Differential technology development: A responsible innovation principle for navigating technology risks

Jonas B. Sandbrink[1,2,*], Hamish Hobbs[1], Jacob L. Swett[3], Allan Dafoe[4,5], Anders Sandberg[1]

[1] *Future of Humanity Institute, University of Oxford, Trajan House, Mill Street, Oxford OX2 0DJ, UK; emails: jonas.sandbrink@trinity.ox.ac.uk, hamish@longview.org; anders.sandberg@philosophy.ox.ac.uk*
[2] *Nuffield Department of Medicine, University of Oxford, Old Road Campus, Roosevelt Dr, Headington, Oxford OX3 7FZ, UK*
[3] *altLabs, Inc., Berkeley, CA 94704 , USA; email: jake@altlabs.tech*
[4] *DeepMind, 5 New Street Square, London, EC4A, UK; email: allandafoe@deepmind.com*
[5] *Centre for the Governance of AI, Trajan House, Mill Street, Oxford OX2 0DJ, UK*

*Correspondence to Jonas B. Sandbrink: jonas.sandbrink@trinity.ox.ac.uk

Highlights

- Existing responsible innovation practices do not consistently leverage risk-reducing interactions between technologies.
- The decarbonisation of the economy highlights how preferentially advancing risk-reducing technologies relative to risk-increasing technologies can mitigate negative technology impacts.
- The principle of "differential technology development" calls for relevant actors to leverage risk-reducing interactions across a technology portfolio by affecting the relative timing of technological developments.
- This principle may inform government research funding priorities and technology regulation, as well as philanthropic research and development funders and corporate social responsibility measures.
- Differential technology development may be particularly promising to mitigate potential catastrophic risks from emerging technologies like synthetic biology and artificial intelligence.

Abstract

Responsible innovation efforts to date have largely focused on shaping individual technologies. However, as demonstrated by the preferential advancement of low-emission technologies, certain technologies reduce risks from other technologies or constitute low-risk substitutes. Governments and other relevant actors may leverage risk-reducing interactions across technology portfolios to mitigate risks beyond climate change. We propose a responsible innovation principle of "differential technology development", which calls for leveraging risk-reducing interactions between technologies by affecting their relative timing. Thus, it may be beneficial to delay risk-increasing technologies and preferentially advance risk-reducing defensive, safety, or substitute technologies. Implementing differential technology development requires the ability to anticipate or identify impacts and intervene in the relative timing of technologies. We find that both are sometimes viable and that differential technology development may still be usefully applied even late in the diffusion of a harmful technology. A principle of differential technology development may inform government research funding priorities and technology regulation, as well as philanthropic research and development funders and corporate social responsibility measures. Differential technology development may be particularly promising to mitigate potential catastrophic risks from emerging technologies like synthetic biology and artificial intelligence.

1. Introduction

The global response to climate change has highlighted that a diverse set of actors can act to shift energy systems towards low emissions alternatives. Decarbonisation has required developing new policy models geared towards influencing the portfolio of available and deployed energy technologies to reduce societal harm (Mowery et al., 2010). To tackle climate change, governments and other relevant actors had to consider the effects of specific technologies in the setting of the broader technology portfolio. Certain technologies, like carbon capture, decrease the negative impacts of complementary high-emissions technologies. Other technologies, like clean energy sources, serve as low-risk substitutes to high emissions technologies. Climate change interventions have demonstrated that actively restructuring the technology portfolio and preferentially advancing risk-reducing technologies is possible. This portfolio-based approach may more generally be useful for mitigating risks from emerging technologies.

It is well established that innovation can produce a range of harms as well as benefits, meaning that the societal impact of technological innovation is determined not just by its speed but also by its direction (Coad et al., 2021). However, the societal impact of new technological innovations is not only determined by their direction, but also by their relative timing. For example, developing cars long before developing seat belts would be expected to result in more vehicle-related deaths. In this paper, we propose a responsible innovation principle of "differential technology development" that explicitly leverages risk-reducing interactions between technologies through their relative timing. Governments and other relevant actors may consider technology interactions to preferentially advance risk-reducing technologies. Differential technology development can help to fill gaps in societal approaches to innovation governance in areas such as biotechnology, where dual use potential is driving a need for novel governance approaches (McLeish and Nightingale, 2007).

In the first half of this article, we introduce the principle of differential technology development and provide definitions and context. First, we analyse existing responsible innovation efforts (section 1.1). Then, we define the principle of differential technology development and suggest its application for catastrophic risk mitigation (section 2). We follow with defining basic terminology for differential technology development (section 3). In the second part of this article, we examine when anticipating or identifying the impacts of technologies is possible (section 4) and how different actors can delay or preferentially advance specific technologies (section 5).

1.1 A history of shaping technology

Attempts to control the pace and direction of technological progress are not new. In one early example, in the 5th century BCE, during the Warring States Period, the Chinese thought leader Mozi condemned military aggression and sought to advance defensive technologies to reduce the incidence of war (Luo and Twiss, 2015, pp. 226–227). In the 1960s, interest in the societal impacts of technologies gave rise to the field of technology assessment, which sought to analyse the short and long-term consequences of the applications of specific new technologies (Banta, 2009). A crucial goal of technology assessment was to shape natural science research and engineering efforts based on broader societal inputs (Guston and Sarewitz, 2002). For instance, starting in the 1980s, the Danish Board of Technology held consensus conferences to consult the public on technology developments (Fisher et al., 2006).

The majority of efforts in the 1990s and 2000s focused on assessing and shaping individual technologies as opposed to a portfolio approach. One such effort, known as value-sensitive design, creates a system for considering human values for a given technological design process (Friedman et al., 2013). Technology assessment and the study of ethical, legal, and social aspects (ELSA) of science and technology feature a similar focus on shaping individual technologies (Schot and Rip, 1997; Zwart et al., 2014). This focus on

individual technologies may have been caused by relevant funding often having been linked to large investments into specific technologies. For instance, the United States National Institutes of Health became the largest bioethics funder when deciding to investigate the ethical, legal, and social implications of the Human Genome Project in 1990 (Zwart et al., 2014). Similarly, the United States National Nanotechnology Initiative reinvigorated the field of technology assessment in the early 2000s (Guston, 2014).

In the 2010s, the broader concept of responsible research and innovation (RRI) began to receive substantial attention and became the centrepiece of European science and innovation policies (Owen et al., 2012). Through responsible innovation, policymakers and academics seek to increase the extent to which research and innovation produce beneficial outcomes for society. Thus, the framing and ambition of RRI has moved beyond the shaping of individual technologies and towards questioning how to prioritise the development of different technologies (Zwart et al., 2014). We build on this work by proposing a general principle that considers how the relative timing of future technologies and their interactions shape their impact.

Climate change mitigation has driven tangible efforts to go beyond the shaping of individual technologies and leverage interactions across the technology portfolio to improve societal and environmental outcomes. The main goal of these efforts has been the preferential advancement of low-carbon technologies and the substitution of fossil fuels (Farmer et al., 2019). To this end, interventions such as carbon pricing and feed-in tariffs for photovoltaics have been used (Green, 2021; Haegel et al., 2017). Acemoglu developed an economic theory of directed technical change, (Acemoglu, 2002) which has been applied to the decarbonisation of the economy (Aghion et al., 2016). In this paper, we generalise these lessons from climate change to argue for a broader principle of differential technology development.

2. Defining differential technology development

2.1 Definition

Technologies can have negative impacts: combustion engines harm the environment and human health, printers allow the counterfeiting of money, and nuclear weapons have created the possibility of widespread destruction. However, other technologies, ranging from electric cars to locks preventing the unauthorised use of nuclear weapons, may reduce these risks. Thus, interactions between technologies are a crucial lever for mitigating negative impacts of future technologies. Advancing risk-reducing technologies before or soon after the advent of risk-increasing technologies would improve societal outcomes. We propose a responsible innovation principle that explicitly leverages these risk-reducing interactions through the relative timing of different technologies. We call this principle *differential technology development*.

> *Differential technology development (DTD):* Leverage risk-reducing interactions between technologies by affecting their relative timing.

This principle calls on relevant actors to preferentially advance risk-reducing technologies and delay risk-increasing technologies. The concept of differential technology development was first articulated by Nick Bostrom (Bostrom, 2014). Bostrom introduces "differential technological development" to explore how the timing of advanced artificial intelligence relative to other technologies may impact associated risks. We propose an application of this principle within the context of responsible innovation.

The principle of differential technology development is characterised by its focus on the interactions between technologies. Responsible innovation already provides frameworks to consider and shape the impacts of particular technologies and projects. Differential technology development provides an additional framework

to consider how to prioritise a portfolio of technology development. For example, government research funding agencies could adopt a principle of differential technology development to guide their overall grantmaking strategy and prioritise between different innovation objectives.

