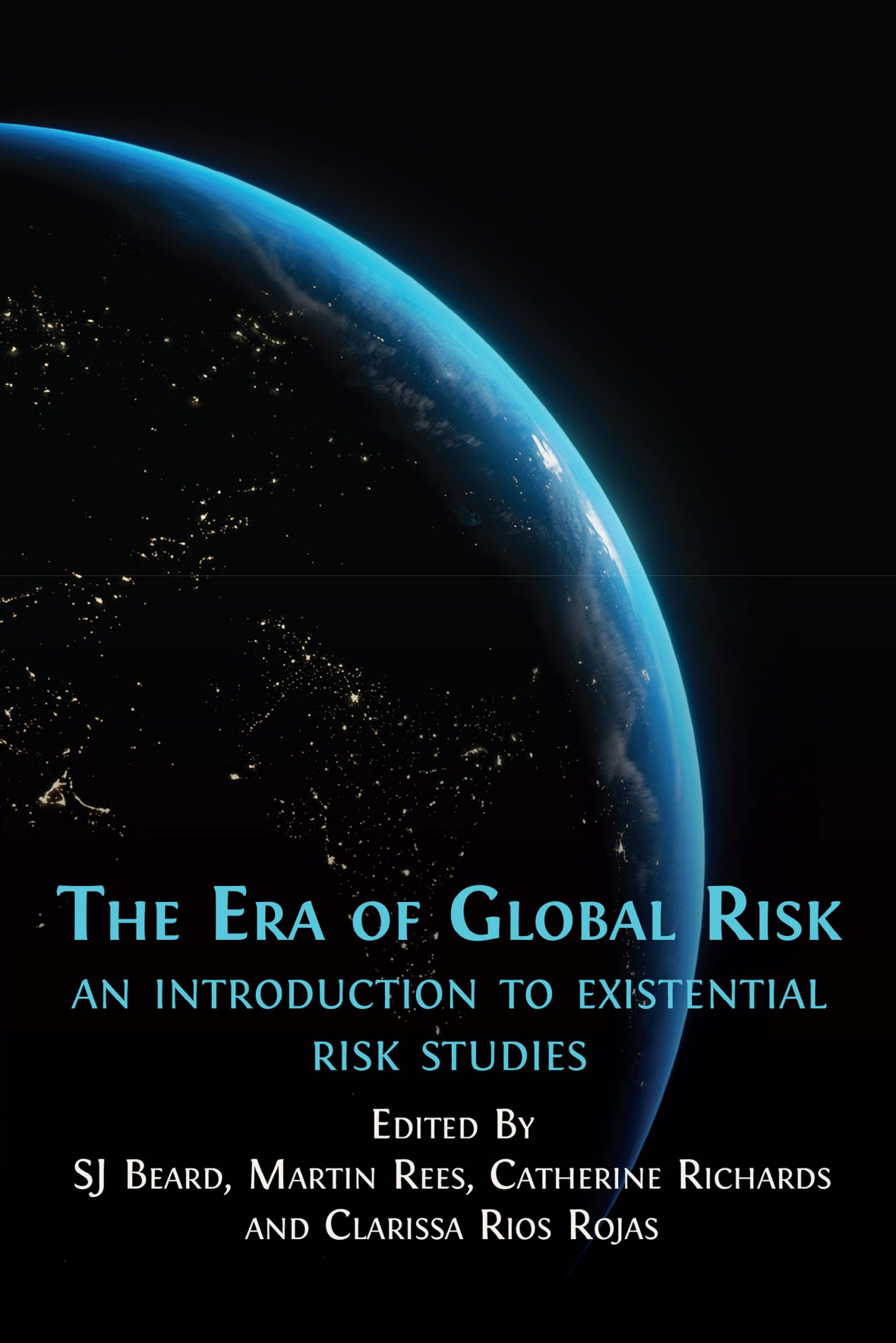

The Era of Global Risk

An Introduction to Existential Risk Studies

Edited By
SJ Beard, Martin Rees, Catherine Richards and Clarissa Rios Rojas

SJ Beard, Martin Rees, Catherine Richards and Clarissa Rios Rojas (eds), *The Era of Global Risk: An Introduction to Existential Risk Studies*. Cambridge, UK: Open Book Publishers, 2023, https://doi.org/10.11647/OBP.0336

Digital material and resources associated with this volume are available at https://doi.org/10.11647/OBP.0336#resources

ISBN Paperback: 978-1-80064-786-2
ISBN Hardback: 978-1-80064-787-9
ISBN Digital (PDF): 978-1-80064-788-6
ISBN Digital ebook (epub): 978-1-80064-789-3
ISBN XML: 978-1-80064-791-6
ISBN HTML: 978-1-80064-792-3
DOI: 10.11647/OBP.0336

Cover image: Anirudh, *Our Planet* (October 14, 2021), https://unsplash.com/photos/Xu4Pz7GI9JY. Cover design by Jeevanjot Kaur Nagpal.

3. Existential Risk and Science Governance

Lalitha S. Sundaram

The study of existential risk has, since its earliest days, been closely linked with scientists (both their work and their concerns).[1] This is easily seen in key moments like the establishment of Pugwash and the creation of the Doomsday Clock, or in publications from prominent scientists in more recent years, such as *Our Final Century*.[2] From nuclear weapons to AI arms races, environmental crises, and, yes, even pandemics, science and technology are deeply implicated in the scholarship of existential risk, whether they are viewed as causes, risk multipliers, or potential mitigating forces (or, indeed, as all three).

In this chapter, I look at how the governance of science might matter for the production and prevention of existential risk in the context of what levers we, as a community, might be ignoring. In particular, I look at the ways in which scientific governance is conventionally framed—seeing science as something extrinsic to be regulated stringently or to be left alone to proceed without interference—and point out some of the shortcomings of this view. Instead, I propose considering scientific governance more broadly as a constellation of socio-technical processes that shape and steer technology, and in doing so, argue that research culture and self-governance not only exist but are central to how science and technology developments play out. I then put forward that there are many more levers at our disposal for ensuring the safe development of potentially very beneficial technologies than narrow views of scientific governance might suggest, and that overlooking them could rob us of much of our arsenal against existential risk. Many of the levers involve self-governance in some way, and I explore how that manifests in

 https://doi.org/10.11647/OBP.0336.03

different disciplines and settings. I end the chapter by proposing some areas where scientists and the existential risk community might together hope to influence those existing modalities.

