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

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

JGM: Very briefly, there are four types of vulnerabilities to civilizational devastation:

Type-1 = eg, 'Easy nukes' in hands of individual terrorists

Type-2a = eg, 'Safe first strike' by some nuclear state

Type-2b = eg, 'Worse global warming', a collective action problem

Type-0 = 'Surprising strangelets', eg a nuclear test inadvertently leading to the ignition of the atmosphere

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.