The Vulnerable World Hypothesis

Nick Bostrom

Future of Humanity Institute, University of Oxford

Research Article

Abstract

Scientific and technological progress might change people's capabilities or incentives in ways that would destabilize civilization. For example, advances in DIY biohacking tools might make it easy for anybody with basic training in biology to kill millions; novel military technologies could trigger arms races in which whoever strikes first has a decisive advantage; or some economically advantageous process may be invented that produces disastrous negative global externalities that are hard to regulate. This paper introduces the concept of a *vulnerable world*: roughly, one in which there is some level of technological development at which civilization almost certainly gets devastated by default, i.e. unless it has exited the 'semi-anarchic default condition'. Several counterfactual historical and speculative future vulnerabilities are analyzed and arranged into a typology. A general ability to stabilize a vulnerable world would require greatly amplified capacities for preventive policing and global governance. The vulnerable world hypothesis thus offers a new perspective from which to evaluate the risk-benefit balance of developments towards ubiquitous surveillance or a unipolar world order.

Policy Implications

- Technology policy should not unquestioningly assume that all technological progress is beneficial, or that complete scientific openness is always best, or that the world has the capacity to manage any potential downside of a technology after it is invented.
- Some areas, such as synthetic biology, could produce a discovery that suddenly democratizes mass destruction, e.g. by empowering individuals to kill hundreds of millions of people using readily available materials. In order for civilization to have a general capacity to deal with "black ball" inventions of this type, it would need a system of ubiquitous real-time worldwide surveillance. In some scenarios, such a system would need to be in place before the technology is invented.
- Partial protection against a limited set of possible black balls is obtainable through more targeted interventions. For example, biorisk might be mitigated by means of background checks and monitoring of personnel in some types of biolab, by discouraging DIY biohacking (e.g. through licencing requirements), and by restructuring the biotech sector to limit access to some cutting-edge instrumentation and information. Rather than allow anybody to buy their own DNA synthesis machine, DNA synthesis could be provided as a service by a small number of closely monitored providers.
- Another, subtler, type of black ball would be one that strengthens incentives for harmful use—e.g. a military technology that makes wars more destructive while giving a greater advantage to the side that strikes first. Like a squirrel who uses the times of plenty to store up nuts for the winter, we should use times of relative peace to build stronger mechanisms for resolving international disputes.

Is there a black ball in the urn of possible inventions?

One way of looking at human creativity is as a process of pulling balls out of a giant urn.[1] The balls represent possible ideas, discoveries, technological inventions. Over the course of history, we have extracted a great many balls – mostly white (beneficial) but also various shades of gray (moderately harmful ones and mixed blessings). The cumulative effect on the human condition has so far been overwhelmingly positive, and may be much better still in the future (Bostrom, 2008). The global population has grown about three orders of magnitude over the last ten thousand years, and in the last two centuries per capita income, standards of living, and life expectancy have also risen.[2]

What we haven't extracted, so far, is a black ball: a technology that invariably or by default destroys the civilization that invents it. The reason is not that we have been particularly careful or wise in our technology policy. We have just been lucky.

It does not appear that any human civilization has been destroyed – as opposed to transformed – by its own inventions.[3] We do have examples of civilizations being destroyed by inventions made elsewhere. For example, the European inventions that enabled transoceanic travel and force projection could be regarded as a black-ball event for the indigenous populations of the Americas, Australia, Tasmania, and some other places. The extinction of archaic hominid populations, such as the Neanderthals and the Denisovans, was probably facilitated by the technological superiority of Homo sapiens. But thus far, it seems, we have seen no sufficiently auto-destructive invention to count as a black ball for humanity.[4]

What if there is a black ball in the urn? If scientific and technological research continues, we will eventually reach it and pull it out. Our civilization has a considerable ability to

Global Policy (2019) 10:4 doi: 10.1111/1758-5899.12718

pick up balls, but no ability to put them back into the urn. We can invent but we cannot un-invent. Our strategy is to hope that there is no black ball.

This paper develops some concepts that can help us think about the possibility of a technological black ball, and the different forms that such a phenomenon could take. We also discuss some implications for policy from a global perspective, particularly with respect to how one should view developments in mass surveillance and moves towards more effectual global governance or a more unipolar world order. These implications by no means settle questions about the desirability of changes in those macrostrategic variables – for there indeed are other strongly relevant factors, not covered here, which would need to be added to the balance. Yet they form an important and under-appreciated set of considerations that should be taken into account in future debates on these issues.

Before getting to the more conceptual parts of the paper, it will be useful to paint a more concrete picture of what a technological black ball could be like. The most obvious kind is a technology that would make it very easy to unleash an enormously powerful destructive force. Nuclear explosions are the most obviously destructive force we have mastered. So let us consider what would have happened if it had been very easy to unleash this force.

A thought experiment: easy nukes

On the morning of 12 September 1933, Leo Szilard was reading the newspaper when he came upon a report of an address recently delivered by the distinguished Lord Rutherford, now often considered the father of nuclear physics (Rhodes, 1986). In his speech, Rutherford had dismissed the idea of extracting useful energy from nuclear reactions as 'moonshine'. This claim so annoyed Szilard that he went out for a walk. During the walk, he got the idea of a nuclear chain reaction – the basis for both nuclear reactors and nuclear bombs. Later investigations showed that making an atomic weapon requires several kilograms of plutonium or highly enriched uranium, both of which are very difficult and expensive to produce. However, suppose it had turned out otherwise: that there had been some really easy way to unleash the energy of the atom – say, by sending an electric current through a metal object placed between two sheets of glass.

So let us consider a counterfactual history in which Szilard invents nuclear fission and realizes that a nuclear bomb could be made with a piece of glass, a metal object, and a battery arranged in a particular configuration. What happens next? Szilard becomes gravely concerned. He sees that his discovery must be kept secret at all costs. But how? His insight is bound to occur to others. He could talk to a few of his physicist friends, the ones most likely to stumble upon the idea, and try to persuade them not to publish anything on nuclear chain reactions or on any of the reasoning steps leading up to the dangerous discovery. (That is what Szilard did in actual history.)

Here Szilard faces a dilemma: either he doesn't explain the dangerous discovery, but then he will not be effective in persuading many of his colleagues to stop publishing; or he tells them the reason for his concern, but then he spreads the dangerous knowledge further. Either way he is fighting a losing battle. The general advance of scientific knowledge will eventually make the dangerous insight more accessible. Soon, figuring out how to initiate a nuclear chain reaction with pieces of metal, glass, and electricity will no longer take genius but will be within reach of any STEM student with an inventive mindset.

Let us roll the tape a little further. The situation looks hopeless, but Szilard does not give up. He decides to take a friend into his confidence, a friend who is also the world's most famous scientist – Albert Einstein. He successfully persuades Einstein of the danger (again following actual history). Now, Szilard has the support of a man who can get him a hearing with any government. The two write a letter to President Franklin D. Roosevelt. After some committee wranglings and report-writing, the top levels of the US government are eventually sufficiently convinced to be ready to take serious action.

What action can the United States take? Let us first consider what actually happened (Rhodes, 1986). What the US government did, after having digested the information provided by Einstein and Szilard, and after having received some further nudging from the British who were also looking into the matter, was to launch the Manhattan Project in order to weaponize nuclear fission as quickly as possible. As soon as the bomb was ready, the US Air Force used it to destroy Japanese population centers. Many of the Manhattan scientists had justified their participation by pointing to the mortal danger that would arise if Nazi Germany got the bomb first; but they continued working on the project after Germany was defeated.[5] Szilard advocated unsuccessfully for demonstrating 'the gadget' over an unpopulated area rather than in a city (Franck et al., 1945). After the war ended, many of the scientists favored the international control of atomic energy and became active in the nuclear disarmament movement; but their views carried little weight, as nuclear policy had been taken out of their hands. Four years later, the Soviet Union detonated its own atomic bomb. The Soviet effort was aided by spies in the Manhattan Project, yet even without espionage it would have succeeded within another year or two (Holloway, 1994). The Cold War followed, which at its peak saw 70,000 nuclear warheads ready to unleash global destruction at a moment's notice, with a trembling finger hovering over the 'red button' on either side (Norris and Kristensen, 2010).[6]

Fortunately for human civilization, after the destruction of Hiroshima and Nagasaki, no other atomic bomb has been detonated in anger. Seventy-three years later, partly thanks to international treaties and anti-proliferation efforts, only nine states possess nuclear weapons. No non-state actor is believed ever to have possessed nuclear weapons.[7]

But how would things have played out if there had been an *easy* way to make nukes? Maybe Szilard and Einstein could persuade the US government to ban all research in

nuclear physics (outside high-security government facilities)? Such a ban on basic science would be subjected to enormous legal and political challenges – the more so as the reason for the ban could not be publicly disclosed in any detail without creating an unacceptable information hazard.[8]

Let us suppose, however, that President Roosevelt could somehow mobilize enough political support to drive through a ban, and that the US Supreme Court could somehow find a way of regarding it as constitutionally valid. We then confront an array of formidable practical difficulties. All university physics departments would have to be closed, and security checks initiated. A large number of faculty and students would be forced out. Intense speculations would swirl around the reason for all these heavy-handed measures. Groups of physics PhD students and faculty banned from their research field would sit around and speculate about what the secret danger might be. Some of them would figure it out. And among those who figured it out, some would feel compelled to use the knowledge to impress their colleagues; and those colleagues would want to tell yet others, to show they were in the know. Alternatively, somebody who opposed the ban would unilaterally decide to publish the secret, maybe in order to support their view that the ban is ineffective or that the benefits of publication outweigh the risks.[9] [10] Careless or disgruntled employees at the government labs would eventually also let slip information, and spies would carry the secret to foreign capitals. Even if, by some miracle, the secret never leaked in the United States, scientists in other countries would independently discover it, thereby multiplying the sources from which it could spread. Sooner or later – probably sooner – the secret would be a secret no more.

In the present age, when one can publish instantaneously and anonymously on the Internet, it would be even more difficult to limit the spread of scientific secrets (Cf. Greenberg, 2012; Swire, 2015).

An alternative approach would be to eliminate all glass, metal, or sources of electrical current (save perhaps in a few highly guarded military depots). Given the ubiquity of these materials, such an undertaking would be extremely daunting. Securing political support for such measures would be no easier than shutting down physics education. However, after mushroom clouds had risen over a few cities, the political will to make the attempt could probably be mustered. Metal use is almost synonymous with civilization, and would not be a realistic target for elimination. Glass production could be banned, and existing glass panes confiscated; but pieces of glass would remain scattered across the landscape for a long time. Batteries and magnets could be seized, though some people would have stashed away these materials before they could be collected by the authorities. Many cities would be destroyed by nihilists, extortionists, revanchists, or even folk who just want to 'see what would happen'.[11] People would flee urban areas. In the end, many places would be destroyed by nuclear fallout, cities would be abandoned, there would be no use of electricity or glass. Possession of proscribed materials, or equipment that could be used to make them, would be harshly punished, such as by on-the-spot execution. To enforce these provisions, communities would be subjected to strict surveillance – informant networks incentivized by big rewards, frequent police raids into private quarters, continuous digital monitoring, and so forth.

That is the optimistic scenario. In a more pessimistic scenario, law and order would break down entirely and societies might split into factions waging civil wars with nuclear weapons, producing famine and pestilence. The disintegration might end only when society has been so reduced that nobody is able any longer to put together a bomb and a delay detonator from stored materials or the scrap of city ruins. Even then, the dangerous insight – once its importance had been so spectacularly demonstrated – would be remembered and passed down the generations. If civilization began to rise from the ashes, the knowledge would lie in wait, ready to pounce as soon as people learned once again how to make sheet glass and electric current generators. And even if the knowledge were forgotten, it would be rediscovered once nuclear physics research was resumed.

We were lucky that making nukes turned out to be hard.

The vulnerable world hypothesis

We now know that one cannot trigger a nuclear explosion with just a sheet of glass, some metal, and a battery. Making an atomic bomb requires several kilograms of fissile material, which is difficult to produce. We pulled out a gray ball that time. Yet with each act of invention, we reach into the urn anew.

Let us introduce the hypothesis that the urn of creativity contains at least one black ball. We can refer to this as the *vulnerable world hypothesis* (VWH). Intuitively, the hypothesis is that there is some level of technology at which civilization almost certainly gets destroyed unless quite extraordinary and historically unprecedented degrees of preventive policing and/or global governance are implemented. More precisely:

> VWH: If technological development continues then a set of capabilities will at some point be attained that make the devastation of civilization extremely likely, unless civilization sufficiently exits the semi-anarchic default condition.

By the 'semi-anarchic default condition' I mean a world order characterized by three features[12] :

1. *Limited capacity for preventive policing*. States do not have sufficiently reliable means of real-time surveillance and interception to make it virtually impossible for any individual or small group within their territory to carry out illegal actions – particularly actions that are very strongly disfavored by > 99 per cent of the population.
2. *Limited capacity for global governance*. There is no reliable mechanism for solving global coordination problems and protecting global commons – particularly in high-stakes

situations where vital national security interests are involved.