Is top-down governance the only way?

One prominent narrative that concerns the intertwining of science and technology with existential risk stems from a number of scientists themselves 'raising the alarm', particularly in the wake of the Manhattan Project and the development and proliferation of nuclear weapons technology following the Second World War. Well-known examples are, of course, Robert Oppenheimer, Leo Szilard, and Enrico Fermi, but other examples include thinkers who were less directly involved in the development of these weapons, but who raised the possibility of anthropogenic civilisational destruction as a subject for serious academic and policy discussion, like Albert Einstein and Bertrand Russell, or even Winston Churchill and H.G. Wells. For this cadre of intellectuals, naturally, given the historical moment through which they lived, the greatest risks emerged from the possibility of technologically enhanced war, with Russell noting in a letter to Einstein that: "although the H-bomb at the moment occupies the centre of attention, it does not exhaust the destructive possibilities of science, and it is probable that the dangers of bacteriological warfare may before long become just as great".[3]

One possible solution that was frequently discussed and advocated by such groups was the establishment of world government, both as a means of preventing future wars that might now prove fatal to humanity and as a way of accelerating the creation and adoption of solutions to pressing global problems. Bertrand Russell, for instance, argued that science was destabilising modern society in the same way that nuclear physicists had recently learnt to destabilise atoms. He argued that science was increasingly disturbing the physical, biological, and psychological basis for societies to the point where "we must accept vast upheavals and appalling suffering" unless four conditions could be established. These were: 1) a single world government with a monopoly on all military force, 2) enough economic equality to eradicate envy between people, 3) a universally low birth rate to ensure a stable world population, and 4) opportunities for everyone to develop individual initiative "in work and play" and to exercise the greatest level of power over themselves.[4]

It is worth noting, however, that Russell saw these conditions as built on one another (to some extent), such that world government is seen as the basis on which to establish equality, sustainability, and individual freedom—rather than the other way around.

It is important to note, however, that despite the strong leanings that Russell, Einstein, and others had towards strong top-down governance (at least on the global stage), the organisation that is (one of their) most significant legacies in the field of existential risk—Pugwash—decidedly does not operate under that model. The Pugwash Conferences that have been organised since the Russell-Einstein Manifesto (see this volume's chapter by Beard and Bronson) "eventually came to be hosted, almost all of them in places other than Pugwash, by national 'Pugwash groups' whose character, institutional affiliations (if any) and methods of work and fundraising varied from country to country",[5] with the Pugwash website itself noting its "rather decentralized organizational structure".[6] Moreover, as noted in the chapter by Beard and Bronson, Pugwash co-existed with several civil movements—such as the Campaign for Nuclear Disarmament and the Committee of 100, established by Russell himself—which justify civil disobedience as a means to give voice to popular fears about existential risks and dissatisfaction with government policy.

Nevertheless, the view of top-down, formal oversight as the only way to achieve governance persists, especially when it comes to governance of existential risks stemming from science. In 2019, Nick Bostrom introduced the "Vulnerable World Hypothesis",[7] whereby technological advances are viewed as balls drawn from a large urn of infinite possibilities (possible technological ideas or advances, that is). The majority of these balls have thus far been beneficial (white) or not catastrophically damaging (grey). However, Bostrom posits, it is only a matter of time before we (human society) draw a "black ball": a "technology that invariably or by default destroys the civilization that invents it".[8] Bostrom's contention is that the primary reason for us not having done so yet is luck, rather than any kind of safeguarding mechanism or policy. In order not to continue relying on such luck (which must, to Bostrom's mind, eventually run out), he describes one way out: that of exiting what he terms the "semi-anarchic default condition".[9] This condition—namely the world as it exists—is one to be overcome if we are to avoid drawing a black ball. The three main

features that Bostrom ascribes to it are a lack of preventative policing, a lack of global governance (in an obvious echo to Russell), and the fact that the actors involved have diverse motivations. Potential mitigations for some types of vulnerability that Bostrom proposes are options for technological curtailment, including differential technological development[10] and preference modification. Here, technologies deemed to be potentially "black ball" can be delayed in their development, or the actors involved could be monitored and their efforts re-focused. Overall, therefore, for Bostrom, a key macro-strategy would involve the strengthening of surveillance, and a global governance super-structure capable of "decisive action".[11] To be clear, Bostrom's argument is not that such a system is a desirable one (indeed, he is at pains to describe many of the potential downsides and concedes it may not be "desirable all things considered"[12]). Nevertheless, within the (hypothetical) context of the semi-anarchical condition of a "Vulnerable World", it is one of the few solutions that Bostrom sees as workable.

All this suggests a dichotomy between models in which science is allowed to develop on its own within existing social institutions (which are viewed as insufficiently overseen, in which case science is presumed to be more dangerous), or else sustained centralised action needs to be taken to reshape these institutions in order to direct scientific research and technological development away from what is risky and make society fit for receiving the benefits of science without being harmed. One way to think about this trade-off is in economic terms: either science (and thus scientists) is left to its own devices participating in an unregulated, *laissez-faire* 'market of ideas' or governments and regulators need to establish beforehand which topics will be most beneficial and safe, and then direct scientists and technology developers to work on those and avoid everything else. This dichotomy "between dirigism and laissez-faire"[13] has been characterised in biotechnology policy in terms of which parts of the biotechnology landscape are likely to yield the most public good, and as the Nuffield Council on Bioethics puts it, we are relying on state intervention on the one hand and market forces on the other.[14] While these ends of the spectrum (and options in between) have usually been considered in terms of their impact on future biotechnologies as a matter of social value, it is not a stretch to see how this could be reframed as a question of risk and safety, and of preventing catastrophic (or even existential) harm. Thus: either risks are foreseen by an authority and

regulated against, or some notion of a market (largely informal, though this could be formalised as insurance)[15] naturally self-adjusts to drive riskier areas of work and practice into abandonment. Plainly, this latter option cannot be of much use when it comes to existential risks, since even the most well-funded insurers could not hope to pay out after humanity has gone extinct, and we might say the same about many of the more severe forms of global catastrophe as well. Extreme risks may prove to be a market failure in every conceivable sense.