Differential technology development does not require perfect prediction of technology impacts. The same approaches applied to the shaping of individual technologies may be used to consider a portfolio of technologies and their interactions. This may sound intractably complex, but in practice, certain technology interactions may be simple. Printing technologies and anti-counterfeiting technologies, cars and seatbelts, and anticoagulants and anticoagulant reversal agents all share simple interactions that dictate risk-reducing effects. The application of differential technology development simply requires these interactions to be considered when deciding on innovation priorities.

Differential technological development is still possible in the early or even late stages of diffusion of a technology. The risk of global warming was not appreciated until long after the diffusion of fossil fuel technologies, but interventions to speed the adoption of clean energy technologies have nonetheless been central to mitigating harms. Adopting a principle of differential technology development could incentivise actors to identify relevant interactions between technologies sooner and act more rapidly on opportunities to positively shape outcomes.

2.2 Application to catastrophic risk mitigation

Catastrophic risk mitigation is particularly amenable to applying the principle of differential technology development because there is often societal consensus around certain goals. Zwart *et al.* argue that diverging opinions on what beneficial outcomes for society look like have hindered responsible research and innovation efforts (Zwart et al., 2014). For instance, public opinion is split on whether to decentralise the monetary system through cryptocurrencies. In contrast, opinions diverge less on the need to reduce negative impacts like greenhouse gas emissions. Thus, mitigating negative impacts to improve societal outcomes may be more actionable than responsible innovation efforts to shape society in a particular direction.

It is already well understood that preventing catastrophic risks from future technologies should be an important priority for responsible innovation efforts. As the power of technologies increases, they can pose catastrophic risks. Advances in synthetic biology and virology create a risk for engineered pandemics much worse than COVID-19 (Schoch-Spana et al., 2017). While artificial intelligence promises benefits across many aspects of society, artificial intelligence-enabled disinformation attacks are already being used to sow disinformation and destabilise democracies and pose risks of destabilising military balances of power (Horowitz, 2018; National Security Commission on Artificial Intelligence, 2021). This paper explores the potential of differential technology development to reduce such risks.

3. Risk-reducing interactions between technologies

In this section, we discuss risk-increasing and risk-reducing technologies that help illustrate the importance of the sequence that different technologies are developed for mitigating risks. These categories serve as a starting point to illustrate opportunities and practical applications of differential technology development.

3.1 Risk-increasing technologies:

“Risk-increasing technologies” may have negative societal impacts by causing insidious harm or through their potential to cause a catastrophe. High-carbon emission technologies insidiously drive global warming and cause harmful air pollution, while the development of nuclear weapons has created the threat of nuclear war and subsequent nuclear winter.

Some risk-increasing technologies, such as biological weapons, are purely offensive and have no civilian use. To mitigate harm from these risk-increasing technologies, we invest in defences such as vaccines (Riedel, 2005). Most risk-increasing technologies feature upsides that drive their development and adoption despite possible negative societal impacts. Certain technologies may, for instance, be "dual-use", featuring both beneficial and harmful applications. This dual-use dilemma is particularly pronounced for synthetic biology (Atlas and Dando, 2006) and artificial intelligence (Brundage et al., 2018) but also applies to many other technologies. For instance, additive manufacturing has many beneficial applications but also facilitates building nuclear weapons (Volpe, 2019).

Other technologies may be able to reduce the risks posed by these risk-increasing, frequently dual-use technologies or achieve the same benefits without creating risks.

3.2 Risk-reducing technologies

3.2.1 Safety technologies

"Safety technologies" reduce or prevent negative societal impacts by modifying risk-increasing technologies. For instance, carbon capture and sequestration may reduce the emissions of coal-fired power plants (Chu, 2009).

For dual-use technologies, safety technologies may reduce the possibility of accidental or deliberate misuse. The development of electronic locks for nuclear weapons, permissive action links (PALs), in the 1960s has reduced the risk of accidental or unauthorised launches (Caldwell, 1987). OpenAI has demonstrated how security-minded user interfaces, referred to as application programming interfaces (APIs), can prevent the misuse of general-purpose machine learning models (Brockman et al., 2020). DNA synthesis screening technology may prevent the harmful application of DNA synthesis machines to generate genetic materials for the illicit creation of pathogens (Esvelt, 2018).

Safety technologies may also be cooperation-monitoring or governance-enabling technologies, such as mechanisms for the verification of the existence of artificial intelligence systems ("AI Verification," n.d.) or tools that enable actors to make trustworthy claims (Brundage et al., 2018).

Minimising the time between developing a risk-increasing technology and a relevant safety technology can mitigate expected societal impacts (Figure 1a). For instance, developing PALs for nuclear weapons as part of the Manhattan Project instead of two decades later would have reduced the window of vulnerability for catastrophic misuse over this period.

3.2.2 Defensive technologies

"Defensive technologies" decrease risks from risk-increasing technologies without modifying these technologies. 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 (Sandbrink and Shattock, 2020). mRNA vaccines and similar vaccine platforms will be crucial to reduce the societal impact of future pandemics of any origin - whether natural or caused by the accidental or deliberate misuse of synthetic biology.

Advancing defensive technologies before relevant risk-increasing technologies prevents a window of vulnerability and thus leads to better expected societal outcomes (Figure 1b). Concretely, if government research funding agencies would prioritise 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.

3.2.3 Substitute and low-risk alternative technologies

"Substitute technologies" achieve similar benefits as a risk-increasing technology while featuring less risk. One prominent example are clean energy technologies, like wind turbines or photovoltaics, which can replace environmentally-harmful fossil fuels. Another example was the development of substitute technologies for the phase-out of ozone-depleting substances, estimated to have otherwise caused two million additional cases of skin cancer each year in 2030 (McKenzie et al., 2019; van Dijk et al., 2013).

There may also be substitutes for dual-use technologies that offer the same benefits with less potential for misuse. For instance, instead of learning to engineer and fine-tune viruses for the delivery of vaccines and therapeutics, non-viral delivery methods could be advanced (Sandbrink et al., 2021). Where viral delivery methods feature substantial advantages, such as for specific gene therapy applications, non-heritable methods of viral modification should be preferentially investigated (Sandbrink and Koblentz, 2022). These non-heritable methods may provide the same benefits but cannot be used to effectively enhance transmissible viruses.

Preferential investigation of low-risk technologies to solve a given challenge would lead to less risk across our technology portfolio (Figure 1c). When risk-increasing technologies are already in use, the advancement of substitute technologies may help to reduce negative societal impacts.

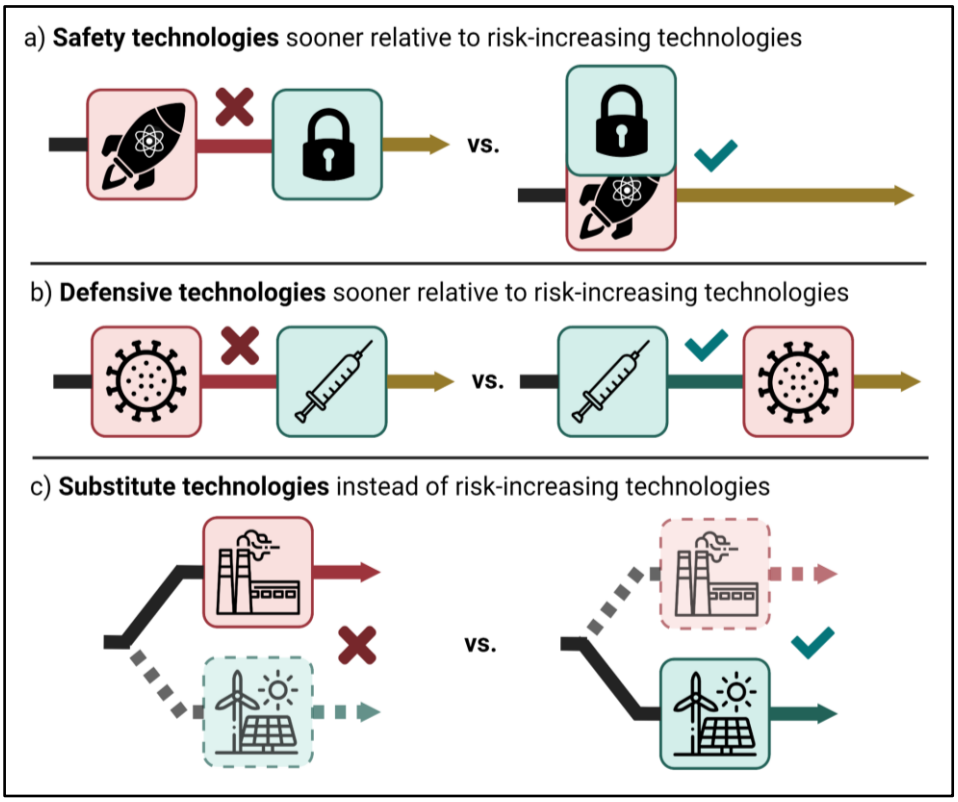

Figure 1: Mechanisms by which differential technology development can reduce negative societal impacts

a) Developing safety technologies with or soon after risk-increasing technologies can reduce negative societal impacts; e.g. electronic locks can prevent the unauthorised use of nuclear weapons. b) Developing defensive technologies before or soon after risk-increasing technologies reduces a window of vulnerability; e.g. society should first develop ways to prevent a pandemic from a particular virus, for instance through vaccinating the population, before disseminating its

blueprint for its synthesis for biotechnology applications. c) Low-risk substitute technologies can replace risk-increasing technologies or make their development unnecessary; e.g. renewable energy sources are replacing the environmentally harmful use of coal for energy production.