3. *Diverse motivations*. There is a wide and recognizably human distribution of motives represented by a large population of actors (at both the individual and state level) – in particular, there are *many* actors motivated, to a substantial degree, by perceived self-interest (e.g. money, power, status, comfort and convenience) and there are *some* actors ('the apocalyptic residual') who would act in ways that destroy civilization even at high cost to themselves.[3]

The term 'devastation of civilization' in the above definition could be interpreted in various ways, yielding different versions of VWH. For example, one could define an existential-risk vulnerable world hypothesis (x-VWH), which would state that at some level of technology, by default, an existential catastrophe occurs, involving the extinction of Earth-originating intelligent life or the permanent blighting of our future potential for realizing value. However, here we will set the bar lower. A key concern in the present context is whether the consequences of civilization continuing in the current semi-anarchic default condition are *catastrophic enough* to outweigh reasonable objections to the drastic developments that would be required to exit this condition. If this is the criterion, then a threshold short of human extinction or existential catastrophe would appear sufficient. For instance, even those who are highly suspicious of government surveillance would presumably favour a large increase in such surveillance *if* it were truly necessary to prevent occasional region-wide destruction. Similarly, individuals who value living in a sovereign state may reasonably prefer to live under a world government *given* the assumption that the alternative would entail something as terrible as a nuclear holocaust. Therefore, we stipulate that the term 'civilizational devastation' in VWH refers (except where otherwise specified) to any destructive event that is at least as bad as the death of 15 per cent of the world population or a reduction of global GDP by > 50 per cent per cent lasting for more than a decade.[13]

It is *not* a primary purpose of this paper to argue that VWH is true. (I regard that as an open question, though it would seem to me unreasonable, given the available evidence, to be at all confident that VWH is *false*.) Instead, the chief contribution claimed here is that VWH, along with related concepts and explanations, is useful in helping us surface important considerations and possibilities regarding humanity's macrostrategic situation. But those considerations and possibilities need to be further analyzed, and combined with other considerations that lie outside the scope of this paper, before they could deliver any definitive policy implications.

A few more clarifications before we move on. This paper uses the word 'technology' in its broadest sense. Thus, in principle, we count not only machines and physical devices but also other kinds of instrumentally efficacious templates and procedures – including scientific ideas, institutional designs, organizational techniques, ideologies, concepts, and memes – as constituting potential technological black balls.[14]

We can speak of vulnerabilities opening and closing. In the 'easy nukes' scenario, the period of vulnerability begins when the easy way of producing nuclear explosions is discovered. It ends when some level of technology is attained that makes it reasonably affordable to stop nuclear explosions from causing unacceptable damage – or that again makes it infeasible to produce nuclear explosions (because of technological regress).[15] If no protective technology is possible (as in, e.g., the case of nuclear weapons it may not be) and technological regress does not occur, then the world becomes permanently vulnerable.

We can also speak of the world being *stabilized* (with respect to some vulnerability) if the semi-anarchic default condition is exited in such a way as to prevent the vulnerability from leading to an actual catastrophe. The ways in which the semi-anarchic default condition would have to be altered in order to achieve stabilization depend on the specifics of the vulnerability in question. In a later section, we will discuss possible means by which the world could be stabilized. For now, we simply note that VWH does not imply that civilization is doomed.

Typology of vulnerabilities

We can identify four types of civilizational vulnerability.

Type-1 ('easy nukes')

The first type is one where, as in the 'easy nukes' scenario, it becomes too easy for individuals or small groups to cause mass destruction:

> *Type-1 vulnerability*: There is some technology which is so destructive and so easy to use that, given the semi-anarchic default condition, the actions of actors in the apocalyptic residual make civilizational devastation extremely likely.

Note that in determining whether a scenario presents a Type-1 vulnerability, there is an inverse relationship between *the ease* with which it becomes possible to cause an incident and *the destructiveness* of incident. The greater the destructiveness of a single incident, the less easy it needs to be to cause such an incident in order for us to diagnose the presence of a Type-1 vulnerability.

Thus, consider a '*very* easy nukes' scenario, in which any halfwit can create an easily portable thermonuclear weapon at the kitchen sink over the course of an afternoon: this would definitely qualify as a civilizational vulnerability. Contrast this with a '*moderately* easy nukes' scenario, in which it takes a five-person team of semi-skilled individuals toiling for an entire year to produce a single bulky few-kiloton device: that might not quite rise to the level of a civilizational vulnerability. It seems possible, in the 'moderately easy nukes' scenario, that the great majority of cities would escape destruction, although the threat posed by a well-

resourced terrorist organization, such as Aum Shinrikyo anno 1995 or Al-Qaeda anno 2001, would increase substantially. However, consider yet another scenario, 'moderately easy bio-doom', in which again it requires a semi-skilled five-person team working for a year to put the black-ball technology into effect, except that this time it is a biological agent, a single point release of which is sufficient to kill billions. In 'moderately easy bio-doom', the threshold for a Type-1 vulnerability *would* be reached. If destroying civilisation required only that a single group succeed with a task at the moderately-easy level, civilization would probably be destroyed within a few years in the semi-anarchic default condition. Indeed, both Aum Shinrikyo and Al-Qaeda sought to obtain nuclear and biological weapons, and would likely have chosen to use them (see e.g. Danzig et al., 2011; Olson, 1999; Mowatt-Larssen and Allison, 2010).

So a Type-1 vulnerability exists if it is either extremely easy to cause a moderate amount of harm or moderately easy to cause an extreme amount of harm.[16] The reason why a black-ball technology that enables only moderate amounts of harm per incident could count as a Type-1 vulnerability is that – if the technology is sufficiently easy to use – a large number of such incidents would be almost certain to occur. Take the scenario where it is easy for an average individual to make a metropolis-busting H-bomb. This is not necessarily a scenario in which a single individual could devastate civilization. Building hundreds of bombs and transporting them to hundreds of cities without getting caught would still be a formidable endeavor even if making a single bomb were fairly easy. The 'easy nukes' scenario nevertheless presents a civilizational vulnerability because it is plausible that there would in fact be hundreds of individuals who would each destroy at least one city under those circumstances.

That this is so almost follows from the law of large numbers combined with the plausible assumption that for any randomly selected person there is some small but appreciable chance that they would be motivated to trigger this kind of destruction – whether out of ideological hatred, nihilistic destructiveness, revenge for perceived injustices, as part of some extortion plot, or because of delusions or mental illness, or perhaps even just to see what would happen. Given the diversity of human character and circumstance, for any ever so imprudent, immoral, or self-defeating action, there is some residual fraction of humans who would choose to take that action. This is especially plausible if the action in question represents a culturally salient affordance – as it everywhere would after one such nuke attack had taken place. In other words, 'easy nukes' is an illustration of a vulnerable world because it looks like the apocalyptic residual has a large enough intersection with the set of empowered actors that one would expect a civilization-devastating amount of destruction to result.

Type-2a ('safe first strike')

A technology that 'democratizes' mass destruction is not the only kind of black ball that could be hoisted out of the urn. Another kind would be a technology that strongly incentivizes powerful actors to use their powers to cause mass destruction. Again we can turn to nuclear history for illustration.

After the invention of the atomic bomb and a short-lived American nuclear monopoly, an arms race ensued between the US and the USSR. The rival superpowers amassed staggering arsenals, topping out at 70,000 nuclear warheads in 1986, more than enough to devastate civilization (Norris and Kristensen, 2010). While public awareness of the perils of the Cold War seems to have faded since its peaceful conclusion in 1991, the academic community – benefiting from the opening of formerly classified archives and the testimony of retired policy makers, officers, and analysts – has uncovered a disconcerting array of practices and incidents which seem to have repeatedly brought the world to the brink.[17] Just how close we came remains a topic of dispute. Some scholars have argued that it was only thanks to a good deal of luck that nuclear holocaust was avoided.[18]

Whether surviving the Cold War required much luck or just a little, we can easily imagine a counterfactual in which the odds of avoiding a nuclear conflagration would be substantially worse. This holds even if we assume that nuclear weapons can be produced only by large technologically advanced states (thus distinguishing the case from the type-1 vulnerability of 'easy nukes'). The counterfactual could involve changes in the technological possibility frontier that would have made the arms race less stable.

For example, it is widely believed among nuclear strategists that the development of a reasonably secure second-strike capability by both superpowers by the mid-1960s created the conditions for 'strategic stability' (Colby and Gerson, 2013). Prior to this period, American war plans reflected a much greater inclination, in any crisis situation, to launch a preemptive nuclear strike against the Soviet Union's nuclear arsenal. The introduction of nuclear submarine-based ICBMs was thought to be particularly helpful for ensuring second-strike capabilities (and thus 'mutually assured destruction') since it was widely believed to be practically impossible for an aggressor to eliminate the adversary's boomer fleet in the initial attack.[19] Other strategies for ensuring a second-strike capability could also be employed, but they had drawbacks. For example, one option, briefly used by the United States, was to have a contingent of long-range nuclear bombers on continuous airborne alert (Sagan, 1995). This program was very costly and increased the risk of accidental or unauthorized attacks. Another option was to build hardened land-based missile silos: in sufficient numbers, these could in principle provide the assurance of a second-strike capability to one side; however, such a large arsenal would then threaten to provide the capacity of a safe first strike against the *other* side, thus again destabilizing any crisis. Road-mobile ICBM launchers, which are harder to attack than silo-based missiles, eventually provided some stabilization when they were deployed by the Soviet Union in 1985, a few years before the end of Cold War (Brower, 1989).

So consider a counterfactual in which a preemptive counterforce strike is more feasible. Imagine some technology that makes it easy to track ballistic missile submarines. We can also imagine that nuclear weapons were a bit more fragile, so that the radius within which a nuclear weapon would be destroyed by the detonation of another nuclear weapon was substantially larger than it actually is.[20] Under those circumstances, it might have been impossible to ensure a second-strike capability. Suppose, further, that technology had been such as to make it very hard to detect missile launches, rendering a launch-on-warning strategy completely unworkable. The crisis instability of the Cold War would then have been greatly amplified. Whichever side struck first would survive relatively unscathed (or might at least have *believed* that it would, since the possibility of a nuclear winter was largely ignored by war planners at the time; Badash, 2001; Ellsberg, 2017).[21] The less aggressive side would be utterly destroyed. In such a situation, mutual fear could easily trigger a dash to all-out war (Schelling, 1960).

Other technological parameter changes could similarly increase the probability of attacks. In the real world, the main 'attraction' of a nuclear first strike is that it would alleviate the fear that one might otherwise oneself become the victim of such a strike; but we can imagine a counterfactual in which there are also *benefits* to nuclear aggression, beyond the removal of a negative. Suppose it were somehow possible to derive great economic gains from initiating a large-scale nuclear assault.[22] It might be hard to see how this could be the case, yet one can imagine some automated manufacturing technology or energy technology making physical resources more valuable; or technology-enabled population growth could again make agricultural land a more vital resource (Drexler, 1986)). Some international relations scholars believe that the net economic benefits of conquest have declined substantially in the post-industrial era and that this decline has been a major contributor to peace.[23] If powerful national economic motives were again added to other causes for war (such as concern for one's own security, disputes over non-economic values, maintenance of national reputation, influence of particularly bellicose special interest groups, inter alia) then armed conflicts might become more common and large-scale nuclear war more likely.

In these examples, the vulnerability arises not from destruction getting easier, but from the actions leading to destruction coming to be supported by stronger incentives. We shall call these Type-2 vulnerabilities. Specifically, a scenario like 'safe first strike', in which some enormously destructive action becomes incentivized, we shall refer to as Type-2a:

> *Type-2a vulnerability*: There is some level of technology at which powerful actors have the ability to produce civilization-devastating harms and, in the semi-anarchic default condition, face incentives to use that ability.

We will see some more examples of Type-2a vulnerabilities below, where the 'civilization-devastating harms' take the form of risk externalities.

Type-2b ('worse global warming')

There is yet another way in which the world could be vulnerable; one that we can illustrate with a counterfactual related to climate change.

In the real world, we observe a secular rise in global mean temperature, widely believed to be driven primarily by human-caused emissions of greenhouse gases such as carbon dioxide, methane, and nitrous oxide (Stocker et al., 2014).Projections vary, depending on the emissions scenario and modelling assumptions, but forecasts that imply an average temperature rise of between 3° C and 4.5° C in 2100 (compared to 2000), in the absence of any significant action to reduce emissions, are quite typical (See Stocker et al. (2014, table 12.2)). The effects of such warming – on sea levels, weather patterns, ecosystems, and agriculture – are usually expected to be net negative for human welfare (See Field et al. (2014, figure 10-1)). Greenhouse gases are emitted by wide range of activities, including in industry, transport, agriculture, and electricity production, and from all around the world, though especially from industrialized or industrializing countries. Efforts to curb emissions have so far failed to achieve much global-scale impact (Friedlingstein et al., 2014)).

Now, we could imagine a situation in which the problem of global warming would be far more dire than it actually seems to be. For example, the transient climate sensitivity (a measure of the medium-term change in mean global surface temperature of the Earth that results from some kind of forcing, such as a doubling of atmospheric CO_2) could have turned out to be much greater than it is (Shindell, 2014). If it had been several times larger than its actual value, we would have been in for a temperature rise of, say, 15° or 20° C instead of 3° – a prospect with far greater civilization-destroying potential than the actual expectation.[24]

We can also imagine other deviations from reality that would have made global warming a worse problem. Fossil fuels could have been even more abundant than they are, and available in more cheaply exploitable deposits, which would have encouraged greater consumption. At the same time, clean energy alternatives could have been more expensive and technologically challenging. Global warming could also have been a worse problem if there were stronger positive feedback loops and nonlinearities, such as an initial phase in which the atmosphere is gradually loaded up with greenhouse gases without much observable or detrimental effect, followed by a second phase in which temperatures shoot up abruptly. To get a truly civilizational threat from global warming, it may also be necessary to stipulate, counterfactually, that mitigation through geoengineering is infeasible.

The vulnerability illustrated by such a 'worse global warming' scenario is different from that of a Type-2a scenario like

'safe first strike'. In a Type-2a vulnerability, some actor has the ability to take some action – such as launching a nuclear first strike – that is destructive enough to devastate civilization. In the 'worse global warming' scenario, no such actor need exist. Instead, in what we will call a Type-2b vulnerability, there is a large number of individually insignificant actors who is each incentivized (under the semi-anarchic default condition) to take some action that contributes slightly to what cumulatively becomes a civilization-devastating problem:

> *Type-2b vulnerability*: There is some level of technology at which, in the semi-anarchic default condition, a great many actors face incentives to take some slightly damaging action such that the combined effect of those actions is civilizational devastation.

