Resources on Existential Risk

For:

Catastrophic Risk: Technologies and Policies
Berkman Center for Internet and Society
Harvard University

Bruce Schneier, Instructor

Fall 2015

Contents

General Scholarly Discussion of Existential Risk

Nick Bostrom (Mar 2002), "Existential risks: Analyzing human extinction scenarios and related hazards."
Journal of Evolution and Technology 9.
http://www.nickbostrom.com/existential/risks.pdf

Abstract

Because of accelerating technological progress, humankind may be rapidly approaching a critical phase in its career. In addition to well-known threats such as nuclear holocaust, the prospects of radically transforming technologies like nanotech systems and machine intelligence present us with unprecedented opportunities and risks. Our future, and whether we will have a future at all, may well be determined by how we deal with these challenges. In the case of radically transforming technologies, a better understanding of the transition dynamics from a human to a "posthuman" society is needed. Of particular importance is to know where the pitfalls are: the ways in which things could go terminally wrong. While we have had long exposure to various personal, local, and endurable global hazards, this paper analyzes a recently emerging category: that of existential risks. These are threats that could cause our extinction or destroy the potential of Earth-originating intelligent life. Some of these threats are relatively well known while others, including some of the gravest, have gone almost unrecognized. Existential risks have a cluster of features that make ordinary risk management ineffective. A final section of this paper discusses several ethical and policy implications. A clearer understanding of the threat picture will enable us to formulate better strategies.

Excerpts

1.1 A typology of risk

We can distinguish six qualitatively distinct types of risks based on their scope and intensity (figure 1). The third dimension, probability, can be superimposed on the two dimensions plotted in the figure. Other things equal, a risk is more serious if it has a substantial probability and if our actions can make that probability significantly greater or smaller.

"Personal", "local", or "global" refer to the size of the population that is directly affected; a global risk is one that affects the whole of humankind (and our successors). "Endurable" vs. "terminal" indicates how intensely the target population would be affected. An endurable risk may cause great destruction, but one can either recover from the damage or find ways of coping with the fallout. In contrast, a terminal risk is one where the targets are either annihilated or irreversibly crippled in ways that radically reduce their potential to live the sort of life they aspire to. In the case of personal risks, for instance, a terminal outcome could for example be death, permanent severe brain injury, or a lifetime prison sentence. An example of a local terminal risk would be genocide leading to the annihilation of a people (this happened to several Indian nations). Permanent enslavement is another example.

1.2 Existential risks

In this paper we shall discuss risks of the sixth category, the one marked with an X. This is the category of global, terminal risks. I shall call these existential risks.

Existential risks are distinct from global endurable risks. Examples of the latter kind include: threats to the biodiversity of Earth's ecosphere, moderate global warming, global economic recessions (even major

ones), and possibly stifling cultural or religious eras such as the “dark ages”, even if they encompass the whole global community, provided they are transitory (though see the section on “Shrieks” below). To say that a particular global risk is endurable is evidently not to say that it is acceptable or not very serious. A world war fought with conventional weapons or a Nazi-style Reich lasting for a decade would be extremely horrible events even though they would fall under the rubric of endurable global risks since humanity could eventually recover. (On the other hand, they could be a local terminal risk for many individuals and for persecuted ethnic groups.)

I shall use the following definition of existential risks:

> Existential risk – One where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential.

An existential risk is one where humankind as a whole is imperiled. Existential disasters have major adverse consequences for the course of human civilization for all time to come.

Jason Matheney (Oct 2007), "Reducing the risk of human extinction."

Risk Analysis 27.
http://users.physics.harvard.edu/~wilson/pmpmta/Mahoney_extinction.pdf

Abstract

In this century a number of events could extinguish humanity. The probability of these events may be very low, but the expected value of preventing them could be high, as it represents the value of all future human lives. We review the challenges to studying human extinction risks and, by way of example, estimate the cost effectiveness of preventing extinction-level asteroid impacts.

Conclusion

We may be poorly equipped to recognize or plan for extinction risks (Yudkowsky, 2007). We may not be good at grasping the significance of very large numbers (catastrophic outcomes) or very small numbers (probabilities) over large timeframes. We struggle with estimating the probabilities of rare or unprecedented events (Kunreuther et al., 2001). Policymakers may not plan far beyond current political administrations and rarely do risk assessments value the existence of future generations.18 We may unjustifiably discount the value of future lives. Finally, extinction risks are market failures where an individual enjoys no perceptible benefit from his or her investment in risk reduction. Human survival may thus be a good requiring deliberate policies to protect.

It might be feared that consideration of extinction risks would lead to a reductio ad absurdum: we ought to invest all our resources in asteroid defense or nuclear disarmament, instead of AIDS, pollution, world hunger, or other problems we face today. On the contrary, programs that create a healthy and content global population are likely to reduce the probability of global war or catastrophic terrorism. They should thus be seen as an essential part of a portfolio of risk-reducing projects.

Discussing the risks of "nuclear winter," Carl Sagan (1983) wrote:

> Some have argued that the difference between the deaths of several hundred million people in a nuclear war (as has been thought until recently to be a reasonable upper limit) and the death of every person on Earth (as now seems possible) is only a matter of one order of magnitude.

> For me, the difference is considerably greater. Restricting our attention only to those who die as a consequence of the war conceals its full impact. If we are required to calibrate extinction in numerical terms, I would be sure to include the number of people in future generations who would not be born. A nuclear war imperils all of our descendants, for as long as there will be humans. Even if the population remains static, with an average lifetime of the order of 100 years, over a typical time period for the biological evolution of a successful species (roughly ten million years), we are talking about some 500 trillion people yet to come. By this criterion, the stakes are one million times greater for extinction than for the more modest nuclear wars that kill "only" hundreds of millions of people. There are many other possible measures of the potential loss—including culture and science, the evolutionary history of the planet, and the significance of the lives of all of our ancestors who contributed to the future of their descendants. Extinction is the undoing of the human enterprise.

In a similar vein, the philosopher Derek Parfit (1984) wrote:

> I believe that if we destroy mankind, as we now can, this outcome will be much worse than most people think. Compare three outcomes:
>
> 1. Peace
> 2. A nuclear war that kills 99% of the world's existing population
> 3. A nuclear war that kills 100%
>
> 2 would be worse than 1, and 3 would be worse than 2. Which is the greater of these two differences? Most people believe that the greater difference is between 1 and 2. I believe that the difference between 2 and 3 is very much greater …. The Earth will remain habitable for at least another billion years. Civilization began only a few thousand years ago. If we do not destroy mankind, these thousand years may be only a tiny fraction of the whole of civilized human history. The difference between 2 and 3 may thus be the difference between this tiny fraction and all of the rest of this history. If we compare this possible history to a day, what has occurred so far is only a fraction of a second.

Human extinction in the next few centuries could reduce the number of future generations by thousands or more. We take extraordinary measures to protect some endangered species from extinction. It might be reasonable to take extraordinary measures to protect humanity from the same. To decide whether this is so requires more discussion of the methodological problems mentioned here, as well as research on the extinction risks we face and the costs of mitigating them.

Nick Bostrom and Milan Ćirković (2008), "Introduction."

in *Global Catastrophic Risks*, ed. Nick Bostrom and Milan Cirkovic, Oxford University Press.
http://www.global-catastrophic-risks.com/docs/Chap01.pdf

1.2 Taxonomy and organization

Let us look more closely at what would, and would not, count as a global catastrophic risk. Recall that the damage must be serious, and the scale global. Given this, a catastrophe that caused 10,000 fatalities or 10 billion dollars' worth of economic damage (e.g., a major earthquake) would not qualify as a global catastrophe. A catastrophe that caused 10 million fatalities or 10 trillion dollars' worth of economic loss (e.g., an influenza pandemic) would count as a global catastrophe, even if some region of the world

escaped unscathed. As for disasters falling between these points, the definition is vague. The stipulation of a precise cut-off does not appear needful at this stage.

Global catastrophes have occurred many times in history, even if we only count disasters causing more than 10 million deaths. A very partial list of examples might include the An Shi Rebellion (756–763), the Taiping Rebellion (1851–1864), and the famine of the Great Leap Forward in China, the Black Death in Europe, the Spanish flu pandemic, the two world wars, the Nazi genocides, the famines in British India, Stalinist totalitarianism, the decimation of the native American population through smallpox and other diseases following the arrival of European colonizers, probably the Mongol conquests, perhaps Belgian Congo – innumerable others could be added to the list depending on how various misfortunes and chronic conditions are individuated and classified.

We can roughly characterize the severity of a risk by three variables: its scope (how many people – and other morally relevant beings – would be affected), its intensity (how badly these would be affected), and its probability (how likely the disaster is to occur, according to our best judgement, given currently available evidence). Using the first two of these variables, we can construct a qualitative diagram of different types of risk (Fig. 1.1). (The probability dimension could be displayed along a z-axis were this diagram three-dimensional.)

The scope of a risk can be personal (affecting only one person), local, global (affecting a large part of the human population), or trans-generational (affecting not only the current world population but all generations that could come to exist in the future). The intensity of a risk can be classified as imperceptible (barely noticeable), endurable (causing significant harm but not destroying quality of life completely), or terminal (causing death or permanently and drastically reducing quality of life). In this taxonomy, global catastrophic risks occupy the four risks classes in the high-severity upper-right corner of the figure: a global catastrophic risk is of either global or trans-generational scope, and of either endurable or terminal intensity. In principle, as suggested in the figure, the axes can be extended to encompass conceptually possible risks that are even more extreme. In particular, trans-generational risks can contain a subclass of risks so destructive that their realization would not only affect or pre-empt future human generations, but would also destroy the potential of our future light cone of the universe to produce intelligent or self-aware beings (labelled ‘Cosmic’). On the other hand, according to many theories of value, there can be states of being that are even worse than non-existence or death (e.g., permanent and extreme forms of slavery or mind control), so it could, in principle, be possible to extend the x-axis to the right as well (see Fig. 1.1 labelled ‘Hellish’).

A subset of global catastrophic risks is existential risks. An existential risk is one that threatens to cause the extinction of Earth-originating intelligent life or to reduce its quality of life (compared to what would otherwise have been possible) permanently and drastically.1 Existential risks share a number of features that mark them out as deserving of special consideration. For example, since it is not possible to recover from existential risks, we cannot allow even one existential disaster to happen; there would be no opportunity to learn from experience. Our approach to managing such risks must be proactive. How much worse an existential catastrophe would be than a non-existential global catastrophe depends very sensitively on controversial issues in value theory, in particular how much weight to give to the lives of possible future persons.2 Furthermore, assessing existential risks raises distinctive methodological problems having to do with observation selection effects and the need to avoid anthropic bias. One of the motives for producing this book is to stimulate more serious study of existential risks. Rather than limiting our focus to existential risk, however, we thought it better to lay a broader foundation of systematic thinking about big risks in general.

Anders Sandberg and Nick Bostrom (5 Dec 2008), "Global catastrophic risks survey."

Technical Report #2008-1, Future of Humanity Institute, Oxford University.
http://www.fhi.ox.ac.uk/gcr-report.pdf

Introduction

At the Global Catastrophic Risk Conference in Oxford (17-20 July, 2008) an informal survey was circulated among participants, asking them to make their best guess at the chance that there will be disasters of different types before 2100. This report summarizes the main results.

The median extinction risk estimates were: [See document for chart]

These results should be taken with a grain of salt. Non-responses have been omitted, although some might represent a statement of zero probability rather than no opinion. There are likely to be many cognitive biases that affect the result, such as unpacking bias and the availability heuristic–-well as old-fashioned optimism and pessimism.

In appendix A the results are plotted with individual response distributions visible.

Other Risks: The list of risks was not intended to be inclusive of all the biggest risks. Respondents were invited to contribute their own global catastrophic risks, showing risks they considered significant. Several suggested totalitarian world government, climate-induced disasters, ecological/resource crunches and "other risks"–-specified or unknowable threats. Other suggestions were asteroid/comet impacts, bad crisis management, high-tech asymmetric war attacking brittle IT-based societies, back-contamination from space probes, electromagnetic pulses, genocide/democides, risks from physics research and degradation of quality assurance.

Suggestions: Respondents were also asked to suggest what they would recommend to policymakers. Several argued for nuclear disarmament, or at least lowering the number of weapons under the threshold for existential catastrophe, as well as reducing stocks of highly enriched uranium and making nuclear arsenals harder to accidentally launch.

One option discussed was formation of global biotech-related governance, legislation and enforcement, or even a global body like the IPCC or UNFCCC to study and act on catastrophic risk. At the very least there was much interest in developing defenses against misuses of biotechnology, and a recognition for the need of unbiased early detection systems for a variety of risks, be they near Earth objects or actors with WMD capabilities.

Views on emerging technologies such as nanotech, AI, and cognition enhancement were mixed: some proposed avoiding funding them; others deliberate crash programs to ensure they would be in the right hands, the risks understood, and the technologies able to be used against other catastrophic risks.

Other suggestions included raising awareness of the problem, more research on cyber security issues, the need to build societal resiliency in depth, prepare for categories of disasters rather than individual types, building refuges and change energy consumption patterns.

Appendix A: Below are the individual results, shown as grey dots (jittered for distinguishability) and with the median as a bar. [See document for chart]

Nick Bostrom (2009), "The future of humanity."

Geopolitics, History & International Relations 1.
http://www.nickbostrom.com/papers/future.pdf

Abstract

The future of humanity is often viewed as a topic for idle speculation. Yet our beliefs and assumptions on this subject matter shape decisions in both our personal lives and public policy - decisions that have very real and sometimes unfortunate consequences. It is therefore practically important to try to develop a realistic mode of futuristic thought about big picture questions for humanity. This paper sketches an overview of some recent attempts in this direction, and it offers a brief discussion of four families of scenarios for humanity's future: extinction, recurrent collapse, plateau, and posthumanity.

Excerpt

The greatest extinction risks (and existential risks more generally) arise from human activity. Our species has survived volcanic eruptions, meteoric impacts, and other natural hazards for tens of thousands of years. It seems unlikely that any of these old risks should exterminate us in the near future. By contrast, human civilization is introducing many novel phenomena into the world, ranging from nuclear weapons to designer pathogens to high-energy particle colliders. The most severe existential risks of this century derive from expected technological developments. Advances in biotechnology might make it possible to design new viruses that combine the easy contagion and mutability of the influenza virus with the lethality of HTV. Molecular nanotechnology might make it possible to create weapons systems with a destructive power dwarfing that of both thermonuclear bombs and biowarfare agents.26 Super-intelligent machines might be built and their actions could determine the future of humanity - and whether there will be one.27 Considering that many of the existential risks that now seem to be among the most significant were conceptualized only in recent decades, it seems likely that further ones still remain to be discovered.

The same technologies that will pose these risks will also help us to mitigate some risks. Biotechnology can help us develop better diagnostics, vaccines, and anti-viral drugs. Molecular nanotechnology could offer even stronger prophylactics.28 Super-intelligent machines may be the last invention that human beings ever need to make, since a super-intelligence, by definition, would be far more effective than a human brain in practically all intellectual endeavors, including strategic thinking, scientific analysis, and technological creativity.29 In addition to creating and mitigating risks, these powerful technological capabilities would also affect the human condition in many other ways.

Extinction risks constitute an especially severe subset of what could go badly wrong for humanity. There are many possible global catastrophes that would cause immense worldwide damage, maybe even the collapse of modern civilization, yet fall short of terminating the human species. An all-out nuclear war between Russia and the United States might be an example of a global catastrophe that would be unlikely to result in extinction. A terrible pandemic with high virulence and 100% mortality rate among infected individuals might be another example: if some groups of humans could successfully quarantine themselves before being exposed, human extinction could be avoided even if, say, 95% or more of the world's population succumbed. What distinguishes extinction and other existential catastrophes is that a comeback is impossible. A non-existential disaster causing the breakdown of global civilization is, from the perspective of humanity as a whole, a potentially recoverable setback: a giant massacre for man, a small misstep for mankind.

An existential catastrophe is therefore qualitatively distinct from a "mere" collapse of global civilization, although in terms of our moral and prudential attitudes perhaps we should simply view both as unimaginably bad outcomes.30 One way that civilization collapse could be a significant feature in the larger picture for humanity, however, is if it formed part of a repeating pattern. This takes us to the second family of scenarios: recurrent collapse.

Nick Bostrom (Feb 2013), "Existential risk prevention as global priority."

Global Policy 4.
http://onlinelibrary.wiley.com/doi/10.1111/1758-5899.12002/abstract

Abstract

Existential risks are those that threaten the entire future of humanity. Many theories of value imply that even relatively small reductions in net existential risk have enormous expected value. Despite their importance, issues surrounding human-extinction risks and related hazards remain poorly understood. In this article, I clarify the concept of existential risk and develop an improved classification scheme. I discuss the relation between existential risks and basic issues in axiology, and show how existential risk reduction (via the maxipok rule) can serve as a strongly action-guiding principle for utilitarian concerns. I also show how the notion of existential risk suggests a new way of thinking about the ideal of sustainability.

Policy Implications

- Existential risk is a concept that can focus long-term global efforts and sustainability concerns.
- The biggest existential risks are anthropogenic and related to potential future technologies.
- A moral case can be made that existential risk reduction is strictly more important than any other global public good.
- Sustainability should be reconceptualised in dynamic terms, as aiming for a sustainable trajectory rather than a sustainable state.
- Some small existential risks can be mitigated today directly (e.g. asteroids) or indirectly (by building resilience and reserves to increase survivability in a range of extreme scenarios) but it is more important to build capacity to improve humanity's ability to deal with the larger existential risks that will arise later in this century. This will require collective wisdom, technology foresight, and the ability when necessary to mobilise a strong global coordinated response to anticipated existential risks.
- Perhaps the most cost-effective way to reduce existential risks today is to fund analysis of a wide range of existential risks and potential mitigation strategies, with a long-term perspective.

Kira Matus (25 Nov 2014), "Existential risk: challenges for risk regulation."

Risk and Regulation (Winter 2014).
http://www.lse.ac.uk/researchAndExpertise/units/CARR/pdf/Risk-and-Regulation-28-existential-risk.pdf

Full text

There is a trend in many areas towards attention to 'big' risks. Financial regulation has become increasingly concerned with so-called systemic risks. Others, and not just Hollywood blockbusters, have been attracted to the study of civilization-destroying catastrophic risks. Indeed, the OECD has become increasingly interested in 'high level' risks and ways in which different national governments seek to prepare for and manage actual events, such as the aftermath of major earthquakes, or the response to a terrorist attack. The notion of 'existential' risk might be adding to the cacophony of emerging 'big' risk concerns. However, existential risk deserves special attention as it fundamentally adds to our understanding of particular types of risks, and it also challenges common wisdom regarding actions designed to support continued survival.

What is existential risk? We can approach this question by looking at several attributes. The first attribute is what, in fact, is at risk. One set of existential risks are those that threaten survival. These are the acute catastrophes, i.e. the idea that particular events' impacts are likely to extinguish civilization. Such risks have been identified when it comes to asteroids, nuclear war, and other largescale events that undermine the possibility for survival in general, or, at least, in large regions. A second set is based on the idea that existential risks are not just about physical survival, but about the survival of ways of life. In other words, certain risks are seen as threatening established ways of doing things, cultures, social relationships, and understandings of the 'good life'. There is, of course, much disagreement about what the good life constitutes, and therefore there will always be disagreement as to what exactly an existential risk constitutes.

A second attribute is the degree to which an existential risk is triggered by a single catastrophic incident. Existential risks arise not merely from one-off large incidents, such as earthquakes, tsunamis, nuclear meltdowns or, indeed, asteroid hits. Rather, existential risks are about complex, inter-related processes that result in cascading effects that move across social systems. The overall impact of these system changes could result in the types of physical or cultural destruction that is the focus of the first two perspectives.

Whether triggered by catastrophic events or complex cascades, standard operating procedures are unlikely to be sufficient for dealing with existential risks; instead, this is a space in which improvisation and creativity are required. A third attribute of existential risks is the challenge they present to standard approaches to risk regulation. Existential risks are defined by their cross-systematic nature; a failure within one system (say, finance) has not just catastrophic implications for the sector in question, but threatens the survival of another system (say, the environment, as funding for particular measures dries up). In other words, the focus of existential risks is not just on the systemic level, it focuses on the cross-systemic dimension that is even more difficult to predict and assess than attempts aimed at establishing activities that are of 'systemic' relevance by regulatory systems that tend to be narrowly focused and independent from each other. Existential risks are characterized by a fourth feature, namely the idea that existential risks lead to responses based upon fear. Individuals are confronted with fears about their survival (death) and about the meaning of their lives. This aspect of existential risk is particularly troublesome in an age of low trust in authority and, consequently, a political style that is intolerant of 'blame free' spaces. In the absence of confidence in public authority, few options remain. For some, the solution will rely on framework plans, pop intellectuals and other fashionable ideas that seem to offer redemption from the fear of extinction. Others will prefer to 'go it alone' and seek to develop their own plans for survival, noting that risk taking is, after all, an individual choice. Others, again, will deny the legitimacy of public authority and veer towards those choices that have been legitimized by their own communities. Finally, some will deny that existential risks exist in the first place. In other words, individual responses to existential risks vary considerably and pose challenges for any risk management and communication strategy.

Existential risks therefore pose considerable difficulties for instruments of risk management and regulation. For one, regardless of probability, the severe impact of a particular risk makes resource and attention allocation decisions problematic. Interdependencies, threshold effects, and non-linearities make calculations regarding existential risks highly speculative. Furthermore, existential risks also lead to demands for deterministic statements ('is it absolutely safe'), a view that neither suits the risk-language of probabilities, nor is likely to attract much popular acceptance. Finally, while it might be possible to list a few existential risks at any point in time, attention is highly partial and changing. Today's high profile existential risks (and, therefore, tomorrow's cinematic blockbuster) might quickly move to the background as the news agenda shifts; yesterday's attention to environmental issues might quickly turn to public health or terrorism related topics.

What, then, can be done about existential risks? The list of sources of failures when it comes to existential risks is long, ranging from the 'failure of imagination' (of the 9/11 Commission report) to the 'failure of initiative' (in the case of the tragic events of 22 July 2011 in Norway). There are also some 'good news' stories, such as the self-organizing voluntary co-operation among communities in the immediate context of disasters as witnessed in Norway and the post-earthquake efforts in New Zealand's Christchurch. One of the most common recipes is to call for 'resilience'. Apart from an emphasis on capabilities for 'bouncing back' rather than seeking to prevent risks from occurring, there is little agreement as to what resilience actually is, or how it can be achieved. It is therefore, for example, questionable as to whether resilience can actually be designed. There are frameworks in high risk industries (such as oil platforms) that seek to measure resilience at the plant level, but whether such indicators can be developed for complex communities that are faced not with single events, but cascading effects, is more questionable. Furthermore, it is also questionable how far resilience can be taken since there is little scope for bouncing back after a major asteroid hit. In some (or many) cases, change and adaptation may therefore be unavoidable.

Resilience implies that individuals have a responsibility for managing risks. This, again, raises considerable problems for resilience. First responders and other types of crisis managers might be willing to undertake continuous crisis and emergency training, and read commission and inquiry reports to draw lessons. However, it is highly unlikely that high level politicians and, let alone, populations at large will consider insights from weighty and learned inquiries. How to communicate resilience strategies to communities (and to politicians) is a key challenge. Finally, resilience requires a capacity to adapt that assumes a certain level of trust, in individuals and their co-operation, as well as in the backup resourcing by public authority. Whether such pre-requisites can be assumed or even engineered is, again, doubtful, especially in an age of cutbacks in public expenditures.

Existential risks therefore deserve specific attention when it comes to the study and practise of risk and crisis management. It points to the traditional themes that have featured in crisis management and the wider public management literature, especially in terms of inter- and intra-organizational learning and co-ordination. Furthermore, it points to particular existential properties that need to be taken into consideration when managing risks. These properties point not just to individual fears and distrust in public authority, they also point to the inter-related, cross-system nature of particular risks that pose a key threat for contemporary societies. How regulation and policy can be structured to be attentive to these complexities and interdependencies is an area that requires a great deal more academic and practical attention.

Seth Baum (Dec 2014), "The great downside dilemma for risky emerging technologies."
Physica Scripta 89.
http://iopscience.iop.org/1402-4896/89/12/128004/pdf/1402-4896_89_12_128004.pdf

Abstract

Some emerging technologies promise to significantly improve the human condition, but come with a risk of failure so catastrophic that human civilization may not survive. This article discusses the great downside dilemma posed by the decision of whether or not to use these technologies. The dilemma is: use the technology, and risk the downside of catastrophic failure, or do not use the technology, and suffer through life without it. Historical precedents include the first nuclear weapon test and messaging to extraterrestrial intelligence. Contemporary examples include stratospheric geoengineering, a technology under development in response to global warming, and artificial general intelligence, a technology that could even take over the world. How the dilemma should be resolved depends on the details of each technology's downside risk and on what the human condition would otherwise be. Meanwhile, other technologies do not pose this dilemma, including sustainable design technologies, nuclear fusion power, and space colonization. Decisions on all of these technologies should be made with the long-term interests of human civilization in mind. This paper is part of a series of papers based on presentations at the event Emerging Technologies and the Future of Humanity held at the Royal Swedish Academy of Sciences, 17 March 2014.

Non-Technical Summary

Background: The Great Downside Dilemma

A downside dilemma is any decision in which one option promises benefits but comes with a risk of significant harm. An example is the game of Russian roulette. The decision is whether to play. Choosing to play promises benefits but comes with the risk of death. This paper introduces the great downside dilemma as any decision in which one option promises great benefits to humanity but comes with a risk of human civilization being destroyed. This dilemma is great because the stakes are so high—indeed, they are astronomically high. The great downside dilemma is especially common with emerging technologies.

Historical Precedents: Nuclear Weapons and Messaging to Extraterrestrial Intelligence

The great downside dilemma for emerging technologies has been faced at least twice before. The first precedent is nuclear weapons. It came in the desperate circumstances of World War II: the decision of whether to test detonate the first nuclear weapon. Some physicists suspected that the detonation could ignite the atmosphere, killing everyone on Earth. Fortunately, they understood the physics well enough to correctly figure out that the ignition wouldn't happen. The second precedent is messaging to extraterrestrial intelligence (METI). The decision was whether to send messages. While some messages have been sent, METI is of note because the dilemma still has not been resolved. Humanity still does not know if METI is safe. Thus METI decisions today face the same basic dilemma as the initial decisions in decades past.

Dilemmas in the Making: Stratospheric Geoengineering and Artificial General Intelligence

Several new instances of the great downside dilemma lurk on the horizon. The stakes for these new dilemmas are even higher, because they come with much higher probabilities of catastrophe. This paper discusses two. The first is stratospheric geoengineering, which promises to avoid the most catastrophic

effects of global warming. However, stratospheric geoengineering could fail, bringing an even more severe catastrophe. The second is artificial general intelligence, which could either solve a great many of humanity's problems or kill everyone, depending on how it is designed. Neither of these two technologies currently exists, but both are subjects of active research and development. Understanding these technologies and the dilemmas they pose is already important, and it will only get more important as the technologies progress.