A broader take on governance

Thinking about governance in this polarised way is very restrictive and ignores a great many realities about science and technology. For a start, science and technology do not exist as separate entities from human activity and society more broadly. Moreover, as an object of governance, they are neither distinct nor static. Science and technology are not monolithic either; their developments do not exist in a vacuum. They are enterprises engaged in by *people*, and so it matters a great deal who those people are. Not only that, it matters how they have been trained, not just in their professions but in how they approach the world and what responsibility they feel they have towards it. And these scientists—these people—do not act in isolation from each other and their communities or from the institutions and wider systems within which they operate either.

So, when it comes to 'governing science' in order to understand, prevent, and mitigate existential risks, instead of reaching immediately for some hegemonic form of extra-community governance, we need to better understand and learn how to shape these wider systems. Indeed, governance is best seen as 'how technologies are shaped and steered' rather than simply, as is too often the case, 'how they are regulated'. We need to think about governance as the group of mechanisms, processes, and communities that structures, guides, and manages technology—and this must include a consideration of systems and networks, as well as norms and culture. This is part of the view taken by Voeneky, who, while considering the legitimacy and efficacy of international law to govern global risk, describes "a multi-layer governance that consists of rules of international law, supranational and national law, private norm

setting, and hybrid forms that combine elements of international or national law and private norm setting".[16]

There are two main failures of considering governance too narrowly (as a single law, policy, or other coercive measure to ensure scientists 'behave responsibly'). The first one is, as noted above, that it draws an artificial boundary around what 'science and technology' is: a boundary that tends to exclude the social and the political, seeing these instead as distinct realms into which technologies are injected or deployed. Instead, we need to acknowledge—and need our governance to acknowledge—that what we are looking at are sociotechnical systems,[17] where cross-pollination exists at every stage. Understanding these different levels of scientific governance is of utmost importance when thinking about existential risks because addressing these types of risk will require a combination of rules, norms, scientific vision, and an appreciation of the ways in which 'the social' and 'the scientific' constantly influence each other. Second, a narrow framing of scientific governance—in part a result of a too-narrow framing of science and technology—positions it solely as a gatekeeper. Viewing governance instead as a mechanism for steering allows us to harness the best of scientific and technological advances, while remaining mindful of the potential risks. However, this is not an easy path to tread; in the remainder of this chapter, I explore some of the reasons why, and propose some ways forward.

Responsibility for understanding and tackling existential risks is dispersed among a great many actors: how much of it is the responsibility of 'the scientific community'? The way 'the scientific community' is organised and how it operates is obviously a vast subject of study in and of itself, and outside the remit of this volume. A simplified (but by no means simple) view is that it involves myriad actors and institutions: governments and funders, the academy, the various institutions that 'house' the research, companies, industries and sectors, professional bodies that set standards and award qualifications, non-governmental organisations, international partnerships and coalitions and, of course, individual scientists and technology developers themselves.

A canonical example of a community's scientific self-governance is the Asilomar Conference on Recombinant DNA, which was prompted by a series of genetics experiments in the early 1970s. Experiments done in the laboratory of Paul Berg at Stanford demonstrated that DNA molecules could be 'recombined' using restriction enzymes to join them

together. Experiments by Stanley Cohen, also at Stanford, and Herbert Boyer at the University of California, San Francisco demonstrated that recombinant (artificially constructed) plasmids (circular molecules of DNA that can replicate independently of chromosomes) could be propagated in bacterial cells. Presentation of this work at a 1973 Gordon Conference, along with informal conversations between colleagues, then prompted the publication of a letter to the heads of the National Academies of Science and the National Academy of Medicine, published in the journal *Science*. This letter was written "on behalf of a number of scientists to communicate a matter of deep concern",[18] namely the newfound ability to recombine DNA from diverse genetic sources. The letter noted that: "Although no hazard has yet been established, prudence suggests that the potential hazard be seriously considered",[19] and called upon the Academies to set up study committees and consider establishing guidelines.

Berg led the publication of another letter in *Science*[20] and *PNAS*[21] where he (along with colleagues from the nascent field) called for scientists to "voluntarily defer" some experiments (essentially to impose a moratorium) until such time as a conference could be convened. It is important to note that no actual harm had yet occurred; no potentially dangerous experiments had even been performed yet (indeed the possibility of direct harm was considered quite remote). In some ways, the actions taken by the researchers were textbook exemplars of the precautionary principle, with Berg's later reflection and recollection of the events noting that "there were no concrete data concerning health risks attributable to recombinant DNA experimentation. Nevertheless, there were also no data absolving the planned experiments of any risk".[22] The moratorium was—as far as it is possible to know—abided by, despite Berg's own acknowledgement at the time that "adherence to [our] major recommendations will entail postponement or possibly abandonment of certain types of scientifically worthwhile experiments".[23, 24] The reasoning behind the moratorium was to allow time to organise an international conference where the issues could be hashed out and guidelines agreed upon: the now-famous International Conference on Recombinant DNA Molecules of 1975, held at the Asilomar Conference Centre.

Three days of discussions at the conference concluded with a consensus of broad guidelines, which were published as a summary

statement.[25] The recommendations included matching "containment levels" under which work would be performed to the appropriate estimated risk assessment of particular experiments; considerations of what organisms were being used; 'good' laboratory procedures to be implemented; and the development of safer 'vectors' and 'hosts'. Moreover, the summary statement also recommended that particular types of experiments be deferred, including those involving DNAs from pathogenic organisms or those "using recombinant DNAs that are able to make products potentially harmful to man, animals, or plants."[26]

And indeed, the recommendations that emerged from the conference formed the basis for much American (and, broadly speaking, worldwide) regulation of the field.

Asilomar was not a perfect process, however. Some contemporary (as well as more recent) criticisms concern the composition of the conference: alongside a few journalists who were under effective embargo, there were some 150 molecular biologists, a handful of (non-practising) lawyers, and a single bioethicist in attendance.[27] This obviously raises questions about the motivations involved: were the scientists merely forestalling more severe external regulation by being very visibly proactive?