What Type-2a and Type-2b have in common is that, in both cases, the damage-capable actors face incentives that would encourage a wide range of normally motivated actors in their situation to pursue the course of action that leads to damage. Global warming would not be a problem if only some small fraction of those actors who can drive cars or chop down a few trees chose to do so; the problem arises only because *many* actors make these choices. And in order for many actors to make those choices, the choices must be supported by incentives that have wide appeal (such as money, status, and convenience). Similarly, if only one in a million actors who could launch a nuclear first strike would actually choose to do so, then it would not be so alarming if there are a handful of actors possessing that capability; but it does get worrisome if launching a nuclear strike is strongly supported by incentives that appeal to normally-motivated actors (such as the motive of preempting a strike by one's adversary). This is in contrast to a Type-1 vulnerability, where the problem arises from the very widespread proliferation of destructive capability. Only an actor with quite unusual values would choose, at great cost and risk to himself, to blow up a city or unleash a doomsday pathogen; the trouble in that case is that if sufficiently many actors possess such a capability, then the subset of them who also have apocalyptic motives is not empty.

Type-0 ('surprising strangelets')

In 1942, it occurred to Edward Teller, one of the Manhattan scientists, that a nuclear explosion would create a temperature unprecedented in Earth's history, producing conditions similar to those in the center of the sun, and that this could conceivably trigger a self-sustaining thermonuclear reaction in the surrounding air or water Rhodes, 1986). The importance of Teller's concern was immediately recognized by Robert Oppenheimer, the head of the Los Alamos lab. Oppenheimer notified his superior and ordered further calculations to investigate the possibility. These calculations indicated that atmospheric ignition would not occur. This prediction was confirmed in 1945 by the Trinity test, which involved the detonation of the world's first nuclear explosive.[25]

In 1954, the US carried out another nuclear test, the Castle Bravo test, which was planned as a secret experiment with an early lithium-based thermonuclear bomb design. Lithium, like uranium, has two important isotopes: lithium-6 and lithium-7. Ahead of the test, the nuclear scientists calculated the yield to be 6 megatons (with an uncertainty range of 4–8 megatons). They assumed that only the lithium-6 would contribute to the reaction, but they were wrong. The lithium-7 contributed more energy than the lithium-6, and the bomb detonated with a yield of 15 megaton – more than double of what they had calculated (and equivalent to about 1,000 Hiroshimas). The unexpectedly powerful blast destroyed much of the test equipment. Radioactive fallout poisoned the inhabitants of downwind islands and the crew of a Japanese fishing boat, causing an international incident.

We may regard it at as lucky that it was the Castle Bravo calculation that was incorrect, and not the calculation of whether the Trinity test would ignite the atmosphere. Counterfactually, if the atmosphere had been susceptible to ignition by a nuclear detonation, and if this fact had been relatively easy to overlook – let us say as easy as it was to overlook the contribution of the lithium-7 in the Castle Bravo test – then the human story (and that of all terrestrial life) would have come to an end in 1945. We can call this scenario 'Castle Bravissimo'.

Whenever we pull a ball from the urn of invention, there could conceivably be a possibility of accidental devastation. Usually, this risk is negligible; but in some cases it could be significant, especially when the technology in question generates some kind of novel perturbation of nature or introduces historically unprecedented conditions. This suggests that we should add to our typology one more category, that of technology-fated accidental civilizational devastation:

> *Type-0 vulnerability*: There is some technology that carries a hidden risk such that the default outcome when it is discovered is inadvertent civilizational devastation.[26]

It is instructive to note, however, that 'Castle Bravissimo' is *not* a perfect illustration of a Type-0 vulnerability. Suppose that careful calculations had shown that there was a 1 per cent probability that a nuclear detonation would ignite the atmosphere and the oceans and thereby extinguish life on Earth. Suppose, further, that it had been known that to resolve the matter further and prove that the chance was zero (or alternatively, that the chance was one) would take another 10 years of meticulous study. It is unclear, under those circumstances, what the leaders of the Manhattan project would have decided. They would presumably have thought it greatly desirable that humanity hold off on developing nuclear weapons for at least another 10 years.[27] On the other hand, they would have feared that Germany might have an advanced bomb project and that Hitler maybe would not pull the brakes because of a 1 per cent risk of destroying the world.[28] They might have concluded that the risk of testing a nuclear bomb was worth taking in order to reduce the probability of Nazi Germany ending up with a nuclear monopoly.

In this version of 'Castle Bravissimo', civilization gets blown up by accident: nobody sought to cause a destructive event. Yet the key actors were locked in a strategic situation that incentivized them to proceed despite the risk. In this respect, the scenario fits as a Type-2a vulnerability; only, the civilization-devastating harm it involves is probabilistic. When nuclear technology becomes possible, powerful actors face incentives, in the semi-anarchic default condition, to use that technology in ways that produce civilization-destroying harms (which here take the form of risk externalities).[29]

Accordingly, in order for us to diagnose a Type-0 vulnerability, we require that a stronger condition be met than merely that the key actors did not intend destruction. We stipulate that 'inadvertent' should here mean that the adverse outcome sprang from bad luck, not coordination failure. In a Type-0 vulnerability, the key actors would, even if they were adequately coordinated, decide to proceed with using the technology, in the belief that the benefits would outweigh costs – but they would be wrong, and the costs would be larger than expected, enough so as to cause civilizational devastation.[30]

Since 'Castle Bravissimo' only ambiguously satisfies this criterion (it being unclear in the original counterfactual to what extent the disaster would have resulted from coordination failure and to what extent from miscalculation/bad luck), it may be useful to introduce a cleaner example of a Type-0 vulnerability. Thus, consider a 'surprising strangelets' scenario in which some modern high-energy physics experiment turns out to initiate a self-catalyzing process in which ordinary matter gets converted into strange matter, with the result that our planet is destroyed. This scenario, and variations thereof in which accelerator experiments generate stable black holes or trigger the decay of a metastable vacuum state, have been analyzed in the literature (Jaffe et al., 2000; Tegmark and Bostrom, 2005). Such outcomes would indeed be very surprising, since analysis indicates that they have a completely negligible chance of occurring. Of course, with *sufficiently* bad luck, a negligible-chance event could occur. But alternatively (and far more likely in this case), the analysis could have a hidden flaw, like the Castle Bravo calculations did; in which case the chance might not be so negligible after all (Ord et al., 2010).[31]

Achieving stabilization

The truth of VWH would be bad news. But it would not imply that civilization will be devastated. In principle at least, there are several responses that could stabilize the world even if vulnerability exists. Recall that we defined the hypothesis in terms of a black-ball technology making civilizational devastation extremely likely *conditional on technological development continuing and the semi-anarchic default condition persisting*. Thus we can theoretically consider the following possibilities for achieving stabilization:

1. Restrict technological development.
2. Ensure that there does not exist a large population of actors representing a wide and recognizably human distribution of motives.
3. Establish extremely effective preventive policing.
4. Establish effective global governance.

We will discuss (3) and (4) in subsequent sections. Here we consider (1) and (2). We will argue they hold only limited promise as ways of protecting against potential civilizational vulnerabilities.

Technological relinquishment

In its general form, technological relinquishment looks exceedingly unpromising. Recall that we construed the word 'technology' broadly; so that completely stopping technological development would require something close to a cessation of inventive activity everywhere in the world. That is hardly realistic; and if it could be done, it would be extremely costly – to the point of constituting an existential catastrophe in its own right (Namely, 'permanent stagnation' (Bostrom, 2013)).

That general relinquishment of scientific and technological research is a non-starter does not, however, imply that *limited* curtailments of inventive activities could not be a good idea. It can make sense to forego particularly perilous directions of advancement. For instance, recalling our 'easy nukes' scenario, it would be sensible to discourage research into laser isotope separation for uranium enrichment (Kemp, 2012). Any technology that makes it possible to produce weapons-grade fissile material using less energy or with a smaller industrial footprint would erode important barriers to proliferation. It is hard to see how a slight reduction in the price of nuclear energy would compensate. On the contrary, the world would probably be better off if it somehow became *harder* and *more expensive* to enrich uranium. What we would ideally want in this area is not technological progress but technological *regress*.

While targeted regress might not be in the cards, we could aim to slow the rate of advancement towards risk-increasing technologies relative to the rate of advancement in protective technologies. This is the idea expressed by the principle of differential technological development. In its original formulation, the principle focuses on existential risk; but we can apply it more broadly to also encompass technologies with 'merely' devastational potential:

> *Principle of Differential Technological Development*. Retard the development of dangerous and harmful technologies, especially ones that raise the level of existential risk; and accelerate the development of beneficial technologies, especially those that reduce the existential risks posed by nature or by other technologies Bostrom, 2002).

The principle of differential technological development is compatible with plausible forms of technological determinism. For example, even if it were ordained that all technologies that *can* be developed *will* be developed, it can still matter *when* they are developed. The order in which they arrive can make an important difference – ideally, protective technologies should come before the destructive

technologies against which they protect; or, if that is not possible, then it is desirable that the gap be minimized so that other countermeasures (or luck) may tide us over until robust protection become available. The timing of an invention also influences what sociopolitical context the technology is born into. For example, if we believe that there is a secular trend toward civilization becoming more capable of handling black balls, then we may want to delay the most risky technological developments, or at least abstain from accelerating them. Even if we suppose that civilizational devastation is unavoidable, many would prefer it to take place further into the future, at a time when maybe they and their loved ones are no longer alive anyway.[32]

Differential technological development doesn't really make sense in the original urn-of-creativity model, where the color of each ball comes as a complete surprise. If we want to use the urn model in this context, we must modify it. We could stipulate, for example, that the balls have different textures and that there is a correlation between texture and color, so that we get clues about the color of a ball before we extract it. Another way to make the metaphor more realistic is to imagine that there are strings or elastic bands between some of the balls, so that when we pull on one of them we drag along several others to which it is linked. Presumably the urn is highly tubular, since certain technologies must emerge before others can be reached (we are not likely to find a society that uses jet planes and flint axes). The metaphor would also become more realistic if we imagine that there is not just one hand daintily exploring the urn: instead, picture a throng of scuffling prospectors reaching in their arms in hopes of gold and glory, and citations.

Correctly implementing differential technological development is clearly a difficult strategic task (Cf. Collingridge, 1980). Nevertheless, for an actor who cares altruistically about long-term outcomes and who is involved in some inventive enterprise (e.g. as a researcher, funder, entrepreneur, regulator, or legislator) it is worth making the attempt. Some implications, at any rate, seem fairly obvious: for instance, don't work on laser isotope separation, don't work on bioweapons, and don't develop forms of geoengineering that would empower random individuals to unilaterally make drastic alterations to the Earth's climate. Think twice before accelerating enabling technologies – such as DNA synthesis machines – that would directly facilitate such ominous developments.[33] But boost technologies that are predominantly protective; for instance, ones that enable more efficient monitoring of disease outbreaks or that make it easier to detect covert WMD programs.

Even if it is the case that all possible 'bad' technologies are bound to be developed eventually, it can still be helpful to buy a little time.[34] However, differential technological development does not on its own offer a solution for vulnerabilities that persist over long periods – ones where adequately protective technologies are much harder to develop than their destructive counterparts, or where destruction has the advantage even at technological maturity.[35]

Preference modification

Another theoretically possible way of achieving civilizational stabilization would be to change the fact that there exists a large population of actors representing a wide and recognizably human distribution of motives. We reserve for later discussion of interventions that would reduce the effective number of independent actors by increasing various forms of coordination. Here we consider the possibility of modifying the distribution of preferences (within a more or less constant population of actors).

The degree to which this approach holds promise depends on which type of vulnerability we have in mind.

In the case of a Type-1 vulnerability, preference modification does not look promising, at least in the absence of extremely effective means for doing so. Consider that some Type-1 vulnerabilities would result in civilizational devastation if there is even a single empowered person anywhere in the world who is motivated to pursue the destructive outcome. With that kind of vulnerability, reducing the number of people in the apocalyptic residual would do nothing to forestall devastation unless the number could be reduced all the way to zero, which may be completely infeasible. It is true that there are other possible Type-1 vulnerabilities that would require a somewhat larger apocalyptic residual in order for civilizational devastation to occur: for example, in a scenario like 'easy nukes', maybe there would have to be somebody from the apocalyptic residual in each of several hundred cities. But this is still a very low bar. It is difficult to imagine an intervention – short of radically re-engineering human nature on a fully global scale – that would sufficiently deplete the apocalyptic residual to entirely eliminate or even greatly reduce the threat of Type-1 vulnerabilities.

Note that an intervention that halves the size of the apocalyptic residual would *not* (at least not through any first-order effect) reduce the expected risk from Type-1 vulnerabilities by anywhere near as much. A reduction of 5 per cent or 10 per cent of Type-1 risk from halving the apocalyptic residual would be more plausible. The reason is that there is wide uncertainty about how destructive some new black-ball technology would be, and we should arguably use a fairly uniform prior in log space (over several orders of magnitude) over the size of apocalyptic residual that would be required in order for civilizational devastation to occur conditional on a Type-1 vulnerability arising. In other words, conditional on some new technology being developed that makes it easy for an average individual to kill at least one million people, it may be (roughly) as likely that the technology would enable the average individual to kill one million people, ten million people, a hundred million people, a billion people, or every human alive.

These considerations notwithstanding, preference modification could be helpful in scenarios in which the set of empowered actors is initially limited to some small definable subpopulation. Some black-ball technologies, when they first emerge from the urn, might be difficult to use and require specialized equipment. There could be a period of several years before such a technology has been perfected to the

point where an average individual could master it. During this early period, the set of empowered actors could be quite limited; for example, it might consist exclusively of individuals with bioscience expertise working in a particular type of lab. Closer screening of applicants to positions in such labs could then make a meaningful dent in the risk that a destructive individual gains access to the biotech black ball within the first few years of its emergence.[36] And that reprieve may offer an opportunity to introduce other countermeasures to provide more lasting stabilization, in anticipation of the time when the technology gets easy enough to use that it diffuses to a wider population.