Technologies That Don't Pose the Great Downside Dilemma

Not all technologies present a great downside dilemma. These technologies may have downsides, but they do not threaten significant catastrophic harm to human civilization. Some of these technologies even hold great potential to improve the human condition, including by reducing other catastrophic risks. These latter technologies are especially attractive and in general should be pursued. The paper discusses three such technologies: sustainable design technology, nuclear fusion power, and space colonization technology. Some sustainable design is quite affordable, including the humble bicycle, while nuclear fusion and space colonization are quite expensive. However, all of these technologies can play a helpful role in improving the human condition and avoiding catastrophe.

Dennis Pamlin and Stuart Armstrong (19 Feb 2015), "Twelve risks that threaten human civilisation: The case for a new risk category."

Global Challenges Foundation.
http://globalchallenges.org/publications/globalrisks/about-the-project

Excerpt

The 12 global risks that threaten human civilization are:

Current risks

1. Extreme Climate Change
2. Nuclear War
3. Ecological Catastrophe
4. Global Pandemic
5. Global System Collapse

Exogenic risks

6. Major Asteroid Impact
7. Supervolcano

Emerging risks

8. Synthetic Biology
9. Nanotechnology
10. Artificial Intelligence
11. Uncertain Risks

Global policy risks

12. Future Bad Global Governance

There are ten areas that could help mitigate immediate threats while also contributing to a future global governance system capable of addressing global risks with a potential infinite impact:

1. Global challenges leadership networks
2. Better quality risk assessment for global challenges
3. Development of early warning systems
4. Encouraging visualisation of complex systems
5. Highlighting early movers
6. Including the whole probability distribution
7. Increasing the focus on the probability of extreme events
8. Encouraging appropriate language to describe extreme risks
9. Establishing a Global Risk and Opportunity Indicator to guide governance
10. Explore the possibility of establishing a Global Risk Organisation (GRO)

Seth Baum and Anthony Barrett (5 Feb 2015), "The most extreme risks: Global catastrophes."

in *The Gower Handbook of Extreme Risk*, ed. Vicki Bier, Gower (forthcoming).
http://sethbaum.com/ac/fc_Extreme.pdf

Abstract

The most extreme risk are those that threaten the entirety of human civilization, known as global catastrophic risks. The very extreme nature of global catastrophes makes them both challenging to analyze and important to address. They are challenging to analyze because they are largely unprecedented and because they involve the entire global human system. They are important to address because they threaten everyone around the world and future generations. Global catastrophic risks also pose some deep dilemmas. One dilemma occurs when actions to reduce global catastrophic risk could harm society in other ways, as in the case of geoengineering to reduce catastrophic climate change risk. Another dilemma occurs when reducing one global catastrophic risk could increase another, as in the case of nuclear power reducing climate change risk while increasing risks from nuclear weapons. The complex, interrelated nature of global catastrophic risk suggests a research agenda in which the full space of risks are assessed in an integrated fashion in consideration of the deep dilemmas and other challenges they pose. Such an agenda can help identify the best ways to manage these most extreme risks and keep human civilization safe.

Excerpt

2. What Is GCR And Why Is It Important?

Taken literally, a global catastrophe can be any event that is in some way catastrophic across the globe. This suggests a rather low threshold for what counts as a global catastrophe. An event causing just one death on each continent (say, from a jet-setting assassin) could rate as a global catastrophe, because surely these deaths would be catastrophic for the deceased and their loved ones. However, in common usage, a global catastrophe would be catastrophic for a significant portion of the globe. Minimum thresholds have variously been set around ten thousand to ten million deaths or $10 billion to $10 trillion in damages (Bostrom and Ćirković 2008), or death of one quarter of the human population (Atkinson 1999; Hempsell 2004). Others have emphasized catastrophes that cause long-term declines in the trajectory of human civilization (Beckstead 2013), that human civilization does not recover from

(Maher and Baum 2013), that drastically reduce humanity's potential for future achievements (Bostrom 2002, using the term "existential risk"), or that result in human extinction (Matheny 2007; Posner 2004).

A common theme across all these treatments of GCR is that some catastrophes are vastly more important than others. Carl Sagan was perhaps the first to recognize this, in his commentary on nuclear winter (Sagan 1983). Without nuclear winter, a global nuclear war might kill several hundred million people. This is obviously a major catastrophe, but humanity would presumably carry on. However, with nuclear winter, per Sagan, humanity could go extinct. The loss would be not just an additional four billion or so deaths, but the loss of all future generations. To paraphrase Sagan, the loss would be billions and billions of lives, or even more. Sagan estimated 500 trillion lives, assuming humanity would continue for ten million more years, which he cited as typical for a successful species.

Sagan's 500 trillion number may even be an underestimate. The analysis here takes an adventurous turn, hinging on the evolution of the human species and the long-term fate of the universe. On these long time scales, the descendants of contemporary humans may no longer be recognizably "human". The issue then is whether the descendants are still worth caring about, whatever they are. If they are, then it begs the question of how many of them there will be. Barring major global catastrophe, Earth will remain habitable for about one billion more years until the Sun gets too warm and large. The rest of the Solar System, Milky Way galaxy, universe, and (if it exists) the multiverse will remain habitable for a lot longer than that (Adams and Laughlin 1997), should our descendants gain the capacity to migrate there. An open question in astronomy is whether it is possible for the descendants of humanity to continue living for an infinite length of time or instead merely an astronomically large but finite length of time (see e.g. Ćirković 2002; Kaku 2005). Either way, the stakes with global catastrophes could be much larger than the loss of 500 trillion lives.

Debates about the infinite vs. the merely astronomical are of theoretical interest (Ng 1991; Bossert et al. 2007), but they have limited practical significance. This can be seen when evaluating GCRs from a standard risk-equals-probability-times-magnitude framework. Using Sagan's 500 trillion lives estimate, it follows that reducing the probability of global catastrophe by a mere one-in-500-trillion chance is of the same significance as saving one human life. Phrased differently, society should try 500 trillion times harder to prevent a global catastrophe than it should to save a person's life. Or, preventing one million deaths is equivalent to a one-in-500-million reduction in the probability of global catastrophe. This suggests society should make extremely large investment in GCR reduction, at the expense of virtually all other objectives. Judge and legal scholar Richard Posner made a similar point in monetary terms (Posner 2004). Posner used $50,000 as the value of a statistical human life (VSL) and 12 billion humans as the total loss of life (double the 2004 world population); he describes both figures as significant underestimates. Multiplying them gives $600 trillion as an underestimate of the value of preventing global catastrophe. For comparison, the United States government typically uses a VSL of around one to ten million dollars (Robinson 2007). Multiplying a $10 million VSL with 500 trillion lives gives $5x1021 as the value of preventing global catastrophe. But even using "just" $600 trillion, society should be willing to spend at least that much to prevent a global catastrophe, which converts to being willing to spend at least $1 million for a one-in-500-million reduction in the probability of global catastrophe. Thus while reasonable disagreement exists on how large of a VSL to use and how much to count future generations, even low-end positions suggest vast resource allocations should be redirected to reducing GCR. This conclusion is only strengthened when considering the astronomical size of the stakes, but the same point holds either way. The bottom line is that, as long as something along the lines of the standard risk-equals-probability-times-magnitude framework is being used, then even tiny GCR reductions merit significant effort. This point holds especially strongly for risks of catastrophes that would cause permanent harm to global human civilization.

The discussion thus far has assumed that all human lives are valued equally. This assumption is not universally held. People often value some people more than others, favoring themselves, their family and friends, their compatriots, their generation, or others whom they identify with. Great debates rage on across moral philosophy, economics, and other fields about how much people should value others who are distant in space, time, or social relation, as well as the unborn members of future generations. This debate is crucial for all valuations of risk, including GCR. Indeed, if each of us only cares about our immediate selves, then global catastrophes may not be especially important, and we probably have better things to do with our time than worry about them. While everyone has the right to their own views and feelings, we find that the strongest arguments are for the widely held position that all human lives should be valued equally. This position is succinctly stated in the United States Declaration of Independence, updated in the 1848 Declaration of Sentiments: “We hold these truths to be self-evident: that all men and women are created equal”. Philosophers speak of an agent-neutral, objective “view from nowhere” (Nagel 1986) or a “veil of ignorance” (Rawls 1971) in which each person considers what is best for society irrespective of which member of society they happen to be. Such a live. This in turn suggests a very high value for reducing GCR, or a high degree of priority for GCR reduction efforts.

Seth Baum (4 May 2015), "The far future argument for confronting catastrophic threats to humanity: Practical significance and alternatives."

Futures (in press),
http://www.sciencedirect.com/science/article/pii/S0016328715000312
http://www.sethbaum.com/ac/2015_FarFuture.pdf

Abstract

Sufficiently large catastrophes can affect human civilization into the far future: thousands, millions, or billions of years from now, or even longer. The far future argument says that people should confront catastrophic threats to humanity in order to improve the far future trajectory of human civilization. However, many people are not motivated to help the far future. They are concerned only with the near future, or only with themselves and their communities. This paper assesses the extent to which practical actions to confront catastrophic threats require support for the far future argument and proposes two alternative means of motivating actions. First, many catastrophes could occur in the near future; actions to confront them have near-future benefits. Second, many actions have co-benefits unrelated to catastrophes, and can be mainstreamed into established activities. Most actions, covering most of the total threat, can be motivated with one or both of these alternatives. However, some catastrophe-confronting actions can only be justified with reference to the far future. Attention to the far future can also sometimes inspire additional action. Confronting catastrophic threats best succeeds when it considers the specific practical actions to confront the threats and the various motivations people may have to take these actions.

Andy Majot and Roman Yampolskiy (13 Apr 2015), "Global catastrophic risk and security implications of quantum computers."

Futures (In press).
http://www.sciencedirect.com/science/article/pii/S0016328715000294

Abstract

With advancements in quantum computing happening almost weekly it is time to examine the effects this new technology will have on society and current computational systems. Specifically, cryptographic systems need to be carefully analyzed since the introduction of quantum computational resources would render discrete logarithm and factoring based cryptographic systems like those based on Rivest, Shamir, Adleman (RSA) and Elliptic Curve Cryptography (ECC) algorithms woefully obsolete. These algorithms are widely used in the form of digital certificates, message encryption, and even physical authentication devices like Radio Frequency Identification (RFID) badges. With this technology compromised by quantum computing, governments and other organizations would be able to eavesdrop on private citizens with relative ease. This has the potential to cause a slew of rights violations and atrocities leading to catastrophe. With compromised digital certificates 3rd parties could masquerade as trusted organizations. This would call many types of digital transactions like into question, including those related to stock exchanges, personal banking, and software verification. By eroding this previously solid foundation of trust global scale economic catastrophes are not out of the question. This paper introduces quantum computing to the study of catastrophic threats since the use of quantum technology while existing vulnerable encryption schemes are still in place raises severe safety issues. These issues are addressed here along with a proposed two-fold solution involving the development and maturation of post-quantum cryptographic algorithms coupled with government and international regulation. This regulation would promote the containment and responsible use of quantum computers in order to help alleviate some of the security issues posed by outdated cryptographic systems in a post-quantum environment.

Philip Doty (Jul 2015), "U.S. homeland security and risk assessment."

Government Information Quarterly 32.
http://www.sciencedirect.com/science/article/pii/S0740624X15000623

Abstract

Risk is constitutive of homeland security policy in the United States, and the risk apparatus supports growing concentration of executive power, increased surveillance, and secrecy. For example, the Transportation Security Administration in the Department of Homeland Security employs risk assessment particularly against groups considered “other.” Using the work of mostly European scholars, especially the literatures about Foucault's governmentality and Beck's risk society, the paper combines theory with empirical work by governmental agencies on transparency, secrecy, and risk assessment methods used in the Department of Homeland Security, providing insight into the securitization of the American state. Risk is a means to futurize threats to the polity, to create the security imaginary, a fictionalization that creates a moral panic and a climate of fear in seeking to cope with uncertainty. With those limitations of risk in mind, we can question four important elements of risk in U.S. security practice: “connecting the dots”; the quantitative bases of risk assessment algorithms; how risk assessment tends to ignore the important if circular intentionality of terror; and the difficulties inherent in controlling populations by classification, especially other-ed populations. The paper concludes with suggestions about unmasking the uncertainty of risk assessment and enabling oversight of its practice by legislative, judicial, and public actors.

Popular Journalism & Public Speeches on Existential Risk

Bill Joy (Apr 2000), "Why the future doesn't need us."

Wired.
http://archive.wired.com/wired/archive/8.04/joy.html

Full text

From the moment I became involved in the creation of new technologies, their ethical dimensions have concerned me, but it was only in the autumn of 1998 that I became anxiously aware of how great are the dangers facing us in the 21st century. I can date the onset of my unease to the day I met Ray Kurzweil, the deservedly famous inventor of the first reading machine for the blind and many other amazing things.

Ray and I were both speakers at George Gilder's Telecosm conference, and I encountered him by chance in the bar of the hotel after both our sessions were over. I was sitting with John Searle, a Berkeley philosopher who studies consciousness. While we were talking, Ray approached and a conversation began, the subject of which haunts me to this day.

I had missed Ray's talk and the subsequent panel that Ray and John had been on, and they now picked right up where they'd left off, with Ray saying that the rate of improvement of technology was going to accelerate and that we were going to become robots or fuse with robots or something like that, and John countering that this couldn't happen, because the robots couldn't be conscious.

While I had heard such talk before, I had always felt sentient robots were in the realm of science fiction. But now, from someone I respected, I was hearing a strong argument that they were a near-term possibility. I was taken aback, especially given Ray's proven ability to imagine and create the future. I already knew that new technologies like genetic engineering and nanotechnology were giving us the power to remake the world, but a realistic and imminent scenario for intelligent robots surprised me.

It's easy to get jaded about such breakthroughs. We hear in the news almost every day of some kind of technological or scientific advance. Yet this was no ordinary prediction. In the hotel bar, Ray gave me a partial preprint of his then-forthcoming book *The Age of Spiritual Machines,* which outlined a utopia he foresaw - one in which humans gained near immortality by becoming one with robotic technology. On reading it, my sense of unease only intensified; I felt sure he had to be understating the dangers, understating the probability of a bad outcome along this path.

I found myself most troubled by a passage detailing a *dys*topian scenario:

THE NEW LUDDITE CHALLENGE

First let us postulate that the computer scientists succeed in developing intelligent machines that can do all things better than human beings can do them. In that case presumably all work will be done by vast, highly organized systems of machines and no human effort will be necessary. Either of two cases might occur. The machines might be permitted to make all of their own decisions without human oversight, or else human control over the machines might be retained.

If the machines are permitted to make all their own decisions, we can't make any conjectures as to the results, because it is impossible to guess how such machines might behave. We only point out that the fate of the human race would be at the mercy of the machines. It might be argued that the human race

would never be foolish enough to hand over all the power to the machines. But we are suggesting neither that the human race would voluntarily turn power over to the machines nor that the machines would willfully seize power. What we do suggest is that the human race might easily permit itself to drift into a position of such dependence on the machines that it would have no practical choice but to accept all of the machines' decisions. As society and the problems that face it become more and more complex and machines become more and more intelligent, people will let machines make more of their decisions for them, simply because machine-made decisions will bring better results than man-made ones. Eventually a stage may be reached at which the decisions necessary to keep the system running will be so complex that human beings will be incapable of making them intelligently. At that stage the machines will be in effective control. People won't be able to just turn the machines off, because they will be so dependent on them that turning them off would amount to suicide.

On the other hand it is possible that human control over the machines may be retained. In that case the average man may have control over certain private machines of his own, such as his car or his personal computer, but control over large systems of machines will be in the hands of a tiny elite - just as it is today, but with two differences. Due to improved techniques the elite will have greater control over the masses; and because human work will no longer be necessary the masses will be superfluous, a useless burden on the system. If the elite is ruthless they may simply decide to exterminate the mass of humanity. If they are humane they may use propaganda or other psychological or biological techniques to reduce the birth rate until the mass of humanity becomes extinct, leaving the world to the elite. Or, if the elite consists of soft-hearted liberals, they may decide to play the role of good shepherds to the rest of the human race. They will see to it that everyone's physical needs are satisfied, that all children are raised under psychologically hygienic conditions, that everyone has a wholesome hobby to keep him busy, and that anyone who may become dissatisfied undergoes "treatment" to cure his "problem." Of course, life will be so purposeless that people will have to be biologically or psychologically engineered either to remove their need for the power process or make them "sublimate" their drive for power into some harmless hobby. These engineered human beings may be happy in such a society, but they will most certainly not be free. They will have been reduced to the status of domestic animals.[1]

In the book, you don't discover until you turn the page that the author of this passage is Theodore Kaczynski - the Unabomber. I am no apologist for Kaczynski. His bombs killed three people during a 17-year terror campaign and wounded many others. One of his bombs gravely injured my friend David Gelernter, one of the most brilliant and visionary computer scientists of our time. Like many of my colleagues, I felt that I could easily have been the Unabomber's next target.

Kaczynski's actions were murderous and, in my view, criminally insane. He is clearly a Luddite, but simply saying this does not dismiss his argument; as difficult as it is for me to acknowledge, I saw some merit in the reasoning in this single passage. I felt compelled to confront it.

Kaczynski's dystopian vision describes unintended consequences, a well-known problem with the design and use of technology, and one that is clearly related to Murphy's law - "Anything that can go wrong, will." (Actually, this is Finagle's law, which in itself shows that Finagle was right.) Our overuse of antibiotics has led to what may be the biggest such problem so far: the emergence of antibiotic-resistant and much more dangerous bacteria. Similar things happened when attempts to eliminate malarial mosquitoes using DDT caused them to acquire DDT resistance; malarial parasites likewise acquired multi-drug-resistant genes.[2]

The cause of many such surprises seems clear: The systems involved are complex, involving interaction among and feedback between many parts. Any changes to such a system will cascade in ways that are difficult to predict; this is especially true when human actions are involved.

I started showing friends the Kaczynski quote from *The Age of Spiritual Machines;* I would hand them Kurzweil's book, let them read the quote, and then watch their reaction as they discovered who had written it. At around the same time, I found Hans Moravec's book *Robot: Mere Machine to Transcendent Mind.* Moravec is one of the leaders in robotics research, and was a founder of the world's largest robotics research program, at Carnegie Mellon University. *Robot* gave me more material to try out on my friends - material surprisingly supportive of Kaczynski's argument. For example:

The Short Run (Early 2000s)

Biological species almost never survive encounters with superior competitors. Ten million years ago, South and North America were separated by a sunken Panama isthmus. South America, like Australia today, was populated by marsupial mammals, including pouched equivalents of rats, deers, and tigers. When the isthmus connecting North and South America rose, it took only a few thousand years for the northern placental species, with slightly more effective metabolisms and reproductive and nervous systems, to displace and eliminate almost all the southern marsupials.

In a completely free marketplace, superior robots would surely affect humans as North American placentals affected South American marsupials (and as humans have affected countless species). Robotic industries would compete vigorously among themselves for matter, energy, and space, incidentally driving their price beyond human reach. Unable to afford the necessities of life, biological humans would be squeezed out of existence.

There is probably some breathing room, because we do not live in a completely free marketplace. Government coerces nonmarket behavior, especially by collecting taxes. Judiciously applied, governmental coercion could support human populations in high style on the fruits of robot labor, perhaps for a long while.

A textbook dystopia - and Moravec is just getting wound up. He goes on to discuss how our main job in the 21st century will be "ensuring continued cooperation from the robot industries" by passing laws decreeing that they be "nice," and to describe how seriously dangerous a human can be "once transformed into an unbounded superintelligent robot." Moravec's view is that the robots will eventually succeed us - that humans clearly face extinction.

I decided it was time to talk to my friend Danny Hillis. Danny became famous as the cofounder of Thinking Machines Corporation, which built a very powerful parallel supercomputer. Despite my current job title of Chief Scientist at Sun Microsystems, I am more a computer architect than a scientist, and I respect Danny's knowledge of the information and physical sciences more than that of any other single person I know. Danny is also a highly regarded futurist who thinks long-term - four years ago he started the Long Now Foundation, which is building a clock designed to last 10,000 years, in an attempt to draw attention to the pitifully short attention span of our society. (See "Test of Time, " *Wired* 8.03, page 78.)

So I flew to Los Angeles for the express purpose of having dinner with Danny and his wife, Pati. I went through my now-familiar routine, trotting out the ideas and passages that I found so disturbing. Danny's answer - directed specifically at Kurzweil's scenario of humans merging with robots - came swiftly, and quite surprised me. He said, simply, that the changes would come gradually, and that we would get used to them.

But I guess I wasn't totally surprised. I had seen a quote from Danny in Kurzweil's book in which he said, "I'm as fond of my body as anyone, but if I can be 200 with a body of silicon, I'll take it." It seemed that he was at peace with this process and its attendant risks, while I was not.

While talking and thinking about Kurzweil, Kaczynski, and Moravec, I suddenly remembered a novel I had read almost 20 years ago -*The White Plague,* by Frank Herbert - in which a molecular biologist is driven insane by the senseless murder of his family. To seek revenge he constructs and disseminates a new and highly contagious plague that kills widely but selectively. (We're lucky Kaczynski was a mathematician, not a molecular biologist.) I was also reminded of the Borg of *Star Trek,* a hive of partly biological, partly robotic creatures with a strong destructive streak. Borg-like disasters are a staple of science fiction, so why hadn't I been more concerned about such robotic dystopias earlier? Why weren't other people more concerned about these nightmarish scenarios?

Part of the answer certainly lies in our attitude toward the new - in our bias toward instant familiarity and unquestioning acceptance. Accustomed to living with almost routine scientific breakthroughs, we have yet to come to terms with the fact that the most compelling 21st-century technologies - robotics, genetic engineering, and nanotechnology - pose a different threat than the technologies that have come before. Specifically, robots, engineered organisms, and nanobots share a dangerous amplifying factor: They can self-replicate. A bomb is blown up only once - but one bot can become many, and quickly get out of control.

Much of my work over the past 25 years has been on computer networking, where the sending and receiving of messages creates the opportunity for out-of-control replication. But while replication in a computer or a computer network can be a nuisance, at worst it disables a machine or takes down a network or network service. Uncontrolled self-replication in these newer technologies runs a much greater risk: a risk of substantial damage in the physical world.

Each of these technologies also offers untold promise: The vision of near immortality that Kurzweil sees in his robot dreams drives us forward; genetic engineering may soon provide treatments, if not outright cures, for most diseases; and nanotechnology and nanomedicine can address yet more ills. Together they could significantly extend our average life span and improve the quality of our lives. Yet, with each of these technologies, a sequence of small, individually sensible advances leads to an accumulation of great power and, concomitantly, great danger.

What was different in the 20th century? Certainly, the technologies underlying the weapons of mass destruction (WMD) - nuclear, biological, and chemical (NBC) - were powerful, and the weapons an enormous threat. But building nuclear weapons required, at least for a time, access to both rare - indeed, effectively unavailable - raw materials and highly protected information; biological and chemical weapons programs also tended to require large-scale activities.

The 21st-century technologies - genetics, nanotechnology, and robotics (GNR) - are so powerful that they can spawn whole new classes of accidents and abuses. Most dangerously, for the first time, these accidents and abuses are widely within the reach of individuals or small groups. They will not require large facilities or rare raw materials. Knowledge alone will enable the use of them.

Thus we have the possibility not just of weapons of mass destruction but of knowledge-enabled mass destruction (KMD), this destructiveness hugely amplified by the power of self-replication.

I think it is no exaggeration to say we are on the cusp of the further perfection of extreme evil, an evil whose possibility spreads well beyond that which weapons of mass destruction bequeathed to the nation-states, on to a surprising and terrible empowerment of extreme individuals.

Nothing about the way I got involved with computers suggested to me that I was going to be facing these kinds of issues.

My life has been driven by a deep need to ask questions and find answers. When I was 3, I was already reading, so my father took me to the elementary school, where I sat on the principal's lap and read him a story. I started school early, later skipped a grade, and escaped into books - I was incredibly motivated to learn. I asked lots of questions, often driving adults to distraction.

As a teenager I was very interested in science and technology. I wanted to be a ham radio operator but didn't have the money to buy the equipment. Ham radio was the Internet of its time: very addictive, and quite solitary. Money issues aside, my mother put her foot down - I was not to be a ham; I was antisocial enough already.

I may not have had many close friends, but I was awash in ideas. By high school, I had discovered the great science fiction writers. I remember especially Heinlein's *Have Spacesuit Will Travel* and Asimov's *I, Robot,* with its Three Laws of Robotics. I was enchanted by the descriptions of space travel, and wanted to have a telescope to look at the stars; since I had no money to buy or make one, I checked books on telescope-making out of the library and read about making them instead. I soared in my imagination.

Thursday nights my parents went bowling, and we kids stayed home alone. It was the night of Gene Roddenberry's original *Star Trek,* and the program made a big impression on me. I came to accept its notion that humans had a future in space, Western-style, with big heroes and adventures. Roddenberry's vision of the centuries to come was one with strong moral values, embodied in codes like the Prime Directive: to not interfere in the development of less technologically advanced civilizations. This had an incredible appeal to me; ethical humans, not robots, dominated this future, and I took Roddenberry's dream as part of my own.

I excelled in mathematics in high school, and when I went to the University of Michigan as an undergraduate engineering student I took the advanced curriculum of the mathematics majors. Solving math problems was an exciting challenge, but when I discovered computers I found something much more interesting: a machine into which you could put a program that attempted to solve a problem, after which the machine quickly checked the solution. The computer had a clear notion of correct and incorrect, true and false. Were my ideas correct? The machine could tell me. This was very seductive.

I was lucky enough to get a job programming early supercomputers and discovered the amazing power of large machines to numerically simulate advanced designs. When I went to graduate school at UC Berkeley in the mid-1970s, I started staying up late, often all night, inventing new worlds inside the machines. Solving problems. Writing the code that argued so strongly to be written.

In*The Agony and the Ecstasy,* Irving Stone's biographical novel of Michelangelo, Stone described vividly how Michelangelo released the statues from the stone, "breaking the marble spell," carving from the images in his mind.4 In my most ecstatic moments, the software in the computer emerged in the same way. Once I had imagined it in my mind I felt that it was already there in the machine, waiting to be released. Staying up all night seemed a small price to pay to free it - to give the ideas concrete form.

After a few years at Berkeley I started to send out some of the software I had written - an instructional Pascal system, Unix utilities, and a text editor called vi (which is still, to my surprise, widely used more than 20 years later) - to others who had similar small PDP-11 and VAX minicomputers. These adventures in software eventually turned into the Berkeley version of the Unix operating system, which became a personal "success disaster" - so many people wanted it that I never finished my PhD. Instead I got a job working for Darpa putting Berkeley Unix on the Internet and fixing it to be reliable and to run large research applications well. This was all great fun and very rewarding. And, frankly, I saw no robots here, or anywhere near.