3.3 Shaping individual technologies

Differential technology development focuses on leveraging interactions between different technologies, but it is also compatible with related efforts to shape individual technologies. Indeed, attempts to implement differential technology development may often benefit from this approach. The shaping of individual technologies may be most important for risk-increasing technologies and technologies that do not fall cleanly into risk-increasing or risk-decreasing categories. This is especially true for general purpose and platform technologies (Bresnahan and Trajtenberg, 1995), which may have both defensive and offensive applications. Some risk-reducing technologies may feature dual-use potential themselves (Sandbrink and Koblentz, 2022). For instance, metagenomic sequencing-based pathogen detection will likely be a crucial defensive technology for preventing pandemics (Consortium, 2021). However, if developed in a way that does not prevent privacy infringements, genomic sequencing may be misused for identifying and tracking individuals or ethnic populations. Therefore, governments should work with technology developers to advance risk-reducing technologies in the most defence-biased manner possible.

Responsible innovation scholars have proposed various strategies for shaping individual technologies through technology assessment, eliciting societal inputs, and concrete interventions (Friedman et al., 2013; Guston and Sarewitz, 2002; Schot and Rip, 1997; Stilgoe et al., 2013). One mechanism applicable to both risk-increasing and risk-reducing technology review is stage-gating. Stage-gating breaks up a larger project into a series of steps with associated decision points (Stilgoe et al., 2013). Thus, stage-gating may help identify risks and opportunities for safety technologies or other forms of interventions to enhance defence bias. Stage-gating was applied to the UK geoengineering SPICE experiment at different stages of preparation, leading initially to its postponement and later its cancellation in 2011 (Stilgoe et al., 2013). Methods and practices similar to stage-gating need to be advanced to assess the dual-use potential of individual technologies and shape their development.

4. Anticipating negative societal impacts

For the principle of differential technology development to be useful, it must be viable in practice. Successfully engaging in differential technology development is a complex task: a government, philanthropic entity, or firm has to anticipate or observe the early impacts of multiple technologies and then steer down one of the better paths. This requires the ability to: 1) anticipate or observe how technologies might interact to produce risks, and 2) intervene in the relative timing of technology development. The possibility of anticipating technology impacts is the subject of this section, while the question of intervention is the subject of section 5.

Collingridge's dilemma of control sums up the challenge of anticipating the impacts of technologies: "When change is easy, the need for it cannot be foreseen; when the need for change is apparent, change has become expensive, difficult, and time-consuming." (Collingridge, 1980). However, the history of technological development suggests that this dilemma is not insurmountable. The following sections outline the conditions under which anticipation of the impact of technologies may be possible.

4.1 When is anticipation possible?

One of the most important historical examples of technology anticipation is that of Leo Szilard (Lanouette, 1992). After discovering the possibility of a nuclear chain reaction in 1933, Szilard anticipated the possibility of a nuclear bomb. Initially, Szilard kept this knowledge hidden. However, in 1939 Szilard's fear of Germany

developing nuclear weapons drove him to write about the possibility of the atomic bomb to United States President Franklin D. Roosevelt. This led to the commencement of the Manhattan Project two years later. In the wake of the German surrender, Szilard unsuccessfully worked to prevent the demonstration of the bomb in Japan and the subsequent nuclear arms race. Szilard's story highlights how security-minded researchers may anticipate the negative societal impacts of novel technologies.

However, also without being at the forefront of scientific discovery, actionable anticipation of societal impacts of technologies may be possible. Actionable anticipation may generally be easier: 1) when risks are linked to defining features, 2) when technology improvements are incremental, or 3) the slower or more controlled technology diffusion is.

4.1.1 Risks linked to defining features

While Szilard could not be confident that a bomb based on the nuclear chain reaction would be feasible, he could predict that if such as weapon was produced, it would be far more energy dense than conventional weapons and thus capable of producing larger explosions. Thus, negative societal impacts of a technology can be anticipated when they are tightly linked to a defining feature of the technology - even if it is radically novel. For example, while we may not know the exact mechanism by which a novel anticoagulant medication will function, we can reliably anticipate that it will increase bleeding risk. Similarly, reviewers of funding proposals can use the goals of scientific experiments to identify associated risks. For instance, gain-of-function experiments to enhance the transmissibility of potential pandemic pathogens can be anticipated to create risks of accidents and misuse (Duprex et al., 2015). As biosecurity risks are commonly linked to defining technology features, the WHO was able to map emerging biotechnologies with dual-use potential (World Health Organisation, 2021).

Attempts of modular innovation may have relatively predictable impacts because they target an established dominant design and simply change one of its components. For example, the development of DNA benchtop synthesis machines will predictably conserve the core DNA synthesis function of industrial DNA synthesisers but increase distributed accessibility and portability. Distributed access to synthetic DNA features significant upsides for life sciences research but can also be anticipated to provide challenges for preventing the illicit synthesis of pathogens.

4.1.2 Incremental improvements

Based on experiences with precursor technologies, the impacts of incrementally improved technologies may be relatively straightforward to assess. The development of increasingly powerful natural language processing machine learning models can be expected to translate into growing usefulness for scams or disinformation campaigns (Caldwell et al., 2020; Seger et al., n.d.).

For incremental technology improvements, it is often possible to observe trends for how improvements affect functionality. The most common example is Moore's Law, which accurately predicted that the number of transistors on a microprocessor chip would double every two years since the 1960s (Schaller, 1997). Moore's Law has informed industrial strategy in the semiconductor industry since the 1990s (Waldrop, 2016). DNA sequencing costs have decreased dramatically, at a greater rate than Moore's Law. While sequencing a whole human genome cost $10 million USD in 2007, in 2021, it was less than $1,000 USD (Wetterstrand, n.d.). With the fall in sequencing costs, we can anticipate the use of sequencing for personalised medicine (Brittain et al., 2017) but also increasing accessibility by less well-resourced actors to use for tracking and identifying ethnic groups or individuals.

4.1.3 Before diffusion of a technology

The window between the invention and widespread adoption of a novel technology may constitute a sensitive intervention point (Farmer et al., 2019) for mitigating the negative societal impacts of novel technologies. Such an intervention may include advancing complementary risk-reducing technologies, particularly safety technologies. For instance, while it would have been useful to have permissive action links (PALs) with the very first nuclear weapon in 1945, it would have been almost as good to develop them before the manufacturing of additional bombs or their proliferation to other countries. This would have been possible given a single actor, the United States Government, initially developed the technology, thus controlling initial diffusion. The United States could have shared permissive action link technology as soon as other countries attained nuclear bombs. Indeed, twenty years later, once it had developed PALs, the United States Government deliberately leaked related information to the Soviet Union (Nye, 1987).

Even uncontrolled diffusion of novel technologies often takes decades and thus offers a window of opportunity to assess impacts and intervene. Potential negative impacts of technologies often correlate with their diffusion and can thus also take a long time to peak. In car safety, for example, the Model T ford was available from 1908, but it was not until 1937 that road fatalities per capita peaked (Federal Highway Administration and Office of Highway Information Management, 2009). Similarly, more than half of CO2 emissions have occurred after 1988, long after technologies reliant upon fossil fuel energy sources were developed during the industrial revolution (Our World in Data, 2022). Of course, as the production and consumption of a technology increases, so does inertia in its trajectory. It becomes integrated into society and different actors become invested in its current uses and its future trajectory. Nevertheless, technology diffusion is generally a gradual process and no strict lock-ins occur at the moment of invention or commercialisation.

Thus, the window between invention and diffusion of a technology may be useful for reflecting on and studying risks, including offering an opportunity for a moratorium to assess impacts and develop safety technologies (see section 5.1.2 for discussion of moratoria). In practice, this could include deliberate investment by governments to identify early impacts of novel technologies.