For Type-2a vulnerabilities, the set of empowered actors is much smaller. Typically what we are dealing with here are states, perhaps alongside a few especially powerful non-state actors. In some Type-2a scenarios, the set might consist exclusively of two superpowers, or a handful of states with special capabilities (as is currently the case with nuclear weapons). It could thus be very helpful if the preferences of even a few powerful states were shifted in a more peace-loving direction. The 'safe first strike' scenario would be a lot less alarming if the actors facing the security dilemma had attitudes towards one another similar to those prevailing between Finland and Sweden. For many plausible sets of incentives that could arise for powerful actors as a consequence of some technological breakthrough, the prospects for a non-devastational outcome would be significantly brightened if the actors in question had more irenic dispositions. Although this seems difficult to achieve, it is not as difficult as persuading almost all the members in the apocalyptic residual to alter their dispositions.

Lastly, consider Type-2b. Recall that such a vulnerability entails that 'by default' a great many actors face incentives to take some damaging action, such that the combined effects add up to civilizational devastation. The incentives for using the black-ball technology must therefore be ones that have a grip on a substantial fraction of the world population – economic gain being perhaps being the prime example of such a near-universal motivation. So imagine some private action, available to almost every individual, which saves each person who takes it a fraction X of his or her annual income, while producing a negative externality such that if half the world's population takes the action then civilization gets devastated. At $X = 0$, we can assume that few people would take the antisocial action. But the greater X is, the larger the fraction of the population that would succumb to temptation. Unfortunately, it is plausible that the value of X that would induce at least half of the population to take the action is small, perhaps less than 1 per cent.[37] While it would be desirable to change the distribution of global preferences so as to make people more altruistic and raise the value of X, this seems difficult to achieve. (Consider the many strong forces already competing for hearts and minds – corporate advertisers, religious organizations, social movements, education systems, and so on.) Even a dramatic increase in the amount of altruism in the world – corresponding, let us say, to a doubling of X from 1 per cent to 2 per cent – would prevent calamity only in a relatively narrow band of scenarios, namely those in which the private benefit of using the destructive technology is in the 1–2 per cent range. Scenarios in which the private gain exceeds 2 per cent would still result in civilizational devastation.

In sum, modifying the distribution of preferences within the set of actors that would be destructively empowered by a black-ball discovery could be a useful adjunct to other means of stabilization, but it can be difficult to implement and would at best offer only very partial protection (unless we assume extreme forms of worldwide re-engineering of human nature).[38]

Some specific countermeasures and their limitations

Beside influencing the direction of scientific and technological progress, or altering destruction-related preferences, there are a variety of other possible countermeasures that could mitigate a civilizational vulnerability. For example, one could try to:

- prevent the dangerous information from spreading;
- restrict access to requisite materials, instruments, and infrastructure;
- deter potential evildoers by increasing the chance of their getting caught;
- be more cautious and do more risk assessment work; and
- establish some kind of surveillance and enforcement mechanism that would make it possible to interdict attempts to carry out a destructive act

It should be clear from our earlier discussion and examples that the first four of these are not general solutions. Preventing information from spreading could easily be infeasible. Even if it could be done, it would not prevent the dangerous information from being independently rediscovered. Censorship seems to be at best a stopgap measure.[39] Restricting access to materials, instruments, and infrastructure is a great way to mitigate *some* kinds of (gray-ball) threats, but it is unavailing for other kinds of threats – such as ones in which the requisite ingredients are needed in too many places in the economy or are already ubiquitously available when the dangerous idea is discovered (such as glass, metal, and batteries in the 'easy nukes' scenario). Deterring potential evildoers makes good sense; but for sufficiently destructive technologies, the existence of an apocalyptic residual renders deterrence inadequate even if every perpetrator were certain to get caught.

Exercising more caution and doing more risk assessment is also a weak and limited strategy. One actor unilaterally deciding to be more cautious may not help much with respect to a Type-2a vulnerability, and would do basically nothing for one of Type-2b or Type-1. In the case of a Type-0 vulnerability, it could help if the pivotal actor were more cautious – though only if the first cautiously tiptoeing actor were not followed by an onrush of incautious actors getting access to the same risky technology (unless the world had somehow, in the interim, been stabilized by other means).[40]

And as for risk assessment, it could lower the risk only if it led to some other countermeasure being implemented.[41]

The last countermeasure in the list – surveillance – does point towards a more general solution. We will discuss it in the next section under the heading of 'preventive policing'. But we can already note that on its own it is not sufficient. For example, consider a Type-2b vulnerability such as 'worse global warming'. Even if surveillance made it possible for a state to perfectly enforce any environmental regulation it chooses to impose, there is still the problem of getting a sufficient plurality of states to agree to adopt the requisite regulation – something which could easily fail to happen. The limitations of surveillance are even more evident in the case of Type-2a vulnerability, such as 'safe first strike', where the problem is that states (or other powerful actors) are strongly incentivized to perform destructive acts. The ability of those states to perfectly control what goes on within their own borders does not solve this problem. What is needed to reliably solve problems that involve challenges of international coordination, is effective global governance.

Governance gaps

The limitations of technological relinquishment, preference modification, and various specific countermeasures as responses to a potential civilizational vulnerability should now be clear. To the extent, therefore, that we are concerned that VWH may be true, we must consider the remaining two possible ways of achieving stabilization:

1. *Create the capacity for extremely effective preventive policing*. Develop the intra-state governance capacity needed to prevent, with extremely high reliability, any individual or small group – including ones that cannot be deterred – from carrying out any action that is highly illegal; and
2. *Create the capacity for strong global governance*. Develop the inter-state governance capacity needed to reliably solve the most serious global commons problems and ensure robust cooperation between states (and other strong organizations) wherever vital security interests are at stake – even where there are very strong incentives to defect from agreements or refuse to sign on in the first place.

The two governance gaps reflected by (1) and (2), one at the micro-scale, the other at the macro-scale, are two Achilles' heels of the contemporary world order. So long as they remain unprotected, civilization remains vulnerable to a potential technological black ball that would enable a strike to be directed there. Unless and until such a discovery emerges from the urn, it is easy to overlook how exposed we are.

In the following two sections, we will discuss how filling in these governance gaps is necessary to achieve a general ability to stabilize potential civilizational vulnerabilities. It goes without saying that there are great difficulties, and also very serious potential downsides, in seeking progress towards (1) and (2). In this paper, we will say little about the difficulties and almost nothing about the potential downsides – in part because these are already rather well known and widely appreciated. However, we emphasize that the lack of discussion about arguments against (1) and (2) should not be interpreted as an implicit assertion that these arguments are weak or that they do not point to important concerns. They would, of course, have to be taken into account in an all-things-considered evaluation. But such an evaluation is beyond the scope of the present contribution, which focuses specifically on considerations flowing from VWH.

Preventive policing

Suppose that a Type-1 vulnerability opens up. Somebody discovers a really easy way to cause mass destruction. Information about the discovery spreads. The requisite materials and instruments are ubiquitously available and cannot quickly be removed from circulation. Of course it is highly illegal for any non-state actor to destroy a city, and anybody caught doing so would be subject to harsh penalties. But it is plausible that more than one person in a million belongs to an undeterrable apocalyptic residual. Though small in relative terms, if each such person creates a city-destroying event, the absolute number is still too large for civilization to endure. So what to do?

If we suddenly found ourselves in such a situation, it may be too late to prevent civilization from being destroyed. However, it is possible to envisage scenarios in which human society would survive such a challenge intact – and the even harder challenge where individuals can single-handedly destroy not just one city but the entire world.

What would be required to stabilize such vulnerabilities is an *extremely* well-developed preventive policing capacity. States would need the ability to monitor their citizens closely enough to allow them to intercept anybody who begins preparing an act of mass destruction.

The feasibility of such surveillance and interception depend on the specifics of the scenario: How long does it take to deploy the black-ball technology destructively? how observable are the actions involved? can they be distinguished from behavior that we don't want to prohibit? But it is plausible that a considerable chunk of the Type-1 vulnerability spectrum could be stabilized by a state that deploys currently available technologies to the fullest extent. And expected advances in surveillance technology will greatly expand the achievable protection.

For a picture of what a really intensive level of surveillance could look like, consider the following vignette:

> *High-tech Panopticon*
>
> Everybody is fitted with a 'freedom tag' – a sequent to the more limited wearable surveillance devices familiar today, such as the ankle tag used in several countries as a prison alternative, the bodycams worn by many police forces, the pocket trackers and wristbands that some parents use to keep track of their children, and, of course, the ubiquitous cell phone (which has been

characterized as 'a personal tracking device that can also be used to make calls').[42] The freedom tag is a slightly more advanced appliance, worn around the neck and bedecked with multidirectional cameras and microphones. Encrypted video and audio is continuously uploaded from the device to the cloud and machine-interpreted in real time. AI algorithms classify the activities of the wearer, his hand movements, nearby objects, and other situational cues. If suspicious activity is detected, the feed is relayed to one of several patriot monitoring stations. These are vast office complexes, staffed 24/7. There, a freedom officer reviews the video feed on several screens and listens to the audio in headphones. The freedom officer then determines an appropriate action, such as contacting the tag-wearer via an audiolink to ask for explanations or to request a better view. The freedom officer can also dispatch an inspector, a police rapid response unit, or a drone to investigate further. In the small fraction of cases where the wearer refuses to desist from the proscribed activity after repeated warnings, an arrest may be made or other suitable penalties imposed. Citizens are not permitted to remove the freedom tag, except while they are in environments that have been outfitted with adequate external sensors (which however includes most indoor environments and motor vehicles). The system offers fairly sophisticated privacy protections, such as automated blurring of intimate body parts, and it provides the option to redact identity-revealing data such as faces and name tags and release it only when the information is needed for an investigation. Both AI-enabled mechanisms and human oversight closely monitor all the actions of the freedom officers to prevent abuse.[43]

Creating and operating the High-tech Panopticon would require substantial investment, but thanks to the falling price of cameras, data transmission, storage, and computing, and the rapid advances in AI-enabled content analysis, it may soon become both technologically feasible and affordable. For example, if the cost of applying this to one individual for 1 year falls to around US$140, then the entire world population could be continuously monitored at a cost of less than 1 per cent of world GDP. At that price, the system would plausibly represent a net saving – even setting aside its use in preventing civilization-scale cataclysms – because of its utility for regular law enforcement. If the system works as advertised, many forms of crime could be nearly eliminated, with concomitant reductions in costs of policing, courts, prisons, and other security systems. It might also generate growth in many beneficial cultural practices that are currently inhibited by a lack of social trust.

If the technical barriers to High-tech Panopticon are rapidly coming down, how about its political feasibility? One possibility is that society gradually drifts towards total social transparency even absent any big shock to the system. It may simply become progressively easier to collect and analyze information about people and objects, and it may prove quite convenient to allow that to be done, to the point where eventually something close to full surveillance becomes a reality – close enough that with just one more turn of the screw it can be turned into High-tech Panopticon.[44] An alternative possibility is that some particular Type-1 vulnerability comes sufficiently starkly into view to scare states into taking extreme measures, such as launching a crash program to create universal surveillance. Other extreme measures that could be attempted in the absence of a fully universal monitoring system might include adopting a policy of preemptive incarceration, say whenever some set of unreliable indicators suggest a greater than1 per cent probability that some individual will attempt a city-destroying act or worse.[45] Political attitudes to such policies would depend on many factors, including cultural traditions and norms about privacy and social control; but they would also depend on how clearly the civilizational vulnerability was perceived. At least in the case of vulnerabilities for which there are several spectacular warning shots, it is plausible that the risk would be perceived very clearly. In the 'easy nukes' scenario, for example, after the ruination of a few great cities, there would likely be strong public support for a policy which, for the sake of forestalling another attack, would involve incarcerating a hundred innocent people for every genuine plotter.[46] In such a scenario, the creation of a High-tech Panopticon would probably be widely supported as an overwhelmingly urgent priority. However, for vulnerabilities not preceded or accompanied by such incontrovertible evidence, the will to robust preventive action may never materialize.

Extremely effective preventive policing, enabled by ubiquitous real-time surveillance, may thus be necessary to stabilize a Type-1 vulnerability. Surveillance is also relevant to some other types of vulnerability, although not so centrally as in the case of Type-1.

In a Type-2b vulnerability, the bad outcome is brought about by the combined actions of a mass of independent actors who are incentivized to behave destructively. But unless the destructive behaviours are very hard to observe, intensification of surveillance or preventive policing would not be needed to achieve stabilization. In 'worse global warming', for instance, it is not essential that individual actions be preempted. Dangerous levels of emissions take time to accumulate, and polluters can be held accountable after the fact; and it is tolerable if a few of them slip through the cracks.

For other Type-2b vulnerabilities, however, enhanced methods of surveillance and social control could be important. Consider 'runaway mob', a scenario in which a mob forms that kills anybody it comes into contact with who refuses to join, and which grows ever bigger and more formidable (Cf. Munz et al., 2009). The ease with which such bad social equilibria can form and propagate, the feasibility of reforming them once they have taken hold, and the toll they exact on human welfare, depend on parameters that could be changed by technological innovations, potentially

for the worse. Even today, many states struggle to subdue organized crime. A black-ball invention (perhaps some clever cryptoeconomic mechanism design) that makes criminal enterprises much more scalable or more damaging in their social effects might create a vulnerability that could only be stabilized if states possessed unprecedented technological powers of surveillance and social control.

