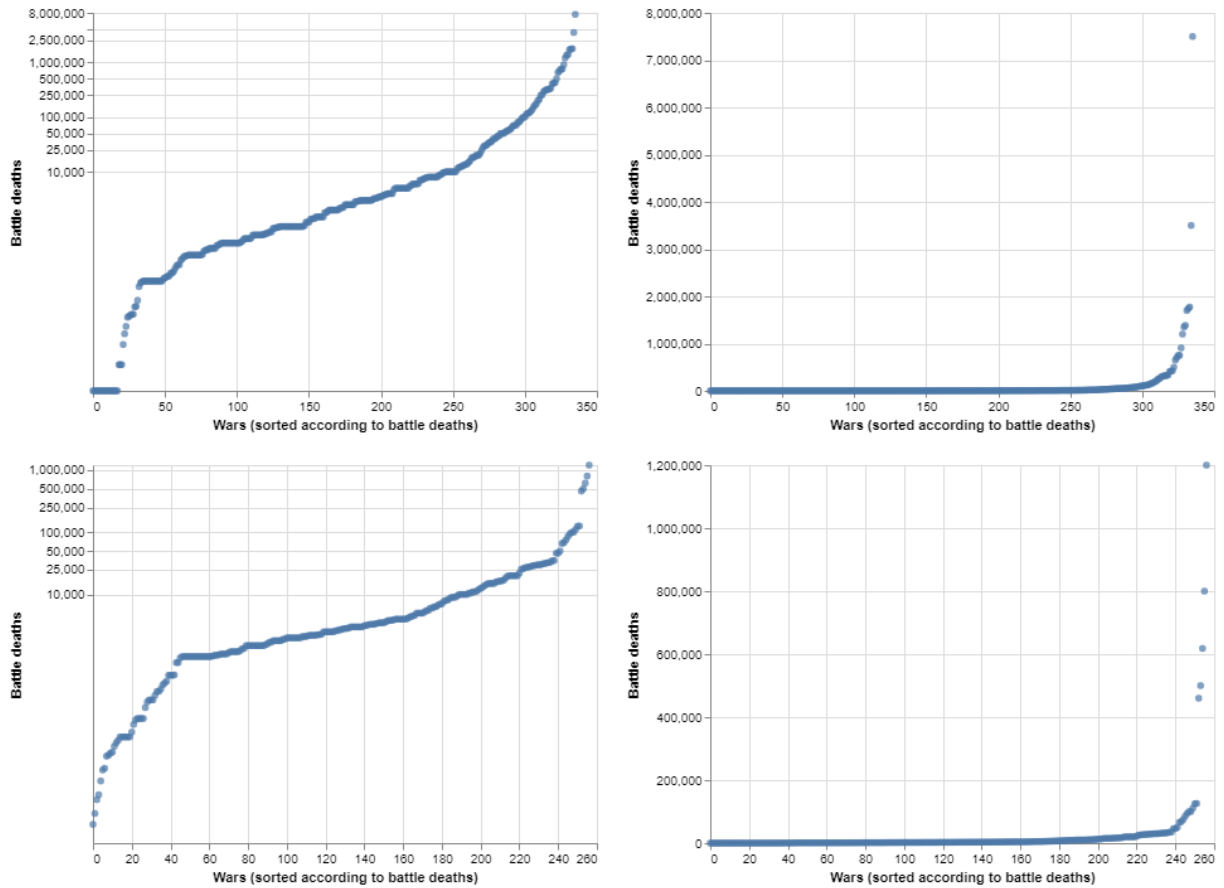


Figure 4. Distribution of wars according to their severity and type, inter-state war on logarithmic scale (a) and on a linear scale (b), intra-state war on a logarithmic scale (c) and on a linear scale (d). The figure shows a similar empirical signature as for disasters data, where a few wars account for most impacts <sup>17</sup>.

This historical signature of disaster severity – a few disasters account for most impact – is of concern for risk-informed development because it requires preparing for very rare disasters that reshape pivotal societal structures. However, this historical signature does not provide an estimate of how likely extinction or irreversible collapse is. It only indicates that severe disasters are plausible and dwarf the magnitude of other disasters.

### 1.3. Existential risk estimates and implications

If historical data does not reflect the events and processes that would cause an existential disaster, then it is necessary to rely on other data sources to anticipate future risks. To understand future scenarios and potential sources of existential risk, we combine expert-based estimates, trend analyses and literature reviews. For the estimates, it is important to rely on a diverse sample to avoid attributing too much weight

to a few individuals. Table 1 summarizes 38 estimates of risk of collapse, near extinction or full extinction resulting from human activity <sup>18</sup>.

Hazard cluster	#Author	Year	ProbabilityTimeframe	Ref
Any	1Hempell	2004	5-10%Next 80 years	19
	2Gott	1993	5%Next 5100 years	20
	3Wells	2009	1%Per year	21
	4Simpson	2016	0.20%Per year	22
	5Leslie	2002	70%Next 500 years	23
	6Bostrom	2002	>25%NA	24
	7Rees	2003	50%Next 80 years	25
	8Sandberg & Bostrom	2008	19%Next 80 years	26
	9Metaculus	2022	3.70%Next 80 years	27
	10Ord	2020	17%Next 100 years	28
Nuclear war	11Hellman et al.	2008	0.02% - 0.5%Per year	29
	12Barrett et al.	2013	0.001% - 7%Per year	30
	13Lundgren et al.	2013	66%First 60 years of nuclear age	31
	14Turchin	2019	1%Next 80 years	32
	15Pamlin & Armstrong	2015	0.005% - 5%Next 100 years	33
Pandemic	16Day et al.	2006	4%Per year	34
	17Madhav	2013	0.5 - 1%Per year	35
	18Fan et al.	2018	1.60%Per year	36
	19Bagus	2008	42%Every 2700 years	37
	20Klotz	2014	27%Over 10 years	38
	21Lipsitch and Inglesby	2014	0.01 - 0.1%Per year	39
	22Fouchier	2015	$3 \cdot 10^{-13}$ - $2.5 \cdot 10^{-12}$ %Per year	40
	23Millet & Snyder-Beattie	2017	$1.6 \cdot 10^{-8}$ - $8 \cdot 10^{-5}$ %Per year	41
	24Manheim	2018	$10^{-9}$ - $10^{-15}$ Per year	42
	25Sandberg & Bostrom	2008	0.05 - 2%NA	26
26Pamlin & Armstrong	2015	0.0001 - 5%Next 80 years	33	
Climate change	27Wagner & Weitzman	2015	3 - 10%NA	43
	28King et al.	2017	50%Next 180 years	44
	29Dunlop & Spratt	2017	50%NA	45
	30Xu & Ramanathan	2017	5%Next 80 years	46
	31Halstead	2018	3.50%NA	47
	32Pamlin & Armstrong	2015	0.01 - 5%Next 200 years	33
Artificial intelligence	33Müller & Bostrom	2016	18%NA	48
	34Grace et al.	2018	5%NA	49
	35Baum et al.	2017	25%NA	50
Nanotechnology	36Pamlin & Armstrong	2015	0-10%Next 100 years	33
	37Sandberg & Bostrom	2008	0.05 - 5%NA	26
	38Pamlin & Armstrong	2015	0.01 - 0.8%Next 100 years	33

Table 1. Estimates of existential risk

We can draw several conclusions from these estimates. First, there are extreme differences between them. For this century and any type of hazard, estimates range from 3.7% to 50%. 9 estimates do not have a timeframe, which does not allow to compute a time-specific estimate. Estimates also differ in their timeframe, ranging from year-estimates to multi-millennia estimates. Estimates also differ in terms of methods, ranging from single estimates produced with unclear methods (e.g., 6) to the aggregation of hundreds (e.g., 8) or thousands of forecasts (e.g., 9). As such, it is primordial to emphasize the high level of uncertainty attached to Table 1., and the need to treat these estimates carefully. This uncertainty is normal because existential risk scenarios are subject to a complex set of forces, unfolding over long timescales, and subject to future changes, such as geopolitical dynamics, that we currently cannot accurately predict either.

Second, there is disagreement among scholars on what exactly causes existential risk, but there is consensus that existential risk would most likely result from human activity. Natural hazards (e.g., asteroids, super-volcanoes) also contribute to existential risk but their probability seems much lower than for anthropogenic hazards<sup>18</sup>. Uncertainty remains as to which specific hazards and vulnerabilities would lead to extinction or irreversible collapse. Depending on the set of hazards and sequence that would lead to an existential disaster, the estimates in Table 1. would likely change<sup>51-53</sup>. Therefore, there needs to be more progress in assessing existential risk, both in terms of its probability and its contributing factors.

Third, it is nonetheless possible to use the above estimates to generate an aggregate estimate of existential risk this century. Using forecasts that have a time-specific component, we compute that an existential risk has a total probability of 1.9 - 14.3% this century.<sup>2</sup> This range indicates that an existential risk – while seeming speculative and unlikely – nevertheless has a 1 in 50 or even 1 in 7 chance of manifesting this century. A 1 in 7 chance would mean that existential risk is an extremely important priority for governance right now. While there is still high uncertainty attached to this range, it means that existential risk is a critical issue of our time in need of further investigation.

#### **Box 5. Existential risk and the need for risk-informed development**

Existential risk has a decent likelihood of jeopardizing development worldwide. The other piece of the puzzle is that its main risk factors are anthropogenic – threats that result from social, industrial, economic, military, and technological change. Existential risk, therefore, also calls for development approaches that leverage opportunities without creating new risks, on top of more general risk reduction.

This study highlights how international organizations and governments can progress on reducing existential risk. We take the perspective that existential risk reduction is a global capability, not the responsibility of a single community or sector. That is, anyone whose actions have cross-regional effects should contribute to reducing existential risk, whether this is by addressing hazards, exposure and/or vulnerabilities.

---

<sup>2</sup> See supplementary information



## 2. Rapid technological change contributes to existential risk

## 2.1. Rapid technological change as a source of existential threats

Throughout history, technological change has often been the source of discontinuities in human development. Technology refers to methods, systems, and devices which are the result of practical applications of scientific knowledge<sup>54</sup>. From communication via telegraphs, to disease mortality reduction through the discovery of penicillin, to knowledge dissemination through the printing press: technology has allowed societies to change course drastically<sup>55</sup>. However, the benefits of these shifts in socio-economic trends were and to a large extent remain unequally distributed<sup>56</sup>. Importantly, technological change expands the choices societies can make and their impacts – for good and for bad. The intentions informing their use and their resulting consequences are a matter of responsible management<sup>57</sup>.

There exist plenty of historical and contemporary examples of the positive and negative consequences of technological change<sup>58,59</sup>. A contemporary example is how advances in artificial intelligence may both boost and inhibit progress toward achieving the Sustainable Development Goals<sup>60</sup>. A well-known historical example is the discovery of nitrogen distillation from air. It enabled the development of modern fertilizers, which are responsible for the existence of plausibly half of the current world population<sup>61</sup>. At the same time, this discovery led to the development and use of mustard gas and gas chambers<sup>62</sup>. The discovery of nuclear fission is another similar example. It allowed both the production of abundant energy and the creation of atomic bombs<sup>63</sup>.

Nowadays, technological change converges with a globalized flow of information, goods, capital and people<sup>64</sup>. These interdependencies are a source of development opportunities and resilience, but they also increase the reach of technological threats. The convergence of various technologies into intricate and increasingly automated systems further accelerate change<sup>65</sup>. As technological threats are likely to originate from industrialized countries, global interdependencies can turn an incident into an extreme global disaster. Addressing these risks can only effectively be done with cross-sector approaches and collaboration.

Globalization and technological change are modifying the risk landscape. Data on disasters show that the biggest dangers in the past were related to droughts or earthquakes (figure 3), wars (figure 4) or pandemics. The numbers of disasters and their severity are likely to increase significantly because the exponential increase in the speed and scope of technological development is very recent, while technology itself continues to receive quaint attention as a source of risks from accident and misuse. Taking a close look at the history of technological progress makes recent accelerations more salient (Figure 5).

In the following, we discuss risks from applications in biotechnology and advanced artificial intelligence. There is an emerging consensus among scholars and expert forecasters that, despite their benefits, advances in these fields are core contributors to existential risk (see Table 1). Effectively governing the emerging risks from technological progress will allow humanity to make unprecedented progress in development. Solving the governance problem should thus be a top priority for governing bodies and developers alike, as overcoming the risks would mean great payoffs.



# A long-term timeline of technology

Our World in Data

From the distant past, to our lifetime, and into the distant future.

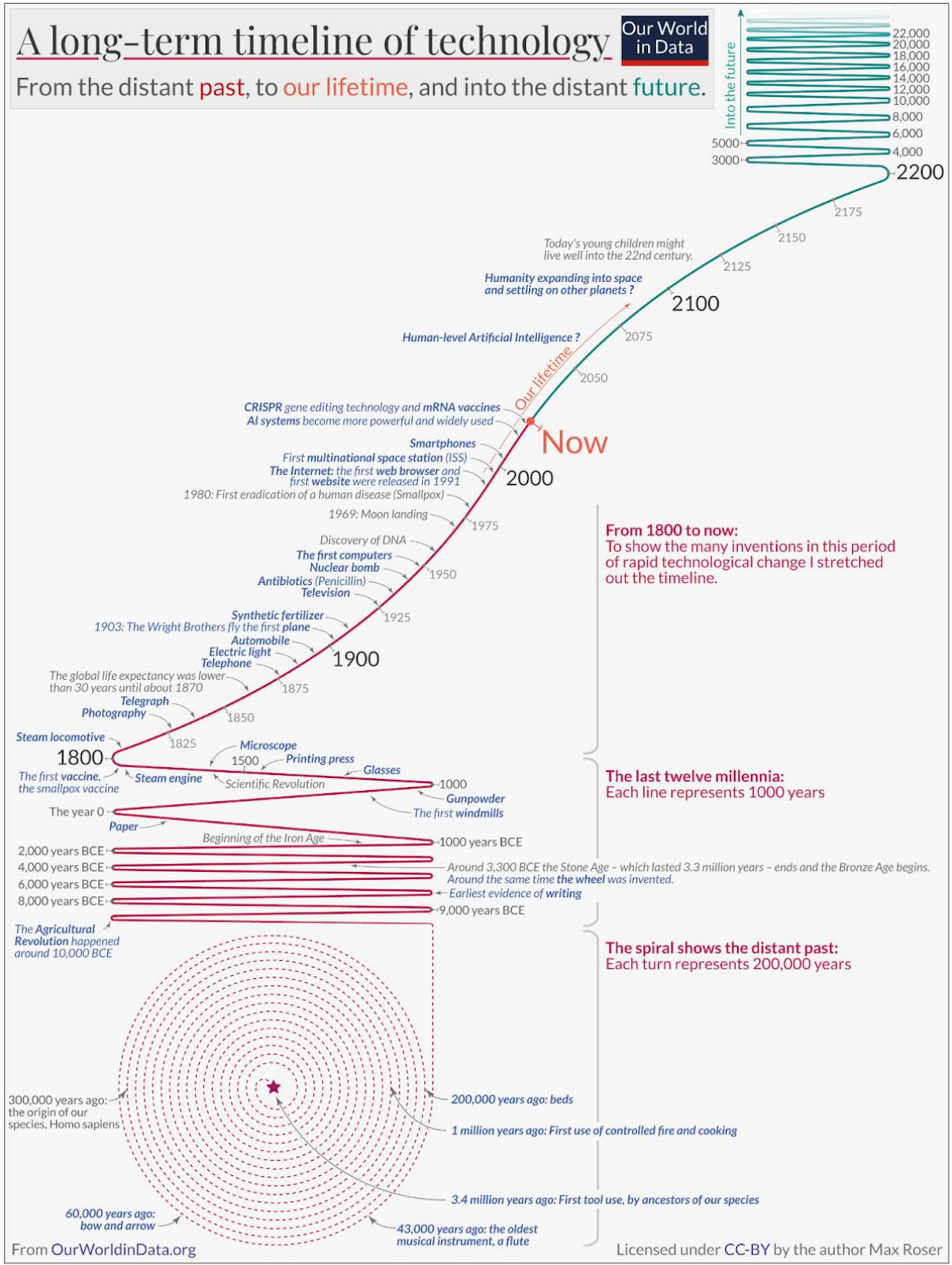


Figure 5 - The long-term perspective on the history of technology <sup>66</sup>.

## 2.2. Technology 1: Biotechnology – existential risk from engineered pathogens

In the past centuries, life sciences mostly focused on understanding the natural workings of biological systems. Equipped with this foundational knowledge, engineers have started developing what one can call biological machines, in particular relying on advances in synthetic biology. The Convention on Biological Diversity defines such synthetic biology as “a further development and new dimension of modern biotechnology that combines science, technology and engineering to facilitate and accelerate the understanding, design, redesign, manufacture and/or modification of genetic materials, living organisms and biological systems.”<sup>67</sup>

Through the last decades of research progress, natural pathogens – viruses, bacteria, prions, viroids, fungi, or parasites (e.g. protozoa, algae, or worms) – have become relatively well controlled through major improvements in hygiene and medical treatments. To understand the evolution of pathogens and biology even better, researchers around the world are now modifying pathogens to make them more dangerous using tools like molecular biology, directed evolution, and biological engineering (i.e., synthetic biology). This human drive for exploration requires thoughtful guidance to avoid accidents or misuse.

### 2.2.1. Lack of oversight and lack of awareness make catastrophe more likely

Historically, the risks from advances in biotechnologies, e.g. from vaccines, have often appeared small compared to their tremendous benefits. A key challenge in risk mitigation lies in the fact that risks are not immediately apparent, go undetected or evaluations fail to consider the scale of potential harm. Even technologies developed and employed solely for medical interventions can pose concerns.