4.1.4 After diffusion of a technology

Differential technology development is also still possible after the diffusion of a technology, once downside risks have been identified. Interventions to advance renewable energy sources and electric cars demonstrate the possibility of reducing negative impacts despite the existing widespread use of high emissions technologies. If governments had a specific policy and regular practice of differential technology development, they could potentially have acted more rapidly to mitigate global warming. Thus, successful anticipation of technology impacts is not necessary for differential technology development, even though it is generally beneficial.

4.2 When is anticipation difficult?

There are limits to anticipation, and certain areas where anticipating societal impacts of technologies may be more challenging.

4.2.1 Higher-order effects

Firstly, higher-order effects can be difficult to predict. This includes the second-order effects of risk-reducing technologies. For instance, the risk-reducing effects of safety technologies depend on their adaptation and related social factors. Even after the development of PALs to prevent the unauthorised launch of nuclear weapons, the US Air Force kept these electronic keys set to only zeros for decades (Ellsberg, 2017).This practice was driven by the military's desire to retain decentralised retaliatory capacity if cut off from central

command. Similarly, developing defensive technologies may have second-order risk-increasing effects in adversarial situations. For instance, the development of a US anti-ballistic missile system was considered dangerous because it might upset the careful balance of mutual deterrence, potentially sparking incentives for a pre-emptive strike or inducing additional investments into nuclear delivery systems (Maas, 2019).

4.2.2 General-purpose technologies

Secondly, the effects of general-purpose technologies that intersect with a wide range of other technologies can be difficult. This is true for the impact of technologies like the computer, internet, or artificial intelligence technologies, which continue to transform human civilisation. However, even in the face of great uncertainty around the effects of general-purpose technology intersections, differential technology development may still be employed for certain applications of such general-purpose technologies. For instance, even if artificial intelligence's future transformative effects are not yet predictable, we can still deploy application programming interfaces to safeguard language models from misuse (Brockman et al., 2020).

4.2.3 Breakthrough discoveries

Lastly, breakthrough discoveries from basic science research are difficult to anticipate. However, as the story of Szilard demonstrates, security-minded researchers have the potential to reflect on how to share their insights responsibly. Furthermore, depending on what basic science research is conducted, it might be possible to predict what insights may be found. For instance, screening novel bacteria for functional components might identify novel tools for molecular biology, similar to how the discovery of thermostable bacteria led to the 1983 invention of Polymerase Chain Reaction (PCR) (Dove, 2018).

4.3 Overcoming costs of anticipation

Although there are many examples where predicting societal impacts of technology has been helpful, anticipation may have costs. Predictions may be incorrect, thus providing false warnings or inspiring the advancement of technologies incorrectly judged to feature little potential for misuse. Furthermore, widespread practices of anticipation and associated consideration of risks might slow the adoption of beneficial novel technologies. These costs are real and merit substantial weight in policy decisions. Nevertheless, the upsides of anticipation and related differential technology development practices likely outweigh the costs when managing extreme risks, given the potential harms involved. Further investigation may find other goals where the costs of anticipation are worth paying.

The cost of anticipation may decrease as it is done more systematically and better methodologies and tools are developed. Multiple methods for anticipating impacts have been proposed: Analogical case studies, scenario exercises, research program mapping, identifying possible negative impacts of technologies and initial products, and researcher interviews (Guston and Sarewitz, 2002). The IARPA FUSE program advances new methodologies for technology foresight and assessment ("IARPA - FUSE," n.d.). Recent work by Zhou et al. 2020 used a deep neural network classifier to forecast emerging technologies. It was able to predict which patented innovations would become classified as emerging technologies with an accuracy of 77%, one year before these technologies were recognised in Gartner's annual graphical summary of emerging technology (Zhou et al., 2020). More generally, in the last decade, novel forecasting techniques have been developed that can help to bound uncertainty and improve insight into the future (Tetlock and Gardner, 2016). Similar advances for predicting possible societal impacts of novel technologies seem achievable.

5. Intervening in the timing of technology development

To apply a principle of differential technology development, actors such as government research funding agencies, government regulatory agencies, technology companies, and philanthropic funders need to be able to intervene in the pace of development of specific technologies. To leverage risk-reducing interactions between technologies, interventions can aim to slow risk-increasing technologies or speed the development of risk-reducing technologies. The following sections discuss how governments and other relevant actors can use different intervention strategies to implement the principle of differential technology development. These sections are not intended as a comprehensive assessment of potential strategies but as an initial indication of the diverse set of potentially promising options for implementation.

5.1 Strategies for delaying risk-increasing technologies

Governments and other actors can sometimes delay the development of specific risk-increasing technologies. Delaying technologies may seem infeasible due to coordination issues or too controversial if it involves slowing down increases in wealth and health. However, this is not necessarily the case. Policy decisions delay the development of technologies all the time. With every decision to advance one technology, resources are diverted from an alternative that could have been developed. This is most apparent in academic funding, where funding is awarded to a small subset of grant proposals.

It can be difficult to see the missing technologies, the paths not taken. For example, policy decisions have led to less innovation in technologies for geoengineering and human genetic engineering but an alternate decision could have led in a very different direction (Stilgoe et al., 2013; Sykora and Caplan, 2017).

As resources can frequently be channelled into alternative technologies, delaying select risk-increasing technologies does not necessarily slow down innovation (Mahdi et al., 2002) - and importantly, this practice can be expected to reduce civilisational risks and thus safeguard innovation in the long term. Concrete strategies to delay risk-increasing technologies until their risks have been assessed and managed are presented below and summarised in Table 1.

5.1.1 Restraint and defunding

Restraint in developing or defunding specific technologies can be a straightforward approach to slowing their development, but can also be complicated by interactions with other actors. Funding for research and development generally leads to more rapid development of related technologies, so reduced funding can directly impact the pace of development (Bolívar-Ramos, 2017). The situation is often complicated by the fact that other developers may continue their efforts, and other actors may step in to fill gaps. For example, former US President Bush restricted the federal funding of stem cell research in 2001 (Taylor, 2005). However, in 2004 the California state government stepped in to fund stem cell research while this research also continued in other countries. Nevertheless, given its significant role in science funding, limited United States federal funding of stem cell research likely still slowed its development. In situations where development and research happen in a distributed manner and are not driven by a single actor, slowing the development of a technology may be possible unilaterally or may require substantial coordination across multiple stakeholders.

5.1.2 Moratoria to assess and address risks

Moratoria of possible risk-increasing technologies may be used to assess risks and advance risk-mitigating measures. In 1974, the international community of researchers working with recombinant DNA engaged in a voluntary moratorium and gathered for the Asilomar conference to evaluate possible public health risks (Berg and Singer, 1995). After the conference, researchers resumed recombinant DNA work with agreed safety measures. Next to such voluntary moratoria, governments and funding bodies can also impose

moratoria on certain research. The United States National Institutes of Health's 2014-2017 moratorium on viral gain-of-function research delayed high-risk research until a new oversight policy was created (Reardon, 2017). Moratoria might be especially useful to delay the development of risk-increasing technologies until safety technologies are in place. One example application might be an industry-led moratorium on the distribution of benchtop DNA synthesis machines until effective DNA synthesis screening has been integrated into all devices. This could prevent the irreversible proliferation of devices that allow illicit access to the materials for the creation of any pathogen.

5.1.3 Bans

In certain cases, banning a technology or its development may be feasible. Bans at the international level require substantial coordination and political will; thus, they are only feasible for technologies that are seen as overwhelmingly negative and not critical for national interests. The Chemical Weapons Convention, Biological Weapons Convention, and Nuclear Non-Proliferation Treaty are examples of international treaties that aim to prevent the development and proliferation of certain military technologies. While enforcement is challenging, each of these treaties has slowed the proliferation of these technologies and has also likely slowed the development of more advanced chemical and biological weapons (Fuhrmann and Lupu, 2016; "The power of treaties," 2013; Wheelis, 2006).

Another example is the Montreal Protocol of 1987, which successfully banned ozone-depleting substances. Global consumption of these substances reduced by 98.5% between 1986 and 2018 (European Environment Agency, 2021). While the ban focused on the production of ozone-depleting substances, it also removed the incentive to continue developing technologies reliant upon ozone-depleting substances and produced a new incentive to develop substitute technologies.

5.1.4 Shifting the technology portfolio: regulation, norms, and information loops

Incentives may be used to encourage a shift away from developing risk-increasing technologies. Governments can use regulatory interventions this way, described by Rip and Kemp as "weed pulling" (Rip and Kemp, 1998). An example is carbon taxation. Countries with higher fuel prices (a proxy for greenhouse gas emissions tax level) have lower levels of innovation in high-emissions technologies and higher levels of innovation in green technologies (Aghion et al., 2016). Similarly, paperwork-heavy regulations may disincentivise relevant research, as is the case for research on dangerous pathogens in the United States (Evans et al., 2021).