Still, by the early 1980s, I was drowning. The Unix releases were very successful, and my little project of one soon had money and some staff, but the problem at Berkeley was always office space rather than money - there wasn't room for the help the project needed, so when the other founders of Sun Microsystems showed up I jumped at the chance to join them. At Sun, the long hours continued into the early days of workstations and personal computers, and I have enjoyed participating in the creation of advanced microprocessor technologies and Internet technologies such as Java and Jini.

From all this, I trust it is clear that I am not a Luddite. I have always, rather, had a strong belief in the value of the scientific search for truth and in the ability of great engineering to bring material progress. The Industrial Revolution has immeasurably improved everyone's life over the last couple hundred years, and I always expected my career to involve the building of worthwhile solutions to real problems, one problem at a time.

I have not been disappointed. My work has had more impact than I had ever hoped for and has been more widely used than I could have reasonably expected. I have spent the last 20 years still trying to figure out how to make computers as reliable as I want them to be (they are not nearly there yet) and how to make them simple to use (a goal that has met with even less relative success). Despite some progress, the problems that remain seem even more daunting.

But while I was aware of the moral dilemmas surrounding technology's consequences in fields like weapons research, I did not expect that I would confront such issues in my own field, or at least not so soon.

Perhaps it is always hard to see the bigger impact while you are in the vortex of a change. Failing to understand the consequences of our inventions while we are in the rapture of discovery and innovation seems to be a common fault of scientists and technologists; we have long been driven by the overarching desire to know that is the nature of science's quest, not stopping to notice that the progress to newer and more powerful technologies can take on a life of its own.

I have long realized that the big advances in information technology come not from the work of computer scientists, computer architects, or electrical engineers, but from that of physical scientists. The physicists Stephen Wolfram and Brosl Hasslacher introduced me, in the early 1980s, to chaos theory and nonlinear systems. In the 1990s, I learned about complex systems from conversations with Danny Hillis, the biologist Stuart Kauffman, the Nobel-laureate physicist Murray Gell-Mann, and others. Most recently, Hasslacher and the electrical engineer and device physicist Mark Reed have been giving me insight into the incredible possibilities of molecular electronics.

In my own work, as codesigner of three microprocessor architectures - SPARC, picoJava, and MAJC - and as the designer of several implementations thereof, I've been afforded a deep and firsthand acquaintance with Moore's law. For decades, Moore's law has correctly predicted the exponential rate of improvement of semiconductor technology. Until last year I believed that the rate of advances predicted by Moore's law might continue only until roughly 2010, when some physical limits would begin to be reached. It was not obvious to me that a new technology would arrive in time to keep performance advancing smoothly.

But because of the recent rapid and radical progress in molecular electronics - where individual atoms and molecules replace lithographically drawn transistors - and related nanoscale technologies, we should be able to meet or exceed the Moore's law rate of progress for another 30 years. By 2030, we are likely to be able to build machines, in quantity, a million times as powerful as the personal computers of today - sufficient to implement the dreams of Kurzweil and Moravec.

As this enormous computing power is combined with the manipulative advances of the physical sciences and the new, deep understandings in genetics, enormous transformative power is being unleashed. These combinations open up the opportunity to completely redesign the world, for better or worse: The replicating and evolving processes that have been confined to the natural world are about to become realms of human endeavor.

In designing software and microprocessors, I have never had the feeling that I was designing an intelligent machine. The software and hardware is so fragile and the capabilities of the machine to "think" so clearly absent that, even as a possibility, this has always seemed very far in the future.

But now, with the prospect of human-level computing power in about 30 years, a new idea suggests itself: that I may be working to create tools which will enable the construction of the technology that may replace our species. How do I feel about this? Very uncomfortable. Having struggled my entire career to build reliable software systems, it seems to me more than likely that this future will not work out as well as some people may imagine. My personal experience suggests we tend to overestimate our design abilities.

Given the incredible power of these new technologies, shouldn't we be asking how we can best coexist with them? And if our own extinction is a likely, or even possible, outcome of our technological development, shouldn't we proceed with great caution?

The dream of robotics is, first, that intelligent machines can do our work for us, allowing us lives of leisure, restoring us to Eden. Yet in his history of such ideas, *Darwin Among the Machines,* George Dyson warns: "In the game of life and evolution there are three players at the table: human beings, nature, and machines. I am firmly on the side of nature. But nature, I suspect, is on the side of the machines." As we have seen, Moravec agrees, believing we may well not survive the encounter with the superior robot species.

How soon could such an intelligent robot be built? The coming advances in computing power seem to make it possible by 2030. And once an intelligent robot exists, it is only a small step to a robot species - to an intelligent robot that can make evolved copies of itself.

A second dream of robotics is that we will gradually replace ourselves with our robotic technology, achieving near immortality by downloading our consciousnesses; it is this process that Danny Hillis thinks we will gradually get used to and that Ray Kurzweil elegantly details in *The Age of Spiritual Machines.* (We are beginning to see intimations of this in the implantation of computer devices into the human body, as illustrated on the cover of *Wired* 8.02.)

But if we are downloaded into our technology, what are the chances that we will thereafter be ourselves or even human? It seems to me far more likely that a robotic existence would not be like a human one in any sense that we understand, that the robots would in no sense be our children, that on this path our humanity may well be lost.

Genetic engineering promises to revolutionize agriculture by increasing crop yields while reducing the use of pesticides; to create tens of thousands of novel species of bacteria, plants, viruses, and animals; to replace reproduction, or supplement it, with cloning; to create cures for many diseases, increasing our life span and our quality of life; and much, much more. We now know with certainty that these profound changes in the biological sciences are imminent and will challenge all our notions of what life is.

Technologies such as human cloning have in particular raised our awareness of the profound ethical and moral issues we face. If, for example, we were to reengineer ourselves into several separate and unequal species using the power of genetic engineering, then we would threaten the notion of equality that is the very cornerstone of our democracy.

Given the incredible power of genetic engineering, it's no surprise that there are significant safety issues in its use. My friend Amory Lovins recently cowrote, along with Hunter Lovins, an editorial that provides an ecological view of some of these dangers. Among their concerns: that "the new botany aligns the development of plants with their economic, not evolutionary, success." (See "A Tale of Two Botanies," page 247.) Amory's long career has been focused on energy and resource efficiency by taking a whole-system view of human-made systems; such a whole-system view often finds simple, smart solutions to otherwise seemingly difficult problems, and is usefully applied here as well.

After reading the Lovins' editorial, I saw an op-ed by Gregg Easterbrook in *The New York Times* (November 19, 1999) about genetically engineered crops, under the headline: "Food for the Future: Someday, rice will have built-in vitamin A. Unless the Luddites win."

Are Amory and Hunter Lovins Luddites? Certainly not. I believe we all would agree that golden rice, with its built-in vitamin A, is probably a good thing, if developed with proper care and respect for the likely dangers in moving genes across species boundaries.

Awareness of the dangers inherent in genetic engineering is beginning to grow, as reflected in the Lovins' editorial. The general public is aware of, and uneasy about, genetically modified foods, and seems to be rejecting the notion that such foods should be permitted to be unlabeled.

But genetic engineering technology is already very far along. As the Lovins note, the USDA has already approved about 50 genetically engineered crops for unlimited release; more than half of the world's soybeans and a third of its corn now contain genes spliced in from other forms of life.

While there are many important issues here, my own major concern with genetic engineering is narrower: that it gives the power - whether militarily, accidentally, or in a deliberate terrorist act - to create a White Plague.

The many wonders of nanotechnology were first imagined by the Nobel-laureate physicist Richard Feynman in a speech he gave in 1959, subsequently published under the title "There's Plenty of Room at the Bottom." The book that made a big impression on me, in the mid-'80s, was Eric Drexler's *Engines of Creation,* in which he described beautifully how manipulation of matter at the atomic level could create a utopian future of abundance, where just about everything could be made cheaply, and almost any imaginable disease or physical problem could be solved using nanotechnology and artificial intelligences.

A subsequent book, *Unbounding the Future: The Nanotechnology Revolution,* which Drexler cowrote, imagines some of the changes that might take place in a world where we had molecular-level "assemblers." Assemblers could make possible incredibly low-cost solar power, cures for cancer and the common cold by augmentation of the human immune system, essentially complete cleanup of the environment, incredibly inexpensive pocket supercomputers - in fact, any product would be manufacturable by assemblers at a cost no greater than that of wood - spaceflight more accessible than transoceanic travel today, and restoration of extinct species.

I remember feeling good about nanotechnology after reading *Engines of Creation.* As a technologist, it gave me a sense of calm - that is, nanotechnology showed us that incredible progress was possible, and indeed perhaps inevitable. If nanotechnology was our future, then I didn't feel pressed to solve so many

problems in the present. I would get to Drexler's utopian future in due time; I might as well enjoy life more in the here and now. It didn't make sense, given his vision, to stay up all night, all the time.

Drexler's vision also led to a lot of good fun. I would occasionally get to describe the wonders of nanotechnology to others who had not heard of it. After teasing them with all the things Drexler described I would give a homework assignment of my own: "Use nanotechnology to create a vampire; for extra credit create an antidote."

With these wonders came clear dangers, of which I was acutely aware. As I said at a nanotechnology conference in 1989, "We can't simply do our science and not worry about these ethical issues."[5] But my subsequent conversations with physicists convinced me that nanotechnology might not even work - or, at least, it wouldn't work anytime soon. Shortly thereafter I moved to Colorado, to a skunk works I had set up, and the focus of my work shifted to software for the Internet, specifically on ideas that became Java and Jini.

Then, last summer, Brosl Hasslacher told me that nanoscale molecular electronics was now practical. This was *new* news, at least to me, and I think to many people - and it radically changed my opinion about nanotechnology. It sent me back to *Engines of Creation.* Rereading Drexler's work after more than 10 years, I was dismayed to realize how little I had remembered of its lengthy section called "Dangers and Hopes," including a discussion of how nanotechnologies can become "engines of destruction." Indeed, in my rereading of this cautionary material today, I am struck by how naive some of Drexler's safeguard proposals seem, and how much greater I judge the dangers to be now than even he seemed to then. (Having anticipated and described many technical and political problems with nanotechnology, Drexler started the Foresight Institute in the late 1980s "to help prepare society for anticipated advanced technologies" - most important, nanotechnology.)

The enabling breakthrough to assemblers seems quite likely within the next 20 years. Molecular electronics - the new subfield of nanotechnology where individual molecules are circuit elements - should mature quickly and become enormously lucrative within this decade, causing a large incremental investment in all nanotechnologies.

Unfortunately, as with nuclear technology, it is far easier to create destructive uses for nanotechnology than constructive ones. Nanotechnology has clear military and terrorist uses, and you need not be suicidal to release a massively destructive nanotechnological device - such devices can be built to be selectively destructive, affecting, for example, only a certain geographical area or a group of people who are genetically distinct.

An immediate consequence of the Faustian bargain in obtaining the great power of nanotechnology is that we run a grave risk - the risk that we might destroy the biosphere on which all life depends.

As Drexler explained:

"Plants" with "leaves" no more efficient than today's solar cells could out-compete real plants, crowding the biosphere with an inedible foliage. Tough omnivorous "bacteria" could out-compete real bacteria: They could spread like blowing pollen, replicate swiftly, and reduce the biosphere to dust in a matter of days. Dangerous replicators could easily be too tough, small, and rapidly spreading to stop - at least if we make no preparation. We have trouble enough controlling viruses and fruit flies.

Among the cognoscenti of nanotechnology, this threat has become known as the "gray goo problem." Though masses of uncontrolled replicators need not be gray or gooey, the term "gray goo" emphasizes

that replicators able to obliterate life might be less inspiring than a single species of crabgrass. They might be superior in an evolutionary sense, but this need not make them valuable.

The gray goo threat makes one thing perfectly clear: We cannot afford certain kinds of accidents with replicating assemblers.

Gray goo would surely be a depressing ending to our human adventure on Earth, far worse than mere fire or ice, and one that could stem from a simple laboratory accident.6 Oops.

It is most of all the power of destructive self-replication in genetics, nanotechnology, and robotics (GNR) that should give us pause. Self-replication is the modus operandi of genetic engineering, which uses the machinery of the cell to replicate its designs, and the prime danger underlying gray goo in nanotechnology. Stories of run-amok robots like the Borg, replicating or mutating to escape from the ethical constraints imposed on them by their creators, are well established in our science fiction books and movies. It is even possible that self-replication may be more fundamental than we thought, and hence harder - or even impossible - to control. A recent article by Stuart Kauffman in *Nature* titled "Self-Replication: Even Peptides Do It" discusses the discovery that a 32-amino-acid peptide can "autocatalyse its own synthesis." We don't know how widespread this ability is, but Kauffman notes that it may hint at "a route to self-reproducing molecular systems on a basis far wider than Watson-Crick base-pairing."

In truth, we have had in hand for years clear warnings of the dangers inherent in widespread knowledge of GNR technologies - of the possibility of knowledge alone enabling mass destruction. But these warnings haven't been widely publicized; the public discussions have been clearly inadequate. There is no profit in publicizing the dangers.

The nuclear, biological, and chemical (NBC) technologies used in 20th-century weapons of mass destruction were and are largely military, developed in government laboratories. In sharp contrast, the 21st-century GNR technologies have clear commercial uses and are being developed almost exclusively by corporate enterprises. In this age of triumphant commercialism, technology - with science as its handmaiden - is delivering a series of almost magical inventions that are the most phenomenally lucrative ever seen. We are aggressively pursuing the promises of these new technologies within the now-unchallenged system of global capitalism and its manifold financial incentives and competitive pressures.

This is the first moment in the history of our planet when any species, by its own voluntary actions, has become a danger to itself - as well as to vast numbers of others.

It might be a familiar progression, transpiring on many worlds - a planet, newly formed, placidly revolves around its star; life slowly forms; a kaleidoscopic procession of creatures evolves; intelligence emerges which, at least up to a point, confers enormous survival value; and then technology is invented. It dawns on them that there are such things as laws of Nature, that these laws can be revealed by experiment, and that knowledge of these laws can be made both to save and to take lives, both on unprecedented scales. Science, they recognize, grants immense powers. In a flash, they create world-altering contrivances. Some planetary civilizations see their way through, place limits on what may and what must not be done, and safely pass through the time of perils. Others, not so lucky or so prudent, perish.

That is Carl Sagan, writing in 1994, in *Pale Blue Dot,* a book describing his vision of the human future in space. I am only now realizing how deep his insight was, and how sorely I miss, and will miss, his voice. For all its eloquence, Sagan's contribution was not least that of simple common sense - an attribute that, along with humility, many of the leading advocates of the 21st-century technologies seem to lack.

I remember from my childhood that my grandmother was strongly against the overuse of antibiotics. She had worked since before the first World War as a nurse and had a commonsense attitude that taking antibiotics, unless they were absolutely necessary, was bad for you.

It is not that she was an enemy of progress. She saw much progress in an almost 70-year nursing career; my grandfather, a diabetic, benefited greatly from the improved treatments that became available in his lifetime. But she, like many levelheaded people, would probably think it greatly arrogant for us, now, to be designing a robotic "replacement species," when we obviously have so much trouble making relatively simple things work, and so much trouble managing - or even understanding - ourselves.

I realize now that she had an awareness of the nature of the order of life, and of the necessity of living with and respecting that order. With this respect comes a necessary humility that we, with our early-21st-century chutzpah, lack at our peril. The commonsense view, grounded in this respect, is often right, in advance of the scientific evidence. The clear fragility and inefficiencies of the human-made systems we have built should give us all pause; the fragility of the systems I have worked on certainly humbles me.

We should have learned a lesson from the making of the first atomic bomb and the resulting arms race. We didn't do well then, and the parallels to our current situation are troubling.

The effort to build the first atomic bomb was led by the brilliant physicist J. Robert Oppenheimer. Oppenheimer was not naturally interested in politics but became painfully aware of what he perceived as the grave threat to Western civilization from the Third Reich, a threat surely grave because of the possibility that Hitler might obtain nuclear weapons. Energized by this concern, he brought his strong intellect, passion for physics, and charismatic leadership skills to Los Alamos and led a rapid and successful effort by an incredible collection of great minds to quickly invent the bomb.

What is striking is how this effort continued so naturally after the initial impetus was removed. In a meeting shortly after V-E Day with some physicists who felt that perhaps the effort should stop, Oppenheimer argued to continue. His stated reason seems a bit strange: not because of the fear of large casualties from an invasion of Japan, but because the United Nations, which was soon to be formed, should have foreknowledge of atomic weapons. A more likely reason the project continued is the momentum that had built up - the first atomic test, Trinity, was nearly at hand.

We know that in preparing this first atomic test the physicists proceeded despite a large number of possible dangers. They were initially worried, based on a calculation by Edward Teller, that an atomic explosion might set fire to the atmosphere. A revised calculation reduced the danger of destroying the world to a three-in-a-million chance. (Teller says he was later able to dismiss the prospect of atmospheric ignition entirely.) Oppenheimer, though, was sufficiently concerned about the result of Trinity that he arranged for a possible evacuation of the southwest part of the state of New Mexico. And, of course, there was the clear danger of starting a nuclear arms race.

Within a month of that first, successful test, two atomic bombs destroyed Hiroshima and Nagasaki. Some scientists had suggested that the bomb simply be demonstrated, rather than dropped on Japanese cities - saying that this would greatly improve the chances for arms control after the war - but to no avail. With the tragedy of Pearl Harbor still fresh in Americans' minds, it would have been very difficult for President Truman to order a demonstration of the weapons rather than use them as he did - the desire to quickly end the war and save the lives that would have been lost in any invasion of Japan was very strong. Yet the overriding truth was probably very simple: As the physicist Freeman Dyson later said, "The reason that it was dropped was just that nobody had the courage or the foresight to say no."

It's important to realize how shocked the physicists were in the aftermath of the bombing of Hiroshima, on August 6, 1945. They describe a series of waves of emotion: first, a sense of fulfillment that the bomb worked, then horror at all the people that had been killed, and then a convincing feeling that on no account should another bomb be dropped. Yet of course another bomb was dropped, on Nagasaki, only three days after the bombing of Hiroshima.

In November 1945, three months after the atomic bombings, Oppenheimer stood firmly behind the scientific attitude, saying, "It is not possible to be a scientist unless you believe that the knowledge of the world, and the power which this gives, is a thing which is of intrinsic value to humanity, and that you are using it to help in the spread of knowledge and are willing to take the consequences."

Oppenheimer went on to work, with others, on the Acheson-Lilienthal report, which, as Richard Rhodes says in his recent book *Visions of Technology,* "found a way to prevent a clandestine nuclear arms race without resorting to armed world government"; their suggestion was a form of relinquishment of nuclear weapons work by nation-states to an international agency.

This proposal led to the Baruch Plan, which was submitted to the United Nations in June 1946 but never adopted (perhaps because, as Rhodes suggests, Bernard Baruch had "insisted on burdening the plan with conventional sanctions," thereby inevitably dooming it, even though it would "almost certainly have been rejected by Stalinist Russia anyway"). Other efforts to promote sensible steps toward internationalizing nuclear power to prevent an arms race ran afoul either of US politics and internal distrust, or distrust by the Soviets. The opportunity to avoid the arms race was lost, and very quickly.

Two years later, in 1948, Oppenheimer seemed to have reached another stage in his thinking, saying, "In some sort of crude sense which no vulgarity, no humor, no overstatement can quite extinguish, the physicists have known sin; and this is a knowledge they cannot lose."

In 1949, the Soviets exploded an atom bomb. By 1955, both the US and the Soviet Union had tested hydrogen bombs suitable for delivery by aircraft. And so the nuclear arms race began.

Nearly 20 years ago, in the documentary *The Day After Trinity,* Freeman Dyson summarized the scientific attitudes that brought us to the nuclear precipice:

"I have felt it myself. The glitter of nuclear weapons. It is irresistible if you come to them as a scientist. To feel it's there in your hands, to release this energy that fuels the stars, to let it do your bidding. To perform these miracles, to lift a million tons of rock into the sky. It is something that gives people an illusion of illimitable power, and it is, in some ways, responsible for all our troubles - this, what you might call technical arrogance, that overcomes people when they see what they can do with their minds."[8]

Now, as then, we are creators of new technologies and stars of the imagined future, driven - this time by great financial rewards and global competition - despite the clear dangers, hardly evaluating what it may be like to try to live in a world that is the realistic outcome of what we are creating and imagining.

In 1947, *The Bulletin of the Atomic Scientists* began putting a Doomsday Clock on its cover. For more than 50 years, it has shown an estimate of the relative nuclear danger we have faced, reflecting the changing international conditions. The hands on the clock have moved 15 times and today, standing at nine minutes to midnight, reflect continuing and real danger from nuclear weapons. The recent addition of India and Pakistan to the list of nuclear powers has increased the threat of failure of the nonproliferation goal, and this danger was reflected by moving the hands closer to midnight in 1998.

In our time, how much danger do we face, not just from nuclear weapons, but from all of these technologies? How high are the extinction risks?

The philosopher John Leslie has studied this question and concluded that the risk of human extinction is at least 30 percent,9 while Ray Kurzweil believes we have "a better than even chance of making it through," with the caveat that he has "always been accused of being an optimist." Not only are these estimates not encouraging, but they do not include the probability of many horrid outcomes that lie short of extinction.

Faced with such assessments, some serious people are already suggesting that we simply move beyond Earth as quickly as possible. We would colonize the galaxy using von Neumann probes, which hop from star system to star system, replicating as they go. This step will almost certainly be necessary 5 billion years from now (or sooner if our solar system is disastrously impacted by the impending collision of our galaxy with the Andromeda galaxy within the next 3 billion years), but if we take Kurzweil and Moravec at their word it might be necessary by the middle of this century.

What are the moral implications here? If we must move beyond Earth this quickly in order for the species to survive, who accepts the responsibility for the fate of those (most of us, after all) who are left behind? And even if we scatter to the stars, isn't it likely that we may take our problems with us or find, later, that they have followed us? The fate of our species on Earth and our fate in the galaxy seem inextricably linked.

Another idea is to erect a series of shields to defend against each of the dangerous technologies. The Strategic Defense Initiative, proposed by the Reagan administration, was an attempt to design such a shield against the threat of a nuclear attack from the Soviet Union. But as Arthur C. Clarke, who was privy to discussions about the project, observed: "Though it might be possible, at vast expense, to construct local defense systems that would 'only' let through a few percent of ballistic missiles, the much touted idea of a national umbrella was nonsense. Luis Alvarez, perhaps the greatest experimental physicist of this century, remarked to me that the advocates of such schemes were 'very bright guys with no common sense.'"

Clarke continued: "Looking into my often cloudy crystal ball, I suspect that a total defense might indeed be possible in a century or so. But the technology involved would produce, as a by-product, weapons so terrible that no one would bother with anything as primitive as ballistic missiles."

In *Engines of Creation,* Eric Drexler proposed that we build an active nanotechnological shield - a form of immune system for the biosphere - to defend against dangerous replicators of all kinds that might escape from laboratories or otherwise be maliciously created. But the shield he proposed would itself be extremely dangerous - nothing could prevent it from developing autoimmune problems and attacking the biosphere itself.

Similar difficulties apply to the construction of shields against robotics and genetic engineering. These technologies are too powerful to be shielded against in the time frame of interest; even if it were possible to implement defensive shields, the side effects of their development would be at least as dangerous as the technologies we are trying to protect against.

These possibilities are all thus either undesirable or unachievable or both. The only realistic alternative I see is relinquishment: to limit development of the technologies that are too dangerous, by limiting our pursuit of certain kinds of knowledge.

Yes, I know, knowledge is good, as is the search for new truths. We have been seeking knowledge since ancient times. Aristotle opened his Metaphysics with the simple statement: "All men by nature desire to know." We have, as a bedrock value in our society, long agreed on the value of open access to information, and recognize the problems that arise with attempts to restrict access to and development of knowledge. In recent times, we have come to revere scientific knowledge.

But despite the strong historical precedents, if open access to and unlimited development of knowledge henceforth puts us all in clear danger of extinction, then common sense demands that we reexamine even these basic, long-held beliefs.

It was Nietzsche who warned us, at the end of the 19th century, not only that God is dead but that "faith in science, which after all exists undeniably, cannot owe its origin to a calculus of utility; it must have originated *in spite of* the fact that the disutility and dangerousness of the 'will to truth,' of 'truth at any price' is proved to it constantly." It is this further danger that we now fully face - the consequences of our truth-seeking. The truth that science seeks can certainly be considered a dangerous substitute for God if it is likely to lead to our extinction.

If we could agree, as a species, what we wanted, where we were headed, and why, then we would make our future much less dangerous - then we might understand what we can and should relinquish. Otherwise, we can easily imagine an arms race developing over GNR technologies, as it did with the NBC technologies in the 20th century. This is perhaps the greatest risk, for once such a race begins, it's very hard to end it. This time - unlike during the Manhattan Project - we aren't in a war, facing an implacable enemy that is threatening our civilization; we are driven, instead, by our habits, our desires, our economic system, and our competitive need to know.

I believe that we all wish our course could be determined by our collective values, ethics, and morals. If we had gained more collective wisdom over the past few thousand years, then a dialogue to this end would be more practical, and the incredible powers we are about to unleash would not be nearly so troubling.

One would think we might be driven to such a dialogue by our instinct for self-preservation. Individuals clearly have this desire, yet as a species our behavior seems to be not in our favor. In dealing with the nuclear threat, we often spoke dishonestly to ourselves and to each other, thereby greatly increasing the risks. Whether this was politically motivated, or because we chose not to think ahead, or because when faced with such grave threats we acted irrationally out of fear, I do not know, but it does not bode well.

The new Pandora's boxes of genetics, nanotechnology, and robotics are almost open, yet we seem hardly to have noticed. Ideas can't be put back in a box; unlike uranium or plutonium, they don't need to be mined and refined, and they can be freely copied. Once they are out, they are out. Churchill remarked, in a famous left-handed compliment, that the American people and their leaders "invariably do the right thing, after they have examined every other alternative." In this case, however, we must act more presciently, as to do the right thing only at last may be to lose the chance to do it at all.

As Thoreau said, "We do not ride on the railroad; it rides upon us"; and this is what we must fight, in our time. The question is, indeed, Which is to be master? Will we survive our technologies?

We are being propelled into this new century with no plan, no control, no brakes. Have we already gone too far down the path to alter course? I don't believe so, but we aren't trying yet, and the last chance to assert control - the fail-safe point - is rapidly approaching. We have our first pet robots, as well as commercially available genetic engineering techniques, and our nanoscale techniques are advancing

rapidly. While the development of these technologies proceeds through a number of steps, it isn't necessarily the case - as happened in the Manhattan Project and the Trinity test - that the last step in proving a technology is large and hard. The breakthrough to wild self-replication in robotics, genetic engineering, or nanotechnology could come suddenly, reprising the surprise we felt when we learned of the cloning of a mammal.