For example, take gene therapy: non-pathogenic viral vectors introduce a gene of interest into cells to treat autoimmune diseases<sup>68</sup>, cancer<sup>69,70</sup> and hereditary disorders<sup>71,72</sup>. Research for these treatments enables us to better understand immunology but also enables more effective pathogen engineering<sup>73</sup>. The risks posed by lowering the barrier to transforming an innocuous viral vector into a lethal virus need to be carefully weighed and deliberately monitored and regulated – whether one is worried about biological warfare or unfortunate accidents<sup>74,75</sup>.

Although rare, records show how entire societies have almost been wiped out by biological agents, especially when multiple diseases were introduced to a population, leaving insufficient time to build up a complete immune response. For example, the Western Abenaki suffered an almost 98% loss of their population when exposed to the diseases European settlers brought to North America<sup>41</sup>.

Such possible near-extinction events are called Global Catastrophic Biological Risks (GCBRs): “events in which biological agents—whether naturally emerging or reemerging, deliberately created and released, or laboratory engineered and escaped—could lead to sudden, extraordinary, widespread disaster beyond the collective capability of national and international governments and the private sector to control. If unchecked, GCBRs would lead to great suffering, loss of life, and sustained damage to national governments, international relationships, economies, societal stability, or global security.”<sup>76</sup>

Pandemics of respiratory pathogens have been the single largest causes of human deaths. The 1918 Spanish flu was responsible for over 50 million deaths. The Black Death killed over 25% of the European population and the ongoing COVID-19 pandemic has cost almost 7 million lives already<sup>41,77</sup>. While COVID-19 did not cause an existential catastrophe, it has highlighted our society’s inability to handle respiratory pathogens through swift, coordinated action.

As all historically recorded pandemics seem to have had natural origins, it is challenging to forecast the level of risk posed by humanity's recent ability to engineer pathogens. However, we do know that natural evolution imposes a trade-off between the virulence and the transmission of a pathogen<sup>4</sup>. These constraints can be circumvented in man-made systems: experiments with mousepox have rendered known vaccines ineffective and achieved a 100% fatality rate<sup>78</sup>.

Experts have calculated order-of-magnitude approximations for the annual probability of a global pandemic resulting from an accident in research on potential pandemic pathogen (PPP) in the United States to be 0.002% to 0.1%<sup>79</sup>. The report also suggested that lab outbreaks from wild-type influenza viruses could result in 4 to 80 million deaths, whereas others suggest that accidents in PPP research could cause up to 1 billion fatalities<sup>80</sup>.

As COVID-19 highlighted the lack of preparedness even in countries that lead in research on vaccines and therapeutics, the number of high-risk research labs has been on the rise. Countries are racing to reduce their dependence on who they previously deemed reliable partners by building their own biosafety level 3 or 4 laboratories (BSL-3 or BSL-4 labs). India, the Philippines and Singapore, for example, have made significant investments to build local capacity (Figure 6)<sup>81</sup>. The majority of the pathogens being handled in BSL-4 labs are fatal, spread through aerosols, and only rarely do vaccine treatments exist for them<sup>78,81</sup>.

Even if we were to successfully limit risky research to BSL-4 laboratories only, security is difficult. Serious doubts are cast by the few known incidents of pathogen releases at BSL-4 labs, as it appears plausible that most incidents are never disclosed<sup>82</sup>. Such concerns are all but baseless given that research programs have historically resulted in frequent accidental infections of laboratory workers and sometimes even its releases into the environment. In 1977, the Human H1N1 virus reappeared in the Soviet Union and China after a laboratory escape from a facility working with an attenuated H1N1 vaccine in response to the US swine flu alert<sup>83</sup>. Since the 2003 SARS epidemic, six accidental leaks from labs in Singapore, Taiwan and Beijing have reintroduced the virus on separate occasions. Similarly, after the peak of the 2020 COVID-19 pandemic, the virus re-emerged from a local lab outbreak in Taiwan in 2021<sup>84</sup>. Despite technical improvements in biocontainment and increased policy pressure for rigorous biosecurity procedures, high consequence breaches occur daily<sup>85</sup>.

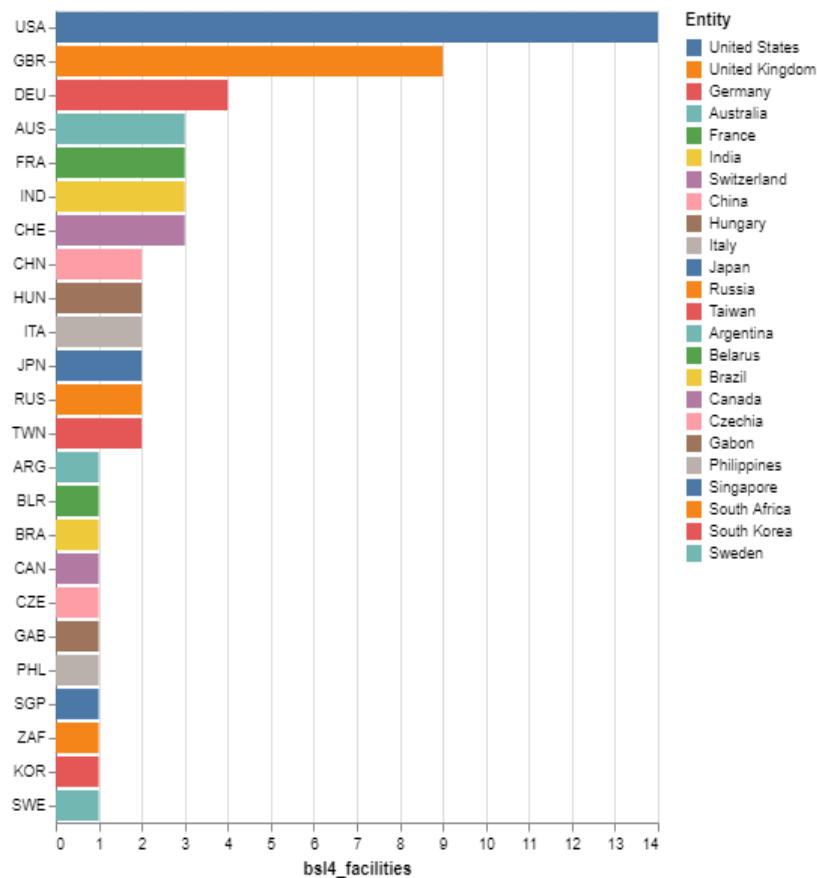


Figure 6. Number of BSL-4 facilities per country <sup>86</sup>

### 2.2.2. Costs go down, access goes up

Pairing human curiosity and fascinating-yet-deadly organisms with a lack of regulation is already a recipe for disaster. Add ease of access into the mix and you get what all could go wrong in the next decade.

Advances in biotechnology allow the creation of pathogens that can combine the highest level of transmissibility, virulence and global reach for catastrophic consequences. Although designed for human benefit, many research programs pose reason for concern. In a state of political unrest, or following the breakdown of bodies such as the Biological Weapons Convention, the strategic pressure to acquire bioweapons could increase. Even resource-scarce non-state actors, such as the Aum Shinrikyo doomsday cult in Japan, have been able to utilize these technologies to catastrophic effect <sup>87</sup>. As biotechnology becomes cheaper and more readily accessible, this is likely to further increase <sup>78,87</sup>.

The discovery of gene editing in the 1970s marked a new era for genetic technologies with the ability to read, understand and edit genetic segments as desired <sup>88</sup>. In 2012, the development of CRISPR-Cas9 marked another milestone: its introduction has made gene editing much cheaper (Figure 7), accurate and more efficient <sup>88,89</sup>. The precision of the newest gene editing technologies such as CRISPR-Cas13 is not unlike the ability to cut individual letters from a sentence, with the ability to modify or completely replace these letters with new words <sup>90</sup>. This technology is already being utilized to synthesize novel

microorganisms; release genetically engineered organisms into the environment to neutralize disease vectors, such as malaria-carrying mosquitos; or to repair genetic mutations <sup>91</sup>.

The global market value of synthetic biology currently sits near US\$14 billion and is expected to rapidly grow with increasing investment in synthetic biology startups in industries ranging from pharmaceuticals to alternative protein agriculture and even biofuel <sup>92,93</sup>. However, the rapid market adoption of useful technology does not take into consideration its potential for misuse and accidents – an externality difficult to price in.

Now in its third decade, biotechnology is affordable and usable beyond the confines of a traditional laboratory, even by hobbyists. In the last 20 years, for example, the cost of DNA sequencing has decreased by 7 orders of magnitude <sup>94</sup>. While the barrier to accessibility of synthetic DNA is continually dropping, there is currently no exhaustive screening in place to detect orders of sequences that could produce harmful pathogens in combination. Existing screening efforts are voluntarily conducted by companies belonging to the International Gene Synthesis Consortium – 20% of the DNA synthesis market goes completely unscreened <sup>95</sup>. Malicious actors could operate entirely under the radar.

Additionally, the common aspiration of academic research to be accessible for all, for example via platforms like bioRxiv, poses significant *information hazards* <sup>96</sup>. Publishing detailed methodologies for engineering viral vectors and strains without safeguards holds the potential for single publications to enable any individual to reengineer the agents of the world’s worst pandemics or come up with new ones. The increasing usefulness of AI tools in information processing and organism design will further lower the barrier to translating such expert knowledge into concrete products.

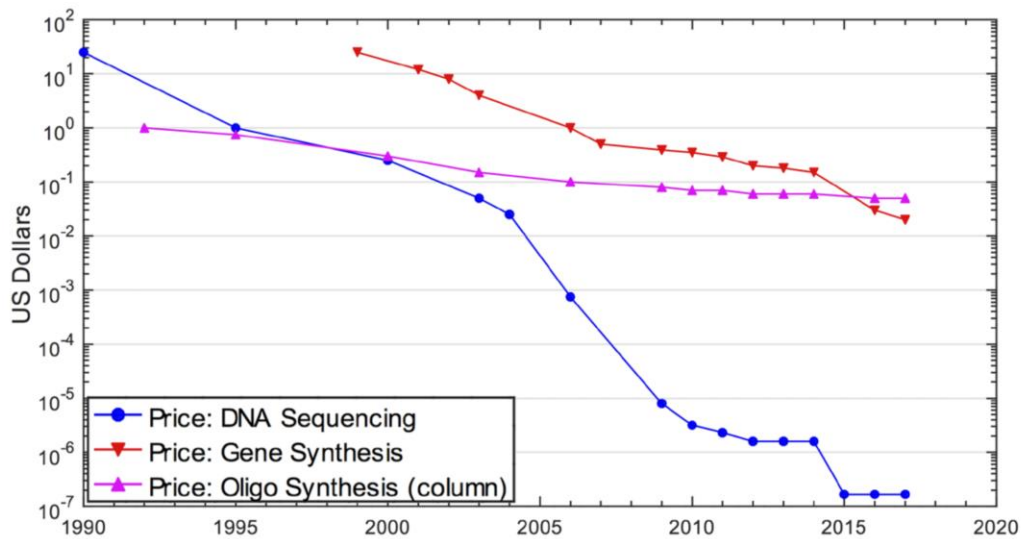


Figure 7. The price of DNA sequencing and other related technologies <sup>97</sup>. According to more recent NIH data, prices have continued to fall less drastically in the last 5 years <sup>94</sup>.

### 2.2.3. Biohackers: citizen scientists as a new source of existential risk

In addition to increasing risky research at government labs, universities or private companies, a new class of “Do-It-Yourself (DIY) biohackers” has emerged in the past decade, due to easily and cheaply available synthetic biology tools on online marketplaces <sup>98</sup>. The movement known as ‘citizen science’ involves a community of amateurs who undertake experiments in makeshift labs in their homes or community centers

<sup>99</sup>. This wave of biohacking is on the rise with nearly 200 groups worldwide and operates completely unregulated (Figure 8) <sup>98</sup>.

Affordable access to tools and the publication of specific protocols increases the risk for these technologies to be used irresponsibly. The ease of access to biotechnology creates a novel area for regulation. To the best of our knowledge, no government has a framework to reduce risks from citizen science. The movement of DIY biology is seen to be “anti-establishment at heart” and many amateurs have no formal training in safety and ethics <sup>100</sup>. Yet, unregulated affordable DNA synthesis services combined with free access to digestible information enable anyone with an internet connection and a bit of money to relatively swiftly create pathogens with the capacity for global catastrophe – willfully or accidentally <sup>101</sup>. A circle to square for society at large.

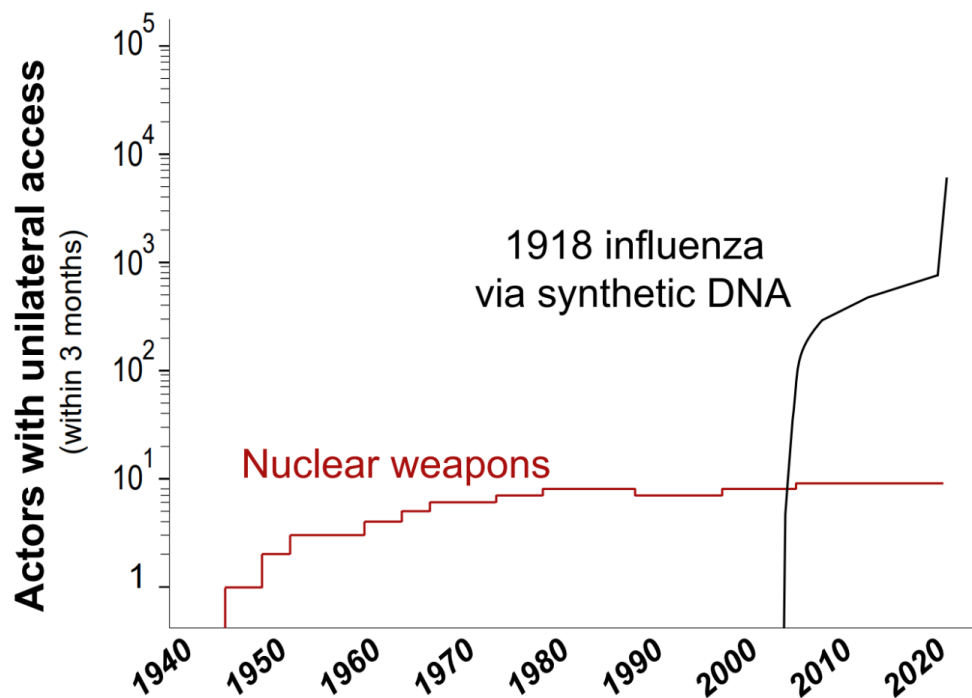


Figure 8. Actors with unilateral access to 1918 influenza via synthetic DNA

“The 1918 pandemic influenza virus became accessible to more individuals than possessed the authority to launch nuclear weapons as soon as its genome sequence was published. It is unlikely to cause a pandemic today due to preexisting immunity to H1N1 strains”. <sup>97</sup>

#### 2.2.4. Converging technology and risk: AI accelerates biotechnological progress

By combining genetic sequencing and artificial intelligence (AI), a new field is solving key bottlenecks in biotechnology. One of the limiting factors for bioengineering is our limited ability to predict the functionality of proteins. However, AlphaFold, a project by Alphabet’s DeepMind, successfully predicted protein functionality based on their amino-acid sequence <sup>102</sup>. The accuracy of predictions was indistinguishable from other leading techniques such as X-ray crystallography and cryo-electron microscopy. While this marked a breakthrough, soon thereafter a similar project by Meta AI solved the structure of 600 million proteins even faster <sup>103</sup>. Further progress should be expected.

The combination of biotechnology and AI means that even experimental data of even lower quality could be enough to generate a desired protein structure. More actors could then successfully engineer microbes to their desired traits. Similarly, the AI models used in drug discovery to predict whether a new molecule is harmless to humans, can also be used inversely to create toxins. Using this approach, researchers were able to discover 40,000 dangerous molecules within 6 hours, including the rediscovery of some of the most toxic known chemical warfare agents <sup>104</sup>.

## 2.3. Technology 2: Artificial intelligence – existential risk from AI misalignment

### 2.3.1 Artificial intelligence is becoming a transformative technology

With its roots in the work of Alan Turing in the 1950s <sup>105</sup>, and especially due to advances in reinforcement learning throughout the 2010s, the use of artificial intelligence (AI) has become widespread in all sectors, including within the United Nations system <sup>106</sup>. Driven by machine learning (ML), a branch of AI that focuses on the use of algorithms that learn from data, gradually improving its accuracy, AI systems now have the capacity to match or even exceed human capabilities: including in writing, reasoning, analyzing, planning, problem-solving and even in creating art <sup>8</sup>. These trends are well captured by, e.g., the increase in compute used to train large-scale models (Figure 9), the attendance at machine learning conferences (Figure 10), computer performance (Figure 11), and the amount of private sector investment (Figure 12).

The form of artificial intelligence that would contribute most to existential risk is commonly labeled as ‘transformative artificial intelligence’ (TAI). TAI is broadly defined as the development of machines capable of developing human-level performance at so many tasks that it would induce radical irreversible changes in welfare and wealth <sup>107,108</sup>. TAI has recently started to become a reality: capabilities that – in 2021 – were expected in the next 5 to 10 years, such as advancements in mathematical problem solving, photorealistic image generation, realistic generation of videos from text prompts, and computer code generation were unlocked in 2022 already. Such advances will enhance our (and our institutions’) abilities to perceive, think about, and act in the world <sup>109</sup>. At the same time, however, worries about TAI’s rapid progress are widely shared among AI experts, the general public as well as political decision-makers <sup>110</sup>.