As regards to Type-2a vulnerabilities, where the problem arises from the incentives facing state powers or other mighty actors, it is less clear how domestic surveillance could help. Historically, stronger means for social control may even have worsened inter-state conflict – the bloodiest inter-state conflicts have depended on the highly effective governance capacities of the modern state, for tax collection, conscription, and war propaganda. It is conceivable that improved surveillance could indirectly facilitate the stabilization of a Type-2a vulnerability, such as by changing sociocultural dynamics or creating new options for making arms-reduction treaties or non-aggression pacts more verifiable. But it seems equally plausible that the net effect of strengthened domestic surveillance and policing powers on Type-2a vulnerabilities would, in the absence of reliable mechanisms for resolving international disputes, be in the opposite direction (i.e. tending to produce or exacerbate such vulnerabilities rather than to stabilize them).

Global governance

Consider again 'safe first strike': states with access to the black-ball technology by default face strong incentives to use it destructively even though it would be better for everybody that no state did so. The original example involved a counterfactual with nuclear weapons, but looking to the future we might get this kind of black ball from advances in biological weapons, or atomically precise manufacturing, or the creation of vast swarms of killer drones, or artificial intelligence, or something else. The set of state actors then confronts a collective action problem. Failure to solve this problem means that civilization gets devastated in a nuclear Armageddon or another comparable disaster. It is plausible that, absent effective global governance, states would in fact fail to solve this problem. By assumption, the problem confronting us here presents special challenges; yet states have frequently failed to solve *easier* collective action problems. Human history is covered head to foot with the pockmarks of war.

With effective global governance, however, the solution becomes trivial: simply prohibit all states from wielding the black-ball technology destructively. In the case of 'safe first strike', the most obvious way to do this would be by ordering that all nuclear weapons be dismantled and an inspection regime set up, with whatever level of intrusiveness is necessary to guarantee that nobody recreates a nuclear capability. Alternatively, the global governance institution itself could retain an arsenal of nuclear weapons as a buffer against any breakout attempt.

To deal with Type-2a vulnerabilities, what civilization requires is a robust ability to achieve global coordination, specifically in matters where state actions have extremely large externalities. Effective global governance would also help with those Type-1 and Type-2b scenarios where some states are reluctant to institute the kind of preventive policing that would be needed to reliably prevent individuals within their territories from carrying out a destructive act.

Consider a biotechnological black ball that is powerful enough that a single malicious use could cause a pandemic that would kill billions of people, thus presenting a Type-1 vulnerability. It would be unacceptable if even a single state fails to put in place the machinery necessary for continuous surveillance and control of its citizens (or whatever other mechanisms are necessary to prevent malicious use with virtually perfect reliability). A state that refuses to implement the requisite safeguards – perhaps on grounds that it values personal freedom too highly or accords citizens a constitutionally inscribed right to privacy – would be a delinquent member of the international community. Such a state, even if its governance institutions functioned admirably in other respects, would be analogous to a 'failed state' whose internal lack of control makes it a safe haven for pirates and international terrorists (though of course in the present case the risk externality it would be imposing on the rest of the world would be far larger). Other states certainly would have ground for complaint.

A similar argument applies to Type-2b vulnerabilities, such as a 'worse global warming' scenario in which some states are inclined to free-ride on the costly efforts of others to cut emissions. An effective global governance institution could compel every state to do its part.

We thus see that while some possible vulnerabilities can be stabilized with preventive policing alone, and some other vulnerabilities can be stabilized with global governance alone, there are some that would require both. Extremely effective preventive policing would be required because individuals can engage in hard-to-regulate activities that must nevertheless be effectively regulated, and strong global governance would be required because states may have incentives *not* to effectively regulate those activities even if they have the capability to do so. In combination, however, ubiquitous-surveillance-powered preventive policing and effective global governance would be sufficient to stabilize most vulnerabilities, making it safe to continue scientific and technological development even if VWH is true.[47]

Discussion

Comprehensive surveillance and global governance would thus offer protection against a wide spectrum of civilizational vulnerabilities. This is a considerable reason in favor of bringing about those conditions. The strength of this reason is roughly proportional to the probability that the vulnerable world hypothesis is true.

It goes without saying that a mechanism that enables unprecedentedly intense forms of surveillance, or a global governance institution capable of imposing its will on any nation, could also have bad consequences. Improved capabilities for social control could help despotic regimes protect themselves from rebellion. Ubiquitous surveillance could

enable a hegemonic ideology or an intolerant majority view to impose itself on all aspects of life, preventing individuals with deviant lifestyles or unpopular beliefs from finding refuge in anonymity. And if people believe that everything they say and do is, effectively, 'on the record', they might become more guarded and blandly conventional, sticking closely to a standard script of politically correct attitudes and behaviours rather than daring to say or do anything provocative that would risk making them the target of an outrage mob or putting an indelible disqualifying mark on their résumé. Global governance, for its part, could reduce beneficial forms of inter-state competition and diversity, creating a world order with single point of failure: if a world government ever gets captured by a sufficiently pernicious ideology or special interest group, it could be game over for political progress, since the incumbent regime might never allow experiments with alternatives that could reveal that there is a better way. Also, being even further removed from individuals and culturally cohesive 'peoples' than are typical state governments, such an institution might by some be perceived as less legitimate, and it may be more susceptible to agency problems such as bureaucratic sclerosis or political drift away from the public interest.[48]

It also goes without saying that stronger surveillance and global governance could have various good consequences aside from stabilizing civilizational vulnerabilities (see also Re, 2016)) ; Bostrom, 2006; cf. Torres, 2018)). More effective methods of social control could reduce crime and alleviate the need for harsh criminal penalties. They could foster a climate of trust that enables beneficial new forms of social interaction and economic activity to flourish. Global governance could prevent interstate wars, including ones that do not threaten civilizational devastation, and reduce military expenditures, promote trade, solve various global environmental and other commons problems, calm nationalistic hatreds and fears, and over time perhaps would foster an enlarged sense of cosmopolitan solidarity. It may also cause increased social transfers to the global poor, which some would view as desirable.

Clearly, there are weighty arguments both for and against moving in these directions. This paper offers no judgment about the overall balance of these arguments. The ambition here is more limited: to provide a framework for thinking about potential technology-driven civilizational vulnerabilities, and to point out that greatly expanded capacities for preventive policing and global governance would be necessary to stabilize civilization in a range of scenarios. Yes, this analysis provides an additional reason in favor of developing those capacities, a reason that does not seem to have been playing a significant role in many recent conversations about related issues, such as debates about government surveillance and about proposed reforms of international and supranational institutions.[49] When this reason is added to the mix, the evaluation should therefore become *more* favourable than it otherwise would have been towards policies that would strengthen governance capacities in these ways. However, whether or not this added reason is sufficiently weighty to tip the overall balance would depend on other considerations that fall outside the scope of this paper.

It is worth emphasizing that the argument in this paper favors certain specific forms of governance capacity strengthening. With respect to surveillance and preventive policing, VWH-concerns point specifically to the desirability of governance capacity that makes it possible to extremely reliably suppress activities that are very strongly disapproved of by a very large supermajority of the population (and of power-weighted domestic stakeholders). It provides support for other forms of governance strengthening only insofar as they help create this particular capacity. Similarly, with respect to global governance, VWH-based arguments support developing institutions that are capable of reliably resolving very high-stakes international coordination problems, ones where a failure to reach a solution would result in civilizational devastation. This would include having the capacity to prevent great power conflicts, suppress arms races in weapons of mass destruction, regulate development races and deployment of potential black-ball technologies, and successfully manage the very worst kinds of tragedy of the commons. It need *not* include the capacity to make states cooperate on a host of other issues, nor does it necessarily include the capacity to achieve the requisite stabilization using only fully legitimate means. While those capacities may be attractive for other reasons, they do not immediately emerge as desiderata simply from taking VWH seriously. For example, so far as VWH is concerned, it would theoretically be satisfactory if the requisite global governance capacity comes into existence via the rise of one superpower to a position of sufficient dominance to give it the ability, in a sufficiently dire emergency, unilaterally to impose a stabilization scheme on the rest of the world.

One important issue that we still need to discuss is that of timing. Even if we became seriously concerned that the urn of invention may contain a black ball, this need not move us to favor establishing stronger surveillance or global governance *now*, if we thought that it would be possible to take those steps *later*, if and when the hypothesized vulnerability came clearly into view. We could then let the world continue its sweet slumber, in the confident expectation that as soon as the alarm goes off it will leap out of bed and undertake the required actions. But we should question how realistic that plan is.

Some historical reflection is useful here. Throughout the Cold War, the two superpowers (and the entire northern hemisphere) lived in continuous fear of nuclear annihilation, which could have been triggered at any time by accident or as the result of some crisis spiralling out of control. The reality of the threat was accepted by all sides. This risk could have been substantially reduced simply by getting rid of all or most nuclear weapons (a move which, as a nice side effect, could also have saved more than ten trillion dollars).[5051] Yet, after several decades of effort, only limited nuclear disarmament and other risk-reduction measures were implemented. Indeed the threat of nuclear annihilation remains with us to this day. In the absence of strong global governance that can enforce a treaty and compel disputants

to accept a compromise, the world has so far been unable to solve this most obvious collective action problem.[52]

But perhaps the reason why the world has failed to eliminate the risk of nuclear war is that the risk was insufficiently great? Had the risk been higher, one could eupeptically argue, then the necessary will to solve the global governance problem would have been found. Perhaps – though it does seem rather shaky ground on which to rest the fate of civilization. We should note that although a technology even more dangerous than nuclear weapons may stimulate a greater will to overcome the obstacles to achieving stabilization, other properties of a black ball could make the global governance problem *more challenging* than it was during the Cold War. We have already illustrated this possibility in scenarios such as 'safe first strike' and 'worse global warming'. We saw how certain properties of a technology set could generate stronger incentives for destructive use or for refusing to join (or defecting from) any agreement to curb its harmful applications.[53]

Even if one felt optimistic that an agreement could *eventually* be reached, the question of timing should remain a serious concern. International collective action problems, even within a restricted domain, can resist solution for a *long* time, even when the stakes are large and indisputable. It takes time to explain why an arrangement is needed and to answer objections, time to negotiate a mutually acceptable instantiation of the cooperative idea, time to hammer out the details, and time to set up the institutional mechanisms required for implementation. In many situations, holdout problems and domestic opposition can delay progress for decades; and by the time one recalcitrant nation is ready to come on board, another who had previously agreed might have changed its mind. Yet at the same time, the interval between a vulnerability becoming clearly visible to all and the point when stabilization measures must be in place could be *short*. It could even be negative, if the nature of the vulnerability leaves room for denialism or if specific explanations cannot be widely provided because of information hazards. These considerations suggest that it is problematic to rely on spontaneous ad hoc international cooperation to save the day once a vulnerability comes into view.[54]

The situation with respect to preventive policing is in some respects similar, although we see a much faster and more robust trend – driven by advances in surveillance technology – towards increasing state capacities for monitoring and potentially controlling the actions of their own citizens than any trend towards effective global governance. At least this is true if we look at the physical realm. In the digital information realm the outlook is somewhat less clear, owing to the proliferation of encryption and anonymization tools, and the frequency of disruptive innovation which makes the future of cyberspace harder to foresee. Sufficiently strong capabilities in physical space would, however, spill over into strong capabilities in the digital realm as well. In High-tech Panopticon, there would be no need for the authorities to crack ciphers, since they could directly observe everything that users type into their computers and everything that is shown on their screens.

One could take the position that we should not develop improved methods of surveillance and social control unless and until a specific civilizational vulnerability comes clearly into view – one that looks sufficiently serious to justify the sacrifice of some types of privacy and the risk of inadvertently facilitating a totalitarian nightmare. But as with the case of international cooperation, we confront a question of timing. A highly sophisticated surveillance and response system, like the one depicted in 'High-tech Panopticon', cannot be conjured up and made fully reliable overnight. Realistically, from our current starting point, it would take many years to implement such a system, not to mention the time required to build political support. Yet the vulnerabilities against which such a system might be needed may not offer us much advance warning. Last week a top academic biolab may have published an article in *Science*; and as you are reading these words, a popular blogger somewhere in the world, in hot pursuit of pageviews, might be uploading a post that explains some clever way in which the lab's result could be used by anybody to cause mass destruction.

In such a scenario, intense social control may need to be switched on almost immediately. In an unfavorable scenario, the lead time could be as short as hours or days. It would then be too late to start developing a surveillance architecture when the vulnerability comes clearly into view. If devastation is to be avoided, the mechanism for stabilization would need to have been put in place beforehand.

What may theoretically be feasible is to develop the *capabilities* for intrusive surveillance and real-time interception in advance, but not initially to *use* those capabilities to anything like their full extent. This would be one way to satisfy the requirement for stabilizing a Type-1 vulnerability (and other vulnerabilities that require highly reliable monitoring of individual actions). By giving human civilization the capacity for extremely effective preventive policing, we would have exited one of the dimensions of the semi-anarchic default condition.

Admittedly, constructing such a system and keeping it in standby mode would mean that some of the downsides of actually instituting intense forms social control would be incurred. In particular, it may make oppressive outcomes more likely:

> "[The] question is whether the creation of a system of surveillance perilously alters that balance too far in the direction of government control . . . We might imagine a system of compulsory cameras installed in homes, activated only by warrant, being used with scrupulous respect for the law over many years. The problem is that such an architecture of surveillance, once established, would be difficult to dismantle, and prove too potent a tool of control if it ever fell into the hands of people who – whether through panic, malice, or a misguided confidence in their own ability to secretly judge the public good – would seek to use it against us (Sanchez, 2013)."

Developing a system for turnkey totalitarianism means incurring a risk, even if one does not intend for the key to be turned.

One could try to reduce this risk by designing the system with appropriate technical and institutional safeguards. For example, one could aim for a system of 'structured transparency' that prevents concentrations of power by organizing the information architecture so that multiple independent stakeholders must give their permission in order for the system to operate, and so that only the specific information that is legitimately needed by some decision-maker is made available to her, with suitable redactions and anonymization applied as the purpose permits. With some creative mechanism design, some machine learning, and some fancy cryptographic footwork, there might be no fundamental barrier to achieving a surveillance system that is at once highly effective at its official function yet also somewhat resistant to being subverted to alternative uses.