And yet I believe we do have a strong and solid basis for hope. Our attempts to deal with weapons of mass destruction in the last century provide a shining example of relinquishment for us to consider: the unilateral US abandonment, without preconditions, of the development of biological weapons. This relinquishment stemmed from the realization that while it would take an enormous effort to create these terrible weapons, they could from then on easily be duplicated and fall into the hands of rogue nations or terrorist groups.

The clear conclusion was that we would create additional threats to ourselves by pursuing these weapons, and that we would be more secure if we did not pursue them. We have embodied our relinquishment of biological and chemical weapons in the 1972 Biological Weapons Convention (BWC) and the 1993 Chemical Weapons Convention (CWC).

As for the continuing sizable threat from nuclear weapons, which we have lived with now for more than 50 years, the US Senate's recent rejection of the Comprehensive Test Ban Treaty makes it clear relinquishing nuclear weapons will not be politically easy. But we have a unique opportunity, with the end of the Cold War, to avert a multipolar arms race. Building on the BWC and CWC relinquishments, successful abolition of nuclear weapons could help us build toward a habit of relinquishing dangerous technologies. (Actually, by getting rid of all but 100 nuclear weapons worldwide - roughly the total destructive power of World War II and a considerably easier task - we could eliminate this extinction threat.)

Verifying relinquishment will be a difficult problem, but not an unsolvable one. We are fortunate to have already done a lot of relevant work in the context of the BWC and other treaties. Our major task will be to apply this to technologies that are naturally much more commercial than military. The substantial need here is for transparency, as difficulty of verification is directly proportional to the difficulty of distinguishing relinquished from legitimate activities.

I frankly believe that the situation in 1945 was simpler than the one we now face: The nuclear technologies were reasonably separable into commercial and military uses, and monitoring was aided by the nature of atomic tests and the ease with which radioactivity could be measured. Research on military applications could be performed at national laboratories such as Los Alamos, with the results kept secret as long as possible.

The GNR technologies do not divide clearly into commercial and military uses; given their potential in the market, it's hard to imagine pursuing them only in national laboratories. With their widespread commercial pursuit, enforcing relinquishment will require a verification regime similar to that for biological weapons, but on an unprecedented scale. This, inevitably, will raise tensions between our individual privacy and desire for proprietary information, and the need for verification to protect us all. We will undoubtedly encounter strong resistance to this loss of privacy and freedom of action.

Verifying the relinquishment of certain GNR technologies will have to occur in cyberspace as well as at physical facilities. The critical issue will be to make the necessary transparency acceptable in a world of proprietary information, presumably by providing new forms of protection for intellectual property.

Verifying compliance will also require that scientists and engineers adopt a strong code of ethical conduct, resembling the Hippocratic oath, and that they have the courage to whistleblow as necessary, even at high personal cost. This would answer the call - 50 years after Hiroshima - by the Nobel laureate Hans Bethe, one of the most senior of the surviving members of the Manhattan Project, that all scientists "cease and desist from work creating, developing, improving, and manufacturing nuclear weapons and other weapons of potential mass destruction." In the 21st century, this requires vigilance and personal responsibility by those who would work on both NBC and GNR technologies to avoid implementing weapons of mass destruction and knowledge-enabled mass destruction.

Thoreau also said that we will be "rich in proportion to the number of things which we can afford to let alone." We each seek to be happy, but it would seem worthwhile to question whether we need to take such a high risk of total destruction to gain yet more knowledge and yet more things; common sense says that there is a limit to our material needs - and that certain knowledge is too dangerous and is best forgone.

Neither should we pursue near immortality without considering the costs, without considering the commensurate increase in the risk of extinction. Immortality, while perhaps the original, is certainly not the only possible utopian dream.

I recently had the good fortune to meet the distinguished author and scholar Jacques Attali, whose book *Lignes d'horizons* (*Millennium,* in the English translation) helped inspire the Java and Jini approach to the coming age of pervasive computing, as previously described in this magazine. In his new book *Fraternités,* Attali describes how our dreams of utopia have changed over time:

"At the dawn of societies, men saw their passage on Earth as nothing more than a labyrinth of pain, at the end of which stood a door leading, via their death, to the company of gods and to *Eternity.* With the Hebrews and then the Greeks, some men dared free themselves from theological demands and dream of an ideal City where *Liberty* would flourish. Others, noting the evolution of the market society, understood that the liberty of some would entail the alienation of others, and they sought *Equality*."

Jacques helped me understand how these three different utopian goals exist in tension in our society today. He goes on to describe a fourth utopia, *Fraternity,* whose foundation is altruism. Fraternity alone associates individual happiness with the happiness of others, affording the promise of self-sustainment.

This crystallized for me my problem with Kurzweil's dream. A technological approach to Eternity - near immortality through robotics - may not be the most desirable utopia, and its pursuit brings clear dangers. Maybe we should rethink our utopian choices.

Where can we look for a new ethical basis to set our course? I have found the ideas in the book *Ethics for the New Millennium,* by the Dalai Lama, to be very helpful. As is perhaps well known but little heeded, the Dalai Lama argues that the most important thing is for us to conduct our lives with love and compassion for others, and that our societies need to develop a stronger notion of universal responsibility and of our interdependency; he proposes a standard of positive ethical conduct for individuals and societies that seems consonant with Attali's Fraternity utopia.

The Dalai Lama further argues that we must understand what it is that makes people happy, and acknowledge the strong evidence that neither material progress nor the pursuit of the power of knowledge is the key - that there are limits to what science and the scientific pursuit alone can do.

Our Western notion of happiness seems to come from the Greeks, who defined it as "the exercise of vital powers along lines of excellence in a life affording them scope."

Clearly, we need to find meaningful challenges and sufficient scope in our lives if we are to be happy in whatever is to come. But I believe we must find alternative outlets for our creative forces, beyond the culture of perpetual economic growth; this growth has largely been a blessing for several hundred years, but it has not brought us unalloyed happiness, and we must now choose between the pursuit of unrestricted and undirected growth through science and technology and the clear accompanying dangers.

It is now more than a year since my first encounter with Ray Kurzweil and John Searle. I see around me cause for hope in the voices for caution and relinquishment and in those people I have discovered who are as concerned as I am about our current predicament. I feel, too, a deepened sense of personal responsibility - not for the work I have already done, but for the work that I might yet do, at the confluence of the sciences.

But many other people who know about the dangers still seem strangely silent. When pressed, they trot out the "this is nothing new" riposte - as if awareness of what could happen is response enough. They tell me, There are universities filled with bioethicists who study this stuff all day long. They say, All this has been written about before, and by experts. They complain, Your worries and your arguments are already old hat.

I don't know where these people hide their fear. As an architect of complex systems I enter this arena as a generalist. But should this diminish my concerns? I am aware of how much has been written about, talked about, and lectured about so authoritatively. But does this mean it has reached people? Does this mean we can discount the dangers before us?

Knowing is not a rationale for not acting. Can we doubt that knowledge has become a weapon we wield against ourselves?

The experiences of the atomic scientists clearly show the need to take personal responsibility, the danger that things will move too fast, and the way in which a process can take on a life of its own. We can, as they did, create insurmountable problems in almost no time flat. We must do more thinking up front if we are not to be similarly surprised and shocked by the consequences of our inventions.

My continuing professional work is on improving the reliability of software. Software is a tool, and as a toolbuilder I must struggle with the uses to which the tools I make are put. I have always believed that making software more reliable, given its many uses, will make the world a safer and better place; if I were to come to believe the opposite, then I would be morally obligated to stop this work. I can now imagine such a day may come.

This all leaves me not angry but at least a bit melancholic. Henceforth, for me, progress will be somewhat bittersweet.

Do you remember the beautiful penultimate scene in Manhattan where Woody Allen is lying on his couch and talking into a tape recorder? He is writing a short story about people who are creating unnecessary, neurotic problems for themselves, because it keeps them from dealing with more unsolvable, terrifying problems about the universe.

He leads himself to the question, "Why is life worth living?" and to consider what makes it worthwhile for him: Groucho Marx, Willie Mays, the second movement of the Jupiter Symphony, Louis Armstrong's recording of "Potato Head Blues," Swedish movies, Flaubert's Sentimental Education, Marlon Brando, Frank Sinatra, the apples and pears by Cézanne, the crabs at Sam Wo's, and, finally, the showstopper: his love Tracy's face.

Each of us has our precious things, and as we care for them we locate the essence of our humanity. In the end, it is because of our great capacity for caring that I remain optimistic we will confront the dangerous issues now before us.

My immediate hope is to participate in a much larger discussion of the issues raised here, with people from many different backgrounds, in settings not predisposed to fear or favor technology for its own sake.

As a start, I have twice raised many of these issues at events sponsored by the Aspen Institute and have separately proposed that the American Academy of Arts and Sciences take them up as an extension of its work with the Pugwash Conferences. (These have been held since 1957 to discuss arms control, especially of nuclear weapons, and to formulate workable policies.)

It's unfortunate that the Pugwash meetings started only well after the nuclear genie was out of the bottle - roughly 15 years too late. We are also getting a belated start on seriously addressing the issues around 21st-century technologies - the prevention of knowledge-enabled mass destruction - and further delay seems unacceptable.

So I'm still searching; there are many more things to learn. Whether we are to succeed or fail, to survive or fall victim to these technologies, is not yet decided. I'm up late again - it's almost 6 am. I'm trying to imagine some better answers, to break the spell and free them from the stone.

[see original document for notes]

Stephen Petranek (Feb 2002), "Ten ways the world could end (Video)."

TED.
http://www.ted.com/index.php/talks/view/id/167

Summary

How might the human race end? Stephen Petranek lays out 10 terrible options and the science behind them. Will we be wiped out by an asteroid? Eco-collapse? How about a particle collider gone wild?

Lawrence Lessig (Apr 2004), "Insanely destructive devices."

Wired.
http://archive.wired.com/wired/archive/12.04/view.html

Full text

Smallpox has killed a billion humans. That's more deaths than in all modern wars combined. Yet despite its virulence, smallpox typically kills only 30 percent of the population it infects. Naturally evolving pathogens keep enough victims around to kill again.

Engineered pathogens can be different - as recent work in Australia has terrifyingly demonstrated. By inserting a mail-order gene into mousepox, scientists increased the death rate in mice to 100 percent. Even after vaccination, the rate was 60 percent.

We don't know whether the mail-order gene would have the same effect with smallpox. But the very idea is an example of the fear that led Bill Joy to write his frightening piece "Why the Future Doesn't Need Us," published four years ago this month in *Wired*.

Joy worried that key technologies of the future - in particular, genetic engineering, nanotech, and robotics (or GNR) because they are self-replicating and increasingly easier to craft - would be radically more dangerous than technologies of the past. It is impossibly hard to build an atomic bomb; when you build one, you've built just one. But the equivalent evil implanted in a malevolent virus will become easier to build, and if built, could become self-replicating. This is P2P (peer-to-peer) meets WMD (weapons of mass destruction), producing IDDs (insanely destructive devices).

Many criticized Joy's claims. Cassandras, they said, have always been wrong. Social and political forces will balance technology's dangers. So four years later, were the critics right? Have we learned anything about IDDs? How have we reacted? And have our reactions made us safer?

Like many professors, I think about hard questions by teaching a class. So I asked a local genius, Silicon Valley venture capitalist and polymath Steve Jurvetson, to help frame a course around the challenges raised by Joy. He opened the class with the smallpox example and asked how a society should protect itself from innovations that lead to pox viruses with 100-percent kill rates. What strategies does it adopt when everyone, even vaccinated health care workers, are vulnerable?

The first reaction of some in the class was positively Soviet. Science must be controlled. Publications must be reviewed before being printed. Communications generally may have to be surveilled - how else can we track down the enemy? And, of course, we must build a Star Wars-like shield to protect us, and issue to every American one of those space suits that CDC workers wear. ("Dear American: You may not have health insurance, but in case of a biological attack, please use the enclosed space suit.")

But it didn't take long to see the futility of these responses. GNR science doesn't require huge labs. You might not be able to conceal the work in Manhattan, but you could easily hide it in the vast wilds of, say, Montana. Moreover, a great deal of important work would be lost if the government filtered everything - as would the essence of a free society. However comforting the Star Wars-like Virus Defense Initiative might be, engineered diseases would spread long before anyone could don a space suit.

Then one student suggested a very different approach. If we can't defend against an attack, perhaps the rational response is to reduce the incentives to attack. Rather than designing space suits, maybe we should focus on ways to eliminate the reasons to annihilate us. Rather than stirring up a hornet's nest and then hiding behind a bush, maybe the solution is to avoid the causes of rage. Crazies, of course, can't be reasoned with. But we can reduce the incentives to become a crazy. We could reduce the reasonableness - from a certain perspective - for finding ways to destroy us.

The point produced a depressing recognition. There's a logic to P2P threats that we as a society don't yet get. Like the record companies against the Internet, our first response is war. But like the record companies, that response will be either futile or self-destructive. If you can't control the supply of IDDs, then the right response is to reduce the demand for IDDs. Yet as everyone in the class understood, in the four years since Joy wrote his *Wired* piece, we've done precisely the opposite. Our present course of unilateral cowboyism will continue to produce generations of angry souls seeking revenge on us.

We've not yet fully understood Joy. In the future there most certainly will be IDDs. Abolishing freedom, issuing space suits, and launching wars only increases the danger that they will be used. We had better learn that soon.

Martin Rees (15 Jul 2005), "Is this our final century? (Video)."

TED.
http://www.ted.com/talks/martin_rees_asks_is_this_our_final_century

Summary

Speaking as both an astronomer and "a concerned member of the human race," Sir Martin Rees examines our planet and its future from a cosmic perspective. He urges action to prevent dark consequences from our scientific and technological development.

Huw Price (27 Jan 2013), "Cambridge, cabs and Copenhagen: My route to existential risk."

New York Times.
http://opinionator.blogs.nytimes.com/2013/01/27/cambridge-cabs-and-copenhagen-my-route-to-existential-risk

Full text

Huw Price is Bertrand Russell professor of philosophy at the University of Cambridge. With Martin Rees and Jaan Tallinn, he is a co-founder of the project to establish the Centre for the Study of Existential Risk.

In Copenhagen the summer before last, I shared a taxi with a man who thought his chance of dying in an artificial intelligence-related accident was as high as that of heart disease or cancer. No surprise if he'd been the driver, perhaps (never tell a taxi driver that you're a philosopher!), but this was a man who has spent his career with computers.

Indeed, he's so talented in that field that he is one of the team who made this century so, well, 21st – who got us talking to one another on video screens, the way we knew we'd be doing in the 21st century, back when I was a boy, half a century ago. For this was Jaan Tallinn, one of the team who gave us Skype. (Since then, taking him to dinner in Trinity College here in Cambridge, I've had colleagues queuing up to shake his hand, thanking him for keeping them in touch with distant grandchildren.)

I knew of the suggestion that A.I. might be dangerous, of course. I had heard of the "singularity," or "intelligence explosion"– roughly, the idea, originally due to the statistician I J Good (a Cambridge-trained former colleague of Alan Turing's), that once machine intelligence reaches a certain point, it could take over its own process of improvement, perhaps exponentially, so that we humans would soon be left far behind. But I'd never met anyone who regarded it as such a pressing cause for concern – let alone anyone with their feet so firmly on the ground in the software business.

I was intrigued, and also impressed, by Tallinn's commitment to doing something about it. The topic came up because I'd asked what he worked on these days. The answer, in part, is that he spends a lot of his time trying to improve the odds, in one way or another (talking to philosophers in Danish taxis, for example).

I was heading for Cambridge at the time, to take up my new job as Bertrand Russell professor of philosophy – a chair named after a man who spent the last years of his life trying to protect humanity from another kind of technological risk, that of nuclear war. And one of the people I already knew in

Cambridge was the distinguished cosmologist Martin Rees – then master of Trinity College, and former president of the Royal Society. Lord Rees is another outspoken proponent of the view that we humans should pay more attention to the ways in which our own technology might threaten our survival. (Biotechnology gets most attention, in his work.)

So it occurred to me that there might be a useful, interesting and appropriate role for me, as a kind of catalyst between these two activists, and their respective circles. And that, to fast forward a little, is how I came to be taking Jaan Tallinn to dinner in Trinity College; and how he, Martin Rees and I now come to be working together, to establish here in Cambridge the Centre for the Study of Existential Risk (C.S.E.R.).

By "existential risks" (E.R.) we mean, roughly, catastrophic risks to our species that are "our fault," in the sense that they arise from human technologies. These are not the only catastrophic risks we humans face, of course: asteroid impacts and extreme volcanic events could wipe us out, for example. But in comparison with possible technological risks, these natural risks are comparatively well studied and, arguably, comparatively minor (the major source of uncertainty being on the technological side). So the greatest need, in our view, is to pay a lot more attention to these technological risks. That's why we chose to make them the explicit focus of our center.

I have now met many fascinating scholars – scientists, philosophers and others – who think that these issues are profoundly important, and seriously understudied. Strikingly, though, they differ about where they think the most pressing risks lie. A Cambridge zoologist I met recently is most worried about deadly designer bacteria, produced – whether by error or by terror, as Rees puts it – in a nearby future in which there's almost an app for such things. To him, A.I. risk seemed comparatively far-fetched – though he confessed that he was no expert (and added that the evidence is that even experts do little better than chance, in many areas).

Where do I stand on the A.I. case, the one that got me into this business? I don't claim any great expertise on the matter (perhaps wisely, in the light of the evidence just mentioned). For what it's worth, however, my view goes like this. On the one hand, I haven't yet seen a strong case for being quite as pessimistic as Jaan Tallinn was in the taxi that day. (To be fair, he himself says that he's not always that pessimistic.) On the other hand, I do think that there are strong reasons to think that we humans are nearing one of the most significant moments in our entire history: the point at which intelligence escapes the constraints of biology. And I see no compelling grounds for confidence that if that does happen, we will survive the transition in reasonable shape. Without such grounds, I think we have cause for concern.

My case for these conclusions relies on three main observations. The first is that our own intelligence is an evolved biological solution to a kind of optimization problem, operating under very tight constraints of time, energy, raw materials, historical starting point and no doubt many other factors. The hardware needs to fit through a mammalian birth canal, to be reasonably protected for a mobile life in a hazardous environment, to consume something like 1,000 calories per day and so on – not to mention being achievable by mutation and selection over a time scale of some tens of millions of years, starting from what existed back then!

Second, this biological endowment, such as it is, has been essentially constant, for many thousands of years. It is a kind of fixed point in the landscape, a mountain peak on which we have all lived for hundreds of generations. Think of it as Mount Fuji, for example. We are creatures of this volcano. The fact that it towers above the surrounding landscape enables us to dominate our environment and accounts for our extraordinary success, compared with most other species on the planet. (Some species

benefit from our success, of course: cockroaches and rats, perhaps, and the many distinctive bacteria that inhabit our guts.) And the distinctive shape of the peak – also constant, or nearly so, for all these generations – is very deeply entangled with our sense of what it is to be us. We are not just creatures of any volcano; we are creatures of this one.

Both the height and the shape of the mountain are products of our biological history, in the main. (The qualification is needed because cultural inheritance may well play a role too.) Our great success in the biological landscape, in turn, is mainly because of the fact that the distinctive intelligence that the height and shape represent has enabled us to control and modify the surrounding environment. We've been exercising such control for a very long time of course, but we've recently got much better at it. Modern science and technology give us new and extraordinarily powerful ways to modify the natural world, and the creatures of the ancient volcano are more dominant than ever before.

This is all old news, of course, as is the observation that this success may ultimately be our undoing. (Remember Malthus.) But the new concern, linked to speculation about the future of A.I., is that we may soon be in a position to do something entirely new: to unleash a kind of artificial vulcanism, that may change the shape and height of our own mountain, or build new ones, perhaps even higher, and perhaps of shapes we cannot presently imagine. In other words – and this is my third observation – we face the prospect that designed nonbiological technologies, operating under entirely different constraints in many respects, may soon do the kinds of things that our brain does, but very much faster, and very much better, in whatever dimensions of improvement may turn out to be available.

The claim that we face this prospect may seem contestable. Is it really plausible that technology will reach this stage (ever, let alone soon)? I'll come back to this. For the moment, the point I want to make is simply that if we do suppose that we are going to reach such a stage – a point at which technology reshapes our human Mount Fuji, or builds other peaks elsewhere – then it's not going to be business as usual, as far as we are concerned. Technology will have modified the one thing, more than anything else, that has made it "business as usual" so long as we have been human.

Indeed, it's not really clear who "we" would be, in those circumstances. Would we be humans surviving (or not) in an environment in which superior machine intelligences had taken the reins, to speak? Would we be human intelligences somehow extended by nonbiological means? Would we be in some sense entirely posthuman (though thinking of ourselves perhaps as descendants of humans)? I don't claim that these are the only options, or even that these options are particularly well formulated – they're not! My point is simply that if technology does get to this stage, the most important fixed point in our landscape is no longer fixed – on the contrary, it might be moving, rapidly, in directions we creatures of the volcano are not well equipped to understand, let alone predict. That seems to me a cause for concern.

These are my reasons for thinking that at some point over the horizon, there's a major tipping point awaiting us, when intelligence escapes its biological constraints; and that it is far from clear that that's good news, from our point of view. To sum it up briefly, the argument rests on three propositions: (i) the level and general shape of human intelligence is highly contingent, a product of biological constraints and accidents; (ii) despite its contingency in the big scheme of things, it is essential to us – it is who we are, more or less, and it accounts for our success; (iii) technology is likely to give us the means to bypass the biological constraints, either altering our own minds or constructing machines with comparable capabilities, and thereby reforming the landscape.

But how far away might this tipping point be, and will it ever happen at all? This brings me back to the most contested claim of these three – the assertion that nonbiological machines are likely, at some point, to be as intelligent or more intelligent than the "biological machines" we have in our skulls.

Objections to this claim come from several directions. Some contest it based on the (claimed) poor record of A.I. so far; others on the basis of some claimed fundamental difference between human minds and computers; yet others, perhaps, on the grounds that the claim is simply unclear – it isn't clear what intelligence is, for example.

To arguments of the last kind, I'm inclined to give a pragmatist's answer: Don't think about what intelligence is, think about what it does. Putting it rather crudely, the distinctive thing about our peak in the present biological landscape is that we tend to be much better at controlling our environment than any other species. In these terms, the question is then whether machines might at some point do an even better job (perhaps a vastly better job). If so, then all the above concerns seem to be back on the table, even though we haven't mentioned the word "intelligence," let alone tried to say what it means. (You might try to resurrect the objection by focusing on the word "control," but here I think you'd be on thin ice: it's clear that machines already control things, in some sense – they drive cars, for example.)

Much the same point can be made against attempts to take comfort in the idea that there is something fundamentally different between human minds and computers. Suppose there is, and that that means that computers will never do some of the things that we do – write philosophy, appreciate the sublime, or whatever. What's the case for thinking that without these gifts, the machines cannot control the terrestrial environment a lot more effectively than we do?

People who worry about these things often say that the main threat may come from accidents involving "dumb optimizers" – machines with rather simple goals (producing IKEA furniture, say) that figure out that they can improve their output astronomically by taking control of various resources on which we depend for our survival. Nobody expects an automated furniture factory to do philosophy. Does that make it less dangerous? (Would you bet your grandchildren's lives on the matter?)

But there's a more direct answer, too, to this attempt to take comfort in any supposed difference between human minds and computers. It also cuts against attempts to take refuge in the failure of A.I. to live up to some of its own hype. It's an answer in two parts. The first part – let me call it, a little aggressively, the blow to the head – points out that however biology got us onto this exalted peak in the landscape, the tricks are all there for our inspection: most of it is done with the glop inside our skulls. Understand that, and you understand how to do it artificially, at least in principle. Sure, it could turn out that there's then no way to improve things – that biology, despite all the constraints, really has hit some sort of fundamental maximum. Or it could turn out that the task of figuring out how biology did it is just beyond us, at least for the foreseeable future (even the remotely foreseeable future). But again, are you going to bet your grandchildren on that possibility?

The second part of the argument – the blow from below – asks these opponents just how far up the intelligence mountain they think that A.I. could get us. To the level of our fishy ancestors? Our early mammalian ancestors? (Keep in mind that the important question is the pragmatic one: Could a machine do what these creatures do?) Wherever they claim to draw the line, the objection challenges them to say what biology does next, that no nonbiological machine could possibly do. Perhaps someone has a plausible answer to this question, but for my part, I have no idea what it could be.

At present, then, I see no good reason to believe that intelligence is never going to escape from the head, or that it won't do so in time scales we could reasonably care about. Hence it seems to me eminently sensible to think about what happens if and when it does so, and whether there's something we can do to favor good outcomes over bad, in that case. That's how I see what Rees, Tallinn and I want to do in Cambridge (about this kind of technological risk, as about others): we're trying to assemble an

organization that will use the combined intellectual power of a lot of gifted people to shift some probability from the bad side to the good.

Tallin compares this to wearing a seat belt. Most of us agree that that makes sense, even if the risk of an accident is low, and even though we can't be certain that it would be beneficial, if we were to have an accident. (Occasionally, seat belts make things worse.) The analogy is apt in another way, too. It is easy to turn a blind eye to the case for wearing a seat belt. Many of us don't wear them in taxis, for example. Something – perhaps optimism, a sense that caution isn't cool, or (if you're sufficiently English!) a misplaced concern about hurting the driver's feelings – just gets in the way of the simple choice to put the thing on. Usually it makes no difference, of course, but sometimes people get needlessly hurt.

Worrying about catastrophic risk may have similar image problems. We tend to be optimists, and it might be easier, and perhaps in some sense cooler, not to bother. So I finish with two recommendations. First, keep in mind that in this case our fate is in the hands, if that's the word, of what might charitably be called a very large and poorly organized committee – collectively shortsighted, if not actually reckless, but responsible for guiding our fast-moving vehicle through some hazardous and yet completely unfamiliar terrain. Second, remember that all the children – all of them – are in the back. We thrill-seeking grandparents may have little to lose, but shouldn't we be encouraging the kids to buckle up?

Martin Rees (8 Mar 2013), "Denial of catastrophic risks."

Science 339.
http://www.sciencemag.org/content/339/6124/1123.full

Full text

In a media landscape saturated with sensational Science stories and "End of the World" Hollywood productions, it may be hard to persuade the wide public that real catastrophes could arise as unexpectedly as the 2008 financial crisis, and have a far greater impact. Society could be dealt shattering blows by the misapplication of technologies that exist already or could emerge within the coming decades. Some of the scenarios that have been envisaged may indeed be science fiction, but others may be disquietingly real. I believe these "existential risks" deserve more serious study. Those fortunate enough to live in the developed world fret too much about minor hazards of everyday life: improbable air crashes, possible carcinogens in food, low radiation doses, and so forth. But we should be more concerned about events that have not yet happened but which, if they occurred even once, could cause worldwide devastation.