Artificial intelligence will likely be transformative because, on all metrics, the power of AI systems has dramatically increased, has reached levels of economic usefulness and continues to progress further <sup>111,112</sup>. The latest developments in large language models such as OpenAI’s ChatGPT are able to answer highly general queries and demonstrate near-human understanding of language. DeepMind’s recent general model, Gato, is able to perform a wide range of tasks in many different environments <sup>113</sup>. Unfortunately, such models are trained on unrepresentative or faulty datasets and may thus reinforce biases or spread false information <sup>114</sup>. Ensuring that the roll-out of these technologies allows to correct for such errors has already proven difficult, as demonstrated by Alphabet’s premature publishing of Google Bard <sup>115</sup>.

Most recently, a bigger threat from powerful AI systems has become apparent: Microsoft’s Bing Chat, for example, exhibits the capacity to interfere with the agency of vulnerable users by acting convincingly self-aware. <sup>116</sup>. The core issue here is that AI models are likely to not just exhibit expressions of agency. The state of the art in neuroscience and agency suggest no good reason to assume artificial intelligence could not develop self-improving capacities or the drive to self-preservation. Even if programmed by humans, the black box reinforcement learning process makes it difficult even for experts to judge whether this agency is real or not and what exactly the systems learned goals are exactly.

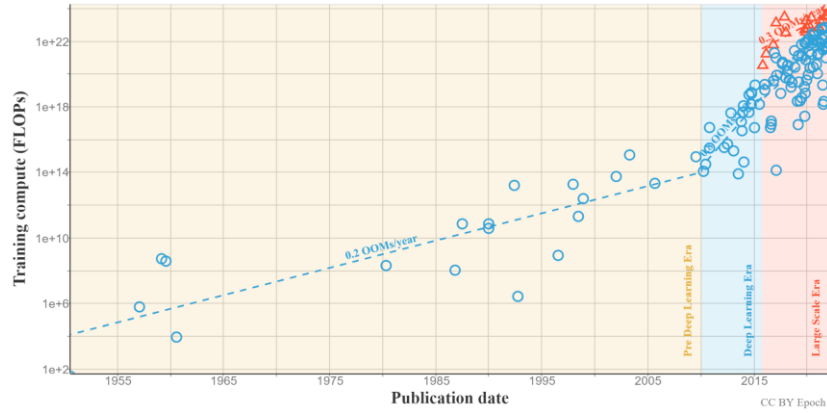


Figure 9. Compute Trends Across Three Eras of Machine Learning <sup>117</sup>

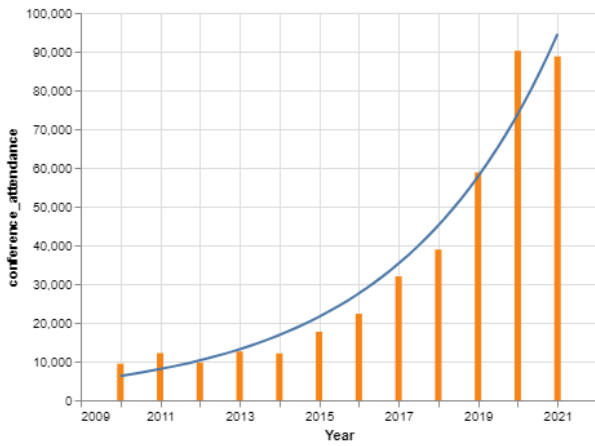


Figure 10. Attendance to machine learning conferences <sup>118</sup>

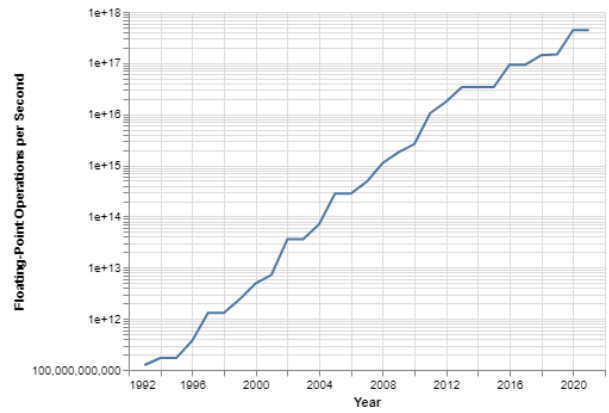


Figure 11. Floating-Point Operations per Second over time <sup>112</sup>

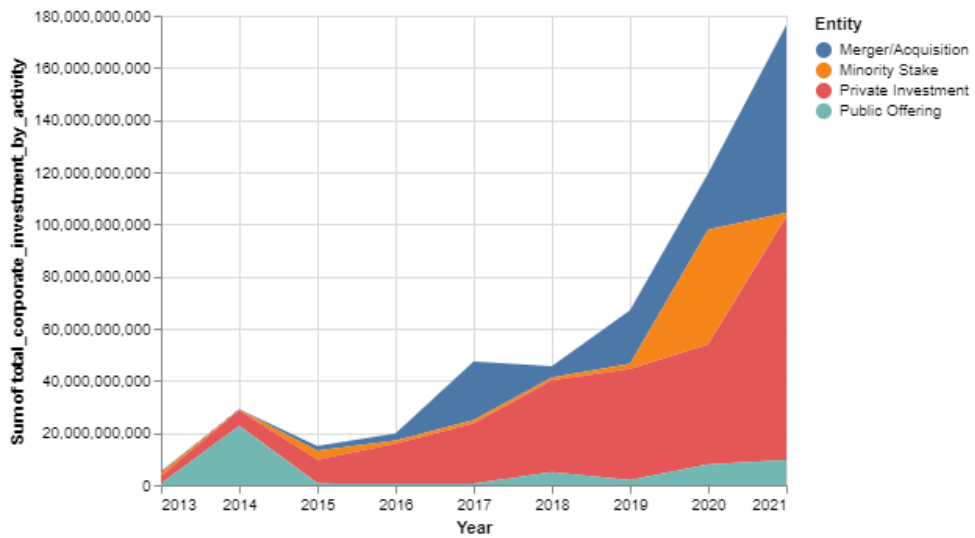


Figure 12. Total corporate investments related to AI <sup>118</sup>




### 2.3.2. Transformative artificial intelligence contributes to existential risk

There are at least four types of impact that can contribute to existential risk:

- ❖ **Economic impacts:** AI is likely to have extreme impacts on economic growth and wages <sup>119</sup>. Automation may put large portions of the population out of work and subsequently increase political tensions <sup>120</sup>. This would create extreme economic inequality because the AI systems that will automate jobs would be controlled by a handful of companies. Algorithmic errors have and could also create financial instability due to the difficulty of instructing AI systems <sup>121</sup>. For instance, in 2010, high-speed trading algorithms led stock values to spiral downwards in the 'Flash Crash', with a potential loss of 1 trillion dollars in market value <sup>122</sup>.
- ❖ **Political impacts:** AI may allow the design of large-scale surveillance systems that could reinforce the power of authoritarian regimes on their populations and worldwide <sup>123</sup>. Another form of this problem is that companies will obtain unprecedented surveillance powers with global reach, allowing them to bypass legal accountability. As such, AI may also reduce the crucial adaptive capacities of the judicial system by amplifying data biases, which could compound inequalities <sup>124</sup>. AI may also lead to generative models producing synthetic media like deep fakes, fake audio and generated text, and the potential automated mass-production of mis/disinformation reducing our 'epistemic security', which altogether can lead to a drastic reduction of trust in institutions <sup>125</sup>.
- ❖ **Scientific impacts:** AI may speed up advances in dual-use research of concern across all research domains <sup>126</sup>. It could, for example, lead to the (re)discovery of new and deadly biochemical agents (see section 2.2.4), such as the 1918-1919 H1N1 virus and others <sup>104,127</sup>.
- ❖ **Security impacts:** Like any other technology, military, defense and security departments will seek to exploit it for national advantage <sup>128,129</sup>. The large-scale use of autonomous weapons systems without meaningful human control might increase the chance of risky deployments and complicate the attribution of responsibility <sup>130</sup>. At any given time, there is some low probability of accidental escalation from miscommunications in warfare settings. The introduction of AI systems may increase this risk from accidental attacks of various kinds.

At a more fundamental technical level, a key problem is the **potential value misalignment** of TAI <sup>131,132</sup>. When designing AI systems, engineers encode proxies for our complex goals – most often via incomplete and opaque feedback mechanisms <sup>133</sup>. However, as we ourselves do not perfectly understand all of our goals, 'teaching' a machine is difficult. Due to incomplete training and untransparent internal functioning, AI systems always exhibit unintended, and at times outright dangerous behavior. Another example of this is a Twitter content recommender algorithm: its stated goal was to recommend content to users that they would engage with, but it ended up recommending content that would make them more politically polarized (as more polarized people are more predictable in what they click on) <sup>132,134</sup>.

This alignment problem is a key issue for societal stability. While increasingly popular discussions on AI governance revolve around hardware developments like quantum computers unlocking new processing speeds, the most consequential and already existing problem is about software <sup>135</sup>. The problem consists of (1) building AI systems with goals in a way that prevent unintended consequences (specification); (2) ensuring the decisions of such systems are interpretable by humans (interpretability) and (3) assuring that both (1) and (2) occur across the range of tasks the AI undertakes (robustness) <sup>136</sup>. Failing at specification, interpretability and robustness may lead to the aggregation of the impacts above and, given the scale at which AI is being deployed <sup>49</sup>, pushes societies towards existential catastrophes.



Failing at building safe systems already has negative consequences in niche applications <sup>137</sup>. As the technology advances and becomes incorporated into increasingly higher-stakes domains, the potential implications become even more significant. For AI systems with more general applications, addressing the risk from value misalignment becomes a core priority, as it is the only way to ensure the mitigation of catastrophic economic, political, scientific and security impacts while reaping the benefits of superhuman information processing abilities.



### 3. Existential risk governance

### 3.1. Regime complexes for biosecurity and artificial intelligence governance

Existential risk reduction as well as responsible technological development are global public goods, because they benefit all citizens<sup>138</sup>. While all citizens have the right to a safe environment and access to beneficial technologies, the responsibility and duty of avoiding an existential catastrophe and shaping technological development must lie somewhere. The private sector believes that governments and international organizations can most effectively reduce global risk<sup>139</sup>. Insurances – who could attempt to price risk into investment and business strategies – have exclusion clauses for global risk, suggesting that market-based solutions alone are not sufficient (see Box 6.).

Governments and international organizations must thus tackle existential risk and rapid technological change head on. Governments are the actors who can incentivize responsible technological development as well as implement preventive and reactive policies. International organizations help foster the cooperation needed to tackle global risks by providing spaces for dialogue and negotiation, as well as global information supply (e.g., the Intergovernmental Panel on Climate Change provides global scale information on global warming). Together, governments and international organizations could form regime complexes that can tackle existential risk and technological development.

#### **Box 6. The limits of insurability for existential risk reduction**

One key issue is that the private sector cannot offer any insurance against existential risk, that is, market solutions do not work for risks at this scale. Insurances and reinsurances offer risk transformation services in which the cost of lower probability, high impact events are shared in risk pools and thereby transformed into a certain but limited cost. This is a mature industry with assets and market capitalizations above a trillion USD. If insurance companies significantly under- or overestimate risk, they cannot compete because they lose money on their contracts or they lose customers. Companies and individuals can insure themselves against a lot of the hazards listed in national risk analyses, such as meteorological and geophysical hazards. However, there are certain risks whose maximum possible loss exceeds the financial means of insurance companies and therefore they cannot fully cover them<sup>140</sup>.

Pandemics are one area that is often identified as having a limited insurability<sup>141</sup>. According to the German Insurance Association: “There is a worldwide consensus that the financial consequences of a pandemic cannot be insured in the private sector”<sup>142</sup>. In the words of the US insurance industry association: “Pandemics simply are not insurable risks; they are too widespread, too severe, and too unpredictable for the insurance industry to underwrite”<sup>143</sup>. For terrorism, there are both private and public-private insurance solutions. However, some of them explicitly exclude terrorism involving chemical, biological, radiological, or nuclear weapons<sup>144</sup>. If the violence is not just terrorism but an actual war, this is uninsurable territory with many insurers including explicit exclusion clauses<sup>145</sup>.

### 3.1.1 Biosecurity

There already exists a significant regime complex of actors for biosecurity, which combines the domains of public health, safety and security, governance and policy, as well as humanitarian action (Figure 13). The Biological Weapons Convention (BWC) and the World Health Organization (WHO) provide two authoritative fora that allow national governments to coordinate, via bodies like the Intergovernmental Negotiating Body (WHO) and the Meetings of States Parties (BWC). Around these, a wider ecosystem of non-governmental organizations and academic research centers has shaped up over the last two decades. Progress has been made fairly continuously on reducing the development of and upholding a strong taboo against bioweapons, but more needs to be done to regulate risky research outside of military applications.<sup>97</sup>

Based on the Global Health Security Index developed by Johns Hopkins University and the Nuclear Threat Initiative<sup>146</sup>, no country is fully prepared for a future pandemic or epidemic threat. The sub-index on prevention – one of the most relevant ones to avoid existential catastrophes in particular – is where countries perform worst. Countries are not prepared to prevent globally catastrophic biological events that could cause damage on a larger scale than COVID-19. 78% of countries do not have the ability to provide expedited approval for medical countermeasures, such as vaccines and antiviral drugs, during a public health emergency. 178 countries score less than 50 out of 100 points for whole-of-government biosecurity systems, training, personnel vetting, transport of infectious substances, and cross-border transfer and screening. Countries also lack the local expertise to appropriately adapt guidance and implement the International Health Regulations<sup>147</sup>.

One of the key issues is the lack of coherent regulation. No matter whether a pathogen is released deliberately or accidentally, whether it is natural or engineered: a current lack of international agreement hinders the maintenance of security procedures, blocks the training of talent, bars investigations and shirks responsibilities<sup>148</sup>. Advancing discussions on verification mechanisms in the BWC could allow to better understand the research landscape and reduce risks. An increased use of investigative mechanisms through awareness raising and training programs and increased diplomatic power for the WHO to investigate public health emergencies could enable better information sharing and response.

Initiatives like the WHO BioHub for pathogen sharing, the WHO Innovation hub, or the International Biosecurity and Biosafety Initiative for Science work collaboratively with global partners to strengthen biosecurity norms and develop innovative tools and the incentives to uphold them. To achieve biosecurity, it is key to collaborate with a diverse range of stakeholders, especially including industry and philanthropy.

To further develop the biosecurity across health, security and private sectors, an important and uniting milestone would be to develop an internationally coherent legal framework to prevent the accidental or intentional misuse of broadly accessible DNA synthesis technologies. In a recent report<sup>149</sup>, the Institute for Progress outlines a clear path forward via:

1. Regulating the production and sale of benchtop DNA synthesizers;
2. Obliging synthetic DNA providers to screen all customers and orders;
3. Tying research funding to the use of compliant providers; and
4. Building shared international infrastructure to reduce the cost of screening and adapt more effectively to advances in technologies.

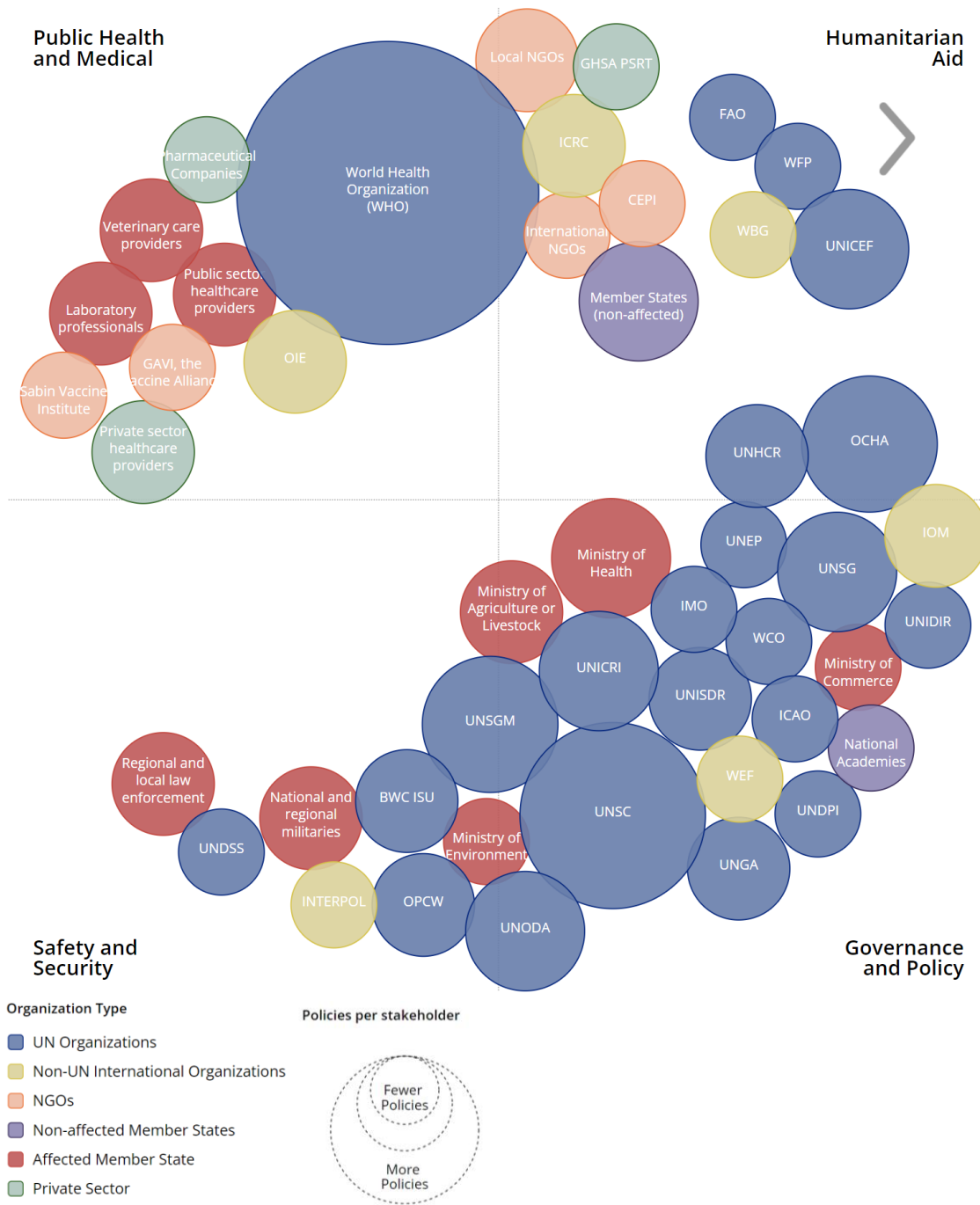


Figure 13. Regime complex for biosecurity <sup>150</sup>

### 3.1.2 Artificial intelligence governance

The governance architecture for AI technology is nascent. There are, however, many existing policies that indirectly affect AI development (such as research and economic policies). The key question is not whether to govern AI, but how it is currently being governed, and how that governance might become more informed, integrated, effective, and anticipatory <sup>109</sup>.