Social norms and pressure may be sufficient to generate incentives to prevent or move away from risk-increasing efforts. Strong social norms against the recreation of dangerous smallpox led to a strong outcry over synthesising the related horsepox virus, despite this not being banned by existing regulations. Climate activism has driven divestment from fossil fuels. While the direct effects of fossil fuel divestment are small, indirect impacts on societal discourse and norms have been substantial (Bergman, 2018).

Indeed, simply awareness of risks may reduce the development of risk-increasing technologies. To highlight the power of such information loops, Donella Meadows tells the story of the United States Toxic Release Inventory (Meadows, 1999). When introduced in 1986, this regulation forced companies to report the release of pollutants publicly. After four years, emissions had dropped by 40% - without any fines, bans, or other interventions (Meadows, 1999). Information loops may similarly be used to encourage the pursuit of less risky lines of life sciences research. Consideration of Replacement, Reduction, and Refinement of animal use in experiments ("the 3Rs") is required in many grant applications to ensure researchers are minimising animal harm. A similar strategy could be deployed for the safety and security risks of life sciences research. If all proposed research projects were categorised by risk level at the funding stage, this might lead to less risky research being funded and conducted. In the absence of a superior promise of one project over another,

researchers and funders should choose the least risky avenue of research. Only if all highly promising low-risk avenues are saturated should researchers and funders engage in more risky projects. Importantly, creating information loops does not rely on government mandates. Non-governmental actors like companies or non-profits may create information loops based on publicly available data.

Table 1: Strategies to delay and assess risk-increasing technologies and research

Strategy	Actors	Example	Reference
Defunding	Government Industry Philanthropy	Limitations in government funding impeded stem cell research	(Taylor, 2005)
Moratoria	Academia Industry Government	1974 voluntary moratorium on recombinant DNA research	(Herzog and Parson, 2016)
Stage-gating	Government Industry Academia NGOs	UK geoengineering SPICE experiment was subject to stage-gated review; first postponed and then cancelled.	(Stilgoe et al., 2013)
Bans	International organisations Government	Montreal protocol reduced production of ozone-depleting substances	(European Environment Agency, 2021), (Fuhrmann and Lupu, 2016)
Regulations	Government	Less innovation in high-emissions technologies through taxation of fossil fuels	(Aghion et al., 2016)
Social norms	Academia NGOs Activists	Social norms against the recreation of eradicated smallpox virus	(Kupferschmidt, 2017)
Advocacy	NGOs Academia Industry	Letter signed by 50 NGOs led to the termination of UK SPICE Geoengineering project	(Stilgoe et al., 2013)
Divestment	NGOs Industry	Fossil fuel divestment has had a significant impact on discourse and norms.	(Bergman, 2018)
Information loops	Government NGOs Industry	Toxic Release Inventory 1986 requiring public reporting of pollutants substantially reduced emissions	(Meadows, 1999)

5.2 Strategies for advancing the development of risk-reducing technologies

Governments, firms, philanthropic funders and other relevant actors are frequently able to advance the development of risk-reducing technologies. Increasing funding or other resources provides a direct pathway

toward advancing a specific technology. However, other strategies may also be effective, such as regulation (Table 2). Advancement of risk-reducing technologies may require less coordination than delaying risk-increasing technologies and may be a viable approach for a range of actors.

5.2.1 Funding, prizes, advance purchase commitments

Governments or other technology funders can accelerate specific technologies by prioritising the funding of related research and development. For example, when firms and universities increase their research spending, this increases their patent output (Bolívar-Ramos, 2017; Ernst, 1998).

Specific strategies may be particularly suitable to advance certain technologies. The United States Defense Advanced Research Projects Agency (DARPA) has successfully developed specific, application-ready technologies from basic research. The agency is credited with key advances in the creation of the internet, synthetic biology and carbon nanotubes (Mervis, 2016). Focused research organisations (FROs) may be particularly well-suited to develop risk-reducing technologies that constitute public goods (Marblestone et al., 2022).

Prizes may be another tool to encourage the development of a specific risk-reducing technology, especially one relying on interdisciplinary research or with unclear solutions. Genetic engineering attribution, the ability to computationally identify laboratory origins of engineered DNA sequences, is a defensive technology that may deter the release of a biological weapon. In 2020, a prize, the Genetic Engineering Attribution Challenge, induced the development of new tools that significantly outperformed state-of-the-art approaches (Crook et al., 2021).

Advance market commitments may be a useful tool for the development of crucial defensive technologies (Monrad et al., 2021). There may be little incentive to develop vaccines or personal protective equipment in the absence of an ongoing emergency. Nevertheless, development and stockpiling are essential for pandemic preparedness. In such cases, commitments to purchase a select number of units or guarantee a market for a product may be used to induce advances. For instance, advanced market commitments have been used to encourage the development of pneumococcal vaccines (Cernuschi et al., 2011).

5.2.2 Regulations can advance risk-reducing technologies

“Regulation is the mother of invention”, as Ruth Ruttenberg noted (Rip and Kemp, 1998). Governments may use regulations to force the development of risk-reducing technologies. This is frequently the case for safety technologies. For instance, regulations on how to handle pathogens reduce research risks through laboratory safety technologies (Pastorino et al., 2017). Similarly, governments could induce the development and adoption of security-sensitive application programming interfaces by mandating their use for machine learning models with significant potential for misuse.

Strategies for advancing substitute technologies have been extensively studied and explored for clean energy and low emissions technologies. A combination of strategies to increase the momentum of niche innovations, weaken existing systems, and strengthen exogenous pressures may help to replace risk-increasing technologies (Geels et al., 2017). Regulations of risk-increasing technologies discussed in 5.1.4, including carbon taxation, incentivise the advancement of substitute technologies (Aghion et al., 2016).

Governments may induce the development of specific technologies through technology-forcing. Clean air standards passed in California in 1988 demanded that 2% of car sales must be zero-emission vehicles (Schot and Rip, 1997). These vehicles did not exist at the time. Thus, this regulation likely contributed to the advancement of electric vehicles.

Niche management, such as feed-in tariffs that enable the competitiveness of substitutes before their commercial viability (Schot and Geels, 2008). For example, subsidies have successfully helped advance photovoltaics (Haegel et al., 2017). Tax incentives for carbon sequestration have likely induced advances in carbon capture and sequestration technologies (Anderson et al., 2021).

5.2.3 Coordination and pre-competitive consortia

Industry coordination may advance risk-reducing technologies and practices. Basic DNA synthesis screening of 80% of the market exists because of voluntary coordination in the International Gene Synthesis Consortium (Diggans and Leproust, 2019; International Gene Synthesis Consortium, 2017). Pre-competitive consortia are a promising method for facilitating cross-sector coordination in biomedicine (Mittleman et al., 2013). Such pre-competitive consortia may be used to develop risk-reducing technologies.

Table 2: Strategies for the preferential advancement of risk-reducing technologies

Strategy	**Actor**	**Example**	**Reference**
Funding and direct development	Government Industry Philanthropy	ARPA-style efforts lead to creation of internet, synthetic biology, carbon nanotubes, clean energy technologies	(Mervis, 2016)
Prizes	Government Industry Philanthropy	Genetic engineering attribution challenge winning entries significantly advanced state-of-the art capabilities	(Crook et al., 2021)
Advanced market commitments	Government Industry Philanthropy	Advance market commitments have been used to induce development of pneumococcal vaccines	(Cernuschi et al., 2011)
Technology forcing	Government	Clean air standards in California in 1988 requiring 2% of car sales to be zero-emission vehicles	(Schot and Rip, 1997)
Niche management	Government	Feed-in tariffs and subsidies for photovoltaics until commercial viability	(Haegel et al., 2017)
Tax incentives	Government	Tax incentives for carbon sequestration to advance carbon capture and storage technologies	(Anderson et al., 2021)
Regulation of risk-increasing technologies	Government	Innovation in low-emissions technologies through taxation of fossil fuels	(Aghion et al., 2016)
Coordination	Industry NGOs Government	Voluntary DNA synthesis screening of companies part of International Gene Synthesis Consortium	(Diggans and Leproust, 2019)

5.3 Challenges for timing interventions

5.3.1 Wrestling deterministic paths

Competition between groups of actors to gain a technological advantage can cause the development of technologies to proceed along more deterministic paths (Dafoe, 2015). In cases of competition to win a market or international competition, the choice of individual actors may be constrained by the competitive dynamic. For example, the United States' development of nuclear weapons was prompted by fears that Nazi Germany would develop them first. If progress on a specific technology is not led by a single or a handful of coordinating actors, delaying or making technologies safer may be difficult. Coordination through international organisations may have the greatest leverage to delay or shape technologies with global power implications. For other technologies, national governments may be able to break competitive dynamics across industry and academia that drive the development of risk-increasing technologies. Next to competition, various other dynamics can also lead to differing degrees of path dependence in technological development, including first-mover advantages, institutional persistence, structural inertia, limits to absorptive capacity, and natural monopolies (Vergne and Durand, 2010).