How likely this is to be achieved in practice is of course another matter, which would require further exploration.[55] Even if a significant risk of totalitarianism would inevitably accompany a well-intentioned surveillance project, it would not follow that pursuing such a project would increase the risk of totalitarianism. A relatively less risky well-intentioned project, commenced at a time of comparative calm, might reduce the risk of totalitarianism by preempting a less-well-intentioned and more risky project started during a crisis. But even if there were some net totalitarianism-risk-increasing effect, it might be worth accepting that risk in order to gain the general ability to stabilize civilization against emerging Type-1 threats (or for the sake of other benefits that extremely effective surveillance and preventive policing could bring).

Conclusions

This paper has introduced a perspective from which we can more easily see how civilization is vulnerable to certain types of possible outcomes of our technological creativity – our drawing a metaphorical black ball from the urn of inventions, which we have the power to extract but not to put back in. We developed a typology of such potential vulnerabilities, and showed how some of them result from destruction becoming too easy, others from pernicious changes in the incentives facing a few powerful state actors or a large number of weak actors.

We also examined a variety of possible responses and their limitations. We traced the root cause of our civilizational exposure to two structural properties of the contemporary world order: on the one hand, the lack of preventive policing capacity to block, with extremely high reliability, individuals or small groups from carrying out actions that are highly illegal; and, on the other hand, the lack of global governance capacity to reliably solve the gravest international coordination problems even when vital national interests by default incentivize states to defect. General stabilization against potential civilizational vulnerabilities – in a world where technological innovation is occurring rapidly along a wide frontier, and in which there are large numbers of actors with a diverse set of human-recognizable motivations – would require that both of these governance gaps be eliminated. Until such a time as this is accomplished, humanity will remain vulnerable to drawing a technological black ball.

Clearly, these reflections provide a pro tanto reason to support strengthening surveillance capabilities and preventive policing systems and for favoring a global governance regime that is capable of decisive action (whether based on unilateral hegemonic strength or powerful multilateral institutions). However, we have not settled whether these things would be desirable all-things-considered, since doing so would require analyzing a number of other strong considerations that lie outside the scope of this paper.

Because our main goal has been to put some signposts up in the macrostrategic landscape, we have focused our discussion at a fairly abstract level, developing concepts that can help us orient ourselves (with respect to long-term outcomes and global desirabilities) somewhat independently of the details of our varying local contexts.

In practice, were one to undertake an effort to stabilize our civilization against potential black balls, one might find it prudent to focus initially on partial solutions and low-hanging fruits. Thus, rather than directly trying to bring about extremely effective preventive policing or strong global governance, one might attempt to patch up particular domains where black balls seem most likely to appear. One could, for example, strengthen oversight of biotechnology-related activities by developing better ways to track key materials and equipment, and to monitor scientists within labs. One could also tighten know-your-customer regulations in the biotech supply sector, and expand the use of background checks for personnel working in certain kinds of labs or involved with certain kinds of experiment. One can improve whistleblower systems, and try to raise biosecurity standards globally. One could also pursue differential technological development, for instance by strengthening the biological weapons convention and maintaining the global taboo on biological weapons. Funding bodies and ethical approval committees could be encouraged to take broader view of the potential consequences of particular lines of work, focusing not only on risks to lab workers, test animals, and human research subjects, but also on ways that the hoped-for findings might lower the competence bar for bioterrorists down the road. Work that is predominantly protective (such as disease outbreak monitoring, public health capacity building, improvement of air filtration devices) could be differentially promoted.

Nevertheless, while pursuing such limited objectives, one should bear in mind that the protection they would offer covers only special subsets of scenarios, and might be temporary. If one finds oneself in a position to influence the macroparameters of preventive policing capacity or global governance capacity, one should consider that fundamental changes in those domains may be the only way to achieve a general ability to stabilize our civilization against emerging technological vulnerabilities.

Notes

For comments, discussion, and critique, I'm grateful to Sonja Alsofi, Stuart Armstrong, Andrew Snyder-Beattie, Chris Anderson, Nick Beckstead, Miles Brundage, Ben Buchanan, Owen Cotton-Barratt, Niel Bowerman, Paul Christiano, Allan Dafoe, Jeff Ding, Eric Drexler, Peter Eckersley, Owain Evans, Thomas Homer-Dixon, Thomas Inglesby, John Leslie, Gregory Lewis, Matthijs Maas, Jason Matheny, Michael Montague, Luke Muehlhauser, Toby Ord, Ben Pace, Richard Re, Anders Sandberg, Julian Savulescu, Stefan Schubert, Carl Shulman, Tanya Singh, Helen Toner, and to the audiences of several workshops and lectures where earlier versions of this work were presented), and to three anonymous referees; and I thank Carrick Flynn, Christopher Galias, Ben Garfinkel, and Rose Hadshar for help with the manuscript and many useful suggestions.
This work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 669751).

1. Obviously, the urn metaphor has important limitations. We will discuss some of them later.
2. The net effect on the conditions of non-human animals is harder to assess. In particular, modern factory farming involves the mistreatment of large numbers of animals.
3. There are, however, examples of 'cultures' or local populations whose demise may have been brought about (at least partially) by their own technological practices, such as Easter Island (Rapa Nui) people and the Ancestral Puebloans in Mesa Verde (Anasazi), who, according to Diamond (2005), cut down their own forests and then suffered environmental collapse.
4. Examples may however be found in other species, if we consider evolutionary adaptations as inventions. For instance, there is a literature exploring how evolutionary dead ends, leading to the extinction of a species or a population, may ensue from advantageous evolutionary changes such as ones involved in specialization (in which adaptation to a narrow niche may entail an irreversible loss of traits needed to survive in a wider range of environments) (Day et al., 2016), the emergence of inbred social systems (among e.g. social spiders) (Aviles and Purcell, 2012), or a switch to selfing (e.g. among flowering plant species transitioning from outcrossing to self-fertilization) (Igic and Busch, 2013).
5. Although most scientists involved in the project favored proposals such as the Baruch Plan, which would have placed nuclear energy under international control, they retained little decision-making power at this point.
6. Metaphorically, of course. But arguably in the metaphor, there should be more than one trembling finger on each side, given the widely delegated command and control (Ellsberg, 2017; Schlosser, 2013).
7. However, within a given state, the number of actors who are empowered to launch nuclear attacks may be quite large. Ellsberg (2017) claims that, for at least a significant portion of the Cold War, the authority to launch nuclear weapons was delegated multiple rungs down the American chain of command. The number of officers with the physical ability to launch nuclear weapons, although not the authority to do so, was also necessarily larger. In the Soviet Union, at one point during the coup against Mikhail Gorbachev in August 1991, all three of the USSR's Chegets ('nuclear briefcases') were in the hands of coup leaders (Sokoski and Tertrais, 2013; Stevenson, 2008).
8. An 'information hazard' is a risk arising from the dissemination of true information, for instance because the information could enable some agents to cause harm. Bostrom (2011) discusses information hazards more generally.
9. They might argue that openness would be a benefit since it would allow more people to work on countermeasures (cf. the debate around gain of function work on flu viruses; (Duprex et al., 2015; Fauci et al., 2011; Sharp, 2005)). They might also argue that, so long as the government continues to justify draconian actions by referencing secret information, it will be dangerously unaccountable to its citizens. A similar belief motivated the American magazine *The Progressive*'s decision in the late 1970s to publish secrets about the hydrogen bomb, even in the face of a legal challenge by the US Department of Energy. The author of the piece, Howard Morland, wrote: 'Secrecy itself, especially the power of a few designated 'experts' to declare some topics off-limits, contributes to a political climate in which the nuclear establishment can conduct business as usual, protecting and perpetuating the production of these horror weapons' Morland (1979, p. 3).
10. Generally, in cases where multiple actors each have some independent probability of taking an action unilaterally, the probability that the action will be taken tends to one as the number of actors increases. When this phenomenon arises for actors with shared goals but discordant judgments, due to randomness in the evidence they are exposed to or the reasoning they carry out, there arises a 'unilateralist's curse' (Bostrom et al., 2016). The curse implies that even a very unwise decision, such as the decision to publish nuclear weapons designs, is likely to be made if enough actors are in a position to take it unilaterally.
11. Many of these same motivations are evident today among 'black hat' hackers who carry out malicious cyber attacks. For instance, as a method of extortion, some anonymous hackers have proven themselves willing to remove cities' abilities to provide vital services to their residents (Blinder and Perlroth, 2018). Motivations for economically damaging cyber attacks have also seemed to include both political ideology and curiosity. Since contemporary cyber attacks are dramatically less destructive than attacks with nuclear weapons, the set of actors that would be willing to use nuclear weapons or threaten their use is surely much smaller than the set of actors willing to engage in malicious hacking. Nevertheless, the social and psychological factors relevant to both cases may be similar.
12. This concept is distinct from that of international anarchy in the field of international relations. The present concept emphasizes that anarchy is a matter of degree and is meant to be relatively neutral as between different schools of thought in IR (cf. Lechner, 2017)). More importantly, it encompasses a lack of governance not just 'at the top' but also 'at the bottom'. That is to say, the semi-anarchic default condition refers to the fact that in the current world order, not only is there a degree of anarchy at the international level, because of lack of global governance or other fully effective means of constraining the actions of state and solving global coordination problems, but there is also a degree of anarchy at the level of individuals (and other sub-state actors) in that even highly functional states currently lack the ability to perfectly regulate the actions of those small actors. For example, despite many states seeking to prevent rape and murder within their territory, rape and murder continue to occur with non-zero frequency. The consequences of this degree of anarchy at the bottom could be vastly magnified if individuals obtained much greater destructive capabilities.
13. For comparison, a death toll of 15 per cent of the present world population is more than double the combined effects of World War I, the Spanish Flu, and World War II as a percentage of global population (and the difference is even bigger in absolute terms). A 50 per cent fall in world GDP is greater than the largest drop in recorded history. During the Great Depression, for example, world GDP fell by an estimated 15 per cent or less, and mostly recovered within a few years (though some models suggest that it has also had a long-lasting depressing effect on trade which has chronically impaired the world economy) (Bolt et al., 2018); Crafts and Fearon, 2010).
14. This paper focuses on *technological* vulnerabilities. There could also be *natural* vulnerabilities that arise independently of the progress of human civilization, such as a violent meteor barrage set to impact our planet at some future date. Some natural vulnerabilities

could be stabilized once our level of technological capability exceeds some threshold (e.g. the ability to deflect meteors). Plausibly, the risk of technological vulnerabilities is greater than the risk of natural vulnerabilities, although the case for this is less clear cut with the severity cutoff of civilizational devastation than it would be if the cutoff were set to existential catastrophe (Bostrom, 2013; Bostrom and Ćirković, 2011).The (big) proviso to this claim (of technological vulnerabilities dominating) is that it presupposes that it is not the case that the world is hemorrhaging value-potential at a significant rate. If instead we evaluate things from a perspective in which what we may term *the bleeding world hypothesis* is true, then it may well be that the default devastation arising from natural (i.e. non-human) processes dominates the equation. The bleeding world hypothesis could hold if, for example: (1) the evaluator cares a lot about existing people (including self and family) and they are naturally dying off at a substantial rate (e.g. from aging), thereby losing both the ability to continue enjoying their lives and the opportunity for vastly greater levels of well-being such as would become possible at technological maturity; (2) the evaluator cares a lot about avoiding suffering that could be avoided with more advanced technology but is occurring currently, piling up disutility; (3) there is some substantial exogenously set rate of civilizational destruction (e.g. natural disasters, random simulation terminations unrelated to our activities Bostrom, 2003), and while we allow time to lapse we incur a cumulative risk of being destroyed before maxing out our technological potential; (4) there are ways, using physics we don't currently understand well, to initiate fast-growing processes of value creation (such as by creating an exponential cascade of baby-universes whose inhabitants would be overwhelmingly happy), and the evaluator cares in a scale-sensitive way about such creation; and (5) other superintelligent constituencies, who are in a position to greatly influence things the evaluator cares about, are impatient for us to reach some advancement, but the value they place on this decays rapidly over time.