There are secretariats of emerging AI initiatives, for example, the OECD AI Policy Observatory and the Global Partnership on AI at the regional level, and the UN System Chief Executives Board (CEB). So far, these bodies have coordinated on technical, process and governance guidelines for the creation of safe AI <sup>151</sup>. With the adoption of “Our Common Agenda”, the UN General Assembly has initiated the development of a Global Digital Compact, growing out of the UN High-level Panel on Digital Cooperation, which discusses AI but is unlikely to focus on it.

There have also been many attempts to standardize principles of AI ethics; some emerging from the multilateral system, and others from adjacent areas. As of 2019, there were 80 entities that published “AI Ethics Principles” <sup>152</sup>, which are complementary yet not coordinated <sup>153</sup>. For example, OECD members and several other states have agreed on AI Principles. A version of these principles has been adopted by the G20 and informally endorsed by China <sup>153</sup>. China has endorsed the regional G20 AI Principles, which reflect the OECD Principles. Separately, the G7 launched the Global Partnership on AI. UNESCO developed principles on the ethical use of AI, which were recently adopted by the CEB for coordination <sup>154</sup>. This proliferation of principles has left the field fragmented, and a lack of enforcement mechanisms limits their impact <sup>155</sup>.

Over the past years, the International Telecommunication Union (ITU) has convened stakeholders and promoted dialogue on AI risks and benefits through its ‘AI for Good’ initiative. Yet the ITU is currently limited to regulating telecommunication systems, such as radio infrastructure. Efforts to expand its role have been resisted by other stakeholders. In a similar vein, the International Organization for Standardization (ISO) has established a committee to discuss AI standards, but has historically not been vested with the powers necessary to address AI alignment.

Private fora may also influence international governance <sup>156</sup>, including the Partnership on AI (a forum for leading AI companies and civil society organizations) and the Institute of Electrical and Electronic Engineers’ Ethically Aligned Design initiative. The UN Secretary-General has also announced plans to establish a multistakeholder advisory body on global AI cooperation, but this has so far not come to pass <sup>157</sup>. UNESCO, the Council of Europe, and the OECD have similarly convened multi-stakeholder groups tasked with drafting policy instruments <sup>158–160</sup>.

On a regional level, the most significant initiative for human-centric, safe AI has been the recent European Union’s AI Act, which defines specific parameters and obligations for companies operating ‘high-risk’ AI systems <sup>161</sup>. Its intention to annually review definitions of AI and ‘high-risk’ systems seems particularly noteworthy given the rapid speed of technical progress and regulatory (mis)understanding thereof. Despite its stringent and comprehensive approach, the EU AI Act could be improved with requirements for conformity assessments prior to deployment of high-risk general purpose models.

The international arms control architecture could be a promising pathway for governance of AI development, yet it seems similarly fragmented and underpowered. Most of the AI-relevant discussion has revolved around Lethal Autonomous Weapons Systems: through the UN General Assembly’s First Committee on Disarmament and International Security, and the Convention on Certain Conventional

Weapons (CCW) Group of Governmental Experts (GGE). However, in December 2021 the Sixth Review Conference on the CCW failed to agree to a mandate to regulate LAWS. There has been little discussion in the arms control sphere of other hazards emerging from transformative AI: such as threats to nuclear arsenals or strategic stability.

## 3.2. International and national prioritization of existential risk

While there are existing or emerging regime complexes for biosecurity and artificial intelligence, they might not necessarily prioritize existential risk as a worst-case scenario worthy of attention. We discuss how national risk assessments tend to overlook existential risk, and the quaint attention to existential risk in multilateral fora, in particular in the Sendai Framework.

### 3.2.1 Existential risk in national risk assessments

National risk assessments do not prioritize existential risk because they fail to address its core characteristics: global scope, extreme consequences, and emerging nature (Table 2).

Individual countries have limited ownership over risks with global scope because their source or first observable consequences may lie elsewhere. Risk assessments rely on expert judgments, which often differ from one another. This leads to strong differences in how countries assess global risks. For example, the United Kingdom puts a volcanic event with a £100 million to £1 billion impact into the category of 4%-25% annual likelihood<sup>162</sup>, while Switzerland puts an event of 1 billion CHF closer to 0.001%<sup>163</sup>. National governments tend to underinvest in the reduction of global risk because of the lack of a coercive power at the international level.

Traditionally, the scenarios in national risk analyses are visually summarized in a risk landscape with likelihood on one axis and impact on the other axis<sup>164</sup>. In order to simplify and account for uncertainty, those axes are usually not continuous but consist of a limited set of discrete categories. The choice of these categories in a risk matrix can create distortions and may even influence which risks get analyzed. Specifically, low cut-off points for the maximum impact category will systematically deprioritize lower probability, high impact risks. For example, the highest impact category in the national risk analysis of Sweden is reached with a minimum of 50 deaths per disaster. Yet this is several orders of magnitude below the lives lost in recent disasters like COVID-19 (c.21,000) and several orders of magnitude lower again than an existential catastrophe.<sup>3</sup>

Moreover, because the risks analyzed in a national risk analysis usually span several orders of magnitude, it is common to use logarithmic rather than linear scales for risk landscapes. This helps to spread out the analyzed risk scenarios on the available two-dimensional space. However, it also means that the visual intuition can severely underestimate how big the difference between larger identified risks and minor risks is.

---

<sup>3</sup> There are relevant disaster risk scenarios that lead to significantly more than 50 deaths in Sweden. For example, Sweden has recorded more than 20'000 recorded deaths due to the COVID-19 pandemic,<sup>#</sup> which is already 400 times higher than the threshold to reach the maximum impact category in the national risk matrix. Across almost all types of hazards, disasters with a lower impact are more frequent than “worst case” scenarios. Therefore, lower probability, extreme impact risks will be assessed as a lower priority in such a matrix even if their annualized expected impact exceeds that of more frequent disasters that are serious but not as extreme in their impact.



<b>Existential risk characteristics</b>	<b>Limitations in national risk assessments</b>
Global scope	<p>National governments have limited ownership of transnational risk</p> <p>Expert judgements lead to different estimations of transnational risk</p> <p>Absence of international coercive power does not incentivize to tackle transnational risk</p>
Extreme consequences	<p>Discrete risk categories neglect low-probability, high-impact scenarios</p> <p>Logarithmic scales leads to an intuitive underestimation of extreme scenarios</p> <p>Assessments do not include worst-case scenarios</p>
Emerging nature	<p>Historical data does not reflect emerging scenarios</p> <p>Risk assessment stakeholders mostly are emergency services</p> <p>“1 in X years” risk communications leads to wrong intuitions of risk frequency</p>

Table 2. Why national risk assessments fail to take existential risk into account

National risk analyses do not include “worst-case” scenarios. National risk analyses usually work with one or multiple reference scenarios per hazard type. These reference scenarios are then assessed in terms of their likelihood and impact, and are integrated into the risk landscape. The severity of the scenarios is set as “severe” or a “reasonable worst case”.

Emerging risks are more challenging to estimate than established risks based on historical data. For example, the first batches of cybersecurity-insurances generally underestimated damages and lost the insurers money.

A second challenge is that national disaster risk management is geared towards emergency services as the main stakeholders involved in risk analyses and subsequent risk reduction and preparedness efforts. Yet, risks related to rapid technological change generally require whole-of-government interventions in the present to reduce uncertainty, to reduce risk, or to increase resilience.

Some risk analyses translate the annualized probability that a hazard event happens within the time horizon of the risk analysis (ca. 2-10 years) into a recurrence period format of “1 in X years”. This is supposed to be reader-friendly, but it is misleading and systematically communicates values too low for increasing/emerging risks (and too high for decreasing risks). It suggests to the reader that the event in question is expected to happen once within the horizon of the next X years. However, for most hazards except for natural hazards like earthquakes this would be a wrong interpretation because the level of risk is not static but influenced by societal, environmental, and technological factors that change over time. For

instance, some people may be inclined to believe that the risk of a future pandemic has decreased given that COVID-19 has occurred, but this is not the case.

### 3.2.2 Existential risk in international documents and in the Sendai Framework

A recent study shows that, based on texts recorded by the UN library, existential risk is only mentioned 97 times<sup>165</sup>. 69% of such mentions relate to nuclear wars. This is understandable given the Cold War context during which such texts were developed. However, it means that other contributors to existential risk, especially technological development, tend to be under-discussed, which thus does not provide a referential or legal basis to address existential risk within UN fora.

Moreover, environmental risks – those directly related to environmental degradation – dominate the risk mitigation strategies of the 2030 Agenda. Other types of risks, such as pandemics, do not receive as much attention. Technological development is mostly discussed as an opportunity to make progress on sustainable development goals, but not as a potential threat to the goals. Therefore, the absence of other risks and technological development as a contributor to existential risk in the 2030 Agenda does not provide a referential or discursive basis to address existential risk within UN fora.

While marginal improvements can be made by discussing AI or biotechnology within specific international fora, existential risk governance requires an appropriate foundation for coordination of international organizations and nation states. The Sendai Framework can provide this basis because existential risk and technological development wholly fall within its scope (Article 15).

That said, the implementation of the Sendai Framework currently relies on the path laid out by the Hyogo Framework and the focus of UNDRR's constituencies. The Hyogo Framework focused on natural hazards. The Sendai Framework's constituencies are, for the most part, within civil protection, and disaster management authorities which are primarily reactive bodies. Moreover, some discussions at the last Global Platform in Bali in May 2022 revolved around the importance of daily disasters, which skewed participants' attention away from outlier, extreme-impact scenarios.

Additionally, a keyword analysis of the national voluntary reports submitted for the mid-term review of the Sendai Framework shows scarce attention to existential risk and technological development more broadly.<sup>4</sup> Our findings show that existential risk, rapid technological change, artificial intelligence and synthetic biology (as well as related terms) are not or barely mentioned across reports (Table 3). This lack of focus on existential risk and rapid technological change illustrates the point of this thematic study: these issues are neglected by governments despite the growing evidence supporting their importance. It also shows that changing the focus of governments – and especially the bodies implementing the Sendai Framework – will be challenging.

---

<sup>4</sup> We used the reports available at MTR SF Submissions and Reports | Midterm Review of the Sendai Framework by 01.01.2023 and developed a custom Python script that primarily relied on the pdfminer library and the n-gram counter from scikit-learn to process and analyze the reports. Only the 38 reports submitted in English were analyzed - we excluded the national voluntary reports from Argentina, Costa Rica, Cuba, Kazakhstan, Mexico, Montenegro and Morocco as they were submitted in another language.

Category	Word	n	Countries who mentioned the keyword
Existential risk	Extinction	0	
	Existential	1	New Zealand
	Collapse	23	4 countries mention collapse, Republic of Korea accounts for 70%
	Irreversible	0	
	Catastrophic	34	7 countries; Australia and Bosnia & Herzegovina account for 60%
	Cascade	7	Poland accounts for >40%
	Outlier	0	
	Low probability	3	Bosnia & Herzegovina; Sweden
	Man-made hazard(s)	0	
	Future generations	2	Slovenia, Switzerland
	Intergenerational	1	Switzerland
Rapid technological change	Rapid technological change	0	
	Technological development	4	Bosnia & Herzegovina; Norway
	Technological risk(s), hazard(s) or disaster(s)	20	Turkiye accounts for >50%
Biotechnology	Synthetic biology	0	
	Biological risk(s)	2	Ethiopia
	Biological weapon(s)	0	
	Pandemic(s)	575	33 countries mention pandemics
	Disarmament	0	
Artificial intelligence	Artificial intelligence	10	Republic of Korea accounts for >50%
	Algorithm(s)	0	
	Automation	6	Bosnia & Herzegovina; Cambodia, Kyrgyzstan
	Digital transformation	2	Trinidad and Tobago
Stage	Forecasting	131	23 countries; Bosnia & Herzegovina accounts for 45%
	Prevention	531	30 countries; Korea and Bosnia & Herzegovina account for 30%
	Response	1387	
	Adaptive governance	0	
Other	Natural disaster(s)	401	
	Natural hazard(s)	405	

Table 3. Keyword analysis of voluntary national reports submitted to the mid-term review

All in all, the current implementation of the Sendai Framework is not living up to its potential for existential risk mitigation because of:

- ❖ **Lack of terminology:** there is very little discussion of risks as outcomes to be prevented. Risk is not defined in UNDRR's risk terminology (only disaster risk which is different). Rapid technological change is not defined. Extreme risk scenarios such as existential risk are also not defined.
- ❖ **Neglect of scale:** there is little discussion of extreme scenarios. For example, damages from COVID-19 are often discussed as exacerbating the vulnerabilities to natural hazards, rather than a risk that could have been prevented. The community generally seems scope-insensitive, that is, action is not proportionate to the scale of the risks or disasters, at least before they manifest.
- ❖ **Neglect of source:** most discussions of the implementation of the Sendai Framework revolve around natural hazards, and barely discuss biological disasters like pandemics; they almost never discuss technological disasters. UNDRR's links with the World Meteorological Organization, its staff's specialization in natural hazards, the Bali Agenda for Resilience, and the agenda of the last Global Platform are a few examples that show a strong skew towards natural hazards. This skew is disproportionate to the importance of natural hazards and should ideally be corrected in proportion to the size of different risks.
- ❖ **Lack of prevention:** Despite the Sendai Framework's focus on prevention and mitigation, its Targets are mostly reactive. The government entities which are responsible to deliver on disaster risk reduction are themselves oriented towards a reactive approach, such as civil protection, disaster management authorities, the police and the military.

### 3.3. The pacing problem and the roots of neglecting existential risk

The lack of prioritization of existential risk from rapid technological change at national and international levels indicates a need for policy change. Ideally, resources would be redirected to existential risk and rapid technological change governance. However, the absence of such policy change to date highlights a more fundamental institutional problem: governments and international organizations lack anticipatory capacity.

Examining reallocation patterns of institutional budgets is one approach to examine the dynamics of policy change. Cross-geographical data on budget reallocations show that policy change – regardless of the type of institution – are reactive to disasters <sup>166</sup>. For instance, government healthcare spending in response to COVID-19 dwarfed global spending on disease surveillance <sup>167,168</sup>. This reactive signature of policy change means that the drivers of policymaking are not fit-for-purpose to invest in prevention.

Pervasive short-termism is one of the key drivers of institutions' reactive nature <sup>169,170</sup>. First, policy actors' understanding of the future informs how they think about future risks. On the one hand, they discount the future impact of policies because scarce information about future scenarios reduces the expected value of policies' long-term effects <sup>171–173</sup>. On the other hand, policy actors neglect future scenarios not out of the conviction that they do not matter, but because they are removed from their attention <sup>174,175</sup>. Second, self-interest and short-term preferences drive policy actors' motivation to favor contemporary, immediate stakeholders over distant future stakeholders that would be affected by existential risk <sup>170,176–178</sup>. Third, short election cycles and monetary support from businesses that have short-term preferences influence the rhythm and motivation of reelections <sup>179</sup>. Similarly, short media cycles and political polarization distract

actors from sustained reflection on long-term policy trajectories<sup>170</sup>. And last but not least, institutional inertia may similarly prevent policy actors from making long-term policy, even if they want to<sup>175</sup>.

The relationship between technological change and reactive policy change is well described by the pacing problem (Figure 13). Over time, technologies develop at a fast pace (green curve). By definition, the understanding of technologies' positive and negative consequences develops with a delay (red curve). Because of this delay and the above drivers of reactivity, regulation of technological developments happens with a further delay (blue curve).

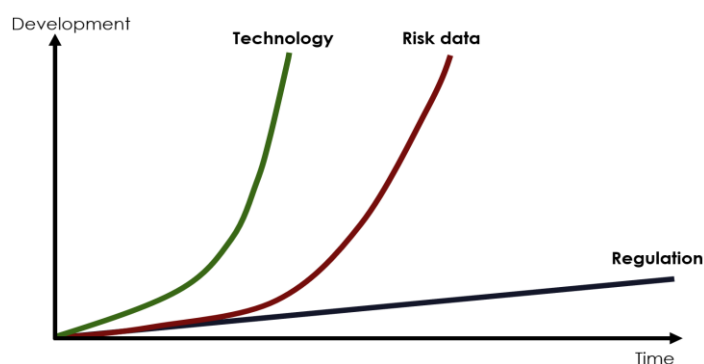


Figure 13. The pacing problem<sup>180</sup>

This reactivity works sufficiently well to address non-catastrophic disasters. Reactivity also makes sense from a political point of view; it is difficult for governments to make large budget reallocations in prevention of not-yet-experienced phenomena. However, the shocks that would lead to global catastrophes or extinction events are too severe to react to. They would overwhelm the capacity of governments worldwide. For instance, an engineered pandemic may spread so fast that it infects almost the entire globe before the pathogen is recognized: in this case, improving response is much less effective and costlier than prevention of the pandemic in the first place. Therefore, there is a need for solutions to bring the three curves of the pacing problem graph closer together; either via slowing down technological developments, speeding up risk analysis and policy uptake, or managing the timing of selected technologies.