Importantly, what would seem to be two obvious and key issues were omitted from the agenda altogether: biological warfare and gene therapy. According to the science historian, Charles Weiner:[28]

> The recombinant DNA issue was defined as a technical problem to be solved by technical means, a technical fix. Larger ethical issues regarding the purposes of the research; long-term goals, including human genetic intervention; and possible abuses of the research were excluded.

Despite this, Asilomar was largely hailed as a success in scientific self-governance, where scientists could demonstrate that they were conscious of the risks their work might incur, and that they could reach a consensus on guidelines—guidelines that would later inform regulation—to minimise these risks. As such, much more than a picturesque Californian conference venue, Asilomar has come to represent the process of pre-emptive scientific self-reflection in the face of emerging technology. Indeed, the "Asilomar Moment" has been invoked numerous times in biological research, but also—in desirable terms—in nanotechnology,[29] geoengineering,[30] and AI, from which we draw our next example.

In a sense, the upholding of the "Asilomar Moment"[31] as a paragon of self-governance simply illustrates the paucity of our understanding when it comes to what constitutes good self-governance, and ignores the many other levers that we have at our disposal. Two examples from contemporary science and technology demonstrate some of the less obvious tools of self-governance and how they can be used.

Contemporary case studies of research culture: Self-governance in action

Asilomar features heavily in another example of self-reflection by technologists, but only as part of a broad set of measures undertaken in the AI community "to shape the societal and ethical implications of AI",[32] actions that Belfield terms "activism". While Belfield draws out other ways in which that activism takes shape (worker organising, for instance), what I will focus on here is that of using the "epistemic community"[33] as an engine for self-reflection and norm-setting.

What Belfield describes echoes some of the historical actions taken by molecular biologists in the 1970s, but the scale is larger and involves more actors. In fact, the community the author defines is almost as broad as the one this chapter considers with its view of what constitutes a "scientific community" (cf. the composition of the 1975 Asilomar Conference). Belfield's AI community "include[s] researchers, research engineers, faculty, graduate students, NGO workers, campaigners and some technology workers more generally—those who would self-describe as working 'on', 'with' and 'in' AI and those analysing or campaigning on the effects of AI".[34]

Some of the actions Belfield describes include the publication of open letters, committees tasked specifically with looking at safety and ethics, and large-scale conferences on the subject, such as the 2015 Puerto Rico and 2017 Asilomar Conference for Beneficial AI. Convened by the Future of Life Institute, the 2017 conference ran over three days, and from it emerged a set of 23 "Asilomar AI Principles". The issues explored during the conference and reflected in the principles were spread across three subsets: Research, Ethics & Values, and Longer-Term Issues. Artificial Intelligence's "Asilomar Moment", despite that venue's totemic importance, is not the only way in which self-governance is apparent in the AI community—indeed, some authors have argued

that lists of principles are not in themselves sufficient to ensure that the field proceeds in "robustly beneficial directions").[35] Belfield describes further actions taken by the community, such as the establishment of research, advocacy, and policy-facing centres that hold AI ethics and safety as their focus, as well as policy proposals that feed concrete input into national and international AI strategies. Other initiatives include the Neural Information Processing Systems Conference's requirement, starting in 2020, for authors to include in their submissions a "broader impact statement" which would address their work's "ethical aspects and future societal consequences".[36] While this approach obviously has its challenges (many of which reflect the complex and interlinked nature of incentives and pressures facing researchers in this—or any other—field), it is clearly an important move towards "effective community-based governance".[37] What we see here, therefore, is an example of those most intimately involved in the development of a technology wanting to have a strong pre-emptive hand in how that technology unfolds: what Baum terms "intrinsic methods"[38] in his discussion of how to ensure that the development of AI proceeds in directions that are safe and societally beneficial. Thus, while the 2017 conference may have self-consciously sought to emulate 1975, it is clear that this epistemic community has reached for—and found—many more governance modalities, and these have largely emerged from within.

Some of the differences between these two Asilomars can be attributed to research culture and how that varies not only across the several decades that separate the two events, but also across disciplines.

Research culture is a powerful force that shapes technology development, but it is also very difficult to study and change. Scientific cultures are also incredibly diverse across fields. In terms of ethics and responsibility, some fields—medicine, for example—have a long history of codified moral 'guideposts', such as the Hippocratic Oath. Within fields, too, the picture is non-homogeneous. In computer science, for instance, while the concept of 'computer ethics' was developed in the early 1940s, it is only in recent years, with public outcry surrounding privacy and the sale of data, that the issue is being given serious consideration by developers, including the work described above. Despite its near omission from the 1975 conference, bioethics has (in the decades since then) dealt extensively with issues such as gene therapy. In recent years—with the emergence of concepts such as Responsible

Research and Innovation[39] [40] alongside the emergence of synthetic biology—attention is now being paid to the responsibilities of 'ordinary scientists at the bench' rather than those purely dealing with the most obviously public-facing parts of biotechnology, such as patient consent or clinical trials. The world of pathogen research is interesting in that, beyond adherence to legal biosafety frameworks and their attendant risk assessments, there appears to be little in the way of work on broader ethical or societal engagement. Instead, issues such as safety and security have usually been raised from outside, from fields such as biosecurity or epidemiology.[41] And, of course, research culture can vary greatly in terms of how individual institutions and even laboratories are run. Given this variety, it is therefore of utmost importance that research culture be better understood, in order to more effectively use it as a means of enculturating responsibility.

An interesting example here is that of the DIY-bio community. DIY-bio can very loosely be defined as the practice of biology, biotechnology, or synthetic biology performed outside of traditional institutions, hence it sometimes being termed 'garage biotechnology' or 'biohacking'. It has been of particular interest to some in the existential risk community, with scenarios of 'lone wolves' accidentally or deliberately engineering pathogens without the oversight that normally comes from working in universities. While it may indeed be true that certain DIY-biologists work totally independently (and these have tended to be the ones garnering the most attention), the overall picture of DIY-bio is quite different. While all DIY-bio laboratories are under their national biosafety and biosecurity regulation, my own research[42] has shown that the organisation of the field as a whole demonstrates a complex ecosystem of laws, norms, and self-governance. For example, there is a DIY-bio Code of Ethics which is repeated and emphasised on several laboratories' websites, and which many of these laboratories require adherence to as a condition of membership. There are also internal Codes of Practice in place that outline more lab-specific expectations. Internal, safety-promoting, and security-promoting practices abound.