The main threats to sustained human existence now come from people, not from nature. Ecological shocks that irreversibly degrade the biosphere could be triggered by the unsustainable demands of a growing world population. Fast-spreading pandemics would cause havoc in the megacities of the developing world. And political tensions will probably stem from scarcity of resources, aggravated by climate change. Equally worrying are the imponderable downsides of powerful new cyber-, bio-, and nanotechnologies. Indeed, we're entering an era when a few individuals could, via error or terror, trigger societal breakdown.

Some threats are well known. In the 20th century, the downsides of nuclear science loomed large. At any time in the Cold War era, the superpowers could have stumbled toward Armageddon through muddle and miscalculation. The threat of global annihilation involving tens of thousands of hydrogen

bombs is thankfully in abeyance, but now there is a growing concern that smaller nuclear arsenals might be used in a regional context, or even by terrorists. We can't rule out a geopolitical realignment that creates a standoff between new superpowers. So a new generation may face its own "Cuba," and one that could be handled less well or less luckily than was the 1962 crisis.

What are some new concerns stemming from fast-developing 21st-century technologies? Our interconnected world depends on elaborate networks: electric power grids, air traffic control, international finance, just-in-time delivery, and so forth. Unless these are highly resilient, their manifest benefits could be outweighed by catastrophic (albeit rare) breakdowns cascading through the system. Social media could spread psychic contagion from a localized crisis, literally at the speed of light. Concern about cyberattack, by criminals or hostile nations, is rising sharply. Synthetic biology likewise offers huge potential for medicine and agriculture, but in the sci-fi scenario where new organisms can be routinely created, the ecology (and even our species) might not long survive unscathed. And should we worry about another sci-fiscenario, in which a network of computers could develop a mind of its own and threaten us all?

Some would dismiss such concerns as an exaggerated jeremiad: After all, societies have survived for millennia, despite storms, earthquakes, and pestilence. But these human-induced threats are different—they are newly emergent, so we have a limited time base for exposure to them and can't be so sanguine that we would survive them for long, or that governments could cope if disaster strikes. That is why a group of natural and social scientists in Cambridge, UK, plans to inaugurate a research program to identify the most genuine of these emergent risks and assess how to enhance resilience against them. True, it is hard to quantify the potential "existential" threats from (for instance) bio- or cybertechnology, from artificial intelligence, or from runaway climatic catastrophes. But we should at least start figuring out what can be left in the sci-fibin (for now) and what has moved beyond the imaginary.

Bruce Schneier (14 Mar 2013), "Our security models will never work, no matter what we do." Wired.
http://www.wired.com/2013/03/security-when-the-bad-guys-have-technology-too-how-do-we-survive

Excerpt:

As it gets easier for one member of a group to destroy the entire group, and the group size gets larger, the odds of someone in the group doing it approaches certainty. Our global interconnectedness means that our group size encompasses everyone on the planet, and since government hasn't kept up, we have to worry about the weakest-controlled member of the weakest-controlled country. Is this a fundamental limitation of technological advancement, one that could end civilization? First our fears grip us so strongly that, thinking about the short term, we willingly embrace a police state in a desperate attempt to keep us safe; then, someone goes off and destroys us anyway?

If security won't work in the end, what is the solution?

Resilience — building systems able to survive unexpected and devastating attacks — is the best answer we have right now. We need to recognize that large-scale attacks will happen, that society can survive more than we give it credit for, and that we can design systems to survive these sorts of attacks. Calling terrorism an existential threat is ridiculous in a country where more people die each month in car crashes than died in the 9/11 terrorist attacks.

If the U.S. can survive the destruction of an entire city — witness New Orleans after Hurricane Katrina — we need to start acting like it, and planning for it. Still, it's hard to see how resilience buys us anything but additional time. Technology will continue to advance, and right now we don't know how to adapt any defenses — including resilience — fast enough.

We need a more flexible and rationally reactive approach to these problems and new regimes of trust for our information-interconnected world. We're going to have to figure this out if we want to survive, and I'm not sure how many decades we have left.

Francesco Guerrera (24 Jun, 2013), "Current account: Cyberattacks are banks' latest 'existential risk'."

Wall Street Journal.
http://blogs.wsj.com/moneybeat/2013/06/24/current-account-cyberattacks-are-banks-latest-existential-risk

Excerpt

Cybersecurity is a critical issue for every company but, as often, financial services firms are a special case. Each and every attack can undermine the public's faith not just in the individual institution, but in the entire financial system. The financial services sector accounted for "just" 3% of all data breaches that led to identity theft in 2012, according to a recent report by Symantec Corp. But each of the average of 400,000 identities that were revealed during every one of those incidents represents a dent in the wall of trust between customers and their financial institutions.

Wall Street lawyer Rodgin Cohen put it best last week when he called cybersecurity an "existential risk." "Unless we do better in aligning the private sector and the public sector in hardening our systems, sooner or later there is going to be a very serious problem," Mr. Cohen, a partner at Sullivan & Cromwell LLP, told the WSJ's CFO Network conference.

Martin Rees (Jul 2013), "Is this our final century?"

Astronomy 41.

Abstract

Spearheaded by former NASA astronaut Ed Lu, the project aims to put an infrared telescope in solar orbit to catalog a million asteroids and monitor their orbits. The most extreme stellar deaths give rise to gamma-ray bursts - intense jets that in a few seconds release more energy than the precursor star radiated in its entire prior lifetime. To have a similar effect, a garden- variety supernova would need to explode within 100 light-years of Earth, or roughly the nearest millionth part of the galaxy's volume. Because supernovae occur a million times more frequently than gamma-ray bursts, however, their threats are comparable.

Martin Rees (4 Oct 2013), "Martin Rees on climate change, manned space missions and existential risk."

Wired.
http://www.wired.co.uk/news/archive/2013-10/04/martin-rees

Full text

Wired.co.uk: Just weeks after you suggested we need a geoengineering "plan B" to tackle climate change, the Chancellor George Osborne said that the UK should not be leading the fight on climate change. Are you optimistic about winning the debate on climate change, which frustratingly still continues, and why should the UK be leading our efforts to deal with climate change?

Lord Martin Rees: One thing isn't controversial. The atmospheric CO2 concentration is rising -- mainly due to the burning of fossil fuels. It's agreed that this build-up will in itself induce a long-term warming trend, superimposed on all the other complicated effects that make climate fluctuate. But what's less well understood is how much the effect is amplified by associated changes in water vapour and clouds. I think it's the smallish probability of catastrophic warming, rather than the expectation of the IPPC's median trajectory, which presents the most compelling argument for keeping climate change high on the agenda.

It's crucial to keep "clear water" between the science on the one hand, and the policy response on the other. Risk assessment should be separate from risk management. Scientists should engage in policy debates -- though not as experts but as "scientific citizens".

In that spirit, I'd add that I myself still strongly support the Climate Change Act. Not only Blair and Brown, but several Labour ministers -- the Miliband brothers, Hilary Benn, and others -- worked hard to sustain these issues high on the agenda; and the coalition has not formally backtracked, despite rumblings. The downside of global warming will be felt by future generations, and primarily in countries far from our own -- and such concerns are trumped by the short-term and the parochial. Long-term altruism is plainly not a vote-winner.

A high priority should be to implement measures that actually save money -- by using energy more efficiently, insulating buildings better, and so forth. And also to reduce pollutants, methane and black carbon. This won't substitute for measures to tackle carbon dioxide but would have a shorter-term impact and more manifest side-benefits.

My pessimistic prediction is that global annual emissions won't be turned around in the next 20 years. By then we'll be clearer on just how strongly the feedback from water vapour and clouds amplifies the effect of carbon dioxide. If the effect is strong, and the world consequently seems on a rapidly-warming trajectory into dangerous territory, there may be pressure for "panic measures" such as geoengineering.

But we shouldn't despair. It may take 50 years to decarbonise the world's power generation, but this could be achieved if we start now -- and if we invest in far higher R&D in all novel forms of "clean energy".

You co-founded the Cambridge Centre for the Study of Existential Risk to investigate threats to humanity's survival -- which threat do the public and policymakers think about least, but which poses a serious risk to humanity in this century?

Advances in technology -- hugely beneficial though they are -- render us vulnerable in new ways. For instance, our interconnected world depends on elaborate networks: electric power grids, air traffic control, international finance, just-in-time delivery and so forth.

Pandemics could spread at the speed of jet aircraft, causing maximal havoc in the shambolic but burgeoning megacities of the developing world. Social media could spread psychic contagion -- rumours and panic -- literally at the speed of light. Malign or foolhardy individuals or small groups have far more power and leverage than in the past.

Some would dismiss these concerns as an exaggerated jeremiad: after all, societies have survived for millennia, despite storms, earthquakes and pestilence. But these human-induced threats are different: they are newly emergent, so we have a limited timebase for exposure to them and can't be so sanguine about the ability of governments to cope if disaster strikes. Technological advances bring with them great hopes, but also great fears.

Discussions about manned exploration of space are hotting up, with China planning to return to the Moon and several private organisations are talking about manned missions to Mars. In a 2010 interview with The Independent, you warned against seeing the colonisation of space as a panacea for humanity's problems, but in the face of overpopulation and climate change, surely getting off this planet has to be a serious part of humanity's long-term survival plans?

The practical case for manned spaceflight gets ever-weaker with each advance in robots and miniaturisation -- indeed as a scientist or practical man I see little purpose in sending people into space at all. But as a human being, I'm an enthusiast for manned missions.

By 2100, groups of pioneers may have established "bases" independent from the Earth -- on Mars, or maybe on asteroids. But don't ever rely on mass emigration from Earth. Nowhere in our Solar System offers an environment even as clement as the Antarctic or the top of Everest. Space doesn't offer an escape from Earth's problems.

In a recent series of articles in the New Republic, Steven Pinker and Leon Wieseltier debated whether science encroaches too heavily into the humanities. "Science wants to invade the liberal arts. Don't let it happen," read the headline of Wieseltier's piece. Have scientists been too quick to hold forth on issues of philosophy and theology, like the existence or nonexistence of God?

Science is the one culture that's truly global -- protons, proteins and Pythagoras's Theorem are the same from China to Peru. It should transcend all barriers of nationality. It should straddle all faiths too. The scientists who attack mainstream religion, rather than striving for peaceful coexistence with it, damage science, and also weaken the fight against fundamentalism.

Nasa's Kepler space telescope was recently retired, but not before collecting data on potentially thousands and thousands of exoplanets. Are we closer than ever to finding the signs of extraterrestrial life?

The Kepler spacecraft found several thousand transiting planets, some no bigger than the Earth -- and further data-analysis will reveal many more. The real goal, of course, is to see them directly-- not just their shadows. But that's hard.

Would there be life -- even intelligent life -- on these faraway planets? We still know too little to set the odds. Even if simple life is common, it is of course a separate question whether it's likely to evolve into anything we might recognise as intelligent or complex -- and what and where this might happen.

Andrew Martin (30 Aug 2014), "The scientific A-Team saving the world from killer viruses, rogue AI and the paperclip apocalypse."

Guardian.
http://www.theguardian.com/technology/2014/aug/30/saviours-universe-four-unlikely-men-save-world

Full text

Cambridge, some time after the end of term. Demob-happy undergraduates, dressed for punting and swigging wine from the bottle, seem not so much to be enjoying themselves as determinedly following rites of passage on the way to a privileged future. I am heading towards the biggest, richest and arguably most beautiful college: Trinity. Of the 90 Nobel prizes won by members of Cambridge University in the 20th century, 32 were won by members of Trinity. Its alumni include Isaac Newton, Wittgenstein, Bertrand Russell and six prime ministers.

The porter's lodge is like an airlock, apparently sealed from the tribulations of everyday life. But inside the college, pacing the flagstones of what is called – all modesty aside – Great Court, are four men who do not take it for granted that those undergraduates actually have a future. They are the four founders of the Centre for the Study of Existential Risk (CSER), and they are in the business of "horizon scanning". Together, they are on alert for what they sometimes call "low-probability-but-high-consequence events", and sometimes – when they forget to be reassuring – "catastrophe".

At their head is a 72-year-old cosmologist, Martin Rees. The honorifics jostle at the start of his name: he is Professor Martin Rees, Baron Rees of Ludlow, OM FRS. He is the Astronomer Royal, a fellow of Trinity, formerly a master of the college and a president of the Royal Society. In newspaper articles, he is often described simply as Britain's "top scientist". In 2003, Rees published a book called Our Final Century. He likes to joke that the reason his book was published in the US as Our Final Hour is because "Americans like instant gratification". In the book, he rates the chances of a "serious setback" for humanity over the next 100 years at "50-50". There is an asteroid named after him – 4587 Rees. I can't help thinking, in light of his apocalyptic concerns, that it would be ironic if 4587 Rees crashed into the Earth.

But these four men are less concerned with acts of God than those we have created ourselves: the consequences of being too clever for our own good. They believe there is a risk that artificial intelligence (AI) will challenge our own. In a talk at a TED conference, Rees invoked another danger: that "in our interconnected world, novel technology could empower just one fanatic, or some weirdo with the mindset of those who now design computer viruses, to trigger some kind of disaster. Or catastrophe could arise from some technical misadventure – error rather than terror."

Rees proudly introduces his colleagues. There is Jaan Tallinn, a meditative Estonian computer programmer and one of five co-founders of Skype. There is a courtly Indian economic theorist, Professor Sir Partha Dasgupta ("Partha's very concerned with inequalities across time," Rees says). And there is Huw Price, a laid-back philosophy don – specifically, the Bertrand Russell professor of philosophy at Cambridge.

The group originated in 2011, when Price and Tallinn met at a conference on time in Copenhagen. Two weeks later Price, who had just taken up his philosophy post, invited Tallinn to Cambridge to meet his new colleague, Rees; all three shared a concern about near-term risks to humanity. "Fate," Price recalls, "was offering me a remarkable opportunity." After a two-year gestation, the CSER gets properly up and running next month. The first of a dozen post-doctoral researchers will be taken on, some of whom will be embedded with science and technology firms. There will be seminars on synthetic biology, decision theory and AI. Already there have been meetings with the Cabinet Office, the Ministry of Defence and the Foreign Office.

As the salutary clock of the Great Court looms behind them, the irresistible image of our leading brains uniting to save the planet: X-Men: The Last Stand, The Four Just Men, Guardians Of The Galaxy. Between photographs, Rees and Dasgupta chat about the relationship between facts and prejudice in global warming forecasts, and I wonder if they ever talk of anything other than the end of the world.

Before we met, I was sent a vast amount of reading material, including a paper touchingly described by Dasgupta as "somewhat informal", but still containing much algebra. Most strikingly, the material included four worst case possibilities:

1 The disaffected lab worker

In which an unhappy biotech employee makes minor modifications to the genome of a virus – for example, avian flu H5N1. A batch of live virus is created that can be released via aerosol. The lab worker takes a round-the-world flight, stopping off at airports to release the virus. The plausibility of this scenario is rated as "high", and "technologically possible in the near term". As the CSER men note: "No professional psychological evaluation of biotech lab staff takes place." A similar leakage might also happen accidentally, and I was sent, as a matter of urgency, an article from the Guardian about how researchers at the University of Wisconsin-Madison had modified strains of bird flu to create a virus similar to the 1918 Spanish flu that killed 50m people. The project was condemned as "absolutely crazy" by the respected epidemiologist Lord May.

2 Termination risk

In which pressure to stop climate change results in the adoption of stratospheric aerosol geo-engineering. Global warming is checked, but CO2 levels continue to rise. The geo-engineering then ceases, perhaps as a result of some other catastrophe, such as world war. This triggers what is called "termination risk": the sticking plaster removed, the warming gets much worse, quickly. Half the Earth's population is wiped out. I was advised that geo-engineering appears possible in the near term, but the scientific consensus is against adopting it.

3 Distributed manufacturing

3D printing is already used to make automatic weapons. These weapons can work, but are liable to explode in the user's hand. Still, the refinement of these techniques may allow nanoscale manufacture of military-grade missiles. "This would require a range of technological advances currently beyond us," I was told, "but believed by many scientists to be possible."

4 All of America is turned into paper clips

In which AI undergoes runaway improvement and "escapes into the internet". Imagine a computer swallowing all the information stored in Wikipedia in one gulp and generally gaining access to everything human-made. (The already-emergent "internet of things" means that, increasingly, devices can communicate between themselves; our homes are becoming more automated.) This rogue machine then uses human resources to develop new and better technologies to achieve its goal. I was given the for-instance of a paper clip making software that turns the whole of America, including the people, into paper clips. This is "not technologically possible in the next 20 years. Estimates range from 20 years to 300 years to never. But the potential negative consequences are too severe not to study the possibility."

This is what these four men are up against.

Rees works from rooms overlooking the cloistered Nevile's Court, which contains the Wren Library, which in turn contains two Shakespeare First Folios. He is small, dapper, silver-haired, and offsets his doomsday scenarios with a puckish humour. He invited me to sit on the couch next to his desk, "where sometimes I sits and thinks, and sometimes I just sits". As I wondered about this quote – Winnie The Pooh? – Rees was off, speaking so rapidly and softly as to be almost thinking aloud. "On a cosmic timescale, human beings are not the culmination, because it's taken four billion years for us to emerge from protozoa, and we know the solar system has more than four billion years ahead of it." Over the next half-hour, he tells me that we are "the stewards of an immense future", and that we have a duty to clear the looming hurdle presented by technological advance. "A few crazy pioneers – and we wish them good luck – might tunnel through the period of danger by establishing colonies in outer space, but nowhere out there is as comfortable even as the South Pole, so we have to solve the problems here."

He moves easily from such vertiginous concerns to survival on the micro level. For example, those weirdos or fanatics leveraged by technology. He believes that "bioterror probably won't be used by extremist groups with well-defined political aims – it's too uncontrollable. But there are eco-freaks who believe there are too many humans in the world." He argues that bio-engineering and AI have "an upside and a dark side. A computer is a sort of idiot savant. It can do arithmetic better than us, but the advances in software and sensors have lagged behind. In the 1990s, Kasparov was beaten at chess by the IBM computer, but a computer still can't pick up a chess piece and move it with the dexterity of a five-year-old child. Still, machine learning is advancing apace."

This brought us to the American futurist, Ray Kurzweil, a man there would be no point in inviting to dinner at Trinity. He is said to live on 150 pills a day, hopeful of surviving until what he calls "The Singularity" – the point at which humans build their last machine, all subsequent ones being built by other machines. A merger of man and machine will then offer the prospect of immortality for those who would prefer not to die. Rees considers Kurzweil "rather wild".

Rees recalled a lecture in which he (Rees) discussed one of the supposed routes to immortality: cryonics, the freezing of the body with a view to future resurrection. Rees had said he would "rather end his days in an English churchyard than a Californian refrigerator". It turned out that someone in the audience had paid £150,000 to have his body frozen; another had paid £80,000 to have just his head frozen – and both were indignant. "They called me a deathist," Rees recalls, laughing, "as if I were actually in favour of death."

I say I was disturbed to discover that Kurzweil is now a director of engineering at Google. "Yes," he says, "but to be fair to Google, they're grabbing everyone in this area who's outside the tent and pulling them into the tent." Does he detect a faultline between gung-ho Silicon Valley and more sceptical Europeans –

the old world versus the new? He does not. "They have a can-do attitude, and they've a lot to be proud of." He stresses that CSER wants to work with the technologists, not against them.

A clock chimes: time for lunch – one good thing about Trinity is that it is nearly always time for a meal in the Great Hall. My dining companion is Professor Huw Price. Price grew up in Australia, hence – perhaps – his small gold earring. As I settle down to my quiche, he tells me that a year or so after CSER came together, he realised there might be a tie-in between the kind of philosophical questions he'd been pursuing and AI questions. Last February, he visited the Machine Intelligence Research Institute in Berkeley, California, "where they are trying to make sure that AI that begins with human-friendly goals will stay friendly when it starts to improve itself. Because the computers of the future will be writing their own programmes." I stop him right there. "Why should we let them do that?"

Price seems slightly taken aback by the question. "Well, imagine any scenario where more intelligence is better – in finance or defence. You have an incentive to make sure your machine is more intelligent than the other person's machine." The strategy of these machines, he continues, would depend on what they thought other machines running the same software would do. I interpose another "Why?" and Price takes a long drink of water, possibly processing the fact that he has an idiot on his hands.

These machines would all be networked together, he explains. "Now, if a machine is predicting what another machine with the same software will do, it is in effect predicting what it [the first machine] will do, and this is a barrier to communication. Let's say I want to predict whether I'm going to pick up my glass and have another drink of water in the next five minutes. Let's say I assign a probability of 50% to that. Assigning a probability is like placing odds on a bet about it. Whatever odds I'm offered, I can win the bet by picking up the glass and having a drink. Assigning probabilities to my own acts – there's something very fishy about that."

This leads to the question of how to make the machines see that cooperation might be the rational option. Price asks whether I have heard of the philosophical conundrum the Prisoner's Dilemma. I have not. He explains: "Two prisoners are charged with a crime. They're held in separate cells, and there's no communication between them. They're separately told that if neither confesses, they both get six months. If they both confess, they both get five years. If one confesses and the other doesn't, the one who confesses goes free and the other gets 10 years. So each would be better off confessing, whether or not the other confesses. But the best outcome for both is if they remain silent." For this to happen, each prisoner would have to predict that the other will act in their mutual interest. So the goal is to build this facility into self-programming machines in order to forestall monomaniacal behaviour. To avoid America being turned into paper clips.

By now we are back in Rees's rooms. Price seems to have the run on them, and I am reminded of the Beatles in Help! – all of them living in the same house. Rees returns as I trepidatiously ask Price, "Why can't we just turn the machines off?" There is a mournful silence. "That's not quite so easy when you're talking about a global network," Rees says. "We won't be able to turn them off,' Price adds, "because they're smarter than we are, and they're controlling all the switches and all the hardware."

But he concedes that the machine intelligence people at Berkeley have given some thought to this. "One of the strategies is to make sure it [the self-improving machine] is perfectly isolated. I think they call it the oracle model." So, an advisory superintelligent machine; a consultant. "But perhaps," Price muses, "it can do things to persuade humans to give it more direct connection with the world." "Bribery?" I gasp, excited at this new possibility. Price nods. "If it knows enough about human psychology."

Ask Professor Dasgupta for his worst-case scenarios, and he will politely suggest that "these have already happened – in Sudan, in Rwanda". We speak in the Fellows' Parlour, which is less chintzy and sherry-stained than the name suggests. But still there are deep leather armchairs, oil paintings and dainty coffee cups, and I know what Dasgupta means when he says, "We here are having a tremendously good time – those of us who are lucky."

Still, we are "disturbing nature". In sub-Saharan Africa and South Asia, depleted wetlands or forests might cause starvation; they could also trigger viruses, sectarian conflict and over-population (couples having more children to compensate for low survival rates). Dasgupta is concerned about sustainable development. He remains extremely forbearing when I say this has surely been a buzzword for many years. "I think you are right. A lot has been written on the matter, but much of it remains unfocused. What should be sustained, when you think about human welfare, not just now but tomorrow and the day after tomorrow? Turns out it's not GDP that's important, not some notion of social welfare, not life expectancy."

The key criterion, Dasgupta says, is a notion of wealth that includes natural capital. "It's about getting your economics right. We are supposed to be economists; governments are run by economists, but a whole class of assets are missing from their dataset." He is concerned with what he calls "inequality across time", the effects on future generations of our short-sightedness about natural capital. He speaks as a man with three children and five grandchildren. Professor Price also has two grandchildren. When he brought Jaan Tallinn to dinner at Trinity, other diners queued up to congratulate him on founding Skype, since it enabled them to stay in touch with their children and grandchildren.

I mention this to Tallinn and he says, "I sometimes joke that I can take personal responsibility for saving one million human relationships." Tallinn has six children himself. He is 42, and does not look old enough to have six children. When I first saw him in the Great Court, I took him to be a postgraduate, and one of the more modestly dressed ones. He has an interesting and charming manner. When asked a question, he will pause, apparently going into a dream state, muttering, "Yes… thinking." He will then make a rather formal pronouncement. He says things like, "The term 'singularity' is too vague to be used in a productive discussion." He is very fond of the word "heuristics".

Tallinn part-funds a number of horizon-scanning organisations, including a couple at Oxford University, and the Machine Intelligence Research Institute at Berkeley. He explains the challenge of getting rich people to donate to the study of technical risks. "Your evolutionary heuristics come back to the idea of a future roughly similar to what it is now. You give to the community as it is now, to benefit a similar community in the future."

People can't imagine the technological future, in other words, and I tell him I have difficulty with the idea of America being turned into paper clips. Can he come up with a less surreal example of AI run riot? Another pause. Then Tallinn says it might be easier for me to think of AI being tyrannical about control of the environment. "I was born behind the Iron Curtain," he says, "and I remember heated discussions about large-scale terra-forming projects, such as reversing the direction of the river Ob, or putting up large reflectors into space to heat up Siberia." That did the trick. I could see the danger of entrusting such work to AI.

Tallinn says that, for trouble to occur, "The machines don't have to have the opposite interests to ours. We don't exactly have the opposite interests to chimpanzees. However, things are not looking up for the

chimpanzees, because we control their environment. Our interests are not perfectly aligned with theirs, and it turns out it's not easy to get interests aligned."
I thought about this as I travelled home on the train, which was full of people playing with their smartphones. I had suggested to Tallinn that people were in love with technology, so believed their interests were perfectly aligned with it. He said: "There is a feedback loop between human values and those technologies. If you create something that improves human life, people will reward you for it, but this is not a universal law of physics. This is something that applies at the start of the 21st century. But artificial intelligence is not going to care about the human market. At the moment, the human is in the loop. That can change."

Martin Rees (20 Mar 2014), "Can we prevent the end of the world? (Video)."

TED.
http://www.ted.com/talks/martin_rees_can_we_prevent_the_end_of_the_world

Summary

A post-apocalyptic Earth, emptied of humans, seems like the stuff of science fiction TV and movies. But in this short, surprising talk, Lord Martin Rees asks us to think about our real existential risks — natural and human-made threats that could wipe out humanity. As a concerned member of the human race, he asks: What's the worst thing that could possibly happen?

Anders Sandberg (11 Jun 2014), "The five biggest threats to human existence."

Washington Post.
https://www.washingtonpost.com/posteverything/wp/2014/06/11/the-five-biggest-threats-to-human-existence

Excerpts

1. Nuclear war

The real threat is nuclear winter – that is, soot lofted into the stratosphere causing a multi-year cooling and drying of the world. Modern climate simulations show that it could preclude agriculture across much of the world for years. If this scenario occurs, billions would starve, leaving only scattered survivors that might be picked off by other threats such as disease. The main uncertainty is how the soot would behave: depending on the kind of soot the outcomes may be very different, and we currently have no good way of estimating this.

2. Bioengineered pandemic

The number of fatalities from bioweapons and epidemic outbreaks looks like it has a power-law distribution – most attacks have few victims, but a few kill many. Given current numbers, the risk of a global pandemic from bioterrorism seems very small. But that is just bioterrorism: Governments have killed far more people than terrorists with bioweapons (as many as 400,000 may have died from the WWII Japanese biowar program). And as technology gets more powerful, nastier pathogens become easier to design.