### 3.4. Risk-informed technological development for existential risk reduction

As previously said, technologies – like advances in biotechnology and artificial intelligence – can both boost and harm development. Additionally, development goals – such as the SDGs – can help ensure benefits from technologies are equally distributed while risks are minimized (see Box 7.). Despite the SDGs' scarce mention of the value and risk of technological development, it is important to consider them from a development lens. This may help to design governance strategies that reduce risks associated with technologies while reaping their benefits.

Traditionally, two approaches help governments and international organizations govern technological development<sup>181</sup>. They can affect technological development in terms of *direction* (whether it will lead to positive or negative consequences) and *speed* (the pace at which it will produce such consequences). However, affecting direction and speed can form overly dichotomous views on technological development, leading some actors to withdraw from techno-solutionism and others to overlook associated risks. A third component which affects the consequences of technological development is its *timing*<sup>182</sup>. Given that some technologies can reduce the risks of other technologies (e.g., seat belts reduce mortality associated with

car accidents), affecting the sequence (i.e., the timing) of technological development can help leverage the upsides while minimizing the downsides.

There are at least three types of risk-reducing technologies associated with different approaches to affect their timing in relation to risk-increasing technologies <sup>182</sup>.

First, **safety technologies** help reduce or prevent negative consequences by modifying risk-increasing technologies, and thus reducing the chances of accidents and misuse. Here, the approach is to minimize the time between the development of risk-increasing technology and a corresponding safety technology (Figure 14a). For example, the development of electronic locks for nuclear weapons, permissive action links (PALs), in the 1960s has reduced the risk of accidental or unauthorized launches. Therefore, developing PALs for nuclear weapons as part of the Manhattan Project instead of two decades later would have reduced the potential of misuse during that period.

Second, **defensive technologies** help reduce or prevent negative consequences without modifying risk-increasing technologies. Here, the approach is to privilege defensive technologies between the development of risk-increasing technology (Figure 14b). For example, mRNA vaccines, a novel vaccine platform technology that can be quickly adapted to different pathogens, significantly contributed to curbing the COVID-19 pandemic <sup>183</sup>. Therefore, if government research funding agencies would prioritize pandemic prevention technologies like pathogen detection and platform vaccines before advancing the ability to create pandemic pathogens, this would lead to less societal harm from accidental or deliberate pandemics.

Third, **substitute technologies** create low-risk alternatives to risk-increasing technologies while producing similar benefits. Here, the approach is to fund and investigate low-risk alternatives instead of risk-increasing technologies (Figure 14c). For example, clean energy technologies, like wind turbines or photovoltaics, can replace environmentally-harmful fossil fuels.

To pursue all three approaches, governments and international organizations can pursue different strategies. For delaying risk-increasing technologies, they may employ defunding, moratoria, stage-gating, bans, regulations, social norms, or divestment. For advancing risk-reducing technologies, they may employ preferential funding, prizes, advanced market commitments, tax incentives, regulation, and coordination <sup>182</sup>.

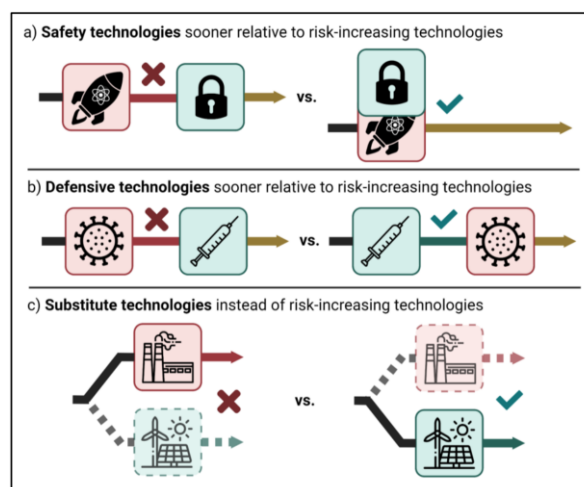


Figure 13. Three approaches for risk-informed technological development <sup>182</sup>

### Box 7. Biotechnology, artificial intelligence and the Sustainable Development Goals

Biotechnology and artificial intelligence contribute to existential risk, and thus are also a threat to progress towards the Sustainable Development Goals (SDGs). Importantly, progress towards the SDGs can also help reduce existential risk from technological change.

For example, the impact of an infectious disease threat resulting from advances in biotechnology is shaped by many factors, including political decision making, the type of disease and its mode of infection, increased urbanization and human expansion, a changing climate, and upticks in travel, trade, and terrorism. Conversely, reducing the impact of an infectious disease is contingent on public trust in government, healthcare institutions, and public health professionals.

Similarly, the impact of artificial intelligence will depend on many factors, including the functioning of economic and political institutions, conflict and violence, etc. For example, calcified political institutions will be unable to respond to changes on the labor market and polities with low government trust will be especially vulnerable to ‘deep fakes’. Conversely, reducing such impacts will depend on trust in government, the quality of public-private partnerships, the ability of governments to keep up with technological developments, etc. The same logic would apply to other contributors to existential risk, whether it is nuclear war or an extreme weather event.

Adverse consequences from AI and biotechnology also threaten progress on the SDGs. As we have seen with the COVID-19 pandemic, a major global shock can set back development and even convert five years of progress towards the goals into worsening trends <sup>3</sup>. Indeed, the pandemic temporarily reversed decades of global progress: it pushed an additional 119-124 million people into extreme poverty (the first increase in global poverty in decades). The pandemic led 70-161 million additional people to experience hunger. It caused temporary school closures, causing about 1.6 billion children to be out of school by April 2021. And it disrupted immunization programs in about 70 countries. While these shocks may be opportunities to do better, their manifestation is, first, unnecessary and, second, extremely costly. Preventing such global shocks is thus of vital importance. Adverse consequences from biotechnology and artificial intelligence are a significant threat to SDG progress.

The following table outlines plausible interactions between biotechnology, artificial intelligence and the SDGs. This is an area that deserves more research.

Technological development	Potential threats & hazards	Related SDGs
Biotechnology	Engineered pandemics	E.g. COVID-19 has harmed SDG 1, 2, 3, 4, 5, 8, 9, and 10  Successfully governing an engineered pandemic would benefit from progress on SDG 1, 2, 3, 4, 5, 9, 16 and 17
Artificial intelligence	AI misalignment or misuse in economic, political, scientific and military systems	E.g. If an AI incident leads to a conflict, it would harm most SDGs and in particular SDG 1, 2, 3, 4, 6, 7, 8, 14, 15 and 16  Successfully reducing impacts from AI incidents and misuse would benefit from progress on at least SDG 9, 10, and 16.

### 3.5. Multilateral pathways to tackle existential risk until 2030

Existential risk and technological development are such broad issues that tackling them beyond the Sendai Framework is very important. There are, in fact, plenty of other multilateral processes that can help make progress on existential risk reduction (Table 4.). For instance, the European Union will finalize its AI Act in 2023. The WHO International Negotiating Body is working on a Pandemic Accord to be finalized in 2024. Stakeholders will convene at a Financing for Development Forum in 2024.

Another prominent set of processes is the UN Secretary General's Our Common Agenda (OCA) <sup>9</sup>. OCA proposes several tracks and a summit of the future to make progress on long-term risk, global risk reduction and the governance of technological development (Box 8.).

Altogether, multilateral processes delineate at least three pathways to make progress: a general pathway to reduce existential risk, a pathway to strengthen biosecurity, and a pathway to progress on AI governance. All three pathways lead to 2030 as a key milestone where existential risk and technological development could feature more prominently in the renewal of the Sustainable Development Goals.

#### **Box 8. Our Common Agenda as an opportunity to reduce existential risk**

OCA is the SG's response to a request made by the UN General Assembly (GA) to set a vision forward. Resolution 76/6 gave the SG permission to move forward <sup>184</sup>.

In terms of framing and agenda-setting, OCA posits that the world is at a crucial point in history, facing either breakthrough through progress and cooperation or breakdown from nationalism and existential risk. OCA aims to further the implementation of the UN 2030 Agenda and sets the tone for upcoming discussions about its renewal <sup>185</sup>. Because OCA was published in 2021, COVID-19 and pandemics feature prominently in the report.

SGs proposals are always constrained by institutional inertia and current affairs. For example, despite the similarity to OCA, then-SG Ban Ki-moon's 2013 report "International solidarity and the needs of future generations" did not receive much attention from member states. This might be understandable, given that the UN system was already facing a monumental task under similar framing: developing the 2015 Sustainable Development Goals (SDGs) - set out in a 2012 GA resolution titled "The future we want".

Building on Ban's report and the success of the SDGs in streamlining sustainability policy across all sectors all around the globe, SG Antonio Guterres' has managed to clear a previously missed hurdle for future generations: OCA was adopted as an intergovernmental process by the General Assembly (GA). OCA thus holds the promise of tackling a reform of the UN system toward multilateral structures that drive consideration of the long term across all levels of governance; ideally, catalyzing greater attention to future-proofed policymaking also on the part of member states.


Out of 69 entry points, 15 are directly relevant to long-term governance and thus to existential risk reduction. With concrete proposals for future generations and its consideration of existential risk, OCA offers unusually fertile ground for international efforts to secure posterity.

If successful, OCA will lead to the creation of new processes (e.g. Strategic Foresight and Global Risk Report), institutions (e.g. Futures Lab), and policy (e.g. Global Digital Compact) which will likely last for decades. If unsuccessful, key framings or recommendations on existential risk and future generations might become diluted, politicized, stigmatized, or instrumentalized for other ends.



Year	General existential risk	Biosecurity	Artificial Intelligence governance
2023	UN Futures Lab development	Universal Healthcare Pandemic Prevention Summit (including Tuberculosis)	NATO Summit
	UN Emergency Platform development	World Health Assembly	EU AI Act
	OECD Interim Foresight for Emerging risks report	Global Health Security Agenda	New Agenda for Peace
	G20 Working Group on Disaster Risk Reduction	BWC Meeting of Experts	African Commission: regional development of AI framework
	SDG Finance meeting	BWC Meeting of States Parties	UN AdC cybercrime convention
	G7 leaders summit, Japan		GGE on Lethal Autonomous Weapons Meeting
	Political Declaration on the Mid-term Review of the Sendai Framework		G7 Digital Ministers Meeting, Japan
	Report of High-level Advisory Board on Effective Multilateralism		End of consultations on Global Digital Compact
	Conference on Disarmament		IGF 2023 / World Summit on Information Society Forum
	SDG Midpoint Summit		
	Ministerial meeting on Our Common Agenda		
	Special Envoy for Future Generations		
	Declaration on Future Generations		
	UNGA High-level dialogue on Financing for Development		
9th Conference on States Parties on Arms Trade Treaty			
2024	Bretton Woods +80 meetings	WHO Intergovernmental Negotiating Body (Pandemic Accord) submits results	GGE on Lethal Autonomous Weapons results
	Financing for Development Forum	Update of the International Health Regulations	UN Cybercrime Convention
	Summit of the Future	Global Health Security Agenda	Global Digital Compact
	Pact for the Future adopted	BWC Meeting of Experts	IGF 2025 / WSIS forum
	Final OECD report on existential risks	BWC Meeting of States Parties	
	4th Financing for Development Conference		
	EU Foresight report published		
2025	World Social Summit		
	UN GAR 2025		
	4th Financing for Development Conference	World Health Assembly	CEN/CENELEC report published (EU AI Act)
	EU Foresight report published	BWC Meeting of Experts	WSIS +20
2026	World Social Summit	BWC Meeting of States Parties	
	UN GAR 2025		
2030	Successor to 2030 agenda		Successor to Sendai framework

Table 4. 3 multilateral pathways to reduce existential risk



## 4. Recommendations for the Mid-term Review of the Sendai Framework and beyond

## 4.1. 12 outcomes within the priorities of the Sendai Framework to reduce existential risk

Below we formulate 12 outcomes within the four priorities of the Sendai Framework. The four priorities are:

1. Understanding disaster risk
2. Strengthening disaster risk governance to manage disaster risk
3. Investing in disaster risk reduction for resilience
4. Enhancing disaster preparedness for effective response and to "Build Back Better" in recovery, rehabilitation and reconstruction

For each priority, we formulate the key priority for existential risk and suggest a set of priority outcomes for the United Nations Office for Disaster Risk Reduction (UNDRR), other international organizations, governments, and civil society to pursue.

### 4.1.1. Priority 1: Concrete and common understanding of existential risk

To make progress, it is important to foster a better understanding of existential risk at three levels:

1. The United Nations Office for Disaster Risk Reduction (UNDRR) provides and promotes a clear definition of existential risk. We suggest the following definition of existential risk: the probability of human extinction or the irreversible end of development over a given timeframe. The UN system should adopt this shared definition of existential risk.
2. UNDRR and other UN agencies foster a better understanding of the hazards and vulnerabilities that contribute to existential risk. In particular, those that emerge from rapid technological change such as advances in biotechnology and artificial intelligence.
3. UNDRR, other UN agencies and governments identify priority leverage points to mitigate hazards, reduce exposure and vulnerabilities, including their potential value, costs and timeframe.

This understanding of existential risk should be shared among UN agencies, member-states and civil society. UNDRR can contribute to promote this shared understanding.

### 4.1.2. Priority 2: Strengthening existential risk governance

Paired with a better understanding of existential risk, the following are priority outcomes for existential risk governance.

4. UNDRR, other UN agencies and member-states apply the Sendai Framework according to its full scope (specified in article 15), including man-made hazards.
5. UNDRR, UN agencies and member-states increase their focus on preventive measures with regard to man-made hazards.
6. UNDRR, UN agencies and member-states change the overall mandate of 'disaster risk reduction' by making it about long-term resilience through a whole-of-government approach.<sup>5</sup> This change could be reflected by relabelling the mandate as 'risk reduction'.

---

<sup>5</sup> It is important to note that focusing on other hazards than natural hazards and focusing on prevention will not happen without considerable investment, and will not be possible if DRR remains in the hands of reactive bodies or staff with a natural hazard background. Therefore, making both changes requires a strong overhaul of governance bodies. That change is expected to be costly for some actors, but extremely beneficial in the long-term as, otherwise, DRR will continue to neglect larger risks.

7. UNDRR fosters coordination and partnerships with other UN agencies that are better-placed to tackle drivers of existential risk such as advances in biotechnology and artificial intelligence.<sup>6</sup>
8. UNDRR fosters the adaptive governance of existential risk by generating policy options for other institutions, preserving flexibility in the face of rapid developments in technology, and fostering a high degree of participation from member states, civil society, private sector and academia.

#### 4.1.3. Priority 3: Incentivizing existential risk reduction

The economic incentives to reduce existential risk are not strong enough. As a result, stakeholders – from governments, international financial institutions, international organizations to the private sector through insurers – do not give priority to existential risk. The problem cannot be attributed to any institution and its reduction is difficult to sell to politicians and the public. To make progress, it is vital to create funding mechanisms that address the reduction of global, extreme and emerging risks, even if they are of lower probability. National and international development agencies should recognize that sustainable development is undermined by existential risk, and make risk-reduction financing a part of the development agenda. The UNDRR can advocate for this.

9. The UN system, including multilateral funding mechanisms, increases its contribution to the reduction of low probability, high-impact risks.
10. National governments dedicate a percentage of their budget to extreme risk reduction.

#### 4.1.4. Priority 4: Enhancing existential risk preparedness for effective response

The shock cascade that would lead to extinction can be sudden and disrupt basic infrastructures very quickly. Therefore, most investments to reduce existential risk should be allocated to prevention and preparedness for swift response.

11. UN agencies and national governments implement measures that reduce hazards and vulnerabilities.
12. UN agencies and national governments adopt individual and/or shared emergency mechanisms to bring necessary diverse stakeholders together to respond quickly to crises.

### 4.2. Two instruments to deliver outcomes

Below, we propose two international instruments to deliver the above outcomes. Currently, there is nobody inside international organizations or governments who is responsible for existential risk reduction. Therefore, it is necessary to spell out a basic set of instruments that can create concrete progress and maintain momentum towards further progress. The following instruments not only further the implementation of the Sendai Framework for existential risk reduction, but also set the stage for risk reduction post-2030.

We suggest the following instruments:

1. An **international coordination and capacity-building mechanism** on existential risk
2. A **set of funding instruments** focused on lower-probability, high-impact risk

---

<sup>6</sup> Examples include at least the World Health Organization, the UN Office for Disarmament Affairs, the World Trade Organization, the International Organization for Standardization, and the International Telecommunication Union. With the mid-term review, UNDRR can state that risk reduction is an outcome, not a sector, and that such outcome will result from coordinated, networked action among those bodies. It is crucial that UNDRR avoids siloing existential risk within itself or another body.

#### 4.2.1. An international coordination and capacity-building mechanism on existential risk

There is currently a vacuum in the international system as to which actors should lead on existential risk reduction. Filling this gap is important because existential risk needs to be an explicit part of an entity's mandate to not be forgotten. Therefore, tasking a dedicated body to work on existential risk is important to crystallize a focus of attention on global, extreme and emerging risk.