What interviews with DIY-bio community members show is that there is often intentionality in how these spaces are set up, so that they "promote a culture of trust, accountability and responsibility".[43] This can include interviews and screening of potential members, policies requiring partnered or group work (which encourages transparency

and discussion), and numerous lines of communication between (biosafety/biosecurity) management and participants. The very fact that these spaces require active participation and engagement from their members in many aspects of management results in a greater degree of sensitisation to concerns about safety and security, but also reputational damage. DIY-bio practitioners therefore have a large incentive not only to behave in a responsible and safe way, but to be seen to be doing so. Moreover, there is a large degree of self-reflection in the community; 2020 saw the publication of a comprehensive "Community Biology Biosafety Handbook"[44] aimed at both established DIY-bio laboratories as well as new ones, in order to "serve as a foundation for establishing biosafety and security practices". The Handbook, which "includes biological, chemical, and equipment safety, but also subjects unique to community labs such as interview practices for screening potential lab members, considerations when working with children at festivals, building tips for creating labs in unconventional spaces, and much more", is a living document, written by several community laboratory leaders, and policy and safety experts (including, for instance, the president of the American Biological Safety Association). Many DIY-bio laboratories also have a strong educational component, which includes education on biosafety and security. Thus, even in a sub-field often assumed to be a 'wild west', there are clearly mechanisms of self-governance and self-regulation at play.

How can we improve science governance?

While governance is traditionally seen as something that happens in a top-down fashion, another way to think about it—one that puts at its centre the *people* involved—is to consider how a research culture is built and changes. Each year brings with it a new crop of practitioners, and so influencing them is an obvious starting point in influencing a field's culture.

Education

A clear route to improving scientific governance is to ensure that scientists see it as part of their job, and this involves education. At the pre-professional level, this can happen in two main ways. First, there

is a need to increase the number of students and scholars researching existential risks *qua* existential risk. The field is growing but remains fairly niche and centred around a small number of elite institutions in wealthy countries. For existential risk to become part of the academic vernacular, it needs to be offered either as part of taught courses or as an option for research at a wider range of institutions globally. Despite recent trends towards multi-disciplinarity, many science and technology researchers find themselves in disciplinary silos, under the 'everyday pressures' of academia or industry. There is a likelihood that their vision of 'risk' is limited only to what their experiments might mean for their own and their labmates' safety. Indeed, unless they work on topics that have explicit forebears in, say, atomic science, they may not ever have heard of existential risk. Increased education around existential risks enlarges the talent pool that does this important research (and may then feed it into policy) and it also serves the function of 'normalising' it in academic discussion, which is one of the indirect goals of pre-professional scientific education that needs attending to. An introduction to existential risk could be tailored to specific scientific disciplines and taught as mandatory modules, thus sensitising future practitioners to the impact of their work from an early stage. This could serve the rising appetite of schoolchildren and university students for engaging with large-scale societal issues, via increased activism and participation in a variety of formal and informal groups and movements. The school climate strikes are an obvious example, but Effective Altruism groups, as well as Student and Young Pugwash, can be leveraged too.

There are reasons for optimism in the realm of formal teaching: for example, while 'engineering ethics' is a relatively well-established part of engineering curricula,[45] other scientific disciplines (such as synthetic biology) are also beginning to offer courses that deal with the societal implications of the science.[46] These types of modules are obvious places to include teaching that develops students' awareness and understanding of existential risk and of acculturating the idea of scientific self-governance. Computer science courses, too, are beginning to include topics such as AI Ethics[47]—again, an ideal route to begin building capacity for self-governance.

A key challenge remains, however, and that is likely most keenly felt when students begin to embark on semi-independent research. At this stage, often during PhD scholarship, while a student may be undertaking

day-to-day research in a self-sufficient way, the overall research themes and directions will, in the main, be set by Principal Investigators (PIs) who will be responding to their own influences and incentives, based on career progression, availability of funding, and research trends. As engaged as a PhD student might be in wider societal impacts, their work will usually *operationally* be dictated by the PI, who may not wish to engage with such impacts. The power differential between PIs and early career researchers cannot be underestimated. As such, part of a PhD student's education in this area will need to be in how to navigate ethical 'grey areas' in a way that feels comfortable for them, but that does not necessarily penalise them in their labs or disadvantage them professionally. One way to deal with this issue is to ensure commitment not only from individual PIs but from institutions: if institutions recognise (as they are beginning to) that ethics and responsibility are valid and valuable as core parts of scientific education, individual recalcitrant PIs can be circumvented. The key, though, is for students wishing to enlarge their view of how their work fits into the world—and this includes thinking about existential risks—to be supported. Recognising differences between disciplines and their research cultures plays a part here too. The experience described above is situated in the context of the natural sciences, under a mostly hierarchical model. Here, the student is far more dependent on the PI for access to resources (financial resources, certainly, but also in terms of access to materials and equipment) and intellectual direction than, say, a student in the humanities might be. In the humanities—philosophy, for example, from which discipline many existential risk researchers hail—the norm is apparently for scholarship at even a relatively junior level (from undergraduate onwards) to be much more self-directed. Even here, however, there will still be institutional norms that a student may find they are expected to adhere to, however implicitly. Either way—and especially as the field of existential risk studies expands to include those from many different academic (and non-academic) backgrounds—it is thus important that assumptions about the autonomy of scholars at this level are questioned.

Professional bodies

One way that this institutional support could be strengthened is through professional bodies and associations or learned societies—organisations that influence more seasoned scientists. These often play a large role in shaping a scientific field, and so can exert a great deal of influence in how it is taught and how its practitioners are trained. While many of these professional organisations do have codes of ethics or statements of responsibility, in the main these tend to be rather inward-facing, setting out guidelines for ethical conduct *within* the discipline (in terms of things like discrimination, plagiarism, or obtaining consent from research participants). Global responsibility needs to be an added dimension.