The more resources are already committed to the development of a specific technology, the more difficult having an impact on the timing of its development is. This is also true for the advancement of risk-reducing technologies. Small actors can have significant effects if no or few other actors are making similar investments. In such cases, a small investment may have a prolonged effect on the level of technology available. Historically, this has proven particularly relevant in cases with little market incentive to develop a publicly beneficial technology. An example is vaccines for neglected or possible future diseases (Monrad et al., 2021).

Interventions to promote the development of the technology can sometimes result in the technology passing a threshold of commercial viability, producing a positive feedback loop that begets further and continued development of the technology by other actors. Government support has successfully led to photovoltaics becoming commercially viable (Haegel et al., 2017; McDonald and Schrattenholzer, 2001), triggering additional investments and reductions in cost. Such feedback loops may render the replacement of marginally risk-increasing technologies cost-effective. The advancement of RNA vaccines and their success against COVID-19 may induce a shift away from viral vector vaccines, which are associated with the development of capabilities with greater potential to be misused (Sandbrink and Koblentz, 2022). Similarly, non-viral delivery methods for gene therapy may eventually become superior to viral delivery methods and thus lead to their replacement.

5.3.2 Directing change

Ensuring that interventions in the development of specific technologies have the desired effect can be difficult. While it generally appears to be possible to advance the development of risk-reducing technologies, this may inadvertently speed up related risk-increasing technologies or create interactions with other technologies that result in negative societal impacts. For example, civilian nuclear power technologies are difficult to develop in a way that does not advance nuclear weapons development.

One part of the difficulty of directing change is the challenge of anticipating higher-order effects of risk-reducing technologies. As discussed in section 4.2, anticipating the effects of general-purpose technologies may be particularly difficult. Advancing the use of metagenomic sequencing for pathogen detection may inadvertently lead to advanced sequencing applications for areas of biotechnology that increase the pandemic risks that pathogen detection hopes to reduce. Interventions to advance or delay specific technologies must be closely coupled to reviewing potential downstream risks and appropriately weigh these risks against the potential benefits..

Another part of the difficulty of directing change is predicting the effects of interventions. For instance, increasing taxation of a risk-increasing technology may inadvertently lead to alternatives with even more or different undesirable consequences. Additional funding for a risk-reducing technology may inadvertently displace other funders. These challenges, while real, are not specific to differential technology development.

5.3.3 Balancing different values

Differential technology development requires governments, philanthropists, or other relevant actors to pursue interventions in pursuit of a goal, such as the goal of mitigating the potential negative impact of a technology. However, any given goal needs to be balanced with other values or goals. Global security needs to be balanced with values of scientific freedom and openness. In certain cases, reducing catastrophic risks may be traded against short-term gains.

Differential technology development does not need to dominate every governance decision to be successfully implemented. Despite the largely universal assumption that a sustainable economy is desirable, this does not mean that adopting a goal of sustainable development necessitates every single policy decision to move in this direction. Rather, it is a cross-cutting consideration of this goal and the aggregate of many decisions that allow such goals to be implemented. Similarly, a broader ambition for differential technology development might carry humanity towards a future with less catastrophic risk. Public engagement is needed to inform how to manage the trade-offs involved in implementation. This is partially ensured through political representatives for government actors but may also call for the engagement of the broader public, including experts from different backgrounds and areas of expertise.

5.3.4 International coordination

Various strategies for implementing differential technology development benefit from the ability to coordinate between governments at an international level. However, differential technology development does not require perfect international, or even national, coordination. Just as there is no single regime to mitigate climate change but a regime complex made up of different international, national, and other societal initiatives (Keohane and Victor, 2011), a combination of initiatives may be sufficient to put differential technology development into practice. Indeed, a more fragmented system may ensure continuous adaptation to evolving circumstances and values, including consideration of local values and culture (Alter and Raustiala, 2018).

6. Conclusion

As novel technologies are becoming increasingly powerful, there is greater urgency for responsible innovation practices to mitigate associated risks. Next to the shaping of individual technologies, the interaction and relative sequence of different technologies is a crucial lever for improving societal outcomes. Thus, actors interested in improving societal outcomes, such as government research funding agencies, government regulators, philanthropic funders, and firms engaged in corporate social responsibility, should consider the principle of differential technology development when allocating resources to research and technology development.

In particular, these actors should consider implementing a principle of differential technology development to manage potential catastrophic risks from technologies. Some technologies appear potentially well suited to the application of this principle for risk reduction, including (Table 3): pandemic prevention investments, DNA synthesis screening, responsible access solutions for machine learning models, and consideration of risks when funding science. The engagement of experts from different fields would likely illuminate further opportunities to usefully apply the principle of differential technology development to mitigate risks.

Table 3: Current and future applications of differential technology development

Negative societal impacts	Intervention
Climate change from fossil fuel use	Replace fossil-fuel and high emissions technologies with low emission alternatives
Pandemic risk from misuse of increasingly powerful biotechnology	Developing better PPE, pathogen detection, vaccines, in particular before widely disseminating ability to create pandemic-capable viruses Preferential advancement of biotechnology solutions with little potential for misuse Advance universal DNA synthesis screening
Quantum computing enabling widespread decryption of sensitive information	Advance new cryptographic methods not easily circumvented through quantum computing
Misuse of artificial intelligence	Advance responsible access solutions like application programming interfaces Advance machine learning models with less potential for misuse
Misaligned artificial intelligence	Prioritise artificial intelligence safety research Create international coordination on preventing misaligned artificial intelligence

Acknowledgements

We are grateful to Michael Aird, Markus Anderljung, Jan Ole Ernst, Ben Garfinkel, Sihao Huang, Matthijs Maas, Cassidy Nelson, and James Wagstaff for useful discussions and comments on the manuscript. Furthermore, we are also grateful for feedback from participants of work-in-progress meetings of the Future of Humanity Institute and Centre for the Governance of AI. We thank Shrestha Rath for help with formatting and organising references. Jonas B. Sandbrink's doctoral research is funded by Open Philanthropy. Hamish Hobbs' contribution to the paper largely occurred while funded as a Research Scholar at the Future of Humanity Institute.

Declaration of competing interest

The authors have no conflicts of interest to declare.

CRediT authorship contribution statement

Jonas B. Sandbrink: Conceptualization, Investigation, Writing - original draft, Writing - review & editing. **Hamish Hobbs**: Conceptualization, Investigation, Writing - original draft, Writing - review & editing. **Jacob L. Swett**: Conceptualization, Writing - review & editing. **Allan Dafoe:** Conceptualization, Writing - review & editing. **Anders Sandberg**: Conceptualization, Writing - review & editing.

Bibliography

Acemoglu, D., 2002. Directed Technical Change. The Review of Economic Studies 69, 781–809. https://doi.org/10.1111/1467-937X.00226

Aghion, P., Dechezleprêtre, A., Hémous, D., Martin, R., Van Reenen, J., 2016. Carbon Taxes, Path Dependency, and Directed Technical Change: Evidence from the Auto Industry. Journal of Political Economy 124, 1–51. https://doi.org/10.1086/684581

AI Verification, n.d. . Center for Security and Emerging Technology. URL https://cset.georgetown.edu/publication/ai-verification/ (accessed 5.3.22).

Alter, K.J., Raustiala, K., 2018. The Rise of International Regime Complexity. Annual Review of Law and Social Science 14, 329–349. https://doi.org/10.1146/annurev-lawsocsci-101317-030830

Anderson, J.J., Rode, D., Zhai, H., Fischbeck, P., 2021. A techno-economic assessment of carbon-sequestration tax incentives in the U.S. power sector. International Journal of Greenhouse Gas Control 111, 103450. https://doi.org/10.1016/j.ijggc.2021.103450

Atlas, R.M., Dando, M., 2006. The dual-use dilemma for the life sciences: perspectives, conundrums, and global solutions. Biosecur Bioterror 4, 276–286. https://doi.org/10.1089/bsp.2006.4.276

Banta, D., 2009. What is technology assessment? International Journal of Technology Assessment in Health Care 25, 7–9. https://doi.org/10.1017/S0266462309090333

Berg, P., Singer, M., 1995. The Recombinant DNA Controversy: Twenty Years Later. Nature Biotechnology 13, 1132–1134.