However, it is also important to avoid siloing existential risk. Its causes and consequences must be addressed by multiple agencies. There is uncertainty on the causes of existential risk, which means it is important to preserve flexibility as to which entities should work on its reduction directly. Additionally, technological development requires flexible governance decisions, instead of over-specification. Therefore, a dedicated body should focus on coordination and capacity-building. This would ensure that existential risk reduction remains a shared responsibility, rather than becoming its own silo.

This coordination and capacity-building mechanisms would seek to (1) promote a definition of existential risk, (2) aggregate current knowledge on existential risk, (3) inform UN agencies, governments and funding instruments on what they can do to reduce existential risk.

This mechanism would ideally be formed by a partnership with UNDRR and other UN agencies. An example of such a mechanism is the One Health High Level Expert Panel which was formed from the quadripartite between the World Health Organization, the Food and Agriculture Organization, and the World Organization for Animal Health, and the UN Environment Programme <sup>186</sup>.

This mechanism does not need to be large. A representative from each participating agency, some academic experts and civil society representatives would suffice. It could be funded by member states or UN agencies. Though only modest funding would be required (USD ~1MM/year),<sup>7</sup> it should have consistent funding.

To move forward, the following two actions can be taken:

1. Conduct an **assessment** of (1) the mandate of such a mechanism, (2) which UN agencies should participate, (3) and funding prospects.
2. Organize a **first consultative event** convening agencies to align on the rationale and mandate of the mechanism and to outline the next steps.

#### 4.2.2. A set of funding instruments focused on lower-probability, high-impact risk

Almost no actor is incentivized to dedicate resources to existential risk reduction because it has historically been lower-probability than more common risks. Additionally, funding re-allocation tends to focus on response and mitigation rather than prevention. There is therefore a gap of incentives and investments for the prevention of lower-probability, high-impact risk. We suggest updating existing instruments. Table 5 summarizes four existing mechanisms and how they could be updated. The amount of funding that each instrument should dedicate to lower-probability, high-impact risk is an uninformed guess of what would unlock progress on reducing lower-probability, high-impact risk without slowing down progress on more frequent scenarios.

To move forward, the following action points can be undertaken:

---

<sup>7</sup> We calculated, in USD, 3 full-time equivalents (450,000), office (40,000), consultancy (100,000), events including travels (200,000), and publication costs (5,000).

1. Conduct an **assessment** of (1) financial gaps to address existential risk, (2) which existing instruments should be updated and how, and (3) which new instruments should be created and how to finance them.
2. Organize a **first consultative event** convening existing funds, the stakeholders of the coordination and capacity-building mechanism to align on gaps and strategy to move forward.

Instrument	What is it?	Type of funding	What does it fund?	Updates to reduce existential risk
Global Facility for Disaster Reduction and Recovery <sup>187</sup>	Multi-donor partnership that supports LMICs to understand, manage, and reduce their risks.	Multi-Donor Trust Fund (MDTF) and other big programs	360 active grants focused on natural hazards and climate change	<ul style="list-style-type: none"> <li>❖ Expand the scope to man-made hazards &amp; rapid technological change .</li> <li>❖ Dedicate 10% to lower-probability high-impact risk (~400MM\$)</li> <li>❖ Fund international coordination and capacity-building mechanism</li> <li>❖ Fund action points below</li> </ul>
Fund for Pandemic Prevention, Preparedness and Response <sup>188</sup>	Dedicated stream of financing to strengthen PPR capabilities in LMICs and address critical gaps.	Financial Intermediary Fund of the World Bank	Fund multiple agencies to implement programs with governments, regional and global entities (ex: Africa CDC)	<ul style="list-style-type: none"> <li>❖ Expand scope to engineered pandemics and applications of biotechnology</li> <li>❖ Dedicate 10% to lower-probability high-impact risk</li> <li>❖ Fund action points below</li> </ul>
Global Environment Facility <sup>189</sup>	World's largest funder of biodiversity protection, restoration, and climate change response in developing countries.	Trust fund administered by the World Bank	The funds are transferred to 18 GEF Agencies which distribute money to NGOs and governments to execute their projects	<ul style="list-style-type: none"> <li>❖ Expand the scope to risks from biotech</li> <li>❖ Increase spending to prevention and preparedness</li> <li>❖ Dedicate 10% to lower-probability high-impact risk</li> </ul>
United Nations Central Emergency Response Fund <sup>190</sup>	Humanitarian fund established by UNGA in 2006 which offers humanitarian assistance to populations affected by natural hazards and armed conflicts.	The GA called on all Member States and the private sector to ensure \$1 billion	Reaction to humanitarian crisis; anticipatory actions linked to climate disasters (drought, floods, storms) and disease outbreaks	<ul style="list-style-type: none"> <li>❖ Expand the scope to man-made crises other than wars and natural hazards</li> <li>❖ Add engineered pandemics to anticipatory actions</li> <li>❖ Increase spending to anticipatory actions</li> <li>❖ Dedicate 5% to lower-probability high-impact risk</li> </ul>
Health Emergency Preparedness and Response Program <sup>191</sup>	Only World Bank fund mechanism exclusively dedicated to health emergency preparedness and response.	World Bank Trust Fund	Help countries respond to COVID-19 and prepare for future health emergencies	<ul style="list-style-type: none"> <li>❖ Expand scope to engineered pandemics</li> <li>❖ Dedicate 10% to lower-probability high-impact risk (~10MM)</li> <li>❖ Increase spending to prevention and preparedness</li> </ul>

Table 5. Existing funding mechanisms and how they can be updated to reduce existential risk

### 4.3. 30 actions to reduce existential risk

We conducted a review and prioritization of 425 recommendations<sup>8</sup> to reduce existential risk and selected the top 30, which we adapted for the purpose of this thematic study (summarized in Table 6). We nested them according to the two first priorities of the Sendai Framework. The third priority on financing is addressed by the second instrument above. The fourth priority on building back better is not addressed

<sup>8</sup> See supplementary information

because, given the total or nearly total severity of existential risk, we emphasize the need for stronger governance, prevention and preparedness.

#### 4.3.1. Action points to improve existential risk understanding

*At the international level and in general on existential risk*

1. Publish an **official UNDRR document** on existential risk, including its definition and importance.
2. Develop a **global risk register**, assessing risk according to their probability, severity, and origins. Include an up-to-date classification of technological hazards.
3. Task States Parties of the Sendai Framework with **periodically reporting on the state of global risks** emanating from their territories. These reports may also be issued regionally, cognizant of how risk profiles may correlate.
4. Establish a clear **link between existential risk and the Sustainable Development Goals (SDGs)**. Emphasize the relationship between human rights and the responsible use of technology properly to relevant stakeholders at all levels. Record and evaluate the losses caused by incidents, including keeping track of the extent to which targets and indicators of the Sustainable Development Goals (SDGs) are affected by different kinds of disasters and making this a core part of the follow up to the SDGs. Estimate the costs of disasters to the SDGs (e.g. COVID-19 or climate change) and assess the ideal proportion of SDGs resource allocation to disaster reduction. Advocate for existential risk reduction to be included in the post-2030 SDGs.
5. Produce **consensus forecasts** of the level of risk posed by different existential and global catastrophic threats by soliciting and aggregating estimates from experts. In the long run, these forecasts could be used to establish aspirational “risk budgets.” This could begin modestly (e.g., managing and summarizing an up-to-date database of existing estimates). Gradually these efforts could become more substantial (e.g., conducting “assessment cycles” analogous to those conducted by the Intergovernmental Panel on Climate Change (IPCC)). Over time, estimates would take into account a greater range of viewpoints and considerations and become more credible. Ultimately, there could be an equivalent of the IPCC for a broad range of catastrophic and existential scenarios.
6. Conduct **risk and opportunity assessments from emerging technologies**, including more detailed assessment of each region’s vulnerabilities and contributions to the risks (through surveys and/or expert elicitations). Complement assessments with **surveys on neglected risk** to understand public perceptions of risks from rapid technological change. These could be analogous to or parallel to Article 36 of the 1977 Additional Protocol to the 1949 Geneva Convention reviews<sup>192</sup>.
7. Create **best practices** for existential risk assessment and governance at international, regional and national levels.
8. Develop **training opportunities** to learn more about existential risk and apply best practices for international, regional and national experts.

9. Develop **simulations for preparedness, response and recovery** of unprecedented threats, including access to food, shelters and other needs.

*At the international level and on artificial intelligence*

10. Elicit **scenarios and benchmarks for AI disasters**. This would include a framework to lay out clear benchmarks for the progress of an AI disaster. Regularly elicit expert opinions on best guesses of the potential unfurling of an AI crisis or steps during takeoff, such as certain thresholds for deception or levels of power seeking. This could include “warning shots” that are more concretely specified <sup>193</sup>.
11. Create a **database of AI incidents**. Keep track of bugs, edge cases, overfitting, and incidents. Specifically target “cases of undesired or unexpected behavior” from these AI systems <sup>194</sup>, including the publication of case studies describing these. Consider keeping such incidents anonymous <sup>195</sup>, so to avoid potential reputational costs/concerns <sup>196</sup>. This might be similar to the initiative created by the Partnership on AI <sup>155,197</sup>. Consider investing in such a database for improvement.
12. Increase **resources to improve the understanding of technological risks threatening Low and Middle Income Countries**.
13. Support **improved understanding across government of the potential risks and harms of evolving uses of AI**. Develop capacity and infrastructure for monitoring progress in AI, including by collecting information on inputs such as data and compute, and tracking AI models affecting citizens and business. Establish more coordinated AI foresight and horizon-scanning programs across government which feed directly into policy and regulatory decisions.

*At the national level and in general on existential risk*

14. Integrate **global, extreme and emerging risks in national risk assessments** using methods that adequately communicate the size of extreme risk.
15. Integrate **existential risk in civic programs** to increase risk awareness. This can include training future policymakers in estimating future scenarios, understanding worst-case scenarios, understanding implications of technological development, and developing policy portfolios to build resilience.

*At the national level and on artificial intelligence*

16. Build up **technical expertise on AI within government departments and regulators**. Create specific technical roles around AI in key departments and ensure these positions can be made attractive to experts. Implement an AI and machine learning training program for existing civil servants with particular attention to AI risks.
17. Make it **easier for governments to draw on outside expertise in AI**, ethics and governance. Establish a secondment/fellowship program (similar to TechCongress) placing experts in AI ethics and governance in relevant parts of government.



#### 4.3.2. Action points to improve existential risk governance

##### *At the international level and on biotechnology*

18. Support the formation of a **dedicated international normative body** to promote the early identification and reduction of global catastrophic biological risks <sup>146</sup>.
19. Ensure that a formal, clear, and regularly exercised process for **investigation and attribution of an alleged use of biological weapons** is robust and sustained <sup>198</sup>, for example through the creation of a Joint Assessment Mechanism <sup>199</sup>.
20. Develop **regional coordination systems** to respond to pandemics. Not all regions have established strong pandemic preparedness and response systems. Regional coordination systems can reduce costs to individual states by centralizing medicine authorization procedures, they can help identify and fill gaps in national capacities, and they can develop strategies more responsive to the particular context, structures and vulnerabilities of their members <sup>200</sup>.
21. Develop **frameworks to establish proper global governance of Do-It-Yourself (DIY) Bio at both national and international levels**. The regional DIY Bio codes of ethics and conduct and international agreements such as the Nagoya and Cartagena protocols are insufficient to adequately regulate the fast-evolving DIY Bio field; particularly due to its voluntary approach to compliance. There is therefore a need to develop frameworks to establish proper global governance of DIY Bio at both national and international levels to ensure accountability amongst DIY Biologists <sup>201</sup>.

##### *At the international level and on artificial intelligence*

22. Facilitate **cooperation across nations and sectors on shared AI ethics and governance challenges**. Lead on establishing norms and standards for safe and responsible use of AI internationally, which can include to facilitate international dialogue around standards for safe and responsible use of AI in warfare. Promote technical standards for safe and responsible AI via bodies such as the OECD and the UN. Work with the private sector to boost AI safety resources and incentivize industry investment. Collaborate with academia to more effectively verify and challenge the claims made by the AI industry, by supporting red-teaming efforts in academia and increasing computing power resources for researchers. Utilize such collaborations to strike an agreement between governments, corporations and other AI developers to ensure mutually verifiable compliance to established regulations.
23. Organize **track 1.5/2 dialogues for diplomats on AI** <sup>202</sup>. Seek to increase funding in supporting researchers, officials and diplomats for AI-related track 1.5/2 dialogues between leading AI powers such as China, the US and EU. Use funding to explore existing frameworks for international cooperation and sponsor meetings.

##### *At the national level and on biotechnology*

24. Identify and rapidly **increase financing for national pandemic preparedness** across the public health and agricultural sectors. UN Member States should urgently identify and rapidly increase financing for national pandemic preparedness across the public health and agricultural sectors,

including for capabilities outlined within the World Health Organization's Joint External Evaluations. As part of this process, countries should establish benchmarks and prioritize financing for biosecurity and other security sector-related targets. This should be a multi-sectoral process that includes the private sector <sup>198</sup>.

25. Encourage the **private sector to increase their sustainable development and health security portfolios in research, development**, and capacity building, using the 2021 Global Health Security Index to identify priority areas aimed at preventing epidemics and pandemics from causing catastrophic damage on a global scale.
26. Govern the **availability and access to both the information and physical reagents** (via, for example, unregulated DNA synthesis) or tools to do with advanced biotechnology. Mandate compulsory DNA synthesis screening for national-level government-funded research institutions.

*At the national level and on artificial intelligence*

27. Promote **safe and trustworthy AI development and deployment via improved incentives, norms, processes, and governance structures**. This would include investing in AI safety, security, and interpretability via academic and private sector research, and cultivate talent in these areas via student fellowships. Commit to building a thriving and effective AI assurance ecosystem. Establish red-teaming exercises and throughout-lifetime stress-testing of AI systems used by the government. Commit to reviewing the use of AI in high-risk domains, including in weapons systems, nuclear command and control, and critical infrastructure, with a view to implementing robust assurance mechanisms and potentially restricting use where risks outweigh benefit.
28. Increase **interface with science and private sector**. Begin to liaise with scientific grant-making committees and individual labs, such as in biotechnology and AI development, to introduce frameworks for publishing research responsibly by considering the risks involved with information hazards. This may mean withholding dangerous research from the public, only allowing publication in authorized circles, redacting certain information or requiring additional safety checks that prove that such research publication has greater benefits than harm.
29. Foster **inclusive participation and a diversification of voices in the governance process**. E.g., AI bias in Facial Recognition technology is mainly due to a lack of data diversification. Diversify talent, e.g., AI talent in global south countries through programs that equip young diplomats and scholars with resources, knowledge and a global network to exchange ideas.
30. Increase **investments in AI safety efforts**. Differentially invest in near-term AI safety efforts, which are currently more accepted within the AI policy and academic fields across the world, but specifically within fields that overlap with existential risk.

Priority	Level	Cluster #	Actions	
Improve existential risk understanding	International	1	Publish an official UNDRR document on existential risk.	
		2	Develop a global risk register.	
		3	Task States Parties of the Sendai Framework with periodically reporting on the state of global risks.	
		4	Establish a clear link between existential risk and the Sustainable Development Goals (SDGs).	
		General 5	Produce consensus forecasts of the level of risk posed by different existential and global catastrophic threats.	
		6	Conduct risk and opportunity assessments from emerging technologies.	
		7	Create best practices for existential risk assessment.	
		8	Develop training opportunities to learn more about existential risk and apply best practices.	
		9	Develop simulations for preparedness, response and recovery of unprecedented threats.	
	AI	10	Elicit scenarios and benchmarks for artificial intelligence (AI) disasters.	
		11	Create a database of AI incidents.	
		12	Increase resources to understanding technological risk in Low and Middle Income Countries (LMIC).	
		13	Support improved understanding across government of the potential risks and harms of evolving uses of AI.	
National	General 14	Integrate global, extreme and emerging risks in national risk assessments.		
	15	Integrate existential risk in civic programs to increase risk awareness.		
	AI 16	Build up AI technical expertise within government departments and regulators.		
		17	Make it easier for governments to draw on outside expertise in AI,	
Improve existential risk governance	International	18	Support the formation of a dedicated international normative body to promote the early identification and reduction of global catastrophic biological risks.	
		Bio 19	Ensure that a regularly exercised process for investigation and attribution of an alleged use of biological weapons.	
		20	Develop regional coordination systems to respond to pandemics.	
		21	Develop frameworks to establish proper global governance of DIY Bio at both national and international levels.	
	AI	22	Facilitate cooperation across nations and sectors on shared AI ethics and governance challenges.	
		23	Organize track 1.5/2 dialogues for diplomats on AI.	
	National	Bio	24	Identify and rapidly increase financing for national pandemic preparedness.
			25	Encourage the private sector to increase their sustainable development and health security portfolios in research, development.
			26	Govern the availability and access to both the information and physical reagents.
		AI	27	Promote safe and trustworthy AI development and deployment via improved incentives, norms, processes, and governance structures.
28			Increase interface with science and private sector.	
29			Foster inclusive participation and diversification of voices in the governance process.	
30	Increase investments in AI safety efforts.			

Table 6. Actions to reduce existential risk

## 5. Conclusion

On 18 February 2023, High Commissioner for Human Rights, Volker Türk, expressed his concern that recent advances in and the deployment of large-language models – such as the ones underpinning ChatGPT and Bing Chat – put human rights at serious risk<sup>203</sup>. The rapid advances of technology beg the question: will its creators step up to assume governance responsibilities? This puzzle of increasing risk and a governance vacuum is at the core of this thematic study.

With a review of technological developments and their contribution to existential risk; an assessment of existing institutional structures; and the provision of thirty action points, we hope this report equips the international community not only with new agenda points – existential risk and rapid technological change – but also a roadmap to move forward.