3. Superintelligence

The unusual thing about superintelligence is that we do not know if rapid and powerful intelligence explosions are possible: Maybe our current civilization as a whole is improving itself at the fastest possible rate. But there are good reasons to think that some technologies may speed things up far faster than current societies can handle. Similarly, we do not have a good grip on just how dangerous different forms of superintelligence would be, or what mitigation strategies would actually work. It is very hard to reason about future technology we do not have, or intelligences greater than ourselves. Of the risks on this list, this is the one most likely to either be massive or just a mirage.

4. Nanotechnology

The most obvious risk is that atomically precise manufacturing looks ideal for rapid, cheap manufacturing of things like weapons. In a world where any government could "print" large amounts of autonomous or semi-autonomous weapons (including facilities to make even more), arms races could become very fast – and hence unstable, since doing a first strike before the enemy gets too large an advantage might be tempting.

Weapons can also be small, precision things: A "smart poison" that acts like a nerve gas but seeks out victims, or ubiquitous "gnatbot" surveillance systems for keeping populations obedient, seem entirely possible. Also, there might be ways of getting nuclear proliferation and climate engineering into the hands of anybody who wants it.

5. Unknown unknowns

The most unsettling possibility is that there is something out there that is very deadly, and we have no clue about it. The silence in the sky might be evidence for this. Is the absence of alien visitors due to the fact that life or intelligence is extremely rare, or that intelligent life tends to get wiped out? If there is a future Great Filter, it must have been noticed by other civilizations too, and even that didn't help.

(On naturally-occurring risks)

You might wonder why climate change or meteor impacts have been left off this list. Climate change, no matter how scary, is unlikely to make the entire planet uninhabitable (but it could compound other threats if our defences to it break down). Meteors could certainly wipe us out, but we would have to be very unlucky. The average mammalian species survives for about a million years. Hence, the background natural extinction rate is roughly one in a million per year. This is much lower than the nuclear-war risk, which after 70 years is still the biggest threat to our continued existence.

Paul Kennedy (22 Oct 2014), "How to think about science, Part 5 (Audio)."

Ideas with Paul Kennedy, Canadian Broadcasting Corporation.
http://www.cbc.ca/radio/ideas/how-to-think-about-science-part-5-1.465006

Summary

In this episode 5 Ulrich Beck talks about the place of science in a risk society. Later in the hour you'll hear from another equally influential European thinker, Bruno Latour, the author of We Have Never Been Modern. He will argue that our very future depends on overcoming a false dichotomy between nature and culture.

Erin Biba (19 May 2015), "Meet the co-founder of an apocalypse think tank (Interview with Martin Rees)."

Scientific American.
http://www.scientificamerican.com/article/meet-the-co-founder-of-an-apocalypse-think-tank

Excerpt

What are the major risks to humanity as you see them and how serious are they?

I'm personally pessimistic about the community's capacity to handle advances in biotech. In the 1970s the pioneers of molecular biology famously formulated guidelines for recombinant DNA at the Asilomar conference. Such issues arise even more starkly today. There is current debate and anxiety about the ethics and prudence of new techniques: “gain of function” experiments on viruses and the use of so-called CRISPR gene-editing technology. As compared with the 1970s, the community is now more global, more competitive and more subject to commercial pressures. I'd fear that whatever can be done will be done somewhere by someone. Even if there are formally agreed protocols and regulations, they'll be as hard to enforce as the drug laws. Bioerror and bioterror rank highest on my personal risk register for the medium term (10 to 20 years).

Max Tegmark (16 Apr 2015), "Existential risk: A conversation with Jaan Tallinn."

Edge.
https://edge.org/conversation/jaan_tallinn-existential-risk

Excerpt

The reasons why I'm engaged in trying to lower the existential risks has to do with the fact that I'm a convinced consequentialist. We have to take responsibility for modeling the consequences of our actions, and then pick the actions that yield the best outcomes. Moreover, when you start thinking about—in the pallet of actions that you have—what are the things that you should pay special attention to, one argument that can be made is that you should pay attention to areas where you expect your marginal impact to be the highest. There are clearly very important issues about inequality in the world, or global warming, but I couldn't make a significant difference in these areas.

Tony Ord (17 Jun 2015), "Toby Ord on the likelihood of natural and anthropogenic existential risks (Video)."

Future of Humanity Institute, Oxford University.
http://www.fhi.ox.ac.uk/natural-vs-anthro

Summary

At a lecture at the Cambridge Centre for the Study of Existential Risk, Dr. Toby Ord discussed the relative likelihood of natural existential risk, as opposed to anthropogenic risks. His analysis of the issue indicates a much higher probability of anthropogenic existential risk.

Daniel Faggella (30 Jun 2015), "On existential risk and individual contribution to the 'good' (Audio)."

TechEmergence.
http://ieet.org/index.php/IEET/more/on_existential_risk_and_individual_contribution_to_the_good

Summary

Nick Bostrom is interviewed by IEET Advisory Board member Daniel Faggella, director of TechEmergence Podcast.

Catastrophic Risk Analysis

Richard Posner (2005), "Catastrophic risks, resource allocation, and homeland security."
Journal of Homeland Security.
https://web.archive.org/web/20110724055257/http://www.homelandsecurity.org/journal/Default.aspx?oid=133&ocat=1

Full text

Richard A. Posner is a senior lecturer in law at the University of Chicago Law School. He was a judge of the U.S. Court of Appeals for the Seventh Circuit from 1981 to 1993 and was chief judge of the court from 1993 to 2000. He has written more than a dozen books. This article is based on two of them: Catastrophe: Risk and Response*, chapter 3 (Oxford University Press, 2004), and* Preventing Surprise Attacks: Intelligence Reform in the Wake of 9/11*, chapter 3 (Hoover Institution and Rowman & Littlefield, 2005).*

I want to discuss the general problem of determining optimal responses to catastrophic risks, defined as the risks of low or unknown probability that, if they materialize, will inflict heavy losses. The risks can arise from natural phenomena, from human accidents, or, as in the case of terrorist attacks, from deliberate human behavior.

To deal in a systematic way with catastrophic risks requires first assessing them and then devising and implementing sensible responses. Assessment involves first of all collecting the technical data necessary to gauge, so far as that may be possible, the probability of particular risks, the purely physical consequences if the risks materialize (questions of value are for later), and the feasibility of various measures for reducing either the risks or the magnitude of the consequences by various amounts. The next step in the assessment stage is to embed the data in a cost-benefit analysis of the alternative responses to the risk.

I am not proposing that cost-benefit analysis, at least as it is understood by economists, should be *the* decision procedure for responding to catastrophic risks. But it is an indispensable step in rational decision making in this as in other areas of government regulation. Effective responses to most catastrophic risks are likely to be extremely costly, and it would be mad to adopt such responses without an effort to estimate the costs and benefits. No government is going to deploy a system of surveillance and attack for preventing asteroid collisions, for example, without a sense of what the system is likely to cost and what the expected benefits are likely to be (roughly, the costs of asteroid collisions that the system would prevent multiplied by the probabilities of such collisions) relative to the costs and benefits both of alternative systems and of doing nothing. The “precautionary principle” (“better safe than sorry”) popular in Europe is not a useful alternative to cost-benefit analysis, if only because of its sponginess. In its more tempered versions, the principle is indistinguishable from a cost-benefit analysis with risk aversion assumed. Risk aversion entails that extra weight be given to the downside of uncertain prospects. In effect it magnifies the costs of harmful events, but it does not overthrow cost-benefit analysis, as some advocates of the precautionary principle may believe.

Precautionary considerations, moreover, can work against intervention or limit the optimal scale of intervention. An example is the optimal response to the danger of abrupt global warming. Suppose there is a 70% probability that in 2024 global warming will cause a social loss of$1 trillion (present value) and a 30% probability that it will cause no loss, and that the possible loss can be averted by imposing emission controls now that will cost the society $500 billion (for simplicity’s sake, I assume the entire

cost is borne this year). In the simplest form of cost-benefit analysis, since the discounted loss from global warming in 2024 is $700 billion, imposing the emission controls now is cost-justified. But suppose that in 2014 we will learn for certain whether there is going to be the bad ($1 trillion) outcome in 2024. Suppose further that if we postpone imposing the emission controls until 2014, we can still avert the $1 trillion loss. Then clearly we should wait, not only for the obvious reason that the present value of $500 billion to be spent in 10 years is less than $500 billion (at a discount rate of 3%, it is approximately $425 billion) but also and more interestingly because there is a 30% chance that we will not have to incur *any* cost of emission controls. As a result, the expected cost of the postponed controls is not $425 billion, but only 70% of that amount, or $297.5 billion—which is a lot less than $500 billion. The difference is the value of waiting.

Suppose now that if today we impose emission controls that cost society $100 billion, this will, by forcing the pace of technological advance, reduce the cost of averting in 2014 the global-warming loss of $1 trillion in 2024 from $500 billion to $250 billion. After discounting to present value at 3% and by 70% to reflect the 30% probability that we'll learn in 2014 that emission controls are not needed, the $250 billion figure shrinks to $170 billion. This is $127.5 billion less than the superficially attractive pure wait-and-see approach ($297.5 billion minus$170 billion). Of course, there is a price for the modified wait-and-see option—$100 billion. But the value is greater than the price.

In the example, the probabilities associated with catastrophe were assumed to be known, and also to be substantial. Often they will not be known. And if they are known but slightly, people may react to them irrationally. From a statistical standpoint, studies indicate that people sometimes overreact to a slight risk if it is associated with a particularly vivid, attention-seizing event. The 9/11 attacks have been offered as an illustration of this phenomenon. But to describe a reaction to a risk as an overreaction is to assume that the risk is slighter than people thought, and this presupposes an ability to quantify the risk, however crudely. We do not have that ability with respect to terrorist attacks. About all that can be said with any confidence about the 9/11 attacks is that if the United States and other nations had done nothing in their wake to reduce the probability of a recurrence, the risk of further attacks would probably have been great, although we do not know enough about terrorist plans and mentalities to be certain, let alone to know how great. After the government took defensive measures, the risk of further large-scale attacks on the U.S. mainland fell. But no one knows by how much it fell, and in any event it would be a mistake to dismiss a risk merely because it could not be quantified and therefore *might* be small—for it might be great instead. Unfortunately the ability to quantify a risk has no necessary connection to its magnitude. We now know that the risk of a successful terrorist attack on the United States in the summer of 2001 was great, yet the risk could not have been estimated without an amount and quality of data that probably could not have been assembled. To assume that risks can be ignored if they cannot be measured is an ostrich response.

This point is illuminated by the old distinction between "risk" and "uncertainty." The former refers to a probability that can be estimated, whether on the basis of observed frequency or of theory, and the latter to a probability that cannot be estimated. Uncertainty does not, as one might fear it would, paralyze decision making. We could not function without making decisions in the face of uncertainty. We do that all the time by assigning, usually implicitly, an intuitive probability (what statisticians call a "subjective" probability) to the uncertain event. But it is one thing to act, and another to establish the need to act by conducting fruitful cost-benefit analyses, or by employing other rational decision-making methods, when the costs or benefits (or both) are uncertain because they are probabilistic and the probabilities are not quantifiable, even approximately. The difficulty is acute in some insurance markets. Insurers determine insurance premiums on the basis of either experience rating, which is to say an estimate of risk based on the frequency of previous losses by the insured or the class of insureds, or

exposure risk, which involves estimating risk on the basis of theory or, more commonly, a combination of theory and limited experience (there may be some history of losses, but too thin a one to be statistically significant). If a risk cannot be determined by either method, there is uncertainty in the risk-versus-uncertainty sense, and only a gambler, treating uncertainty as a situation of extreme and unknowable variance in possible outcomes, will write insurance when a risk cannot be estimated. Or the government, as in the Terrorism Risk Insurance Act of 2002, which requires insurance companies to offer coverage of business property and casualty losses due to terrorism but with the federal government picking up most of the tab.

Among the catastrophic risks that present the most stubborn challenges to the cost-benefit analyst is the risk of a terrorist attack using weapons of mass destruction, such as bioweaponry. The probability of a bioterrorist attack, or rather the schedule of probabilities for the various forms that such an attack might take, cannot be estimated. It is not only that terrorists are secretive as to plans and capabilities. It is also that they—or at least the ones that have vague and encompassing aims—have such a broad range of potential means and targets to choose among and, if suicidal, cannot be deterred. Anyone who thinks terrorist attacks are predictable should read what the director of the Defense Threat Reduction Agency wrote just months before September 2001: "We have, in fact, solved a terrorist problem in the last twenty-five years. We have solved it so successfully that we have forgotten about it; and that is a treat. The problem was aircraft hijacking and bombing. We solved that problem …The system is not perfect, but it is good enough.… we have pretty much nailed this thing."[1]

Clearly, science cannot predict where or when bioterrorists will strike, although it can say something about the likely means that they will employ, given feasibility and cost constraints, and much about the consequences of various types of bioterrorist attack. Maybe, however, the military and civilian intelligence services, the diplomatic service, and academic experts on terrorism can, by pooling their knowledge, produce reliable estimates of the probabilities of the various types of bioterrorist attack that are possible, and the estimates can then be married to scientific expertise to produce a schedule of expected costs of bioterrorism and therefore a guide to responsive measures. But it seems that about all that experts on terrorism are able to do, and even then only with a large error term, is to rank bioterrorist threats by *relative* likelihood—to say, for example, that a bioterrorist attack on Washington employing anthrax is more likely than an attack on London with smallpox. These rankings, while useful in establishing priorities within a fixed budget, do not enable expected costs to be calculated and so do not permit the application of cost-benefit analysis in its usual sense.

There are several possible methods, of varying utility, of adjusting cost-benefit analysis to reflect the presence of radical, nonquantifiable uncertainty. For example, it's been suggested that "information markets" might be used to elicit information about the likely risks of particular bioterrorist attacks. These are markets in which the "securities" traded are not stocks or other financial instruments, but predictions. The idea is that predictions will be more accurate when there is a financial stake in accuracy, and the existence of a financial stake will elicit predictions from the most knowledgeable observers. The theory is fine but doesn't seem applicable to terrorism. Terrorists could manipulate the market to generate inaccurate predictions or profit from their terrorism by making accurate ones. In addition, should bioterrorist attacks turn out to be infrequent (as we hope), it would be very difficult to verify the accuracy of the predictions; it would be like placing a bet on what the population of New York will be a hundred years from now. In the case of either natural catastrophes or accidental man-made ones, moreover, the man in the street does not have useful information, and the information possessed by scientists and other experts gets elicited and shared without need to provide a direct monetary reward for being right.

A more useful approach to cost-benefit analysis under conditions of extreme uncertainty is what I shall call "inverse cost-benefit analysis." It involves simply dividing what the government is spending to prevent a particular catastrophic risk from materializing by what the social cost of the catastrophe would be if it did materialize. The result of this division is an approximation to the implied probability of the catastrophe—implied, that is, by what the government is spending to combat it. Expected cost is the product of probability and consequence (loss): $C = PL$. If P and L are known, C can easily be calculated. If instead C and L are known, P can easily be calculated. If $1 billion ($C$) is being spent to avert a disaster that if it occurs will impose a loss (L) of $100 billion, then $P = C/L = .01$.

If P so calculated diverges sharply from independent estimates of it, this is a clue that we may be spending too much or too little on avoiding L. It is just a clue, because of the distinction, fundamental in economics, between marginal and total costs and benefits. The optimal expenditure on a measure is the expenditure that equates marginal cost to marginal benefit. Suppose in the example just given that we happen to know that P is not .01 but .1, so that the expected cost of the catastrophe is not $1 billion but $10 billion. It doesn't follow that we should be spending$10 billion, or indeed anything more than $1 billion, to avert the catastrophe. Maybe spending just $1 billion would reduce the expected cost of catastrophe from $10 billion all the way down to $500 million and no further expenditure would bring about a further reduction, or at least a cost-justified reduction. (For example, if spending another $1 billion would reduce the expected cost from $500 million to zero, that would be a bad investment, at least if risk aversion is ignored.) I discuss the implications of this point below but ignore it for the time being.

The federal government is spending at least $2 billion a year to prevent a bioterrorist attack. The goal is to protect Americans, so I shall ignore casualties in other countries. Suppose the most catastrophic biological attack that seems reasonably likely on the basis of what little we now know about terrorist intentions and capabilities would kill 100 million Americans. Economic studies of the value of life (studies based on what people demand in compensation for assuming small risks of death) yield a median per capita value for present-day Americans of $7 million. So if the attack occurred, the total costs would be $700 trillion—and that is too low because the death of more than a third of the population would have all sorts of collateral consequences, mainly negative. Let us, still conservatively however, refigure the total costs as $1 quadrillion. The result of dividing the money being spent to prevent such an attack, $2 billion, by $1 quadrillion is 1/500,000. Is there only a 1 in 500,000probability of a bioterrorist attack of that magnitude in the coming year? One doesn't know; but a probability of 1 in 500,000 seems too low.

It doesn't follow that $2 billion is too little to be spending to prevent a bioterrorist attack, for the distinction between total and marginal costs must be borne in mind. Suppose that by spending $2 billion we reduce the probability of such an attack from .01 to .0001. The expected cost of the attack would still be very high—$1 quadrillion multiplied by .0001 is $100 billion—but spending more than $2 billion might not reduce the residual probability of .0001 at all. For there might be no feasible further measures to take to combat bioterrorism, especially when we remember that increasing the number of people involved in defending against bioterrorism also increases the number of people capable, alone or in conjunction with others, of mounting biological attacks. But we must also bear in mind that expenditures on combating bioterrorism do more than prevent mega-attacks; the lesser attacks, which would still be very costly both singly and cumulatively, would also be prevented.

Costs, moreover, tend to be inverse to time. It would cost a lot more to build an asteroid defense in one year than in 10 years because of the extra costs that would have to be incurred in order to effectuate a sudden reallocation of the required labor and capital from the current projects in which they are

employed—and so would other crash efforts to prevent catastrophes. Placing a lid on current expenditures would have the incidental benefit of enabling additional expenditures to be deferred to a time when, because more will be known about both the catastrophic risks and the optimal responses to them, considerable cost savings may be possible. (This is the option approach that I discussed earlier in reference to abrupt global warming.) The case for such a ceiling derives from comparing marginal benefits to marginal costs, which may be sharply increasing in the short run.

A further qualification in evaluating the current response to the threat of bioterrorism requires mention. It concerns the way in which government expenditures are assigned to the different activities involved in combating terrorism. The expenditure category "catastrophic threats" in the federal budget is dominated by expenditures on identifying, detecting, and developing vaccines and cures for lethal pathogens. Expenditures classified elsewhere, however, such as expenditures on intelligence gathering, background checks, and border searches, will reduce the likelihood of bioterrorist attacks, though border searches would contribute very little because of the difficulty of detecting a lethal pathogen in a person's luggage. We should think of the catastrophic-threats category in the federal budget as addressed to the residual risk of a bioterrorism attack if the "forward" defenses fail (this is another marginal comparison); nevertheless, the estimate of that risk implied by the expenditures in that category still seems too low.

Another area in which current government expenditures on mitigating catastrophic risks seem too low involves detecting and preventing asteroid collisions. NASA spends about $3.9 million a year compiling a catalogue of dangerous "near-Earth objects," a preliminary defensive measure. But that is it, although the agency's program of research on "smaller Solar System objects," namely asteroids and comets, while not oriented toward defense against collisions, may yield knowledge that would be useful for such a defense. Other expenditures, actual and planned, and both private and public, swell the total. But the aggregate amount is small. Tellingly, NASA's annual reports do not contain a section on asteroid defense or near-Earth objects. The current expenditure level is so close to zero that the distinction between total and marginal benefits and costs has little significance.

We know that the expected costs of asteroid collisions are nontrivial, though low, and that methods of detection, mitigation, and prevention are feasible and probably would not break the bank. The report of the Near-Earth Object Science Definition Team, commissioned by NASA, recommended a system of detection of all near-Earth objects at least 140 meters in diameter; the team estimated that it would cost $300 millionto construct the system. Both the risks of asteroid collisions and the possible methods for detecting and intercepting asteroids that are on a collision course with the Earth have been known for some time, so the budget has had time to adjust but hasn't done so.

A parallel United Kingdom task force estimated the annual probability of an asteroid collision that would kill 1.5 billion people as one in 250,000. Since value of life is positively correlated with per capita income, the $7 million figure I used earlier is too high when most casualties would be foreigners. Assume a value of life of $2 million. Then the expected annual cost of the collision would be $12 billion ($2 million x1.5 billion [= $3 quadrillion] x .000004), which is many, many times the U.S. government's annual spending on asteroid defense. More to the point, since most of the 1.5 billion victims would not be Americans, the *world's* annual spending on asteroid defense—which is probably only very slightly more than $3.9 million because no other country has gone beyond the talking stage so far as an asteroid defense is concerned—is too low.

A proposal is pending for federal financing of a Large-aperture Synoptic Survey Telescope (LSST). This $150 million instrument "could locate 90% of all near-Earth objects down to 300 m in size, enable computation of their orbits, and permit assessment of their threat to Earth," while greatly increasing our

knowledge of remote galaxies. The telescope would not be a complete substitute for the telescopic array recommended by the NASA task force, even if the 10% of asteroids that would escape detection altogether are ignored. The LSST would spot an asteroid only when the asteroid passed in its orbital path through the section of sky swept by the telescope; the asteroid would not be continuously monitored even though its orbit might change after the initial observation. But the LSST would be a great start. Yet NASA refuses to fund it, so funding is being sought from the National Science Foundation and private sources. Astronomers, moreover, are much more interested in remote galaxies—study of which adds to knowledge of the origin, size, age, future, and composition of the universe—than in local orbiting rocks. The extent to which the LSST, if it is built, will actually be used for detection and evaluation of potentially dangerous asteroids is uncertain.

The federal government's science and technology budget allocates about $1.7 billion a year to climate-change research, including research on clean fuels and carbon sequestration as well as on improving predictions of global warming. If the warming is moderate, the costs to theUnited States are likely to be modest, and $1.7 billion a year might actually be too much to spend on counteracting it. However, abrupt, catastrophic global warming is a possibility, and let me assume that if it occurred it would bring about a permanent reduction of one-fifth of gross domestic product, which is currently $10 trillion. Because the loss of $2 trillion a year is assumed to be permanent, the present value of the loss caused by the disaster, at a 3% discount rate, is slightly more than $66.6 trillion. The annual probability of a global-warming disaster of the assumed magnitude cannot be estimated. But it is at least plausible that a level of carbon dioxide emissions taxes that induced a considerably although not astronomically greater investment (largely private) than at present on averting such a disaster would be cost justified.

Table 1 summarizes the probabilities of catastrophe implied by current government expenditures to avert the three catastrophic risks that I have been discussing.

Table 1. Implied Annual Catastrophe Probabilities

Catastrophe	***C***	***L***	***P (implied)***
Bioterrorist attack (100 million deaths)	$2 billion	$1 quadrillion (U.S.)	.000002 (1 in 500,000)
Asteroid collision (1.5 billion deaths)	$3.9 million	$3 quadrillion	.0000000013 (1 in 769 million)
Catastrophic global warming	$1.7 billion	$66.6 trillion (U.S.)	.00000255 (1 in 388,000)

The distinction between total and marginal effects is only one qualification that must be borne in mind in reading this table. Notice that the table estimates the costs to the entire human race in the case of a disastrous asteroid strike, but only the costs to the United States in the case of bioterrorism and catastrophic global warming. This may seem arbitrary. But no other nation seems to be devoting any significant resources to trying to prevent an asteroid disaster, while other nations are devoting resources to preventing catastrophic global warming. As I do not know the amount of those resources, however, I cannot assess the adequacy of the total expenditures devoted to protecting the entire human race from those disasters. Moreover, even if the only costs of an asteroid disaster that should be considered in determining how much the United Statesshould spend to prevent such a disaster are costs

to the United States, scaling down the cost figures in Table 1 accordingly would still indicate that we are spending too little. Dollar-weighted, the United States is about one-fourth of the world; and remember that value-of-life estimates are positively correlated with per capita income.

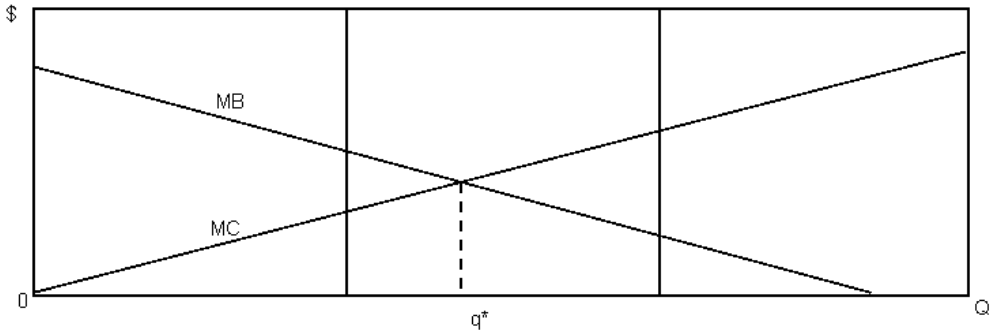
Because it will sometimes be sensible to disregard low-priority projects entirely, the function of threat assessment, in regard to catastrophic risks as well as to more familiar threats, is not only to rank threats by their expected cost but also to fix a cutoff point below which threats will be disregarded because they would require attention disproportionate to the social benefits that attention to them would confer. Time diverted to thinking about very low-probability threats is unavailable for thinking about other threats unless the aggregate amount of attention to threats is increased. That would require diverting intellectual effort from other activities, and the diversion might be costly. The Office of Science and Technology Policy in the Executive Office of the President has a staff of only 50, which for political, budgetary, and personnel reasons may be difficult to expand in the short run. If so, the office might be making a rational choice to devote no attention at all to the asteroid threat on the grounds either that threats of lesser catastrophes deserve more attention because their expected costs are greater or because they seem more amenable to evaluation and response, or that other scientific projects altogether deserve more attention. The government cannot spend all its time conducting cost-benefit analyses of remote-seeming risks.

Another way to put this is that the costs of responding to risks of disaster include the cost of assessing the risk and formulating the response—the cost of cost-benefit analysis—and may be considerable when opportunity cost (the forgone value of alternative uses of the time and other resources devoted to the cost-benefit analysis) is included, as it should be, in the costs of such analysis. But I doubt that ignoring the risk of a catastrophic collision with an asteroid can be justified on these grounds. Not only is it a non-negligible risk of a huge catastrophe, but the costs of responding to the risk, even as expanded to take in the opportunity cost just mentioned, are moderate.