However, there is also a need for professional engineering and science organisations to examine their own relationship to the major sectors driving existential risk, such as fossil fuel and (nuclear) arms industries. They must be transparent in how they accept funding and sponsorship events—especially educational ones—and in how they seek to remove themselves from these relationships. In the UK, for example, several such professional organisations have been criticised as being less-than-transparent in their financial interactions with these industries.[48]

This is somewhat at odds with another important role that professional societies can play: the interface between practising scientists/technology developers and global governance mechanisms, such as arms control treaties or the Biological and Toxin Weapons Convention (BTWC). Instruments such as these are often seen as unconnected to scientists' everyday practice, but sensitising practitioners to the relevance of these high-level discussions will highlight the obligations that are incurred and help them recognise the role that governance at the highest levels can play. The global bodies that oversee these international agreements and treaties can also play a role, by actively seeking input from practising scientists, not just from security or governance experts. For example, the BTWC implementation office could issue calls for institutions to nominate a diverse group of scientists to attend the yearly Meeting of Experts, making sure that it is not composed of the same group of scientists (typically those who are involved in research already widely thought of as 'risky') who engage with these issues year after year.

Policy engagement

Having scientists who are encouraging policy engagement also helps us tackle another aspect of scientific governance that has proven difficult. Research culture is, as we have seen, a complex domain, and it becomes even more so when it interacts with formal, top-down governance mechanisms. While a scientist or technology expert may be '*an* authority' in their domain, they are not necessarily '*in* authority' when it comes to questions of governance. Rather, this is done by policymakers, who may not have the necessary scientific understanding to do so effectively. This is especially tricky when it comes to governing emerging technologies (such as those we tend to associate with existential risk) because of the uncertainties involved. It is the 'Collingridge dilemma'[49] writ large: the balance of uncertainty and 'ability to govern' have an inverse relationship so that, almost perversely, technologies are easier to govern and shape when less is known about them and their impacts. And so, policy engagement from a diverse group of scientists is necessary at every stage, to ensure that multiple points of view are taken into account when feeding into governance, thereby hopefully bypassing the false choice of over- or under-regulation mentioned above entirely.

Collective action

Established scientists and technology developers can exert influence and support younger practitioners by being more open and vocal about their own commitments to ethical science and the prevention and mitigation of existential risks. In many instances, it appears that established scientists and engineers hold private concerns about existential risks that they are uneasy to voice for several reasons. So, providing a *collective* means of expressing concern and pledging action can be a useful way of eliciting more honest communication about the scale and seriousness of the problems humanity faces. For example, Scientists for Global Responsibility's Science Oath for the Climate[50] encourages their members (scientists, engineers, and other academics) to sign an oath to take professional and personal action, and to speak out publicly. This Oath astutely recognises the hesitancy that some scientists might face in doing so, and emphasises the connection between personal action and the ability of groups to influence and change systems.

Collective action is also of utmost importance when it comes to tackling another facet of governance: funding and complicity. More funding for work on existential risks is obviously a critical factor, but equally important is *where* that funding comes from. At present, business[51] and the military[52] dominate as funders and performers of research, but there is still a lot of money in the public sector, and universities are obviously places where much existential risk research is performed. However, an important part of scientists' responsibilities is to be judicious in what funding is accepted and what partnerships are entered into. As well as actively pursuing research into understanding and preventing existential risks, there must be scope for curbing the influence of organisations and sectors that are responsible for causing existential risks. But these organisations and sectors often have the resources to be attractive partners for scientists, and a strong incentive to do so: a kind of green-washing, or 'ethical-washing'. For instance, in the UK, the private consortium Atomic Weapons Establishment funnels over £8.5 million to over 50 universities as part of its Technical Outreach programme. Similarly, one of the biggest science and engineering fairs (unironically titled The Big Bang) gains most of its sponsorship from a number of weapons companies.[53] At a minimum, all these types of relationships need to be made transparent and strong ethical safeguards enacted.

Public outreach

Scientists can also use their voices to communicate and engage with society. Existential risk as part of a public agenda requires buy-in from that public, not only in ensuring that existential risks remain high on that agenda, but by helping to curb the undue influence of highly problematic industries which, as we have seen, can undermine both the spirit and practice of 'responsible science'. Public influence matters, not least because many of those who hold power are (at least in democracies) ultimately still accountable to that public, and will respond to pressure from their constituencies. Again, this will require effort from scholars of existential risk, and from scientists and technology developers from the wider STEM(M) community. In light of the COVID-19 pandemic, there is likely to be an increased appetite for topics related to resilience and preparedness—more 'realistic' scenarios than the usual Hollywood

zombie stories. Existential risk scholars should capitalise on this to work with science communicators, taking cues from the fields of disaster communication and environmental psychology to craft messages that inspire action, rather than hopelessness.

In closing

We have thus far discussed several ways in which 'the scientific community' can engage with governance—many levers that can be used in preventing a mitigating existential risk—but there are other elements that affect cultural values around science and technology innovation. Scientific communities exist within nation-states (though international collaborations and coalitions are, of course, common). As a result, ultimately, the research culture that an individual scientist finds herself in will be shaped in large part by the type of regime in which she finds herself. What place is there for distributed scientific governance—which includes the elements we discussed above—in a political regime that requires a particular technological direction to be taken? As we noted before, there is a difference between being 'an authority' and 'in authority'.

It is commonplace to distinguish between democracies and authoritarian governments according to their decision-making procedures and the presence, or absence, of elections as an opportunity to replace key decision-makers within them. However, we can also distinguish both democracies and authoritarian governments from totalitarian governments, whose main purpose is to break down the division between public and private, to erase the capability for freedom of speech and freedom of thought. As Immanuel Kant ("How much and how correctly would we think if we did not think as it were in community with others to whom we communicate our thoughts, and who communicate theirs with us!"[54]) and John Stuart Mill (in Chapter 2 of *On Liberty*[55]) show us, thought and speech are very clearly linked, and when you break down the division between the public and the private sphere so that these freedoms no longer really exist in any meaningful sense, you start to see where technology can be designed to maintain the status quo, even if the institutional machinery of elections remain in place.

Even in democratic societies, some totalitarian *tendencies* can have a similar effect. For example, AI technology may be used to further particular political ends in the name of protection and security. Another example is related to surveillance, which is a well-known tool that totalitarian governments have long used in the name of protecting their citizens; advanced technologies are only making this easier. Not only that, surveillance has (as we explored earlier in this chapter) been considered quite seriously as that elusive one-shot solution to the issue of scientific governance—or, at the very least, a serious contender for the 'least-bad'[56] option.