15. The world could remain vulnerable after profound technological regress, for instance if many prefabricated nukes remain even after civilization regresses past the point of becoming incapable of manufacturing new ones.
16. It is important to our original 'easy nukes' scenario that each nuclear use requires the efforts of only one individual or of a small group. Although it might require the combined efforts of hundreds of actors to devastate civilization in that scenario – after all, ruining one city or one metropolitan area is not the same as ruining a civilization – these hundreds of actors need not coordinate. This allows the apocalyptic residual to come into play.
17. Baum et al. (2018) provide an up-to-date list of nuclear accidents and occasions on which the use of nuclear weapons was considered. Sagan (1995) provides a more thorough account of dangerous practices throughout the Cold War. Schlosser (2013) examines near-accidents, focusing in particular on one incident which resulted in a non-nuclear detonation of a Titan-II ICBM.
18. Although there have been few scholarly attempts to assess the degree of luck involved in avoiding this outcome, one recent estimate, drawing from a dataset of near-miss instances, places the probability of the US and USSR avoiding nuclear war below 50 per cent (Lundgren, 2013). This is consistent with the views of some officials with insider knowledge of nuclear crises, such as President John F. Kennedy, who expressed the belief that, in hindsight, the Cuban missile crisis had between a one-in-two and a one-in-three chance of leading to nuclear war. Nonetheless, a number of prominent international security scholars, such as Kenneth Waltz and John Mueller, hold that the probability of nuclear war has been consistently very low (Mueller, 2009; Sagan and Waltz, 2012).
19. Perhaps believed erroneously. According to a former Commander of the US Pacific Fleet, there was a period during the Cold War when antisubmarine surveillance became extremely effective: '[The US] could identify by hull number the identity of Soviet subs, and therefore we could do a body count and know exactly where they were. In port or at sea. . . . so I felt comfortable that we had the ability to do something quite serious to the Soviet SSBN force on very short notice in almost any set of circumstances.' (quoted in Ford and Rosenberg, 2005, p. 399).
20. In fact, advances in remote sensing, data processing, AI, drones, and nuclear delivery systems are now threatening to undermine nuclear deterrence, especially for states with relatively small and unsophisticated nuclear arsenals (Lieber and Press, 2017).
21. Not really unscathed, of course: radioactive fallout would affect allies and to a degree the homeland; the economic repercussions would wreak havoc on markets and usher in a worldwide depression. Still, it would be far preferable to being the target of the assault (especially if we set aside nuclear winter).
22. Another possibility is that there would be political gains, such as an increased ability to engage in nuclear coercion against third parties after having demonstrated a willingness to use nuclear weapons.
23. Brooks (1999); Gartzke (2007); Gartzke and Rohner (2011). For a dissenting view, see Liberman (1993).
24. Human civilization could probably never have arisen if the Earth's climate had been that sensitive to carbon dioxide, since past CO_2 levels (4,000 ppm during the Cambrian period compared to about 410 pmm today) would then presumably have seriously disrupted the evolution of complex life. A less remote counterfactual might instead involve some compound that does not occur in significant quantities in nature but is produced by human civilization, such as chlorofluorocarbons. CFCs have been phased out via the Montreal Protocol because of their destructive effect on the ozone layer, but they are also very potent greenhouse gases on a per kilogram basis. So we could consider a counterfactual in which CFCs had been industrially useful on a far greater scale than they were, but with dramatic delayed cumulative effects on global climate.
25. The report commissioned by Oppenheimer ends: 'One may conclude that the arguments of this paper make it unreasonable to expect that the N + N reaction could propagate. An unlimited propagation is even less likely. However, the complexity of the argument and the absence of satisfactory experimental foundation make further work on the subject highly desirable' (Konopinski et al., 1946).
26. Type-0 could be viewed as the limiting case of a Type-1: it refers to a vulnerability that requires zero ill-intentioned actors in order for civilizational devastation to result – only normally responsible actors who are willing to proceed with using a technology after an ordinary amount of scrutiny has been given to the new technology.
27. And if 10 years, why not permanently.
28. In fact, an account by Albert Speer, the German Minister of Armaments, suggests that Werner Heisenberg discussed the possibility of a runaway chain reaction with Hitler and that this possibility may have further dampened Hitler's enthusiasm for pursuing the bomb (Rhodes, 1986).
29. A real-world version of this kind of Type-2a vulnerability, in which key actors face strategic incentives to take actions that create unwanted risks for civilization, could arise in the context of a race to develop machine superintelligence. In unfavorable circumstances, competitive dynamics could present a leading developer with the choice between launching their own AI before it is safe or relinquishing their lead to some other developer who is willing to take greater risks (Armstrong et al., 2016).
30. For a discussion of how a rational planner would balance consumption growth with safety in various models where growth-inducing innovation also carries a risk of introducing innovations that reduce lifespan, see Jones (2016).
31. Even the 'surprising strangelets' scenario may be confounded by coordination problems, though to a lesser degree than 'Castle Bravo/Trinity test'. The people deciding on science funding allocations may have different priorities than the public that is providing

the funding. They might, for example, place a higher value on satisfying intellectual curiosity, relative to the value placed on keeping risks low and providing near-term material benefits to the masses. Principal-agent problems could then result in more funding for particle accelerators than such experiments would get in the absence of coordination problems. Prestige contests between nations – which might in part be viewed as another coordination failure – may also be a driver of basic science funding in general and high-energy physics in particular.

32. The people alive at the time when the devastation occurs might prefer that it had taken place earlier, before they were born, so that it would all be over and done with and they wouldn't be affected. Their preferences seem to run into a non-identity problem, since if a civilizational devastation event had taken place before they were conceived they would almost certainly not have come into existence (Parfit, 1987).
33. More broadly, many refinements in biotechnological tools and techniques, which make it easier for amateur DIY biohackers to accomplish what previously could only be done by well-resourced professional research labs, come under suspicion from this perspective. It is very questionable whether the benefits of DIY biohacking (glow-in-the-dark house plants?) are worth proliferating the ability to turn bioengineering to potentially risky or malicious purposes to an expanded set of relatively unaccountable actors.
34. The counterargument that 'if I don't develop it, somebody else will; so I might as well do it' tends to overlook the fact that a given scientist or developer has at least some marginal impact on the expected timing of the new discovery. If it really were the case that a scientist's efforts could make *no* difference to when the discovery or invention is made, it would appear that the efforts are a waste of time and resources, and should be discontinued for that reason. A relatively small shift in when some technological capability becomes available (say, one month) could be important in some scenarios (such as if the dangerous technology imposes a significant risk per month until effective defenses are developed and deployed).
35. By 'technological maturity' we mean the attainment of capabilities affording a level of economic productivity and control over nature close to the maximum that could feasibly be achieved (in the fullness of time) (Bostrom, 2013).
36. Access control in bioscience has grown in importance since the 2001 'Amerithrax' incident. In the United States, institutions handling dangerous pathogens are obliged to assess suitability for employees who will have access, which are also vetted by federal agencies (Federal Select Agent Program, 2017), and similar approaches are recommended to countries developing their biosecurity infrastructure (Centre for Biosecurity and Biopreparedness, 2017). The existing regime suffers from two shortcomings: first, there is no global coordination, so bad actors could 'shop around' for laxer regulatory environments; second, the emphasis remains on access to biological materials (e.g. samples of certain microorganisms), whereas biological information and technology is increasingly the principal object of security concern (Lewis et al., 2019).
37. A value of X substantially less than 1 per cent seems consistent with how little most people give to global charity. It is possible, however, that an act-omission distinction would make people willing to accept a substantially larger personal sacrifice in order not to contribute to a global bad than they would in order to contribute a global good.
38. Note, however, that a positive shift in the preference distribution – even if insufficient to avert catastrophe by simply making some individual actors not choose the destructive option – could have important *indirect* effects. For example, if a large number of people became slightly more benevolently inclined, this might shift society into a more cooperative equilibrium that would support stronger governance-based stabilization methods such as the ones we discuss below (cf. 'moral enhancements'; Persson and Savulescu, 2012).
39. At a global level, we find a patchwork of national classification schemes and information control systems. They are generally designed to protect military and intelligence secrets, or to prevent embarrassing facts about regime insiders from being exposed to the public, not to regulate the spread of scientific or technological insights. There are some exceptions, particularly in the case of technical information that bears directly on national security. For instance, the Invention Secrecy Act of 1951 in the United States gives defense agencies the power to bar the award of a patent and order that an invention be kept secret; though an inventor who refrains from seeking patent protection is not subject to these strictures (Parker and Jacobs 2003). Nuclear inventions are subject to the 'born secret' provision of the Atomic Energy Act of 1946, which declares all information concerning the design, development, and manufacture of nuclear weapons – regardless of origin – classified unless it has been officially declassified (Parker and Jacobs 2003). Other legal tools, such as export controls, have also been used in attempts to stem the flow of scientific information. The (unsuccessful) efforts of multiple US government agencies to block the publication and use of strong encryption protocols developed in the 1970s and 1980s provide one notable example (Banisar, 1999).Voluntary self-censorship by the scientific community has been attempted on very rare occasions. Leo Szilard had some partial successes in convincing his physicist colleagues to refrain from publishing on aspects of nuclear fission (before the start of the Manhattan Project and the onset of official secrecy), though he encountered opposition from some scientists who wanted their own work to appear in journals or who felt that openness was a sacred value in science. More recently, there were some attempts at scientific self-censorship in relation to avian flu research (Gronvall, 2013). In this case, the efforts may have been not only ineffectual but counterproductive, inasmuch as the controversy sparked by open debate about whether certain results should be published drew more attention to those results than they would have received if publication had proceeded unopposed – the so-called 'Streisand effect'.Overall, attempts at scientific self-censorship appear to have been fairly half-hearted and ineffectual. (I say *appears*, because of how things unfolded in the publicly known episodes where censorship was attempted. But truly successful attempts to suppress scientific information wouldn't necessarily show up in the public record.) Even if a few journal editors could agree on standards for how to deal with papers that pose information hazards, nothing would prevent a frustrated author from sending her manuscript to another journal with lower standards or to publish it on her personal Internet page. Most scientific communities have neither the culture, nor the incentives, nor the expertise in security and risk assessment, nor the institutional enforcement mechanisms that would be required for dealing effectively with infohazards. The scientific ethos is rather this: every ball must be extracted from the urn as quickly as possible and revealed to everyone in the world immediately; the more this happens, the more progress has been made; and the more you contribute to this, the better a scientist you are. The possibility of a black ball does not enter into the equation.
40. In any case, it is unclear whether we would really want to be more cautious in general. It might be desirable (from various evaluative perspectives) to encourage greater caution specifically in situations where there could be extreme global downsides. Yet exhortations to exercise voluntary caution and restraint in these causes may not be very effective if the reason for the normatively excessive risk-taking is a coordination problem: the risk-taker gaining some private benefit (e.g. profit or prestige) while generating a global risk externality. In such cases, therefore, the solution may require a strengthening of global governance capacity.

41. It is also possible for risk assessment work to increase the level of risk, by generating information hazards (Bostrom, 2011).
42. The Orwellian-sounding name is of course intentional, to remind us of the full range of ways in which such a system could be applied.
43. Implementation details are for illustration only. For example, similar functionality could be provided by mixed reality eyeglasses instead of a necklace. Versions of the device could be designed that would provide many benefits to the user along with its surveillance function. In theory, some of the monitoring could be crowd-sourced: when suspicious activity is detected by the AI, the video feed is anonymized and sent to a random 100 citizens, whose duty is to watch the feed and vote on whether it warrants further investigation; if at least 10 per cent of them think it does, the (non-anonymized) feed gets forwarded to the authorities.
44. Examples of 'conveniences' that will plausibly drive more intrusive surveillance include various kinds of consumer applications and economically useful or profitable monitoring (e.g. for ad targeting, price discrimination, etc.); the ability to prevent various things that cause public outrage, such as child abuse or small-scale terrorism; and, especially for authoritarian regimes, the ability to suppress political opposition.
45. A milder version of the policy might merely debar such weak suspects from accessing the equipment and materials necessary to produce the destructive effect. The extent to which this might suffice depends on the details of the scenario.
46. A partial implementation of the High-tech Panopticon might replace incarceration in this scenario, in which only those on some long list of 'individuals of heightened concern' were required to wear the freedom tags.
47. Of course, it is theoretically possible that either of these remedies would raise rather than lower civilization's total vulnerability to a potential black ball, for example, if adequate global coordination made extremely effective national policing less likely, or vice versa. The character of the regimes that would tend to arise under conditions of stronger preventive policing or global governance could also differ from those in the status quo in ways that would increase or decrease some civilizational vulnerabilities. For example, surveillance-empowered world leaders might be more or less prone to taking foolish decisions that increase Type-0 vulnerabilities.
48. A special case of Type-2a vulnerability is one in which some set of regimes jointly achieve the devastational threshold by harming their own populations. Suppose, for instance, that one held an extremely pessimistic view of political leaders, and thought that they would be willing to kill an extremely large fraction of their own populations if doing so would help them hold on to power or gain more resources. Whereas today such a genocidal initiative would usually be counterproductive from the leader's point of view (because it would spark revolts and crash the economy), one could imagine a different technological environment in which these restraints would be loosened – for example, if a highly centralized AI police force could reliably suppress any resistance and if robots could easily replace human workers. (Fortunately, it would appear that in many scenarios where these things becomes technologically feasible, the ruler's incentives for genocidal actions would also be weakened: the hypothesized AI police force would presumably enable the ruler to maintain power without killing off large parts of the population, and automation of the economy would greatly increase wealth so that a smaller fraction of national income would suffice to give all citizens a high standard of living.)
49. For example, surveillance debates often focus on the tradeoffs between the privacy interests of individuals and public demand for security against small-scale terrorist attacks. (Even terrorist incidents that are usually regarded as large, such as the 9/11 attacks, are insignificantly small-scale by the standards used in this paper.)
50. According to Schwartz (1998), the nuclear arms race during the Cold War cost 5.8 trillion (1996 dollars) in American expenditures alone, which is equivalent to 9.3 trillion in 2018 dollars. This estimate is quite comprehensive and covers the fuel cycle, weapons, delivery systems, decommissioning, etc. If we add the expenditures of other countries, we can conclude that human civilization spent well in excess of 10 trillion dollars on developing and maintaining a capacity to destroy itself with nuclear arms. Most of the cost was incurred in a period when world GDP was substantially lower than it is today. Even larger amounts were spent on *non-nuclear* military capabilities (over 20 trillion in 2018 dollars in the US alone). It is possible that the nuclear expenditures saved money on balance by reducing non-nuclear military spending. Both nuclear and non-nuclear military spending reflects the failure of human civilization to solve global coordination.
51. *Substantially* rather than *entirely*, since even if all nuclear weapons were dismantled, new ones might be created.
52. An agreement for total nuclear disarmament might, of course, have to involve some provisions about conventional forces and other matters as well, so as not to endanger strategic stability.
53. One might look at other historical examples to obtain a larger reference class. The world's efforts so far with respect to combating global warming do not inspire confidence in its ability to deal expeditiously with even more difficult global collective action problems. On the other hand, the problem of ozone depletion was successfully addressed with the Montreal Protocol.
54. Unilateral imposition may be faster, but it requires that some actor has the capability to impose its will single-handedly on the rest of the world. If one actor has such an overwhelming power advantage, a form of de facto (weak or latent) global governance is presumably already in place.
55. For example, a well-intentioned project may be subverted in its implementation; or it might turn out to have bugs or institutional design flaws that become apparent only after a period of normal operation. Even if the system itself functions precisely as intended and remains uncorrupted, it might inspire the creation of other surveillance systems that do not have the same democratic safeguards.