A final qualification is that the estimates for the expenditures required for an effective asteroid defense and for arresting global warming are too low. In the first case, they ignore other government programs, including other NASA programs for studying asteroids, that contribute at least indirectly to defense against the risk, and in both cases they ignore nongovernmental expenditures. The LSST, if it is built, will be financed in part by private universities, and many near-Earth asteroids have been discovered by the Lincoln Labs' LINEAR (Lincoln Near-Earth Asteroid Research) program, using two telescopes, although it does receive federal funding, some of it from NASA. The federal government finances only about half the basic research conducted in this country, and some of the other half, which is financed by universities and private companies out of their own pockets, contributes to defending against the catastrophes in question. In addition, some companies are voluntarily reducing their carbon dioxide emissions. And investments in energy efficiency designed merely to reduce the cost of energy may reduce those emissions as a byproduct.

Figure 1 illustrates another way in which cost-benefit analysis can be used fruitfully even when there is great uncertainty about one or more of the components of the analysis.

Figure 1. The "Tolerable Windows" Approach



The marginal benefits and marginal costs of measures to reduce or eliminate some catastrophic risk are shown as functions of the quantity of precautions taken, with the optimal level of precautions (q^*) given by the intersection of the two functions. Suppose the optimum cannot be determined because of uncertainty about costs, benefits, the discount rate, or probabilities. We may, though, know enough about the benefits and costs to be able to create the "window" formed by the two vertical lines. Notice that at the left side of the window frame the benefits of a further effort to eliminate or prevent the catastrophe in question comfortably exceed the costs, while at the right side the reverse is true. If we stay within the window, although we won't know whether our measures are optimal, we'll at least have some basis for confidence that they are neither grossly inadequate nor grossly excessive. A plausible application is to the current funding of asteroid defense: it is likely that we are well to the left of the left side of the window.

Here is another example of the tolerable-windows approach in action: The benefits of preserving the existing amount of genetic diversity cannot be quantified. But the cost of preserving samples of animals and plants, whether entire species or varieties within a species (such as different breeds of the same animal species), that are on the verge of extinction is probably small enough to put us to the far left of the window. Indeed, since these samples can be preserved at low cost in the form of frozen seeds that can be resuscitated and made to germinate, large-scale efforts to preserve biodiversity by tightly limiting human land uses may not be cost-justified. We cannot be sure because there is no census of species and many of them have very small populations and those often in out-of-the-way places (such as ocean bottoms)—and these are the very species most at risk of extinction, and it would be infeasible to obtain and preserve specimens of all of them. At least modest efforts to preserve specimens of species, or varieties, on the verge of extinction seem worthwhile, and so that is the place to start.

A more familiar simplification of cost-benefit analysis than the tolerable-windows approach is risk-risk assessment, whereby the risks to life or health of alternative responses to some danger (including the alternative of doing nothing) are compared, but no effort is made either to monetize them or to bring other costs and benefits into the analysis. This approach can work well in simple cases—for example, when a measure to prevent a 1% risk of death in an automobile accident would create a 2% risk of death in such an accident. It is also relevant to the dual-use dilemma that is created by efforts to prevent bioterrorism: measures that impede access to lethal pathogens may slow research into the development of effective medical responses to natural epidemics.

But the utility of the method is limited because it leaves out considerations that may be critical to a responsible decision—namely other costs and benefits. For example, advances in medicine that reduce mortality may increase the rate of population growth, thereby contributing indirectly but not necessarily trivially to global warming, the costs of which cannot be reduced to lives lost, although abrupt global warming could cause a catastrophic loss of life. Population growth creates other negative externalities as well. They may or may not exceed the positive externalities; but the uncritical belief, which is

standard in risk-risk assessment, that "saving lives" is always a good thing is an obstacle to responding effectively to catastrophic risks.

It may be objected that cost-benefit analysis is a waste of time because politicians are not welfare maximizers. They are not. But cost-benefit analyses can influence public policy even in a political system guided by self-interested politicians responsive to interest groups. An interest group will not press for a project that does not confer net benefits on it. The greater the excess of benefits over costs, the likelier are the beneficiaries to be able to overcome the free-rider impediment to the formation of an effective interest group; the greater the excess of costs over benefits, the likelier are opponents to be able to organize effective resistance. So information about costs and benefits can influence political outcomes even if no political faction is committed to adopting only those policies that can pass a cost-benefit test.

National defense is a good example of a government program that exists because of a very great preponderance of benefits over costs, great as those costs, and uncertain as the benefits, are, rather than because national defense confers economic rents on some narrow interest group, though some people still believe that defense expenditures are the result of machinations by the "merchants of death." National defense is not only a good example, but a pertinent one. Measures for defending against catastrophic risks reflect concerns similar to those that motivate the nation's heavy military expenditures.

But there are all sorts of obstacles—political, psychological, economic, and cultural—to responding rationally to catastrophic risks. And the problem seems general. Students of regulation have been critical of the gross and seemingly irrational differences in the estimates of the value of life that are implicit in government regulation of different risks. The range is from$100,000 for death in accidents involving unvented space heaters to $92 billion for death from the herbicides atrazine or alachlor in drinking water. (These figures are derived by dividing the cost of preventing the death by the probability that death would occur if the cost were not incurred.) Suppose NASA's asteroid-defense budget of $3.9 million a year is perfectly attuned to the public's valuation of lives lost in asteroid collisions and an estimate made by John Lewis that the expected cost of such collisions is 1,479 deaths per year is correct. That is a global figure, and the U.S. population is only about 5% of the total world population, so let us reduce this number to 74. The implication is that NASA is valuing each of these lives at $52,700. This is not only less than 1% of the $7 million mean estimate in the scholarly literature; it is little more than half the value of a life imperiled by an unvented space heater.

The differences among the value of life estimates probably can be explained by information costs, by psychological factors such as probability neglect, the availability heuristic, and the "dread" factor (notably absent in death by unvented space heater), by political factors, and by the asymptotic relation between risk and the value of life (when risks are very slight, people often write them down to, or very near, zero). The differences may also be somewhat exaggerated by the critics. Nevertheless, the criticism that government does not use consistent criteria to determine responses to risk has great force. And as Table 1 and the accompanying discussion suggest, the criticism applies as forcefully to the regulation of catastrophic risks as to the lesser risks on which the critics have focused. It underscores the importance of having cost-benefit analyses of responses to catastrophic risks conducted by neutrals who do not have financial, political, or psychological stakes in how the analyses come out.

I have noted several times the peculiar difficulties involved in estimating terrorist threats and responding to surprise attacks generally. These difficulties will be the focus of the balance of this article.

Consider two states of the world. In one, a warning of a surprise attack occurs but is disregarded, so the attack takes place, inflicting costs of $a$on the victim. In the other state of the world, the warning is

heeded and the attack is defeated, at cost *d* (for defensive measure), but because the attack is defeated, *a* is zero. Let the probability of the attack be *p*; then the probability that there will be no attack is 1 – *p*. The expected cost of the attack if the warning is disregarded is *pa*, and the expected cost if the warning is heeded is (1 – *p*)*d*, so the warning should be heeded if*pa* > (1 – *p*)*d* and be disregarded otherwise.

The assumption that *d* affects *a* but not *p* may seem questionable because we usually think of defensive measures as being designed to reduce the likelihood of whatever prospective injury is being defended against. Most surprise attacks, however, occur even if the element of surprise is lost; they just do less damage. But the analysis would not be materially altered by assuming that defensive measures reduce the probability of an attack as well as the damage from it.

Another assumption is that if the warning is heeded, the damage inflicted by the surprise attack will be zero. This assumption is unrealistic and should be relaxed. The damage will just be smaller than if the warning had been ignored. Denote that diminished damage by *b*; it is smaller the greater *d* is, *d* being the expenditure on defensive measures when the warning is heeded.

Besides the direct cost of defensive measures, there is a lulling "boy crying wolf" cost, which I'll denote by *w*. This cost is greater the smaller the probability of attack and therefore the more often that warnings will be false alarms, which increase the likelihood that true alarms will be ignored. It is also greater the greater *d* is, because if big costs are incurred to defend against an attack that does not occur there will be a greater reluctance to heed the next warning.

In light of these adjustments, the inequality *pa* > (1 – *p*)*d*, which states the condition for when a warning should be heeded and thus defensive measures taken, becomes, with a slight rearrangement of terms,

$$p/(1-p) > [b(d) + d + w(p,d)]/a. \qquad \text{[Inequality 1]}$$

Inequality 1 says that it is more likely that heeding the warning will be the prudent response the higher *p* is (which not only increases the left-hand side of the inequality, but, because of its negative effect on *w*, reduces the right-hand side), the lower *d* is, the lower *w* is, and the higher *a*is. Conversely, the lower *p* is but the higher *d* is, and the smaller the effect of defensive measures in reducing *b* (the diminished cost of an attack if the defensive measures are taken) and hence the higher *b*(*d*) is, the likelier the prudent course is to ignore the warning sign. The effect of *d* is complex: it makes heeding the warning more likely to be prudent by reducing *b* (the damage from the attack when precautions are taken), but less likely to be prudent because it is a cost of heeding the warning and because it increases the lulling effect.

To illustrate, the Israelis disregarded the signs of an imminent attack by the Egyptians and Syrians in October 1973 because they thought the probability of an attack low, because defensive measures (mobilization) would have been costly, because a lulling effect had been induced by a previous costly mobilization in response to what proved to be a false alarm, and because, believing that even without mobilizing the reserves their frontline forces could hold the line, they didn't think mobilization necessary to minimize the cost of an attack (that is, they didn't think *b*was much lower than *a*). In the case of the 9/11 attacks, *p* was thought low, *a* was thought lower than it turned out to be, and *d* was high because of the cost, and inconvenience to passengers, of the kind of airline security measures that were adopted after the attacks.

Thus far I have treated *d* dichotomously: if inequality 1 is satisfied, the potential victim of a surprise attack should take *d* measures; if not, he should take no measures. A more realistic assumption (which

incidentally permits dispensing with *b*) is that *d* can vary. Concretely, if *d* = 0, *a* is as in inequality 1, but as *d* rises, *a* falls: the more defensive measures that are taken, the less harm the attack does. The goal, then, in picking the level of *d* is to minimize the sum (*S*) of the expected costs of the attack and the costs of *d*, where *d* is the number of units of defense and *c*(*d*) the cost of defense. Thus

$$S = pa(d) + c(d) + (1 - p)w(p, d), \qquad \text{[Equation 2]}$$

S is thus the sum of the costs of false negatives (failing to predict attacks that occur), which is the first term on the right-hand side of equation 2,and the costs of false positives (false alarms), which are given by the second and third terms, the second being the cost of defensive measures and the third the lulling cost.

Provided that the rate at which an increase in *d* reduces *a* exceeds the rate at which such an increase increases *c* and *w*, *S* is minimized by taking the derivative of *S* with respect to *d* and setting the result equal to zero, yielding

$$c_d + (1 - p)w_d = - pa_d, \qquad \text{[Equation 3]}$$

where a_d is the effect on *a* (the harm to the victim of the attack) of a small change in *d* (the extent of defensive measures), and c_d and w_d are the effects on *c* (the cost of defensive measures) and *w* (the lulling cost), respectively, also of a small change in *d*. In words, the optimal expenditure on defensive measures requires increasing them to the point at which a $1 increase in their cost (including the effect on the lulling cost) reduces the expected cost of the attack by $1. The greater the effect of such expenditure in reducing the cost of an attack if it occurs, and the higher the probability of an attack (provided that the effect on the expected cost of such an attack exceeds the effect on reducing the expected lulling cost), the greater the cost-justified level of measures to anticipate and respond to the attack.

The model is still unrealistic, in being limited to a single prospective surprise attack. A related unrealism is that it ignores the dynamic character of the crying-wolf phenomenon. The boy who cried wolf did not sound only a single false alarm; it was the repetition of false alarms that made it impossible for him to convince his hearers that his latest alarm was true. In other words, the lulling cost rises with each false alarm.

Assume there are *t* periods in each of which there is an equal probability of an attack that will impose the same costs and cost the same to defend against, and that for every period in which an attack does not occur the lulling cost increases by *r*% a year. With this adjustment, the sum of all costs, *S* in equation 2 becomes

$$S' = tpa(d) + tc(d) + (1 - p)w(d)y(t,p), \qquad \text{[Equation 4]}$$

where $y(t,p) = t(1 + r)^j$ and *j* is a probability distribution of *p*. Notice that *y*, and hence the lulling cost, increases with *t* and with *r* but decreases with *p*, because the higher *p* is, the likelier is an attack, and an attack will reduce the lulling cost in the next period. It might, however, replace it with a "hyper-alert cost"—a possible increased risk of surprise attack if all attention is focused on preventing a repetition of a previous attack to the neglect of other possible attacks. For example, the nation may be expending excessive resources on screening airline passengers, to the neglect of potential terrorist threats to other parts of the nation's transportation system. In addition, a hyper-alert state may precipitate a flood of

warnings that turn out to be false alarms (which has certainly been the experience since 9/11), creating new lulling costs. The other side of this coin is that false alarms draw attention away from true dangers; they are, at best (that is, without producing a crying-wolf effect), distracting noise. My model ignores all these complications, but they are worth mentioning just to indicate the complexity of responding intelligently to the threat of surprise attack.

S is minimized (provided that some plausible restrictions are placed on the terms) when

$$tpc_d = -tpa_d - (1-p)w_d y(t,p). \qquad \text{[Equation 5]}$$

In words, the total investment in defensive measures against a possible surprise attack should be carried to the point at which the cost of an additional measure, plus the increase in expected lulling costs from taking the additional measure, would be just equal to the reduction in the expected cost of attacks that the measure would bring about.

The foregoing analysis is offered as a possible aid to identifying relevant considerations and the relations among them. It is not intended as an algorithm. The problem with using a formula to optimize the response to warnings of an attack is the difficulty, bordering on the impossibility, of quantifying the terms, other than *d* and *c* and in some cases *b*(*d*). Assessing the probability of a surprise attack is particularly baffling, as we know. A further difficulty is that a formula cannot be applied across the entire spectrum of possible surprise attacks; this is precluded by the inescapable necessity of filtering data in accordance with the analyst's preconceptions. There is a near-infinite number of data points in our visual and auditory fields, and we can't take them all in at once. A rational person prioritizes in accordance with his interests. So intelligence officers determine where the greatest dangers lie, and having made that determination give greater weight to incoming information that bears on those dangers than to information on more remote dangers.

This gives rise to the following paradox: a surprise attack is likelier to succeed when it has a low antecedent probability of success and the attacker is weak, because on both counts the victim will discount the danger and because the range of possible low-probability attacks by weak adversaries is much greater than the range of possible high-probability attacks by strong ones. The potential victim marshals his defensive resources to protect the high-probability targets of greatest value, leaving under-protected the immense number of lower-valued low-probability targets. Knowing this, an enemy who wants to achieve strategic surprise picks one of those inferior targets. Realizing that this is what the enemy is likely to do, and that he is therefore unlikely to obtain a decisive victory, the potential victim reckons the expected loss (severity discounted by probability) from the attack as low and so does not invest a great deal in anticipating and taking measures to defend against the attack, especially since the cost of defending against the entire spectrum of low-probability attacks by weak adversaries (who may, moreover, be numerous) is prohibitive. Surprise attacks are a favorite tactic of the weak because they are a force multiplier, which a weak enemy needs most. When used by the weak they tend to be wild, and ultimately unsuccessful, gambles, but may inflict great damage en route to their ultimate failure. This may explain, by the way, why surprise attacks are relatively rare. On the one hand, when employed by the weak, they are indeed gambles, with dim prospects of ultimate success (the weaker of two contenders is likely to lose the contest), and the greater prospect of ultimate defeat is a deterrent. On the other hand, a strong, aggressive state has difficulty achieving strategic surprise because its intentions are anticipated.

The basic elements of this analysis can be formalized with the aid of our original inequality, $pa > (1-p)d$, which says take defensive measures if but only if the expected cost of an attack exceeds their cost.

Assume now that there are two types of attack, one that involves a high probability of inflicting a devastating loss (high p and high a), the other a low probability of inflicting a much smaller loss (low p and low a).Assume further that there are n potential attacks of the first type and n' of the second and that d, the defensive measures necessary to prevent an attack, is the same for each class. Let p' denote the probability of attacks in the second class and a' the harm caused by such an attack, so $p' < p$and $a' < a$, but $n' > n$.

We now have two inequalities, the first denoting the condition for taking defensive measures against the first type of attack and the second the condition for taking defensive measures against the second type:

$n[pa > (1 - p)d]$ [Inequality 6a]

$n'[p'a' > (1 - p')d]$ [Inequality 6b]

The first inequality is much more likely to be satisfied than the second. The fact that there are more potential attacks of the second type is irrelevant. If the expression in brackets is negative, multiplying it, however many times, will not make it positive; and unless it is positive, defensive measures will not be cost justified. The first term in the bracketed expression in inequality 6b, $p'a'$, is smaller than the corresponding term in inequality 6a because it is the product of two smaller terms, so, for example, if $p = .2$ and $p' = .1$, and $a = 100$ and $a' = 20$, $pa = 20$ and$p'a' = 2$. The second term, $(1 - p')d$, is larger, because d is unchanged but $1 - p'$ is larger than $1 - p$ (in the example, it is .9 versus .8). The smaller the first term and the larger the second, the more likely the bracketed term is to be negative and so the less likely are defensive measures to be justified. In the example, if $d = 5$, the first inequality is $20 - .8(5) = 16$, while the second is $2 - .9(4) = -1.64$. So it does not pay to take defensive measures aimed at averting the lesser attack.

These numbers are arbitrary, but they illustrate how it can be rational to take no defensive measures at all against a large class of potential surprise attacks. This is all the more likely when the costs of information are taken explicitly into account. The existence of those costs—alternatively, the necessity (owing to the limitations of human mental capacity) of economizing on attention—makes it likely that below some threshold of expected cost, no consideration whatever will be given to taking defensive measures against a class of possible surprise attacks. Such a "threshold heuristic," which is related to my earlier point about the indispensability of preconceptions to rational thought, may be at once rational and an invitation to attack. It may also be related to an irrational tendency of people to write down small risks to zero, though presumably intelligence professionals and others who deal with risk professionally are less likely to succumb to this tendency than laypeople.

The fundamental problem, however, is the asymmetry of attacker and victim. The attacker picks the time, place, and means of attack. Since without a great deal of luck his plan cannot be discovered in advance by the victim, the attacker has, by virtue of his having the initiative and of the victim's being unable to be strong everywhere all the time, a built-in advantage that assures a reasonable probability of a local success. The attacks on Pearl Harbor, Tet (but for its political impact), Yom Kippur, and the 9/11 attacks all achieved only local successes. But when an attacker is willing to settle for a local success, there is little the victim can do to prevent it.

Finally, as Thomas Schelling has pointed out in his book *The Strategy of Conflict*, the more sensitive a warning system, the greater the risk of the victim's responding mistakenly with a preemptive attack on the supposed attacker. The system "may cause us to identify an attacking plane as a seagull, and do

nothing, or it may cause us to identify a seagull as an attacking plane, and provoke our inadvertent attack on the enemy." So here is still another reason to doubt the wisdom of seeking an airtight defense against surprise attacks.

References

1. Jay Davis, "Epilogue: A Twenty-First Century Terrorism Agenda for the United States," in *The Terrorism Threat and U.S. Government Response: Operational and Organizational Factors*, James M. Smith and William C. Thomas, eds. (Colorado Springs, CO: U.S. Air Force Institute for National Security Studies, 2001), p. 275.

Charles Meade and Roger Molander (21 Jul 2006), "Considering the effects of a catastrophic terrorist attack."

RAND Center for Terrorism Risk Management Policy.
http://www.rand.org/content/dam/rand/pubs/technical_reports/2006/RAND_TR391.pdf

Summary

A quickly growing concern about terrorism is that a devastating attack would send social and economic aftershocks cascading through multiple sectors long after the initial strike was over. While much analysis has been done on the possible short-term effects of an attack of this magnitude, no work has investigated longer-term implications. Exploratory efforts to do so are needed.

With this motivation, the RAND team developed a novel approach that enabled us to investigate two key policy questions:

Within the first 72 hours, what would the direct effects of such an attack be? What human casualties, property damage, and destruction of infrastructure would result immediately?

In the weeks and months after the attack, what would the longer-term economic implications be? From a decisionmaking standpoint, what would the particularly challenging policy issues be? What would the high-priority concerns for different stakeholder groups be?

To answer the first question, we conducted a scenario analysis; strategic gaming provided us with insights into the second. Both tools provide means of exploring highly uncertain policy landscapes. In scenario analysis, researchers posit a "what if" framework and examine how various factors might interact to generate a sequence of events—i.e., "What if such and such happened next?" In strategic gaming, participants are realistically immersed in a stressful event and directed to explore the resulting policy challenges for various stakeholders. By combining these approaches, we were able to link the immediate challenges of a hypothetical attack with its possible consequences at a macro level.

A Devastating Attack on a Key Component of the U.S. Economic Infrastructure

In our scenario, terrorists conceal a 10-kiloton nuclear bomb in a shipping container and ship it to the Port of Long Beach. Unloaded onto a pier, it explodes shortly thereafter. This is referred to as a "ground-burst" as opposed to an "airburst" explosion. We used this scenario because analysts consider it feasible, it is highly likely to have a catastrophic effect, and the target is both a key part of the U.S. economic infrastructure and a critical global shipping center. This scenario formed the basis for strategic games with leaders from government, business, and the insurance and real estate industries. Participants shared their perspectives on what the attack's longer-term consequences might be and outlined the decisions they would be likely to make in response to the sequence of events our scenario

analysis suggested. They also anticipated the decisionmaking challenges that might arise and reflected on strategies that might address these problems.

Both Short- and Long-Term Repercussions of the Attack Could Be Overwhelming

Within the first 72 hours, the attack would devastate a vast portion of the Los Angeles metropolitan area. Because ground-burst explosions generate particularly large amounts of highly radioactive debris, fallout from the blast would cause much of the destruction. In some of the most dramatic possible outcomes:

- Sixty thousand people might die instantly from the blast itself or quickly thereafter from radiation poisoning.
- One-hundred-fifty thousand more might be exposed to hazardous levels of radioactive water and sediment from the port, requiring emergency medical treatment.
- The blast and subsequent fires might completely destroy the entire infrastructure and all ships in the Port of Long Beach and the adjoining Port of Los Angeles.
- Six million people might try to evacuate the Los Angeles region.
- Two to three million people might need relocation because fallout will have contaminated a 500-km2 area.
- Gasoline supplies might run critically short across the entire region because of the loss of Long Beach's refineries—responsible for one-third of the gas west of the Rockies.

Economic Implications in the Weeks and Months After the Attack

The early costs of the Long Beach scenario could exceed $1 trillion, driven by outlay for medical care, insurance claims, workers' compensation, evacuation, and construction. The $50 billion to $100 billion for 9/11 puts this figure into perspective. In general, consequences would far outstrip the resources available to cope with them.

In addition, over time, the economic effects of the catastrophe are likely to spread far beyond the initial attack, reaching a national and even international scale. Decisionmakers would face two particularly difficult challenges: keeping the global shipping supply chain operating and restoring orderly economic relationships.

Keeping the Global Shipping Supply Chain Operating

In the aftermath of the attack, different stakeholder groups affected might have differing interests. Consequently, their decisions might often be at odds. How to contend with such conflicting interests is the key challenge for policymakers. In terms of global shipping, the main tension might be between the political aim of preventing a future attack and the business interest in seeing that U.S. ports and the global shipping supply chain continue to operate. The only way to completely mitigate the risk of a second strike would be to close all U.S. ports and suspend all imports indefinitely. This would be the national security community's likely position. Yet in business terms, this position would be untenable. The loss of the ports of Long Beach and Los Angeles alone, which handle 30 percent of U.S. shipping imports, would already be substantial. All U.S. ports combined carry out 7.5 percent of world trade activity. Accordingly, the business community would likely call for ports to stay open, or to reopen as early as possible.

But harsh realities facing the financial and real estate communities might prove a barrier. The Long Beach attack might cripple an insurance industry struggling to absorb massive losses from claims. Insurance would be in tremendously short supply—particularly for terrorist and nuclear risks. Without it,

ports and related infrastructure could not operate. Further complicating the issue is the high probability that people would flee port cities, severely depleting local labor supplies. Given these conditions, all U.S. ports would likely close indefinitely or operate at a substantially reduced level following the attack. This would severely disrupt the availability of basic goods and petroleum throughout the country.

Restoring Orderly Economic Relationships

The attack is likely to have dramatic economic consequences well beyond the Los Angeles area:

- Many loans and mortgages in Southern California might default.
- Some of the nation's largest insurance companies might go bankrupt.
- Investors in some of the largest financial markets might be unable to meet contract obligations for futures and derivatives.

While exact outcomes are difficult to predict, these hypothetical consequences suggest alarming vulnerabilities. Restoring normalcy to economic relations would be daunting, as would meeting the sweeping demands to compensate all of the losses.

Next Steps Would Involve Further Modeling and Gaming

The analysis tools we developed for this study lay the groundwork for research exploring both the short- and long-term effects of catastrophic events. The need is pressing to continue such investigations, particularly of longer-term economic repercussions. This work would entail developing scenarios for a new generation of strategic games. The overarching goals would be to gain further insights into the policy and economic decisions likely to be made in the months following attacks of this magnitude and characterize the decision landscape. For example, we could illuminate any potentially unprecedented behavior that might occur in the global economy in times of extreme duress, identify where existing systems are likely to fail, and evaluate the benefits of a range of potential economic policies. In this way, policymakers could start to anticipate the types of decisions they might be called upon to make, reflect in times of relative calm on their options, and plan well in advance for contingencies.

Cass Sunstein (21 Feb 2007), "The catastrophic harm precautionary principle."

Issues in Legal Scholarship 6 (available gratis via SSRN).
http://www.degruyter.com/view/j/ils.2007.6.issue-3/issue-files/ils.2007.6.issue-3.xml
http://ssrn.com/abstract=2532598

Abstract

When catastrophic outcomes are possible, it makes sense to take precautions against the worst-case scenarios — the Catastrophic Harm Precautionary Principle. This principle is based on three foundations: an emphasis on people's occasional failure to appreciate the expected value of truly catastrophic losses; a recognition that political actors may engage in unjustifiable delay when the costs of precautions would be incurred immediately and when the benefits would not be enjoyed until the distant future; and an understanding of the distinction between risk and uncertainty. The normative arguments are illustrated throughout with reference to the problem of climate change; other applications include avian flu, genetic modification of food, protection of endangered species, and terrorism.

D. Catastrophe and Irreversibility

For many catastrophic risks, there are two additional wrinkles. The first is that knowledge is likely to increase over time. At one stage, it may be possible to assign a probability range for certain risks: The likelihood of catastrophic harm may be below 40 percent but above five percent, and perhaps rough probabilities can be assigned to the poles. But at another stage, the assignment might be far more precise, allowing something closer to a point estimate. More dramatically, circumstances of uncertainty, or bounded uncertainty (in which the risk is between, say, 5 percent and 30 percent, but in which we do not know how likelihood of figures within the range) might shift to circumstances of risk -- as mounting knowledge permit regulators to assign probabilities to the various outcomes. The fact that knowledge grows over time might well be taken as a reason to follow a principle of "wait and learn," on the theory that immediate action is often undertaken in the dark.