One could think of this as a twisted reading of—and tacit approval from—Mill, for whom "the only purpose for which power can be rightfully exercised over any member of a civilized community, against his will, is to prevent harm to others".[57] What greater harm can there be, after all, than existential risk? And indeed, so might greater awareness and sensitisation to catastrophic risks make this exertion of power more palatable, though the true nature of the bargain being struck—with technology as its facilitator—remains obscured. It is tempting to think this might be a case of neutral technologies being (mis-)applied to politically charged problems, as indeed existential risks can be, but as we have explored in this chapter neither technology nor its creators are neutral. This goes beyond political factions and knee-jerk thinking that authoritarianism must only have 'bad' solutions and that democracies must only have 'good' ones. Instead, we need to look more closely at priorities (and prioritisation): how science is directed in service of those priorities and what the pitfalls may be—and whether we are willing to live with them.

In any case, as I have argued in this chapter, any 'one-shot solution' is unlikely to be an effective nor practicable way of approaching scientific governance—at least not in the long term. Just like the Madisonian challenge of balancing factions with pursuing "great and aggregate interests",[58] scientific governance also faces difficulties that mean that an almost federated system of governance is necessary for it to work. The field of 'science and technology' is too broad and too diverse, the actors face numerous, often conflicting, sets of incentives for a single top-down approach to suffice, even when the ultimate aim (of 'Responsible Research and Innovation') is the same. We need to reflect seriously on how science

and technology developments are socially and politically inflected, how important power is as a determinant of action, and how collective action might be used as a means of prompting change. Scientific governance will not be achieved by merely making stricter rules. The best hope, as ever, is more education and more thoughtful engagement with the many systems that make up the scientific enterprise. What we have outlined in this chapter are some options for engaging with and influencing research culture which, as we have seen, can be a key determinant in self-governance, and thus in overall governance and the promotion of safer and more socially valuable scientific and technological goods.

Notes and References

1 Beard, S.J. and Phil Torres, 'Ripples on the great sea of life: A brief history of existential risk studies', *SSRN Electronic Journal* (2020), https://doi.org/10.2139/ssrn.3730000

2 Rees, Martin J., *Our Final Century: A Scientist's Warning: How Terror, Error, and Environmental Disaster Threaten Humankind's Future in This Century—on Earth and Beyond.* Heinemann (2003).

3 Russell, Bertrand, *The Selected Letters of Bertrand Russell*, ed. by Nicholas Griffin. Routledge (2001), p. 489.

4 Russell, Bertrand, 'Can a scientific society be stable?', *BMJ*, *2*(4640) (1949), pp.1307–11. https://doi.org/10.1136/bmj.2.4640.1307

5 Robinson, Julian Perry, 'Contribution of the Pugwash movement to the international regime against chemical and biological weapons', *10th Workshop of the Pugwash Study Group on the Implementation of the Chemical and Biological Weapons Conventions* (1998). http://www.sussex.ac.uk/Units/spru/hsp/documents/pugwash-hist.pdf.

6 Pugwash, 'About Pugwash', *Pugwash Conferences on Science and World Affairs* (2013), https://pugwash.org/about-pugwash/

7 Bostrom, Nicholas, 'The Vulnerable World Hypothesis', *Global Policy*, *10*(4) (2019), pp.455–76. https://doi.org/10.1111/1758-5899.12718

8 Bostrom (2019), p.455.

9 Bostrom (2019).

10 Bostrom, Nicholas, 'Existential risks: Analyzing human extinction scenarios and related hazards', *Journal of Evolution and Technology*, *9* (2022). https://ora.ox.ac.uk/objects/uuid:827452c3-fcba-41b8-86b0-407293e6617c

11 Ibid, p.470.

12 Ibid.

13 Rip, Arie and Anton J. Nederhof, 'Between dirigism and laissez-faire: Effects of implementing the science policy priority for biotechnology in the Netherlands', *Research Policy*, *15*(5) (1986), pp.253–68. https://doi.org/10.1016/0048-7333(86)90025-9

14 Nuffield Council on Bioethics, *Emerging Biotechnologies: Technology, Choice and the Public Good* (2012).

15 Farquhar, Sebastien, Owen Cotton-Barratt, and Andrew Snyder-Beattie, 'Pricing externalities to balance public risks and benefits of research', *Health Security*, *15*(4) (2017), pp.401–08. https://doi.org/10.1089/hs.2016.0118

16 Voeneky, Silja, 'Human rights and legitimate governance of existential and global catastrophic risks', in Gerald L. Neuman and Silja Voeneky (eds), *Human Rights, Democracy, and Legitimacy in a World of Disorder*. Cambridge University Press (2018). https://doi.org/10.1017/9781108355704

17 Jasanoff, Sheila and Sang-Hyun Kim, *Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power*. University of Chicago Press (2015).

18 Singer, Maxine and Dieter Soll, 'Guidelines for DNA hybrid molecules', *Science (New York, N.Y.)*, *181*(4105) (1973), p.1114. https://doi.org/10.1126/science.181.4105.1114

19 Ibid (1973).

20 Berg, Paul et al., 'Potential biohazards of recombinant DNA molecules', *Science*, *185*(4148) (1974), p303. https://doi.org/10.1126/science.185.4148.303

21 'Potential biohazards of recombinant DNA molecules', *Proceedings of the National Academy of Sciences*, *71*(7) (1974), pp.2593–94. https://doi.org/10.1073/pnas.71.7.2593

22 Berg, Paul, 'Moments of discovery', *Annual Review of Biochemistry*, *77*(1) (2008), pp.15–44. https://doi.org/10.1146/annurev.biochem.76.051605.153715

23 Berg, Paul et al. (1974).

24 Interestingly, in her examination of international law and the governance of existential risks, Voekeny notes that "a moratorium on a specific

kind of research that is based on the consensus of the relevant scientific community and backed up by the relevant scientific journals—which will not publish experiments that violate the moratorium—can be even more effective than a prohibition based on an international treaty that is implemented, top-down, by States parties."

25 Berg, Paul et al., 'Summary statement of the Asilomar conference on recombinant DNA molecules', *Proceedings of the National Academy of Sciences of the United States of America*, *72*(6) (1975), pp.1981–84.