References

Armstrong, S., Bostrom, N. and Shulman, C. (2016) 'Racing to the Precipice: A Model of Artificial Intelligence Development', *AI & Society*, 31 (2), pp. 201–06.

Aviles, L. and Purcell, J. (2012) 'The Evolution of Inbred Social Systems in Spiders and Other Organisms: From Short-term Gains to Long-term Evolutionary Dead Ends?', *Advances in the Study of Behavior*, 44 (1), pp. 99–133.

Badash, L. (2001) 'Nuclear Winter: Scientists in the Political Arena', *Physics in Perspective*, 3 (1), pp. 76–105.

Banisar, D. (1999) 'Stopping Science: The Case of Cryptography', *Health Matrix*, 9(2), pp. 253–87.

Baum, S., de Neufville, R. and Barrett, A. (2018) 'A Model for the Probability of Nuclear War', *Global Catastrophic Risk Institute Working Paper 18–1*. Available from: https://doi.org/10.2139/ssrn.3137081 [Accessed 31 July 2019].

Blinder, A. and Perlroth, N. (2018) 'A Cyberattack Hobbles Atlanta, and Security Experts Shudder', *New York Times* Available from: https://www.nytimes.com/2018/03/27/us/cyberattack-atlanta-ransomware.html

Bostrom, N. (2002) 'Existential Risks – Analyzing Human Extinction Scenarios and Related Hazards',*Journal of Evolution and Technology*, 9 (1), pp. 1–30.

Bostrom, N. (2003) 'Are You Living in a Computer Simulation?', *Philosophical Quarterly*, 53 (211), pp. 243–55.

Bostrom, N. (2006) 'What is a Singleton',*Linguistic and Philosophical Investigations*, 5 (2), pp. 48–54.

Bostrom, N. (2008) 'Why I Want to be a Posthuman When I Grow Up', in B. Gordijn and R. Chadwick (eds), *Medical Enhancement and Posthumanity*. New York: Springer, pp. 107–37.
Bostrom, N. (2011) 'Information Hazards: A Typology of Potential Harms from Knowledge', *Review of Contemporary Philosophy*, 10(1), pp. 44–79.
Bostrom, N. (2013) 'Existential Risk Prevention as Global Priority',*Global Policy*, 4 (1), pp. 15–31.
Bostrom, N. and Ćirković, M. M. (eds) (2011) *Global Catastrophic Risks*. Oxford: Oxford University Press, pp. 1–29.
Bostrom, N., Douglas, T. and Sandberg, A. (2016) 'The Unilateralists Curse and the Case for a Principle of Conformity', *Social Epistemology*, 30 (4), pp. 350–71.
Brooks, S. G. (1999) 'The Globalization of Production and the Changing Benefits of Conquest', *Journal of Conflict Resolution*, 43 (5), pp. 646–70.
Brower, M. (1989) 'Targeting Soviet Mobile Missiles: Prospects and Implications', *Survival*, 31 (5), pp. 433–45.
Centre for Biosecurity and Biopreparedness (2017) An Efficient and Practical Approach to Biosecurity. Available from: https://www.biosecurity.dk/fileadmin/user_upload/PDF_FILER/Biosecurity_book/An_efficient_and_Practical_approach_to_Biosecurity_web1.pdf [Accessed 31 July 2019].
Colby, E. A. and Gerson, M. S. (eds.) (2013) *Strategic Stability: Contending Interpretations*. Carlisle Barracks: U.S. Army War College Press.
Collingridge, D. (1980) *The Social Control of Technology*. New York: St. Martin's Press.
Crafts, N. and Fearon, P. (2010) 'Lessons from the 1930s Great Depression', *Oxford Review of Economic Policy*, 26 (3), pp. 285–317.
Danzig, R., Sageman, M., Leighton, T., Hough, L., Yuki, H., Kotani, R. et al. (2011) Aum Shinrikyo: Insights Into How Terrorists Develop Biological and Chemical Weapons, 2nd Edn. Washington, D.C.: Center for a New American Security.
Day, E. H., Hua, X. and Bromham, L. (2016) 'Is Specialization an Evolutionary Dead End? Testing for Differences in Speciation, Extinction and Trait Transition Rates across Diverse Phylogenies of Specialists and Generalists', *Journal of Evolutionary Biology*, 29 (6), pp. 1257–67.
Diamond, J. (2005) *Collapse: How Societies Choose to Fail or Succeed*. New York: Penguin.
Drexler, K. E. (1986) *Engines of Creation*. New York: Anchor.
Duprex, W. P., Fouchier, R. A., Imperiale, M. J., Lipsitch, M. and Relman, D. A. (2015) 'Gain-of-function Experiments: Time for a Real Debate', *Nature Reviews Microbiology*, 13 (1), pp. 58–64.
Ellsberg, D. (2017) *The Doomsday Machine: Confessions of a Nuclear War Planner*. New York: Bloomsbury Publishing USA.
Fauci, A. S., Nabel, G. J. and Collins, F. S. (2011) 'A Flu Virus Risk Worth Taking.' The Washington Post, Available from: https://www.washingtonpost.com/opinions/a-flu-virus-risk-worth-taking/2011/12/30/gIQAM9sNRP_story.html [Accessed 31 July 2019].
Federal Select Agent Program ((2017) Suitability Assessment Program Guidance. Available from: https://www.selectagents.gov/resources/Suitability_Guidance.pdf [Accessed 31 July 2019].
Field, C. B.,Barros, V. R., Mastrandrea, M. D., Mach, K. J., Abdrabo, M. A.-K., Adger, W. N. et al. (2014) *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Working Group II Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press.
Ford, C. A. and Rosenberg, D. A. (2005) 'The Naval Intelligence Underpinnings of Reagan's Maritime Strategy', *Journal of Strategic Studies*, 28 (2), pp. 379–409.
Franck, J., Hughes, D. J., Nickson, J. J., Rabinowitch, E., Seaborg, G. T., Stearns, J. C. and Szilard, L. (1945) 'Report of the Committee on Political and Social Problems', *Manhattan Project 'Metallurgical Laboratory', University of Chicago, U.S. National Archives, Record Group 77, Records of the Chief of Engineers, Manhattan Engineer District, Harrison-Bundy File, folder #76*.
Friedlingstein, P., Andrew, R. M., Rogelj, J., Peters, G. P., Canadell, J. G., Knutti, R. et al. (2014) 'Persistent Growth of CO2 Emissions and Implications for Reaching Climate Targets', *Nature Geoscience*, 7 (10), pp. 709–15.
Gartzke, E. (2007) 'The Capitalist Peace', *American Journal of Political Science*, 51 (1), pp. 166–91.
Gartzke, E. and Rohner, D. (2011) 'The Political Economy of Imperialism, Decolonization and Development', *British Journal of Political Science*, 41 (3), pp. 525–56.
Greenberg, A. (2012) *This Machine Kills Secrets: How Wikileaks, Hacktivists, and Cypherpunks are Freeing the World's Information*. New York: Dutton.
Gronvall, G. K. (2013) *H5n1: A Case Study for Dual-Use Research*. New York: Council on Foreign Relations.
Holloway, D. (1994) *Stalin and the Bomb: The Soviet Union and Atomic Energy, 1939–1956*. New Haven: Yale University Press.
Igic, B. and Busch, J. W. (2013) 'Is Self-fertilization an Evolutionary Dead End?', *New Phytologist*, 198 (2), pp. 386–97.
Inklaar, R., de Jong,H., Bolt, J. and van Zanden, J. L. (2018) 'Rebasing 'Maddison': New Income Comparisons and the Shape of Long-run Economic Development', *GGDC Research Memorandum GD-174*, Groningen: Groningen Growth and Development Center.
Jaffe, R. L., Busza, W., Wilczek, F. and Sandweiss, J. (2000) 'Review of Speculative 'Disaster Scenarios' at RHIC', *Reviews of Modern Physics*, 72 (4), pp. 1125–1140.
Jones, C. I. (2016) 'Life and Growth', *Journal of Political Economy*, 124 (2), pp. 539–78.
Kemp, R. S. (2012) *'SILEX and Proliferation', Bulletin of the Atomic Scientists*. Available from: https://thebulletin.org/2012/07/silex-and proliferation/ [Accessed 31 July 2019] .
Konopinski, E. J., Marvin, C. and Teller, E. (1946) *'Ignition of the Atmosphere with Nuclear Bombs', Report LA-602*. Los Alamos, NM: Los Alamos Laboratory.
Lechner, S. (2017) *'Anarchy in International Relations', Oxford Research Encyclopedia of International Studies* [online], Oxford: Oxford University Press. Available from: https://oxfordre.com/internationalstudies/view/10.1093/acrefore/9780190846626.001.0001/acrefore-9780190846626-e-79?print=pdf [Accessed 31 July 2019].
Lewis, G., Millet, P., Sandberg, A., Snyder-Beattie, A., and Gronvall, G.(2019) 'Information Hazards in Biotechnology', *Risk Analysis*, 39(5), pp. 975–81.
Liberman, P. (1993) 'The Spoils of Conquest', *International Security*, 18 (2), pp. 125–153.
Lieber, K. A. and Press, D. G. (2017) 'The New Era of Counterforce: Technological Change and the Future of Nuclear Deterrence', *International Security*, 41 (4), pp. 9–49.
Lundgren, C. (2013) 'What are the Odds? Assessing the Probability of a Nuclear War', *The Nonproliferation Review*, 20 (2), pp. 361–74.
Morland, H. (1979) 'The H-bomb Secret: How We Got It - Why We're Telling It', *The Progressive*, 43 (11), pp. 3–12.
Mowatt-Larssen, R. and Allison, G. T. (2010) *Al Qaeda Weapons of Mass Destruction Threat: Hype or Reality?*. Cambridge, MA: Belfer Center for Science and International Affairs.
Mueller, J. (2009) *Atomic Obsession: Nuclear Alarmism from Hiroshima to al-Qaeda*. Oxford: Oxford University Press.
Munz, P., Hudea, I., Imad, J. and Smith, R. J. (2009) ''When Zombies Attack!: Mathematical Modelling of an Outbreak of Zombie Infection', in J.M., Tchuenche and C., Chiyaka (eds.), *Infectious Disease Modelling, Research Progress*. New York: Nova Science Publishers, pp. 133–50.
Norris, R. S. and Kristensen, H. M. (2010) 'Global Nuclear Weapons Inventories, 1945–2010', *Bulletin of the Atomic Scientists*, 66 (4), pp. 77–83.
Olson, K. B. (1999) 'Aum Shinrikyo: Once and Future Threat?', *Emerging Infectious Diseases*, 5 (4), pp. 413–16.
Ord, T., Hillerbrand, R. and Sandberg, A. (2010) 'Probing the Improbable: Methodological Challenges for Risks with Low Probabilities and High Stakes', *Journal of Risk Research*, 13 (2), pp. 191–205.

Parfit, D. (1987) *Reasons and Persons*., Oxford: Clarendon Press.
Parker, E. R. and Jacobs, L. G. (2003) 'Government Controls of Information and Scientific Inquiry', *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science*, 1 (2), pp. 83–95.
Persson, I. and Savulescu, J. (2012) *Unfit for the Future: The Need for Moral Enhancement*. Oxford: Oxford University Press.
Re, R. M. (2016) 'Imagining Perfect Surveillance.' *UCLA Law Review Discourse*, 64(1), pp. 264–92.
Rhodes, R. (1986) *The Making of the Atomic Bomb*. London: Simon and Schuster.
Sagan, S. D. (1995) *The Limits of Safety: Organizations, Accidents, and Nuclear Weapons*. Princeton, NJ: Princeton University Press.
Sagan, S. D. and Waltz, K. N. (2012) *The Spread of Nuclear Weapons: An Enduring Debate*. New York: WW Norton.
Sanchez, J. (2013) 'A Reply to Epstein and Pilon on NSA's Metadata Program', Cato at Liberty. Available from: https://www.cato.org/blog/reply-epstein-pilon-nsas-metadata-program [Accessed 31 July 2019].
Schelling, T. C. (1960) *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.
Schlosser, E. (2013) *Command and Control: Nuclear Weapons, the Damascus Accident, and the Illusion of Safety*. New York: Penguin Press.
Schwartz, S. I. (1998) *Atomic Audit: The Costs and Consequences of US Nuclear Weapons Since 1940*. Washington, D.C.: Brookings Institution Press.
Sharp, P. A. (2005) '1918 Flu and Responsible Science', *Science*, 310 (5745), p. 17.
Shindell, D. T. (2014) 'Inhomogeneous Forcing and Transient Climate Sensitivity', *Nature Climate Change*, 4 (4), pp. 274–77.
Sokoski, H. D. and Tertrais, B. (2013) *Nuclear Weapons Security Crisis: What Does History Teach?*. Washington, D.C.: Nonproliferation Policy Education Center.
Stevenson, J. (2008) *Thinking Beyond the Unthinkable: Harnessing Doom from the Cold War to the War on Terror*. New York: Viking.
Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M. M. B., Allen, S. K., Boschung, J. et al. (2014) *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. New York: Cambridge University Press.
Swire, P. (2015) *The Declining Half-Life of Secrets and the Future of Signals Intelligence*. Washington, D.C.: New America.
Tegmark, M. and Bostrom, N. (2005) 'Astrophysics: Is a Doomsday Catastrophe Likely?', *Nature*, 438 (7069), p. 754.
Torres, P. (2018) 'Superintelligence and the Future of Governance: On Prioritizing the Control Problem at the End of History.' in R. Yampolskiy (ed.), *Artificial Intelligence Safety and Security*. Milton: Chapman and Hall/CRC, pp. 357–74.

Author Information

Nick Bostrom is a Professor at Oxford University, where he directs the Future of Humanity Institute, which includes the Center for the Governance of AI and research groups working on other macrostrategic challenges for humanity.