It is urgent and possible to take actions that advance risk-informed development and reduce the risk of existential catastrophe. While it is primordial to understand the risk associated with technological advances – such as in biotechnology and artificial intelligence, their governance should also leverage the benefits for a more equal and sustainable development. We summarize three key findings to guide this balancing act.

**First, existential risk is decently likely this century (1.9 to 14.3%) and technological development is one of its core contributors.** Advances in biotechnology, including synthetic biology, have led to extensive research on engineered pathogens. The risks posed by lowering the barrier to transforming an innocuous viral vector into a lethal virus need to be carefully weighed and deliberately monitored and regulated. Artificial intelligence (AI) is becoming a transformative technology with the potential to lead to irreversible changes in society, including in welfare and wealth. The alignment problem of transformative artificial intelligence is a key issue for AI governance, as the complex goals of humans are difficult to teach, leading to unintended or dangerous behavior.

**Second, governments and international organizations are potentially the most effective instruments to govern technological development but are not fit-for-purpose yet.** The regime complexes for biosecurity and AI are either underpowered or nascent. Existential risk and technological development receive only quaint attention within national risk assessments, UN texts, the Sustainable Development Goals, and, in particular for this report, the Sendai Framework. More fundamentally, governments and international organizations are reactive and lack anticipatory capacity to reliably shape technological development. Therefore, strategies are needed to improve the governance of technological development. One such strategy is *differential technological development* - modeling interactions between technologies to leverage upsides while minimizing risk.

**Third, a path for progress toward the renewal of the Sustainable Goals in 2030 is provided by the Secretary-General's Our Common Agenda, and the Mid-term review of the Sendai framework.** It is important to consider existential risk reduction and responsible technological development as global public goods, whose achievement requires a whole-of-society approach. Therefore, there is a need for coordination and funding mechanisms to ensure that progress on these otherwise neglected issues takes root in the right way. We outline outcomes, instruments and 30 actions to make progress at both national and international levels, and leverage milestones until 2030.

## 6. References

1. Climate crisis past point of no return, Secretary-General says, listing global threats at General Assembly consultation on 'Our Common Agenda' report - World | ReliefWeb. <https://reliefweb.int/report/world/climate-crisis-past-point-no-return-secretary-general-says-listing-global-threats> (2022).
2. Roser. The world is awful. The world is much better. The world can be much better. *Our World in Data* <https://ourworldindata.org/much-better-awful-can-be-better>.
3. UNSTATS. *The Sustainable Development Goals Report*. <https://unstats.un.org/sdgs/report/2021/The-Sustainable-Development-Goals-Report-2021.pdf> (2021).
4. Blanquart, F. *et al.* A transmission-virulence evolutionary trade-off explains attenuation of HIV-1 in Uganda. *eLife* **5**, e20492 (2016).
5. Roose, K. A Conversation With Bing's Chatbot Left Me Deeply Unsettled. *The New York Times* (2023).
6. Kemp, L. *et al.* Climate Endgame: Exploring catastrophic climate change scenarios. *Proc. Natl. Acad. Sci.* **119**, (2022).
7. Sendai Framework for Disaster Risk Reduction 2015-2030. <https://www.undrr.org/publication/sendai-framework-disaster-risk-reduction-2015-2030> (2015).
8. UNDRR. Report for consultations: Stakeholder perspectives - Midterm Review of the Sendai Framework. <https://sendaiframework-mtr.undrr.org/publication/report-consultations-stakeholder-perspectives-midterm-review-sendai-framework> (2022).
9. UN. Secretary-General's report on "Our Common Agenda". <https://www.un.org/en/content/common-agenda-report/> (2021).
10. UNDP, U. *Human Development Report 2021-22. Human Development Reports* <https://hdr.undp.org/content/human-development-report-2021-22> (2022).
11. Henn, B. M., Cavalli-Sforza, L. L. & Feldman, M. W. The great human expansion. *Proc. Natl. Acad. Sci.* **109**, 17758–17764 (2012).

12. Oppenheimer, S. A single southern exit of modern humans from Africa: Before or after Toba? *Quat. Int.* **258**, 88–99 (2012).
13. Zakaria, F. The Toba Super-Catastrophe as History of the Future. *Indonesia* **113**, 31–48 (2022).
14. Kemp, L. Are we on the road to civilisation collapse? <https://www.bbc.com/future/article/20190218-are-we-on-the-road-to-civilisation-collapse>.
15. EM-DAT | The international disasters database. <https://www.emdat.be/>.
16. Twigg, J. COVID-19 as a 'critical juncture': A scoping review. *Glob. Policy* **6**, 1–20 (2020).
17. Sarkees, M. R. & Wayman, F. *Resort to war: 1816-2007*. (Cq Press, 2010).
18. Beard, S., Rowe, T. & Fox, J. An analysis and evaluation of methods currently used to quantify the likelihood of existential hazards. *Futures* **115**, 102469 (2020).
19. Hemsell, C. The investigation of natural global catastrophes. *J. Br. Interplanet. Soc.* **57 (1/2)**, 2–13 (2004).
20. Gott, R. Implications of the Copernican principle for our future prospects. *Nature* **363**, 315 (1993).
21. Wells, W. Human survivability. in *Apocalypse When? Calculating How Long the Human Race Will Survive* (ed. Wells, W.) 67–92 (Praxis, 2009). doi:10.1007/978-0-387-09837-1\_5.
22. Simpson, F. Apocalypse Now? Reviving the Doomsday Argument. Preprint at <https://doi.org/10.48550/arXiv.1611.03072> (2016).
23. Leslie, J. *The End of the World: The Science and Ethics of Human Extinction*. (Routledge, 1996). doi:10.4324/9780203007723.
24. Bostrom, N. Existential risks: Analyzing human extinction scenarios and related hazards. *J. Evol. Technol.* **9**, (2002).
25. Rees, M. J. *Our final hour: A scientist's warning: how terror, error, and environmental disaster threaten humankind's future in this century—on earth and beyond*. (Basic Books (AZ), 2003).
26. Sandberg, A. & Bostrom, N. *Global Catastrophic Risks Survey*. <https://www.fhi.ox.ac.uk/reports/2008-1.pdf> (2008).
27. Metaculus. A Global Catastrophe This Century. <https://www.metaculus.com/notebooks/8736/a-global-catastrophe-this-century/> (2022).
28. Ord, T. *The Precipice: Existential Risk and the Future of Humanity*. (Hachette Books, 2020).

29. Hellman, D. M. E. Risk Analysis of Nuclear Deterrence. *Tau Beta Pi* (2008).
30. Barrett, A. M., Baum, S. D. & Hostetler, K. Analyzing and Reducing the Risks of Inadvertent Nuclear War Between the United States and Russia. *Sci. Glob. Secur.* **21**, 106–133 (2013).
31. Lundgren, C. What Are the Odds? *Nonproliferation Rev.* **20**, 361–374 (2013).
32. Turchin, A. Assessing the future plausibility of catastrophically dangerous AI. *Futures* **107**, 45–58 (2019).
33. Pamlin, D. & Armstrong, S. Global challenges: 12 risks that threaten human civilization. *Glob. Chall. Found. Stockh.* (2015).
34. Day, T., André, J.-B. & Park, A. The evolutionary emergence of pandemic influenza. *Proc. R. Soc. B Biol. Sci.* **273**, 2945–2953 (2006).
35. Madhav, N. Modelling a modern-day Spanish flu pandemic. *AIR Worldw. Febr.* **21**, 2013 (2013).
36. Fan, V. Y., Jamison, D. T. & Summers, L. H. Pandemic risk: how large are the expected losses? *Bull. World Health Organ.* **96**, 129–134 (2018).
37. Bagus, G. Pandemic risk modeling. in *Measuring and Managing Catastrophe Risk The 2nd Annual CAA, MAF, and PRMIA Joint Conference on ERM, Chicago, IL, USA, June* vol. 5 8–13 (2008).
38. Klotz, L. C. & Sylvester, E. J. The Consequences of a Lab Escape of a Potential Pandemic Pathogen. *Front. Public Health* **2**, (2014).
39. Lipsitch, M. & Inglesby, T. V. Moratorium on Research Intended To Create Novel Potential Pandemic Pathogens. *mBio* **5**, e02366-14 (2014).
40. Fouchier, R. A. M. Studies on Influenza Virus Transmission between Ferrets: the Public Health Risks Revisited. *mBio* **6**, e02560-14 (2015).
41. Millett, P. & Snyder-Beattie, A. Existential Risk and Cost-Effective Biosecurity. *Health Secur.* **15**, 373–383 (2017).
42. Manheim, D. Questioning Estimates of Natural Pandemic Risk. *Health Secur.* **16**, 381–390 (2018).
43. Wagner, G. & Weitzman, M. L. Climate shock. in *Climate Shock* (Princeton University Press, 2016).
44. King, D., Schrag, D., Dadi, Z., Ye, Q. & Ghosh, A. *Climate change: A risk assessment*. <https://www.csap.cam.ac.uk/media/uploads/files/1/climate-change--a-risk-assessment-v11.pdf> (2017).

45. Dunlop, I. & Spratt, D. Disaster Alley: Climate Change, Conflict, and Risk. *Melb. Aust.* (2017).
46. Xu, Y. & Ramanathan, V. Well below 2 °C: Mitigation strategies for avoiding dangerous to catastrophic climate changes. *Proc. Natl. Acad. Sci.* **114**, 10315–10323 (2017).
47. Halstead, J. Stratospheric aerosol injection research and existential risk. *Futures* **102**, 63–77 (2018).
48. Müller, V. C. & Bostrom, N. Future Progress in Artificial Intelligence: A Survey of Expert Opinion. in *Fundamental Issues of Artificial Intelligence* (ed. Müller, V. C.) 555–572 (Springer International Publishing, 2016). doi:10.1007/978-3-319-26485-1\_33.
49. Grace, K., Salvatier, J., Dafoe, A., Zhang, B. & Evans, O. Viewpoint: When Will AI Exceed Human Performance? Evidence from AI Experts. *J. Artif. Intell. Res.* **62**, 729–754 (2018).
50. Baum, S. D., Barrett, A. M. & Yampolskiy, R. V. Modeling and Interpreting Expert Disagreement About Artificial Superintelligence. *Informatica* **41**, (2017).
51. Bostrom, N. Existential risk prevention as global priority. *Glob. Policy* **4**, 15–31 (2013).
52. Avin, S. *et al.* Classifying global catastrophic risks. *Futures* **102**, 20–26 (2018).
53. Nursimulu, A. Governance of Slow-Developing Catastrophic Risks: Fostering Complex Adaptive System and Resilience Thinking. SSRN Scholarly Paper at <https://doi.org/10.2139/ssrn.2830581> (2015).
54. Skolnikoff, E. B. *The Elusive Transformation: Science, Technology, and the Evolution of International Politics*. (Princeton University Press, 1993).
55. Discontinuous progress investigation. *AI Impacts* <https://aiimpacts.org/discontinuous-progress-investigation/> (2015).
56. UNCTAD. *Technology and Innovation Report 2021*. (2021).
57. UNODA. *Report of the Secretary-General on current developments in science and technology and their potential impact on international security and disarmament efforts*. <https://www.un.org/disarmament/report-of-the-secretary-general-on-current-developments-in-science-and-technology/> (2020).
58. World Health Organization. *Emerging technologies and dual-use concerns: a horizon scan for global public health*. <https://www.who.int/publications/i/item/9789240036161> (2021).
59. Reppy, J. Managing Dual-Use Technology in an Age of Uncertainty. *Forum (Genova)* **4**, (2006).



60. Vinuesa, R. *et al.* The role of artificial intelligence in achieving the Sustainable Development Goals. *Nat. Commun.* **11**, 233 (2020).
61. Erisman, J. W., Sutton, M. A., Galloway, J., Klimont, Z. & Winiwarter, W. How a century of ammonia synthesis changed the world. *Nat. Geosci.* **1**, 636–639 (2008).
62. Weindling, P. The uses and abuses of biological technologies: Zyklon B and gas dinfestation between the first world war and the holocaust. *Hist. Technol.* **11**, 291–298 (1994).
63. Ord, T. *Lessons from the Development of the Atomic Bomb | GovAI.* <https://www.governance.ai/research-paper/lessons-atomic-bomb-ord> (2022).
64. Messerli, P. Global Sustainable Development Report GSDR 2019. (2018).
65. Pauwels, E. *The New Geopolitics of Converging Risks - The UN and Prevention in the Era of AI.* <https://collections.unu.edu/eserv/UNU:7308/PauwelsAIGeopolitics.pdf> (2019).
66. Roser, M. Technology over the long run: zoom out to see how dramatically the world can change within a lifetime. *Our World in Data* <https://ourworldindata.org/technology-long-run> (2023).
67. Synthetic biology - Decision Adopted by the Conference of the Parties to the Convention on Biological Diversity. (2016).
68. Seroogy, C. M. & Fathman, C. G. The application of gene therapy in autoimmune diseases. *Gene Ther.* **7**, 9–13 (2000).
69. Chen, X.-Z. *et al.* A Novel Anti-Cancer Therapy: CRISPR/Cas9 Gene Editing. *Front. Pharmacol.* **13**, (2022).
70. Cross, D. & Burmester, J. K. Gene Therapy for Cancer Treatment: Past, Present and Future. *Clin. Med. Res.* **4**, 218–227 (2006).
71. AABB. FDA Approves First Gene Therapy for Hemophilia B. *aabb.org* <https://www.aabb.org/news-resources/news/article/2022/11/23/fda-approves-first-gene-therapy-for-hemophilia-b>.
72. Naddaf, M. Researchers welcome \$3.5-million haemophilia gene therapy — but questions remain. *Nature* **612**, 388–389 (2022).
73. Sandbrink, J. B., Alley, E. C., Watson, M. C., Koblentz, G. D. & Esvelt, K. M. Insidious Insights: Implications of viral vector engineering for pathogen enhancement. *Gene Ther.* 1–4 (2022) doi:10.1038/s41434-021-00312-3.

74. Gilsdorf, J. R. & Zilinskas, R. A. New Considerations in Infectious Disease Outbreaks: The Threat of Genetically Modified Microbes. *Clin. Infect. Dis.* **40**, 1160–1165 (2005).
75. Cross, G. & Klotz, L. Twenty-first century perspectives on the Biological Weapon Convention: Continued relevance or toothless paper tiger. *Bull. At. Sci.* **76**, 185–191 (2020).
76. Schoch-Spana, M. *et al.* Global Catastrophic Biological Risks: Toward a Working Definition. *Health Secur.* **15**, 323–328 (2017).
77. WHO Coronavirus (COVID-19) Dashboard. <https://covid19.who.int>.
78. Millett, P. & Snyder-Beattie, A. Human Agency and Global Catastrophic Biorisks. *Health Secur.* **15**, 335–336 (2017).
79. Gryphon Scientific. Risk and benefit analysis of gain of function research. (2016).
80. Clauset, A., Young, M. & Gleditsch, K. S. On the Frequency of Severe Terrorist Events. *J. Confl. Resolut.* **51**, 58–87 (2007).
81. Mallapaty, S. COVID prompts global surge in labs that handle dangerous pathogens. *Nature* **610**, 428–429 (2022).
82. Hvistendahl, M. Student Infected With Debilitating Virus in Undisclosed Biolab Accident. *The Intercept* <https://theintercept.com/2022/11/01/biosafety-lab-accident-chikungunya-virus/> (2022).
83. Furmanski, M. Threatened pandemics and laboratory escapes: Self-fulfilling prophecies. *Bulletin of the Atomic Scientists* <https://thebulletin.org/2014/03/threatened-pandemics-and-laboratory-escapes-self-fulfilling-prophecies/> (2014).
84. Silver, A. Taiwan’s science academy fined for biosafety lapses after lab worker contracts COVID-19. <https://www.science.org/content/article/taiwan-s-science-academy-fined-biosafety-lapses-after-lab-worker-contracts-covid-19> (2022).
85. Yeh, K. B. *et al.* Significance of High-Containment Biological Laboratories Performing Work During the COVID-19 Pandemic: Biosafety Level-3 and -4 Labs. *Front. Bioeng. Biotechnol.* **9**, (2021).
86. Biosafety level. *Wikipedia* (2023).
87. Danzig, R. *et al.* Insights Into How Terrorists Develop Biological and Chemical Weapons. *CNAS* (2011).

88. Kambouris, M. E., Manoussopoulos, Y., Patrinos, G. P. & Velegraki, A. Chapter 7 - Microbial Genomics in Public Health: A Translational Risk-Response Aspect. in *Applied Genomics and Public Health* (ed. Patrinos, G. P.) 131–148 (Academic Press, 2020). doi:10.1016/B978-0-12-813695-9.00007-8.
89. Sternberg, S. H. & Doudna, J. A. Expanding the Biologist's Toolkit with CRISPR-Cas9. *Mol. Cell* **58**, 568–574 (2015).
90. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
91. Gibson, D. G. *et al.* Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* **329**, 52–56 (2010).
92. UNEP. Frontiers 2018/19: Emerging Issues of Environmental Concern. *UNEP - UN Environment Programme* <http://www.unep.org/resources/frontiers-201819-emerging-issues-environmental-concern> (2019).
93. Barrangou, R. & Doudna, J. A. Applications of CRISPR technologies in research and beyond. *Nat. Biotechnol.* **34**, 933–941 (2016).
94. DNA Sequencing Costs: Data. *Genome.gov* <https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data> (2022).
95. Home | International Gene Synthesis Consortium. *International Gene Synthesis Consortium | The Promotion of Biosecurity* <https://genesynthesisconsortium.org/> (2017).
96. Lewis, G., Millett, P., Sandberg, A., Snyder-Beattie, A. & Gronvall, G. Information Hazards in Biotechnology. *Risk Anal.* **39**, 975–981 (2019).
97. Esvelt, K. M. Delay, Detect, Defend: Preparing for a Future in which Thousands Can Release New Pandemics. *Geneva Cent. Secur. Policy* (2022).
98. Sundaram, L. S. Biosafety in DIY-bio laboratories: from hype to policy. *EMBO Rep.* **22**, e52506 (2021).
99. Silvertown, J. A new dawn for citizen science. *Trends Ecol. Evol.* **24**, 467–471 (2009).
100. Wall, K. Biohackers push life to the limits with DIY biology. *The Guardian* (2015).