26 Berg, Paul et al (1975).

27 Dworkin, Roger B., 'Science, society, and the expert town meeting: Some comments on Asilomar', *Southern California Law Review*, *51*(6) (1978), pp.1471–82.

28 Weiner, Charles, 'Drawing the line in genetic engineering: Self-regulation and public participation', *Perspectives in Biology and Medicine*, *44*(2) (2001), pp.208–20. https://doi.org/10.1353/pbm.2001.0039

29 Toumey, Chris, 'An Asilomar for nanotech', *Nature Nanotechnology*, *9* (2014), pp.495–96. https://doi.org/10.1038/nnano.2014.139

30 Kintisch, Eli, '"Asilomar 2" takes small steps toward rules for geoengineering', *Science*, *328*(5974) (2010), pp.22–23. https://doi.org/10.1126/science.328.5974.22

31 Petsko, Gregory A., 'An Asilomar moment', *Genome Biology*, *3*(10) (2002), comment1014.1-comment1014.3 https://doi.org/10.1186/gb-2002-3-10-comment1014

32 Belfield, Haydn, 'Activism by the AI community: Analysing recent achievements and future prospects', in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (presented at the AIES '20: AAAI/ACM Conference on AI, Ethics, and Society, ACM, 2020), pp.15–21. https://doi.org/10.1145/3375627.3375814

33 Haas, Peter M., 'Introduction: epistemic communities and international policy coordination', *International Organization*, *46*(1) (1992), pp.1–35.

34 Belfield (2020).

35 Whittlestone, Jess et al., 'The role and limits of principles in AI ethics: Towards a focus on tensions', in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (presented at the AIES '19: AAAI/ACM Conference on AI, Ethics, and Society, ACM, 2019), pp.195–200. https://doi.org/10.1145/3306618.3314289

36 'NeurIPS 2020', https://nips.cc/Conferences/2020/CallForPapers

37 Prunkl, Carina E.A. et al., 'Institutionalizing ethics in AI through broader impact requirements', Nature Machine Intelligence, 3(2) (2021), pp.104–10. https://doi.org/10.1038/s42256-021-00298-y

38 Baum, S.D., 'On the promotion of safe and socially beneficial artificial intelligence', *AI and Society*, *32*(4) (2017), pp.543–51. https://doi.org/10.1007/s00146-016-0677-0

39 Owen, R., P. Macnaghten, and J. Stilgoe, 'Responsible research and innovation: From science in society to science for society, with society', *Science and Public Policy*, *39*(6) (2012), pp.751–60. https://doi.org/10.1093/scipol/scs093; Alix, Jean-Pierre, 'RRI: Buzzword or vision of modern science policy?', *EuroScientist Journal* (2016). https://www.euroscientist.com/rri-new-buzzword-vision-modern-science-policy/

40 Owen, R., P. Macnaghten, and J. Stilgoe (2012).

41 Lipsitch, Marc and Thomas V. Inglesby, 'Moratorium on research intended to create novel potential pandemic pathogens', *MBio*, *5*(6) (2014), e02366–14. https://doi.org/10.1128/mBio.02366-14; Koblentz, Gregory D., 'Dual-use research as a wicked problem', *Frontiers in Public Health*, *2* (2014), p.113. https://doi.org/10.3389/fpubh.2014.00113

42 Sundaram, Lalitha S., 'Biosafety in DIY-bio laboratories: From hype to policy', *EMBO Reports*, *22*(4) (2021), e52506. https://doi.org/10.15252/embr.202152506

43 Sundaram (2021).

44 'Community biology biosafety handbook', *Genspace.* https://www.genspace.org/community-biology-biosafety-handbook

45 Royal Academy of Engineering, *Engineering Ethics Toolkit—Engineering Professors Council.* https://epc.ac.uk/resources/toolkit/ethics-toolkit/

46 Imperial College London, 'Systems and synthetic biology MRes | Study | Imperial College London', *Imperial College London.* https://www.imperial.ac.uk/study/courses/postgraduate-taught/systems-synthetic-biology/; University of Edinburgh, 'DPT: Systems and synthetic biology (MSc) (PTMSCSSBIO1F)', *University of Edinburgh.* http://www.drps.ed.ac.uk/21-22/dpt/ptmscssbio1f.htm.

47 Stavrakakis, Ioannis et al., 'The teaching of computer ethics on computer science and related degree programmes: A European survey', *International Journal of Ethics Education*, *7*(1) (2022), pp.101–29. https://doi.org/10.1007/s40889-021-00135-1

48 Parkinson, Stuart and Philip Wood, *Irresponsible Science? | SGR: Responsible Science* (Scientists for Global Responsibility, October 2019). https://www.sgr.org.uk/publications/irresponsible-science

49 Collingridge, David, *The Social Control of Technology*. St Martin's Press (1980).

50 'Why do we need the climate oath?', *Scientists for Global Responsibility.* https://www.sgr.org.uk/projects/why-do-we-need-climate-oath

51 Office for National Statistics (ONS), 'Gross domestic expenditure on research and development, UK: 2020', (2022). https://www.ons.gov.uk/economy/governmentpublicsectorandtaxes/researchanddevelopmentexpenditure/bulletins/ukgrossdomesticexpenditureonresearchanddevelopment/2020#cite-this-statistical-bulletin

52 Kuiken, Todd, *Wilson Center: US Trends in Synthetic Biology Research Funding* (September 2015). https://www.wilsoncenter.org/publication/us-trends-synthetic-biology-research-funding

53 Parkinson and Wood (2019).

54 Wood, Allen W., 'What does it mean to orient oneself in thinking? (1786)', in *Religion and Rational Theology*, by Immanuel Kant, ed. by Allen W. Wood and George di Giovanni, 1st edn. Cambridge University Press (1996), pp.1–18 (p.16). https://doi.org/10.1017/CBO9780511814433.003

55 Mill, John Stuart, *On Liberty*. Cambridge University Press (2012).

56 Rees (2003).

57 Mill (2012), p. 22

58 Hamilton, Alexander, James Madison, and John Jay, *The Federalist With Letters of "Brutus"*, ed. by Terence Ball. Cambridge University Press (2003), p.45. https://doi.org/10.1017/CBO9780511817816