Bergman, N., 2018. Impacts of the Fossil Fuel Divestment Movement: Effects on Finance, Policy and Public Discourse. Sustainability 10, 2529. https://doi.org/10.3390/su10072529

Bolívar-Ramos, M.T., 2017. The relation between R&D spending and patents: The moderating effect of collaboration networks. Journal of Engineering and Technology Management 46, 26–38. https://doi.org/10.1016/j.jengtecman.2017.11.001

Bostrom, N., 2014. Superintelligence: Paths, Dangers, Strategies, Illustrated edition. ed. OUP Oxford, Oxford.

Bresnahan, T.F., Trajtenberg, M., 1995. General purpose technologies 'Engines of growth'? Journal of Econometrics 65, 83–108. https://doi.org/10.1016/0304-4076(94)01598-T

Brittain, H.K., Scott, R., Thomas, E., 2017. The rise of the genome and personalised medicine. Clin Med (Lond) 17, 545–551. https://doi.org/10.7861/clinmedicine.17-6-545

Brockman, G., Murati, M., Welinder, P., OpenAI, 2020. OpenAI API. OpenAI API. URL https://openai.com/blog/openai-api/

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G.C., Steinhardt, J., Flynn, C., hÉigeartaigh, S.Ó., Beard, S., Belfield, H., Farquhar, S., Lyle, C., Crootof, R., Evans, O., Page, M., Bryson, J., Yampolskiy, R., Amodei, D., 2018. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation (No. arXiv:1802.07228). arXiv. https://doi.org/10.48550/arXiv.1802.07228

Caldwell, D., 1987. Permissive action links: A description and proposal. Survival 29, 224–238. https://doi.org/10.1080/00396338708442358

Caldwell, M., Andrews, J.T.A., Tanay, T., Griffin, L.D., 2020. AI-enabled future crime. Crime Science 9, 14. https://doi.org/10.1186/s40163-020-00123-8

Cernuschi, T., Furrer, E., Schwalbe, N., Jones, A., Berndt, E.R., McAdams, S., 2011. Advance market

commitment for pneumococcal vaccines: putting theory into practice. Bull World Health Organ 89, 913–918. https://doi.org/10.2471/BLT.11.087700

Chu, S., 2009. Carbon Capture and Sequestration. Science 325, 1599–1599. https://doi.org/10.1126/science.1181637

Coad, A., Nightingale, P., Stilgoe, J., Vezzani, A., 2021. Editorial: the dark side of innovation. Industry and Innovation 28, 102–112. https://doi.org/10.1080/13662716.2020.1818555

Collingridge, D., 1980. The social control of technology. Frances Pinter, London.

Consortium, T.N.A.O., 2021. A Global Nucleic Acid Observatory for Biodefense and Planetary Health.

Crook, O.M., Warmbrod, K.L., Lipstein, G., Chung, C., Bakerlee, C.W., McKelvey Jr., T.G., Holland, S.R., Swett, J.L., Esvelt, K.M., Alley, E.C., Bradshaw, W.J., 2021. Analysis of the first Genetic Engineering Attribution Challenge. arXiv:2110.11242 [cs].

Diggans, J., Leproust, E., 2019. Next Steps for Access to Safe, Secure DNA Synthesis. Front. Bioeng. Biotechnol. 7. https://doi.org/10.3389/fbioe.2019.00086

Dove, A., 2018. PCR: Thirty-five years and counting. Technology Feature.

Duprex, W.P., Fouchier, R.A.M., Imperiale, M.J., Lipsitch, M., Relman, D.A., 2015. Gain-of-function experiments: time for a real debate. Nat Rev Microbiol 13, 58–64. https://doi.org/10.1038/nrmicro3405

Ellsberg, D., 2017. The Doomsday Machine: Confessions of a Nuclear War Planner, 1st edition. ed. Bloomsbury Publishing.

Ernst, H., 1998. Industrial research as a source of important patents. Research Policy 27, 1–15. https://doi.org/10.1016/S0048-7333(97)00029-2

Esvelt, K.M., 2018. Inoculating science against potential pandemics and information hazards. PLoS Pathog 14. https://doi.org/10.1371/journal.ppat.1007286

European Environment Agency, 2021. Consumption of ozone-depleting substances [WWW Document]. European Environment Agency. URL https://www.eea.europa.eu/ims/consumption-of-ozone-depleting-substances

Evans, S.W., Greene, D., Hoffmann, C., Lunte, S., 2021. Stakeholder Engagement Workshop on the Implementation of the United States Government Policy for Institutional Oversight of Life Sciences Dual Use Research of Concern: Workshop Report (SSRN Scholarly Paper No. ID 3955051). Social Science Research Network, Rochester, NY. https://doi.org/10.2139/ssrn.3955051

Farmer, J.D., Hepburn, C., Ives, M.C., Hale, T., Wetzer, T., Mealy, P., Rafaty, R., Srivastav, S., Way, R., 2019. Sensitive intervention points in the post-carbon transition. Science 364, 132–134. https://doi.org/10.1126/science.aaw7287

Federal Highway Administration, Office of Highway Information Management, 2009. Motor Vehicle Traffic Fatalities 1900-2007.

Fisher, E., Mahajan, R.L., Mitcham, C., 2006. Midstream Modulation of Technology: Governance From Within. Bulletin of Science, Technology & Society 26, 485–496. https://doi.org/10.1177/0270467606295402

Friedman, B., Kahn, P.H., Borning, A., Huldtgren, A., 2013. Value Sensitive Design and Information Systems, in: Doorn, N., Schuurbiers, D., van de Poel, I., Gorman, M.E. (Eds.), Early Engagement and New Technologies: Opening up the Laboratory, Philosophy of Engineering and Technology. Springer Netherlands, Dordrecht, pp. 55–95. https://doi.org/10.1007/978-94-007-7844-3_4

Fuhrmann, M., Lupu, Y., 2016. Do Arms Control Treaties Work? Assessing the Effectiveness of the Nuclear Nonproliferation Treaty. int stud q 60, 530–539. https://doi.org/10.1093/isq/sqw013

Geels, F.W., Sovacool, B.K., Schwanen, T., Sorrell, S., 2017. Sociotechnical transitions for deep decarbonization. Science 357, 1242–1244. https://doi.org/10.1126/science.aao3760

Green, J.F., 2021. Does carbon pricing reduce emissions? A review of ex-post analyses. Environ. Res. Lett. 16, 043004. https://doi.org/10.1088/1748-9326/abdae9

Guston, D.H., 2014. Understanding 'anticipatory governance.' Soc Stud Sci 44, 218–242. https://doi.org/10.1177/0306312713508669

Guston, D.H., Sarewitz, D., 2002. Real-time technology assessment. Technology in Society, American Perspectives on Science and Technology Policy 24, 93–109. https://doi.org/10.1016/S0160-791X(01)00047-1

Haegel, N.M., Margolis, R., Buonassisi, T., Feldman, D., Froitzheim, A., Garabedian, R., Green, M., Glunz, S., Henning, H.-M., Holder, B., Kaizuka, I., Kroposki, B., Matsubara, K., Niki, S., Sakurai, K., Schindler, R.A., Tumas, W., Weber, E.R., Wilson, G., Woodhouse, M., Kurtz, S., 2017. Terawatt-scale photovoltaics: Trajectories and challenges. Science 356, 141–143. https://doi.org/10.1126/science.aal1288

Herzog, M.M., Parson, E.A., 2016. Moratoria for Global Governance and Contested Technology: The Case of Climate Engineering, UCLA Public Law & Legal Theory, Series Open Access Policy Deposits. UCLA: School of Law.

Horowitz, M.C., 2018. Artificial Intelligence, International Competition, and the Balance of Power. Texas National Security Review 1.

IARPA - FUSE [WWW Document], n.d. URL https://www.iarpa.gov/research-programs/fuse (accessed 5.10.22).

International Gene Synthesis Consortium, 2017. Harmonized screening protocol v2.0. International Gene Synthesis Corporation.

Keohane, R.O., Victor, D.G., 2011. The Regime Complex for Climate Change. Perspectives on Politics 9, 7–23. https://doi.org/10.1017/S1537592710004068

Kupferschmidt, K., 2017. How Canadian researchers reconstituted an extinct poxvirus for $100,000 using mail-order DNA. Science Insider. https://doi.org/doi:10.1126/science.aan7069

Lanouette, W., 1992. Genius in the shadows: a biography of Leo Szilard : the man behind the bomb. Charles Scribner's Sons ; Maxwell Macmillan International, New York : Oxford.