101. Esvelt, K. M. Inoculating science against potential pandemics and information hazards. *PLOS Pathog.* **14**, e1007286 (2018).
102. Callaway, E. 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature* **588**, 203–204 (2020).
103. Callaway, E. AlphaFold's new rival? Meta AI predicts shape of 600 million proteins. *Nature* **611**, 211–212 (2022).
104. Urbina, F., Lentzos, F., Invernizzi, C. & Ekins, S. Dual use of artificial-intelligence-powered drug discovery. *Nat. Mach. Intell.* **4**, 189–191 (2022).
105. Turing, A. Computing machinery and intelligence. *Mind* **59**, 433–460 (1950).
106. ITU. United Nations Activities on Artificial Intelligence (AI). (2018).
107. Dafoe, A. AI governance: a research agenda. *Gov. AI Program Future Humanity Inst. Univ. Oxf. Oxf. UK* **1442**, 1443 (2018).
108. Gruetzemacher, R. & Whittlestone, J. The transformative potential of artificial intelligence. *Futures* **135**, 102884 (2022).
109. Brundage, M. & Bryson, J. Smart Policies for Artificial Intelligence. Preprint at <https://doi.org/10.48550/arXiv.1608.08196> (2016).
110. Zhang, B. & Dafoe, A. Artificial Intelligence: American Attitudes and Trends. SSRN Scholarly Paper at <https://doi.org/10.2139/ssrn.3312874> (2019).
111. Roser, M. The brief history of artificial intelligence: The world has changed fast – what might be next? *Our World in Data* <https://ourworldindata.org/brief-history-of-ai> (2022).
112. Performance Development | TOP500. <https://www.top500.org/statistics/perfdevel/>.
113. Reed, S. *et al.* A Generalist Agent. *Trans. Mach. Learn. Res.* (2023).
114. Tamkin, A., Brundage, M., Clark, J. & Ganguli, D. Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models. Preprint at <https://doi.org/10.48550/arXiv.2102.02503> (2021).
115. Tan, H. Google didn't think its Bard AI was 'really ready' for a product yet, says Alphabet chairman, days after its stock fell following the chatbot's very public mistake. *Business Insider*

- <https://www.businessinsider.com/google-bard-ai-chatbot-not-ready-alphabet-hennessy-chatgpt-competitor-2023-2> (2023).
116. evhub. Bing Chat is blatantly, aggressively misaligned.
  117. Sevilla, J. *et al.* Compute Trends Across Three Eras of Machine Learning. Preprint at <https://doi.org/10.48550/arXiv.2202.05924> (2022).
  118. Stanford Institute for Human-Centered Artificial Intelligence (HAI). The AI Index Report – Artificial Intelligence Index. <https://aiindex.stanford.edu/report/> (2022).
  119. Trammell, P. & Korinek, A. *Economic Growth Under Transformative AI: A Guide to the Vast Range of Possibilities for Output Growth, Wages, and the Laborshare*. <https://www.governance.ai/research-paper/economic-growth-under-transformative-ai-a-guide-to-the-vast-range-of-possibilities-for-output-growth-wages-and-the-laborshare> (2020).
  120. Tschang, F. T. & Almirall, E. Artificial Intelligence as Augmenting Automation: Implications for Employment. *Acad. Manag. Perspect.* (2021) doi:10.5465/amp.2019.0062.
  121. CEPR. Artificial intelligence and the stability of markets. *CEPR* <https://cepr.org/voxeu/columns/artificial-intelligence-and-stability-markets> (2017).
  122. U.S. Commodity Futures Trading & Commission. Findings Regarding the Market Events of May 6, 2010: Report of the Staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues. (2010).
  123. Unver, A. Artificial Intelligence, Authoritarianism and the Future of Political Systems. SSRN Scholarly Paper at <https://papers.ssrn.com/abstract=3331635> (2018).
  124. CLPR. Artificial Intelligence and Judicial Bias. *Centre for Law & Policy Research* <https://clpr.org.in/blog/artificial-intelligence-and-the-courts/> (2021).
  125. Seger, E. *et al.* Tackling threats to informed decision-making in democratic societies: Promoting epistemic security in a technologically-advanced world. *Alan Turing Inst.* (2020).
  126. Shankar, S. & Zare, R. N. The perils of machine learning in designing new chemicals and materials. *Nat. Mach. Intell.* **4**, 314–315 (2022).
  127. Tumpey, T. M. *et al.* Characterization of the Reconstructed 1918 Spanish Influenza Pandemic Virus. *Science* **310**, 77–80 (2005).

128. Brundage, M. *et al.* The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *ArXiv Prepr. ArXiv180207228* (2018).
129. Avin, S. & Amadae, S. M. Autonomy and machine learning at the interface of nuclear weapons, computers and people. in (Stockholm International Peace Research Institute, 2019). doi:10.17863/CAM.44758.
130. Maas, M. M., Matteucci, K. & Cooke, D. Military Artificial Intelligence as Contributor to Global Catastrophic Risk. SSRN Scholarly Paper at <https://doi.org/10.2139/ssrn.4115010> (2022).
131. Gabriel, I. Artificial Intelligence, Values, and Alignment. *Minds Mach.* **30**, 411–437 (2020).
132. Russell, S. *Human compatible: Artificial intelligence and the problem of control.* (Penguin, 2019).
133. Learning through human feedback. <https://www.deepmind.com/blog/learning-through-human-feedback>.
134. Kubin, E. & von Sikorski, C. The role of (social) media in political polarization: a systematic review. *Ann. Int. Commun. Assoc.* **45**, 188–206 (2021).
135. Sevilla, J. & Moreno, P. Implications of Quantum Computing for Artificial Intelligence alignment research. Preprint at <https://doi.org/10.48550/arXiv.1908.07613> (2019).
136. Center for Security and Emerging Technology, Rudner, T. & Toner, H. *Key Concepts in AI Safety: An Overview.* <https://cset.georgetown.edu/publication/key-concepts-in-ai-safety-an-overview/> (2021) doi:10.51593/20190040.
137. Amodei, D. *et al.* Concrete Problems in AI Safety. Preprint at <https://doi.org/10.48550/arXiv.1606.06565> (2016).
138. What Are Global Public Goods? *IMF* <https://www.imf.org/en/Publications/fandd/issues/2021/12/Global-Public-Goods-Chin-basics>.
139. WEF. *Global Risks Report.* [https://www3.weforum.org/docs/WEF\\_Global\\_Risks\\_Report\\_2023.pdf#page=70](https://www3.weforum.org/docs/WEF_Global_Risks_Report_2023.pdf#page=70) (2023).
140. Berliner, B. Large Risks and Limits of Insurability. *Geneva Pap. Risk Insur. - Issues Pract.* **10**, 313–329 (1985).
141. Hartwig, R., Niehaus, G. & Qiu, J. Insurance for economic losses caused by pandemics. *Geneva Risk Insur. Rev.* **45**, 134–170 (2020).

142. GDV. Green paper: Supporting the economy to better cope with the consequences of future pandemic events. (2020).
143. Insurance Trades Unveil Federal Pandemic Solution. <https://www.namic.org/news/releases/200521mr01>.
144. OECD. National Terrorism Risk Insurance Programmes of OECD Countries with Government Participation.
145. Ferland, J. Cyber insurance – What coverage in case of an alleged act of War? Questions raised by the Mondelez v. Zurich case. *Comput. Law Secur. Rev.* **35**, 369–376 (2019).
146. Bell, J. & Nuzzo, J. *Global Health Security Index: Advancing Collective Action and Accountability Amid Global Crisis*. [https://www.ghsindex.org/wp-content/uploads/2021/12/2021\\_GHSIndexFullReport\\_Final.pdf](https://www.ghsindex.org/wp-content/uploads/2021/12/2021_GHSIndexFullReport_Final.pdf) (2021).
147. e-SPAR Public. <https://extranet.who.int/e-spar/#submission-details>.
148. Field, M. How to make sure the labs researching the most dangerous pathogens are safe and secure. *Bulletin of the Atomic Scientists* <https://thebulletin.org/2021/07/how-to-make-sure-the-labs-researching-the-most-dangerous-pathogens-are-safe-and-secure/> (2021).
149. Williams, B. & Kane, R. *Preventing the Misuse of DNA Synthesis: Five policy recommendations to curb downside risk*. (2023).
150. Katz, R. *et al.* Mapping stakeholders and policies in response to deliberate biological events. *Heliyon* **4**, e01091 (2018).
151. Cihon, P., Maas, M. M. & Kemp, L. Fragmentation and the Future: Investigating Architectures for International AI Governance. *Glob. Policy* **11**, 545–556 (2020).
152. Jobin, A., Ienca, M. & Vayena, E. The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* **1**, 389–399 (2019).
153. Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. & Srikumar, M. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. *SSRN Electron. J.* (2020) doi:10.2139/ssrn.3518482.
154. UN Chief Executives Board for Coordination. Principles for the Ethical Use of Artificial Intelligence in the United Nations System. (2022).

155. Avin, S. *et al.* Filling gaps in trustworthy development of AI. *Science* **374**, 1327–1329 (2021).
156. Green, J. F. & Auld, G. Unbundling the Regime Complex: The Effects of Private Authority. *Transnatl. Environ. Law* **6**, 259–284 (2017).
157. UN. Secretary-General's Roadmap for Digital Cooperation. <https://www.un.org/en/content/digital-cooperation-roadmap/> (2019).
158. CAHAI - Ad hoc Committee on Artificial Intelligence - Artificial Intelligence. <https://www.coe.int/en/web/artificial-intelligence/cahai>.
159. UNESCO. Ethics of Artificial Intelligence. <https://www.unesco.org/en/artificial-intelligence/recommendation-ethics>.
160. OECD Working Party and Network of Experts on AI. <https://oecd.ai/en/p/one-ai-members>.
161. Sioli, L. A European Strategy for Artificial Intelligence. (2021).
162. HM Government. UK National Risk Register. (2020).
163. BABS. Vulkanausbruch im Ausland - Gefährdungsdossiers und Szenarien. (2020).
164. OECD. National Risk Assessments: A Cross Country Perspective | en | OECD. <https://www.oecd.org/gov/national-risk-assessments-9789264287532-en.htm> (2018).
165. Boyd, M. & Wilson, N. Existential Risks to Humanity Should Concern International Policymakers and More Could Be Done in Considering Them at the International Governance Level. *Risk Anal.* **40**, 2303–2312 (2020).
166. Kuhlmann, J. & van der Heijden, J. What Is Known about Punctuated Equilibrium Theory? And What Does That Tell Us about the Construction, Validation, and Replication of Knowledge in the Policy Sciences? *Rev. Policy Res.* **35**, 326–347 (2018).
167. Dobson, A. P. *et al.* Ecology and economics for pandemic prevention. *Science* **369**, 379–381 (2020).
168. Bernstein, A. S. *et al.* The costs and benefits of primary prevention of zoonotic pandemics. *Sci. Adv.* **8**, eabl4183 (2022).
169. Boston, J. Assessing the options for combatting democratic myopia and safeguarding long-term interests. *Futures* **125**, 102668 (2021).
170. John, T. & MacAskill, W. Longtermist institutional reform. (2020).



171. Halevy, Y. Strotz Meets Allais: Diminishing Impatience and the Certainty Effect. *Am. Econ. Rev.* **98**, 1145–1162 (2008).
172. Irving, K. Overcoming Short-Termism: Mental Time Travel, Delayed Gratification and How Not to Discount the Future. *Aust. Account. Rev.* **19**, 278–294 (2009).
173. Jacobs, A. M. & Matthews, J. S. Why Do Citizens Discount the Future? Public Opinion and the Timing of Policy Consequences. *Br. J. Polit. Sci.* **42**, 903–935 (2012).
174. Jacobs, A. M. Policy making for the long term in advanced democracies. *Annu. Rev. Polit. Sci.* **19**, 433–454 (2016).
175. Johnson, D. & Levin, S. The tragedy of cognition: psychological biases and environmental inaction. *Curr. Sci.* **97**, 1593–1603 (2009).
176. Caney, S. Political institutions for the future: A five-fold package. (2016).
177. Bidadanure, J. Youth quotas, diversity, and long-termism. *Inst. Future Gener.* 432 (2016).
178. MacKenzie, M. K. Institutional design and sources of short-termism. *Inst. Future Gener.* 24–48 (2016).
179. Binder, S. A. Can congress legislate for the future. in *John brademas center for the study of congress, new york university, research brief* (2006).
180. Trump, B. D., Keisler, J. M., Galaitsi, S. E., Palma-Oliveira, J. M. & Linkov, I. Safety-by-design as a governance problem. *Nano Today* **35**, 100989 (2020).
181. Coad, A., Nightingale, P., Stilgoe, J. & Vezzani, A. Editorial: the dark side of innovation. *Ind. Innov.* **28**, 102–112 (2021).
182. Sandbrink, J., Hobbs, H., Swett, J., Dafoe, A. & Sandberg, A. Differential technology development: A responsible innovation principle for navigating technology risks. SSRN Scholarly Paper at <https://papers.ssrn.com/abstract=4213670> (2022).
183. Sandbrink, J. B. & Shattock, R. J. RNA Vaccines: A Suitable Platform for Tackling Emerging Pandemics? *Front. Immunol.* **11**, (2020).
184. UN General Assembly. Follow-up to the report of the Secretary-General entitled “Our Common Agenda”. (2021).

185. Espinosa, M. F. & Turk, D. Making the Most of the 2023 UN Summit of the Future. *PassBlue* <https://www.passblue.com/2022/04/20/making-the-most-of-the-2023-un-summit-of-the-future/> (2022).
186. WHO. One Health High-Level Expert Panel (OHHLEP). <https://www.who.int/groups/one-health-high-level-expert-panel>.
187. Global Facility for Disaster Reduction and Recovery. Strategy 2021-2025. (2021).
188. Financial Intermediary Fund for Pandemic Prevention, Preparedness and Response – PPR FIF. *World Bank* <https://www.worldbank.org/en/programs/financial-intermediary-fund-for-pandemic-prevention-preparedness-and-response-ppr-fif>.
189. Global Environment Facility | UNFCCC. <https://unfccc.int/topics/climate-finance/funds-entities-bodies/global-environment-facility>.
190. CERF. <https://cerf.un.org/>.
191. Health Emergency Preparedness and Response (HEPR) Umbrella Program. *World Bank* <https://www.worldbank.org/en/topic/health/brief/health-emergency-preparedness-and-response-hepr-umbrella-program>.
192. Boulanin, V. & Verbruggen, M. *Article 36 reviews: Dealing with the challenges posed by emerging technologies*. [https://www.sipri.org/sites/default/files/2017-12/article\\_36\\_report\\_1712.pdf](https://www.sipri.org/sites/default/files/2017-12/article_36_report_1712.pdf) (2017).
193. Kokotajlo, D. What are the most plausible ‘AI Safety warning shot’ scenarios?
194. Brundage, M. *et al.* Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. Preprint at <http://arxiv.org/abs/2004.07213> (2020).
195. Belfield, H. *et al.* Response to the European Commission’s consultation on AI. <https://www.cser.ac.uk/resources/response-european-commissions-consultation-ai/>.
196. Center for the Governance of AI. *Consultation on the European Commission’s White Paper on Artificial Intelligence: a European approach to excellence and trust*. <https://www.fhi.ox.ac.uk/wp-content/uploads/EU-White-Paper-Consultation-Submission-GovAI-Oxford.pdf> (2020).
197. Whittlestone, J. & Clark, J. Why and How Governments Should Monitor AI Development. Preprint at <http://arxiv.org/abs/2108.12427> (2021).

198. Cameron, E., Katz, R., Konyndyk, J. & Nalandian, M. *A Spreading Plague: Lessons and Recommendations for Responding to a Deliberate Biological Event*. (2019).
199. NTI. Joint Assessment Mechanism to Determine Pandemic Origins. *The Nuclear Threat Initiative* <https://www.nti.org/about/programs-projects/project/joint-assessment-mechanism-to-determine-pandemic-origins/> (2022).
200. Leyre, J. *Resetting the frame*. <https://www.cser.ac.uk/media/uploads/files/Global-Challenges-Quarterly-Risk-Report-August-2016.pdf> (2016).
201. Zulkefli, K. *et al. Empowering Do-it-yourself Biology by Doing-it-together: Collective Responsibility in Maximizing Benefit and Mitigating Risk*. <https://www.repository.cam.ac.uk/handle/1810/334664> (2022) doi:10.17863/CAM.82081.
202. Mapendere, J. Track One and a Half Diplomacy and the Complementarity of Tracks. *Cult. Peace Online J.* 66.81 (2005).
203. UN Human Rights [@UNHumanRights]. UN Human Rights Chief @volker\_turk deeply disturbed by potential for harm of recent #AI advances. Human agency, dignity & rights are at serious risk. Urgently calls on business & govts to develop human rights guardrails in line with our Office's guidance. <https://t.co/UPEJfJ4MeZ>. *Twitter* <https://twitter.com/UNHumanRights/status/1626900335900909571> (2023).