Luo, B., Twiss, S.B. (Eds.), 2015. Chinese Just War Ethics: Origin, development, and dissent, War, conflict and ethics. Routledge/Taylor & Francis Group, London.

Maas, M.M., 2019. How viable is international arms control for military artificial intelligence? Three lessons from nuclear weapons. Contemporary Security Policy 40, 285–311. https://doi.org/10.1080/13523260.2019.1576464

Mahdi, S., Nightingale, P., Berkhout, F., 2002. A review of the impact of regulation on the chemical industry. University of Sussex.

Marblestone, A., Gamick, A., Kalil, T., Martin, C., Cvitkovic, M., Rodriques, S.G., 2022. Unblock research bottlenecks with non-profit start-ups. Nature 601, 188–190. https://doi.org/10.1038/d41586-022-00018-5

McDonald, A., Schrattenholzer, L., 2001. Learning rates for energy technologies. Energy Policy 29, 255–261. https://doi.org/10.1016/S0301-4215(00)00122-1

McKenzie, R., Bernhard, G., Liley, B., Disterhoft, P., Rhodes, S., Bais, A., Morgenstern, O., Newman, P., Oman, L., Brogniez, C., Simic, S., 2019. Success of Montreal Protocol Demonstrated by Comparing High-Quality UV Measurements with "World Avoided" Calculations from Two Chemistry-Climate Models. Sci Rep 9, 12332. https://doi.org/10.1038/s41598-019-48625-z

McLeish, C., Nightingale, P., 2007. Biosecurity, bioterrorism and the governance of science: The increasing convergence of science and security policy. Research Policy 36, 1635–1654. https://doi.org/10.1016/j.respol.2007.10.003

Meadows, D., 1999. Leverage Points: Places to Intervene in a System. Sustainability Institute.

Mervis, J., 2016. What makes DARPA tick? Science 351, 549–553. https://doi.org/10.1126/science.351.6273.549

Mittleman, B., Neil, G., Cutcher-Gershenfeld, J., 2013. Precompetitive consortia in biomedicine—how are we doing? Nat Biotechnol 31, 979–985. https://doi.org/10.1038/nbt.2731

Monrad, J.T., Sandbrink, J.B., Cherian, N.G., 2021. Promoting versatile vaccine development for emerging pandemics. npj Vaccines 6, 1–7. https://doi.org/10.1038/s41541-021-00290-y

Mowery, D.C., Nelson, R.R., Martin, B.R., 2010. Technology policy and global warming: Why new policy models are needed (or why putting new wine in old bottles won't work). Research Policy 39, 1011–1023. https://doi.org/10.1016/j.respol.2010.05.008

National Security Commission on Artificial Intelligence, 2021. Final Report. Washington, DC.

Nye, J.S., 1987. Nuclear learning and U.S.–Soviet security regimes. International Organization 41, 371–402. https://doi.org/10.1017/S0020818300027521

Our World in Data, 2022. Cumulative CO_2 emissions by world region [WWW Document]. Our World in Data. URL https://ourworldindata.org/grapher/cumulative-co2-emissions-region (accessed 8.17.22).

Owen, R., Macnaghten, P., Stilgoe, J., 2012. Responsible research and innovation: From science in society to science for society, with society. Science and Public Policy 39, 751–760. https://doi.org/10.1093/scipol/scs093

Pastorino, B., de Lamballerie, X., Charrel, R., 2017. Biosafety and Biosecurity in European Containment Level 3 Laboratories: Focus on French Recent Progress and Essential Requirements. Frontiers in Public Health 5.

Reardon, S., 2017. US government lifts ban on risky pathogen research. Nature 553, 11–11. https://doi.org/10.1038/d41586-017-08837-7

Riedel, S., 2005. Smallpox and biological warfare: a disease revisited. Proc (Bayl Univ Med Cent) 18, 13–20.

Rip, A., Kemp, R., 1998. Technological change. Human choice and climate change: Vol. II, Resources and Technology 327–399.

Sandbrink, J.B., Koblentz, G.D., 2022. Biosecurity risks associated with vaccine platform technologies. Vaccine 40, 2514–2523. https://doi.org/10.1016/j.vaccine.2021.02.023

Sandbrink, J.B., Shattock, R.J., 2020. RNA Vaccines: A Suitable Platform for Tackling Emerging Pandemics? Front. Immunol. 11, 608460. https://doi.org/10.3389/fimmu.2020.608460

Sandbrink, J.B., Watson, M.C., Hebbeler, A.M., Esvelt, K.M., 2021. Safety and security concerns regarding transmissible vaccines. Nature Ecology & Evolution 5, 405–406. https://doi.org/10.1038/s41559-021-01394-3

Schaller, R.R., 1997. Moore's law: past, present and future. IEEE Spectr. 34, 52–59.

https://doi.org/10.1109/6.591665

Schoch-Spana, M., Cicero, A., Adalja, A., Gronvall, G., Kirk Sell, T., Meyer, D., Nuzzo, J.B., Ravi, S., Shearer, M.P., Toner, E., Watson, C., Watson, M., Inglesby, T.V., 2017. Global Catastrophic Biological Risks: Toward a Working Definition. Health Secur 15, 323–328. https://doi.org/10.1089/hs.2017.0038

Schot, J., Geels, F.W., 2008. Strategic niche management and sustainable innovation journeys: theory, findings, research agenda, and policy. Technology Analysis & Strategic Management 20, 537–554. https://doi.org/10.1080/09537320802292651

Schot, J., Rip, A., 1997. The past and future of constructive technology assessment. Technological Forecasting and Social Change 54, 251–268. https://doi.org/10.1016/S0040-1625(96)00180-1

Seger, E., Avin, S., Pearson, G., Briers, M., Ó Heigeartaigh, S., Bacon, H., n.d. Tackling threats to informed decisionmaking in democratic societies. The Alan Turing Institute, 2020.

Stilgoe, J., Owen, R., Macnaghten, P., 2013. Developing a framework for responsible innovation. Research Policy 42, 1568–1580. https://doi.org/10.1016/j.respol.2013.05.008

Sykora, P., Caplan, A., 2017. The Council of Europe should not reaffirm the ban on germline genome editing in humans. EMBO Rep 18, 1871–1872. https://doi.org/10.15252/embr.201745246

Taylor, P.L., 2005. The gap between law and ethics in human embryonic stem cell research: overcoming the effect of U.S. federal policy on research advances and public benefit. Sci Eng Ethics 11, 589–616. https://doi.org/10.1007/s11948-005-0028-x

Tetlock, P.E., Gardner, D., 2016. Superforecasting: The Art and Science of Prediction. Random House.

The power of treaties, 2013. . Nature 501, 5–5. https://doi.org/10.1038/501005a

van Dijk, A., Slaper, H., den Outer, P.N., Morgenstern, O., Braesicke, P., Pyle, J.A., Garny, H., Stenke, A., Dameris, M., Kazantzidis, A., Tourpali, K., Bais, A.F., 2013. Skin Cancer Risks Avoided by the Montreal Protocol-Worldwide Modeling Integrating Coupled Climate-Chemistry Models with a Risk Model for UV. Photochem Photobiol 89, 234–246. https://doi.org/10.1111/j.1751-1097.2012.01223.x

Vergne, J.-P., Durand, R., 2010. The Missing Link Between the Theory and Empirics of Path Dependence: Conceptual Clarification, Testability Issue, and Methodological Implications. Journal of Management Studies 47, 736–759. https://doi.org/10.1111/j.1467-6486.2009.00913.x

Volpe, T.A., 2019. Dual-use distinguishability: How 3D-printing shapes the security dilemma for nuclear programs. Journal of Strategic Studies 42, 814–840. https://doi.org/10.1080/01402390.2019.1627210

Waldrop, M.M., 2016. The chips are down for Moore's law. Nature 530, 144–147. https://doi.org/10.1038/530144a

Wetterstrand, K., n.d. DNA Sequencing Costs: Data [WWW Document]. National Human Genome Research Institute. URL https://www.genome.gov/sequencingcostsdata (accessed 3.28.22).

Wheelis, A., 2006. The way we are, 1st ed. ed. W.W. Norton, New York.

World Health Organisation, 2021. Emerging technologies and dual-use concerns: a horizon scan for global public health.

Zhou, Y., Dong, F., Liu, Y., Li, Z., Du, J., Zhang, L., 2020. Forecasting emerging technologies using data augmentation and deep learning. Scientometrics 123, 1–29. https://doi.org/10.1007/s11192-020-03351-6

Zwart, H., Landeweerd, L., van Rooij, A., 2014. Adapt or perish? Assessing the recent shift in the

European research funding arena from ‘ELSA’ to ‘RRI.’ Life Sciences, Society and Policy 10, 11. https://doi.org/10.1186/s40504-014-0011-x