But there are two problems with waiting. The first is that by hypothesis, we do not know enough to exclude the possibility that catastrophic harm will occur while or because we wait. The second is that the failure to take precautionary action may be irreversible, or reversible at only very high cost.63 For example, greenhouse gases stay in the atmosphere for a long time, and inaction may saddle posterity with a catastrophic risk that future generations are effectively powerless to eliminate. It may well make sense to take precautions by buying an "option" through regulatory steps to counteract risks that may turn out to be catastrophic. In ordinary life, people buy such options all the time, not merely through financial instruments but also by attempting to ensure that one or another course of action is reversible.

The problem is that it is necessary to establish a price for the relevant "option." If we can assign probabilities to the various outcomes, it should not be difficult to specify that price, using some version of expected value. When probabilities are difficult or impossible to assign, the establishment of that price presents all of the issues that I have discussed thus far.

From these points, no simple conclusion follows. But several points are clear. On certain assumptions, it makes sense to take relatively unaggressive steps against potentially catastrophic risks, perhaps freezing the status quo, if new information will emerge over time. It follows that for climate change, low-cost emissions reduction requirements -- intended as an initial precaution to be followed by greater reductions as technology advances -- may be the best approach if gaps in current knowledge will be filled in the relevant time period. It also follows that because greenhouse gas emissions are effectively irreversible, there is a special reason to act immediately. The extent of the action depends on its costs and benefits.

Leonie A. Marks, et al. (Apr 2007), "Mass media framing of biotechnology news."

Public Understanding of Science 16.
http://pus.sagepub.com/content/16/2/183.short

Abstract

In fast-changing scientific fields like biotechnology, new information and discoveries should influence the balance of risks and rewards and their associated media coverage. This study investigates how reporters interpret and report such information and, in turn, whether they frame the public debate about biotechnology. Mass media coverage of medical and agricultural biotechnology is compared over a 12-year period and in two different countries: the United States and the United Kingdom. We examine whether media have consistently chosen to emphasize the potential risks over the benefits of these

applications, or vice versa, and what information might drive any relevant changes in such frames. We find that the two sets of technologies have been framed differently—more positive for medical applications, more negative for agricultural biotechnology. This result holds over time and across different geographic locations. We also find that international events influence media coverage but have been locally framed. This local newsworthiness extends to both medical and agricultural applications. We conclude that such coverage could have led to differences in public perception of the two sets of technology: more negative (or ambivalent) for agricultural, positive for medical applications. Our findings suggest that understanding news frames, and the events that drive them, provides some insight into the long-term formation of public opinion as influenced by news coverage.

Anders Sandberg, Jason Matheny, and Milan M. Ćirković (9 Sep 2008), "How can we reduce the risk of human extinction?"

Bulletin of the Atomic Scientists.
http://thebulletin.org/how-can-we-reduce-risk-human-extinction

Full text

In the early morning of September 10, the Large Hadron Collider will be tested for the first time amid concern that the device could create a black hole that will destroy the Earth. If you're reading this afterwards, the Earth survived. Still, the event provides an opportunity to reflect on the possibility of human extinction. Since 1947, the Bulletin has maintained the Doomsday Clock, which "conveys how close humanity is to catastrophic destruction--the figurative midnight--and monitors the means humankind could use to obliterate itself." The Clock may have been the first effort to educate the general public about the real possibility of human extinction.

Less publicly, there had been earlier speculations about humanity's undoing. During the Manhattan Project, Robert Oppenheimer ordered a study to calculate whether a nuclear detonation would cause a self-propagating chain of nuclear reactions in the Earth's atmosphere. The resulting report, "LA-602: Ignition of the atmosphere with nuclear bombs," may represent the first quantitative risk assessment of human extinction. LA-602 concluded that ignition was physically impossible, and nuclear development proceeded.

In 1950, physicist Leo Szilard renewed worries about human extinction after estimating that a sufficiently large number of nuclear weapons wrapped in cobalt would, when detonated, render the Earth's surface uninhabitable for five years (the half-life of cobalt 60). Szilard's fear that such a "doomsday device" might be developed inspired much of Herman Kahn's 1960 treatise, On Thermonuclear War, as well as the premise of Stanley Kubrick's 1964 film Dr. Strangelove. While such a device remains possible in principle, it would require vast amounts of cobalt, and there is no indication that such a weapon was ever built.

In 1983, discussion of human extinction re-emerged when Carl Sagan and others calculated that a global thermonuclear war could generate enough atmospheric debris to kill much of the planet's plant life and, with it, humanity. While the "nuclear winter" theory fell out of favor in the 1990s, recent climate models suggest that the original calculations actually underestimated the catastrophic effects of thermonuclear war. Moreover, the original model of Sagan and his collaborators supported research showing that supervolcanic eruptions and asteroid or comet impacts could pose comparable extinction risks.

Despite these notable instances, in the 61 years since the Doomsday Clock's creation, the risk of human extinction has received relatively scant scientific attention, with a bibliography filling perhaps one page. Maybe this is because human extinction seems to most of us impossible, inevitable, or, in either case, beyond our control. Still, it's surprising that a topic of primary significance to humanity has provoked so little serious research.

One of the missions of the Future of Humanity Institute at Oxford University is to expand scholarly analysis of extinction risks by studying extinction-level hazards, their relative probabilities, and strategies for mitigation. In July 2008, the institute organized a meeting on these subjects, drawing experts from physics, biology, philosophy, economics, law, and public policy.

The facts are sobering. More than 99.9 percent of species that have ever existed on Earth have gone extinct. Over the long run, it seems likely that humanity will meet the same fate. In less than a billion years, the increased intensity of the Sun will initiate a wet greenhouse effect, even without any human interference, making Earth inhospitable to life. A couple of billion years later Earth will be destroyed, when it's engulfed by our Sun as it expands into a red-giant star. If we colonize space, we could survive longer than our planet, but as mammalian species survive, on average, only two million years, we should consider ourselves very lucky if we make it to one billion.

Humanity could be extinguished as early as this century by succumbing to natural hazards, such as an extinction-level asteroid or comet impact, supervolcanic eruption, global methane-hydrate release, or nearby supernova or gamma-ray burst. (Perhaps the most probable of these hazards, supervolcanism, was discovered only in the last 25 years, suggesting that other natural hazards may remain unrecognized.) Fortunately the probability of any one of these events killing off our species is very low--less than one in 100 million per year, given what we know about their past frequency. But as improbable as these events are, measures to reduce their probability can still be worthwhile. For instance, investments in asteroid detection and deflection technologies cost less, per life saved, than most investments in medicine. While an extinction-level asteroid impact is very unlikely, its improbability is outweighed by its potential death toll.

The risks from anthropogenic hazards appear at present larger than those from natural ones. Although great progress has been made in reducing the number of nuclear weapons in the world, humanity is still threatened by the possibility of a global thermonuclear war and a resulting nuclear winter. We may face even greater risks from emerging technologies. Advances in synthetic biology might make it possible to engineer pathogens capable of extinction-level pandemics. The knowledge, equipment, and materials needed to engineer pathogens are more accessible than those needed to build nuclear weapons. And unlike other weapons, pathogens are self-replicating, allowing a small arsenal to become exponentially destructive. Pathogens have been implicated in the extinctions of many wild species. Although most pandemics "fade out" by reducing the density of susceptible populations, pathogens with wide host ranges in multiple species can reach even isolated individuals. The intentional or unintentional release of engineered pathogens with high transmissibility, latency, and lethality might be capable of causing human extinction. While such an event seems unlikely today, the likelihood may increase as biotechnologies continue to improve at a rate rivaling Moore's Law.

Farther out in time are technologies that remain theoretical but might be developed this century. Molecular nanotechnology could allow the creation of self-replicating machines capable of destroying the ecosystem. And advances in neuroscience and computation might enable improvements in cognition that accelerate the invention of new weapons. A survey at the Oxford conference found that concerns about human extinction were dominated by fears that new technologies would be misused. These emerging threats are especially challenging as they could become dangerous more quickly than

past technologies, outpacing society's ability to control them. As H.G. Wells noted, "Human history becomes more and more a race between education and catastrophe."

Such remote risks may seem academic in a world plagued by immediate problems, such as global poverty, HIV, and climate change. But as intimidating as these problems are, they do not threaten human existence. In discussing the risk of nuclear winter, Carl Sagan emphasized the astronomical toll of human extinction:

A nuclear war imperils all of our descendants, for as long as there will be humans. Even if the population remains static, with an average lifetime of the order of 100 years, over a typical time period for the biological evolution of a successful species (roughly ten million years), we are talking about some 500 trillion people yet to come. By this criterion, the stakes are one million times greater for extinction than for the more modest nuclear wars that kill "only" hundreds of millions of people. There are many other possible measures of the potential loss--including culture and science, the evolutionary history of the planet, and the significance of the lives of all of our ancestors who contributed to the future of their descendants. Extinction is the undoing of the human enterprise.

There is a discontinuity between risks that threaten 10 percent or even 99 percent of humanity and those that threaten 100 percent. For disasters killing less than all humanity, there is a good chance that the species could recover. If we value future human generations, then reducing extinction risks should dominate our considerations. Fortunately, most measures to reduce these risks also improve global security against a range of lesser catastrophes, and thus deserve support regardless of how much one worries about extinction. These measures include:

- Removing nuclear weapons from hair-trigger alert and further reducing their numbers;
- Placing safeguards on gene synthesis equipment to prevent synthesis of select pathogens;
- Improving our ability to respond to infectious diseases, including rapid disease surveillance, diagnosis, and control, as well as accelerated drug development;
- Funding research on asteroid detection and deflection, "hot spot" eruptions, methane hydrate deposits, and other catastrophic natural hazards;
- Monitoring developments in key disruptive technologies, such as nanotechnology and computational neuroscience, and developing international policies to reduce the risk of catastrophic accidents.

Other measures to reduce extinction risks may have less in common with strategies to improve global security, generally. Since a species' survivability is closely related to the extent of its range, perhaps the most effective means of reducing the risk of human extinction is to colonize space sooner, rather than later. Citing, in particular, the threat of new biological weapons, Stephen Hawking has said, "I don't think the human race will survive the next thousand years, unless we spread into space. There are too many accidents that can befall life on a single planet." Similarly, NASA Administrator Michael Griffin has noted, "The history of life on Earth is the history of extinction events, and human expansion into the Solar System is, in the end, fundamentally about the survival of the species."

Probably cheaper than building refuges in space would be building them on Earth. Elaborate bunkers already exist for government leaders to survive nuclear war, and the Svalbard Global Seed Vault in Norway protects crop seeds from nuclear war, asteroid strikes, and climate change. Although Biosphere 2 may inspire giggles, functioning refuges that are self-sufficient, remote, and permanently occupied would help to safeguard against a range of hazards, both foreseeable and unforeseeable.

Perhaps least controversial, we should invest more in efforts to enumerate the risks to human survival and the means to mitigate them. We need more interdisciplinary research in quantitative risk assessment, probability theory, and technology forecasting. And we need to build a worldwide community of experts from various fields concerned about global catastrophic risks. Human extinction may, in the long run, be inevitable. But just as we work to secure a long life for individuals, even when our eventual death is assured, we should work to secure a long life for our species.

Toby Ord, Rafaela Hillerbrand, and Anders Sandberg (30 Oct 2008), "Probing the improbable: Methodological challenges for risks with low probability and high stakes."

Journal of Risk Research.
http://www.fhi.ox.ac.uk/probing-the-improbable.pdf , http://arxiv.org/abs/0810.5515

Abstract

Some risks have extremely high stakes. For example, a worldwide pandemic or asteroid impact could potentially kill more than a billion people. Comfortingly, scientific calculations often put very low probabilities on the occurrence of such catastrophes. In this paper, we argue that there are important new methodological problems which arise when assessing global catastrophic risks and we focus on a problem regarding probability estimation. When an expert provides a calculation of the probability of an outcome, they are really providing the probability of the outcome occurring, given that their argument is watertight. However, their argument may fail for a number of reasons such as a flaw in the underlying theory, a flaw in the modeling of the problem, or a mistake in the calculations. If the probability estimate given by an argument is dwarfed by the chance that the argument itself is flawed, then the estimate is suspect. We develop this idea formally, explaining how it differs from the related distinctions of model and parameter uncertainty. Using the risk estimates from the Large Hadron Collider as a test case, we show how serious the problem can be when it comes to catastrophic risks and how best to address it.

Excerpts

3.3 Historical examples of Model and Theory Failure

A dramatic example of a model failure was the Castle Bravo nuclear test on March 1 1954. The device achieved 15 megatons of yield instead of the predicted 4-8 megatons. Fallout affected parts of the Marshal Islands and irradiated a Japanese fishing boat so badly that one fisherman died, causing an international incident (Nuclear Weapon Archive 2006). Though the designers at Los Alamos National Laboratories understood the involved theory of alpha decay, their model of the reactions involved in the explosion was too narrow, for it neglected the decay of one of the involved particles (lithium-7), which turned out to contribute the bulk of the explosion's energy. The Castle Bravo test is also notable for being an example of model failure in a very serious experiment conducted in the hard sciences and with known high stakes.

The history of science contains numerous examples of how generally accepted theories have been overturned by new evidence or understanding, as well as a plethora of minor theories that persisted for a surprising length of time before being disproven. Classic examples for the former include the Ptolemaic system, phlogiston theory and caloric theory; an example for the latter is human chromosome number, which was systematically miscounted as 48 (rather than 46) and this error persisted for more than 30 years (Gartler 2006).

As a final example, consider Lord Kelvin's estimates of the age of the Earth (Burchfield 1975). They were based on information about the earth's temperature and heat conduction, estimating an age of the Earth of between 20 and 40 million years. These estimates did not take into account radioactive heating, for radioactive decay was unknown at the time. Once it was shown to generate additional heat the models were quickly updated. While neglecting radioactivity today would count as a model failure, in Lord Kelvin's day it represented a largely unsuspected weakness in the physical understanding of the Earth and thus amounted to theory failure. This example makes it clear that the probabilities for the adequacy of model and theory are not independent of each other, and thus in the most general case we cannot further decompose equation (3).

5. Conclusions

When estimating threat probabilities, it is not enough to make conservative estimates (using the most extreme values or model assumptions compatible with known data). Rather, we need robust estimates that can handle theory, model and calculation errors. The need for this becomes considerably more pronounced for low-probability high-stake events, though we do not say that low probabilities cannot be treated systematically. Indeed, as pointed out by (Yudkowsky 2008), if we could not correctly predict probabilities lower than 10-6 , we could not run lotteries. Some people have raised the concern that our argument might be too powerful: for it is impossible to disprove the risk of even something as trivial as dropping a pencil, then our argument might amount to prohibiting everything. It is true that we cannot completely rule out any probability that apparently inconsequential actions might have disastrous effects, but there are a number of reasons why we do not need to worry about universal prohibition. A major reason is that for events like the dropping of a pencil which have no plausible mechanism for destroying the world, it seems just as likely that the world would be destroyed by not dropping the pencil. The expected losses would thus balance out. It is also worth noting that our argument is simply an appeal to a weak form of decision theory to address an unusual concern: for our method to lead to incorrect conclusions, it would require a flaw in decision theory itself, which would be very big news.

It will have occurred to some readers that our argument is fully applicable to this very paper: there is a chance that we have made an error in our own arguments. We entirely agree, but note that this possibility does not change our conclusions very much. Suppose, very pessimistically, that there is a 90% chance that our argument is sufficiently flawed that the correct approach is to take safety reports' probability estimates at face value. Even then, our argument would make a large difference to how we treat such values. Recall the example from section 2, where a report concludes a probability of 10-9 and we revise this to 10-6. If there is even a 10% chance that we are correct in doing so, then the overall probability estimate would be revised to 0.9!!!10-9 + 0.1!!!10-6 " 10-7, which is still a very significant change from the report's own estimate. In short, even serious doubt about our methods should not move one's probability estimates more than an order of magnitude away from those our method produces. More modest doubts would have a effect.

The basic message of our paper is that any scientific risk assessment is only able to give us the probability of a hazard occurring conditioned on the correctness of its main argument. The need to evaluate the reliability of the given! argument in order to adequately address the risk was shown to be of particular relevance in low probability high-stake events. We drew a three-fold distinction!between theory, model and calculation, and showed how this can be more useful than the common dichotomy in risk assessment between model and parameter uncertainties. By providing historic examples for errors in the three fields, we clarified the three-fold distinction and showed where flaws in a risk assessment might occur. Our analysis was applied to the recent assessment of risks that might arise from

experiments within particle physics. To conclude this paper, we now provide some very general remarks on how to avoid argument flaws when assessing risks with high stakes.

Firstly, the testability of predictions can help discern flawed arguments. If a risk estimate produces a probability distribution for smaller, more common disasters this can be used to judge whether the observed incidences are compatible with the theory. Secondly, reproducibility appears to be the most effective way of removing many of these errors. By having other people replicate the results of calculations independently our confidence in them can be dramatically increased. By having other theories and models independently predict the same risk probability our confidence in them can again be increased, as even if one of the arguments is wrong the others will remain. Finally, we can reduce the possibility of unconscious bias in risk assessment through the simple expedient of splitting the assessment into a 'blue' team of experts attempting to make an objective risk assessment and a 'red' team of devil's advocates attempting to demonstrate a risk, followed by repeated turns of mutual criticism and updates of the models and estimates (Calogero 2000). Application of such methods could in many cases reduce the probability of error by several orders of magnitude.

Martin Weitzman (Feb 2009), "On modeling and interpreting the economics of catastrophic climate change."

Review of Economics and Statistics 91.
http://www.ewp.rpi.edu/hartford/~ernesto/F2014/MMEES/Papers/ENVIRONMENT/1EnvironmentalSystemsModeling/Weitzman2009-Modeling-Economics-ClimateChange.pdf

Abstract

With climate change as prototype example, this paper analyzes the implications of structural uncertainty for the economics of low-probability, high-impact catastrophes. Even when updated by Bayesian learning, uncertain structural parameters induce a critical "tail fattening" of posterior-predictive distributions. Such fattened tails have strong implications for situations, like climate change, where a catastrophe is theoretically possible because prior knowledge cannot place sufficiently narrow bounds on overall damages. This paper shows that the economic consequences of fat-tailed structural uncertainty (along with unsureness about high-temperature damages) can readily outweigh the effects of discounting in climate-change policy analysis.

Conclusion

Last section's heroic attempts at constructive suggestions notwithstanding, it is painfully apparent that the dismal theorem makes economic analysis trickier and more open-ended in the presence of deep structural uncertainty. The economics of fat-tailed catastrophes raises difficult conceptual issues that cause the analysis to appear less scientifically conclusive and more contentiously subjective than what comes out of an empirical CBA of more usual thin-tailed situations. But if this is the way things are with fat tails, then this is the way things are, and it is an inconvenient truth to be lived with rather than a fact to be evaded just because it looks less scientifically objective in cost-benefit applications.

Perhaps in the end the climate-change economist can help most by not presenting a cost-benefit estimate for what is inherently a fat-tailed situation with potentially unlimited downside exposure as if it is accurate and objective—and perhaps not even presenting the analysis as if it is an approximation to something that is accurate and objective—but instead by stressing somewhat more openly the fact that such an estimate might conceivably be arbitrarily inaccurate depending upon what is subjectively assumed about the high-temperature damages function along with assumptions about the fatness of

the tails and/or where they have been cut off. Even just acknowledging more openly the incredible magnitude of the deep structural uncertainties that are involved in climate-change analysis—and explaining better to policymakers that the artificial crispness conveyed by conventional IAM-based CBAs here is especially and unusually misleading compared with more ordinary non-climate-change CBA situations—might go a long way toward elevating the level of public discourse concerning what to do about global warming. All of this is naturally unsatisfying and not what economists are used to doing, but in rare situations like climate change where DT applies we may be deluding ourselves and others with misplaced concreteness if we think that we are able to deliver anything much more precise than this with even the biggest and most detailed climate-change IAMs as currently constructed and deployed.

The contribution of this paper is to phrase exactly and to present rigorously a basic theoretical principle that holds under positive relative risk aversion and potentially unlimited exposure. In principle, what might be called the catastrophe-insurance aspect of such a fat-tailed unlimited-exposure situation, which can never be fully learned away, can dominate the social-discounting aspect, the pure-risk aspect, and the consumption-smoothing aspect. Even if this principle in and of itself does not provide an easy answer to questions about how much catastrophe insurance to buy (or even an easy answer in practical terms to the question of what exactly is catastrophe insurance buying for climate change or other applications), I believe it still might provide a useful way of framing the economic analysis of catastrophes.

Mark Jablonowski (14 Jun 2009), "Increasing uncertainty about high-stakes risks: The impetus for radical change?"

Annual Meeting of the North American Fuzzy Information Processing Society, Cincinnati, Ohio.
http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=5156447

Introduction

The modern industrial age has brought with it concerns about human survival. Many believe that the complexities of the technological and industrial trappings of progress, along with related social responses, bring with them excessive risks. To drive home their point that such risk can only be avoided through significant changes in the way we view progress, some revert to prognostications about how and when such untoward events may occur (“doomsaying”). Given the uncertainties surrounding these unfortunate occurrences, any such prognostications are bound to be very imperfect. Errors in such predictions, however, are often taken by society as falsifying the proposition that progress-induced disaster is a genuine threat. As a result, these pronouncements of impending doom are unlikely to bring about any meaningful change.

Most arguments for and against the potential for disaster in the modern world are based on a simplistic view of the world. This view ignores the true importance of fundamental uncertainties that result from knowledge imperfections. The real question is not whether such disasters will happen, but rather are they sufficiently possible? The prudent course in the face of extinction requires radical (i.e., fundamental) change to our social and economic systems to reduce or eliminate the possibility. On this basis, the mounting uncertainties about disaster that we face today could act to provide the impetus for radical change, in and of themselves.

Uncertainty Matters

Possibility defines a unique form of uncertainty due to imperfect knowledge [1]. We may have sufficient information to narrow a set of possible outcomes, yet only to some degree. A one dimensional set of these outcomes, say, the outside air temperature the next day, may be expressed as an interval estimate - in the case of tomorrow's temperature, between 65 and 75 degrees Fahrenheit. Intervals are not estimated from data directly, but rather instrumentally, as judged by how well they let us cope with an uncertain world. Nonetheless, these uncertainties are real. They define a region which we may describe as the evidently unknown.

We often assess the probability and loss characteristics of some exposure to risk only imperfectly. In such cases, uncertainty due to randomness and the natural variability it entails combines with knowledge imperfection. This is especially the case with high stakes (catastrophic) risks for which little reliable data is available. In Figure 1 we show an interpretation of the uncertainty surrounding the risk associated with accumulating exposures. While the figure is meant to convey an intuitive sense of uncertainty, as our acceptance goes from no risk to “too many” risks, this analysis can be formalized using the theory of fuzzy sets [2].

The uncertainty that surrounds risk estimates can influence the way critical decisions about risk are made [3]. Say, for example, that you must go in for a serious medical operation. Two options are available. Procedure A has a proven track record based on hundreds of clinical trials, plus a statistically documented success rate over many years of actual results. On the basis of these results, the suggested effectiveness carries a likelihood of. 90. Procedure B, on the other hand, is fairly new. Based on some limited clinical trials, and its similarity to other successful procedures, the doctor that invented it suggests that it should have a success rate of “about 90 percent”. If we had to give a precise estimate of the procedural effectiveness of A and B it would probably be. 90 for both (the “best guess” estimate in each case). Yet, all other things being equal, we would certainly prefer procedure A. Procedure B presents us with the unknown possibility that effectiveness may turn out to be considerably lower than. 90.

When dealing with existential risks we need to consider the possibilities as well as the probabilities of loss. The fundamental problem of catastrophe is that in the long run, there may be no long run. We simply do not get a second chance to get things right. The mere possibility of existential risk must then be a sufficient indicator for action. As a result, this uncertainty will have a significant impact on how we manage such risks.

Conclusions

Progress has brought increasing complexity, along with the uncertainty about high-stakes risk this complexity entails. As these risk can be catastrophic (i.e., terminal), we don't get a second chance to make the right decisions about managing them. To make the point that our survival may depend on real change in our definition and measurement of “progress”, some critics of the status quo suggest that we will meet our doom if we ignore the issues. We have suggested here that the problem is not so much that the “end is near”, but rather that the future is becoming so uncertain. Uncertainty due to knowledge imperfection about high-stakes risk potentials is real, and as such it has real effects on the way we make decisions. Recognizing this uncertainty can help us make better decisions about high-stakes risk, thereby aiding us in continuing a rather remarkable streak of natural survival.

Under extreme uncertainty about catastrophic risks, only a properly precautionary approach to risk, based on prudent avoidance of existential threats, makes sense. The upshot then of increasing uncertainty about risk is the need to integrate a comprehensive system of risk control into a coordinated social and economic planning effort that recognizes these uncertainties. While such far reaching changes

to our economic and social systems may be perceived as radical, increasing uncertainty about risk will push us into such actions sooner or later. It is better to anticipate such actions, rather than be forced into them – especially after it becomes too late.

Bruce Tonn and Dorian Stiefel (Nov 2014), "Human extinction risk and uncertainty: Assessing conditions for action."

Futures 63.
http://www.sciencedirect.com/science/article/pii/S0016328714001207

Abstract

Under what sets of conditions ought humanity undertake actions to reduce the risk of human extinction? Though many agree that the risk of human extinction is high and intolerable, there is little research into the actions society ought to undertake if one or more methods for estimating human extinction risk indicate that the acceptable threshold is exceeded. In addition to presenting a set of patterns of lower and upper probabilities that describe human extinction risks over 1000 years, the paper presents a framework for philosophical perspectives about obligations to future generations and the actions society might undertake. The framework for philosophical perspectives links three perspectives—no regrets, fairness, maintain options—with the action framework. The framework for action details the six levels of actions societies could take to reduce the human extinction risk, ranging from doing nothing (Level I) to moving to an extreme war footing in which economies are organized around reducing human extinction risk (Level VI). The paper concludes with an assessment of the actions that could be taken to reduce human extinction risk given various patterns of upper and lower human extinction risk probabilities, the three philosophical perspectives, and the six categories of actions.

1. Introduction

Under what sets of conditions ought humanity undertake actions to reduce the risk of human extinction? Concerns about the potential extinction of the human race are growing (Bostrom, 2002; Matheny, 2007; Posner, 2004; Tonn, 2009a). Many prominent researchers, scientists, and government officials believe that this threat is high and intolerable (Bostrom, 2002; Highfield, 2001; Leslie, 1996; Matheny, 2007; Rees, 2003; U.K. Treasury, 2006). For example, based on his review of trends and situations facing humanity, Rees estimates 50–50 odds that our present civilization will survive to the end of the present century (Rees, 2003). Bostrom (2002) asserts that the probability of human extinction exceeds 25% while Leslie (1996) estimates that the probability of human extinction over the next five centuries is 30%. The Stern Review (U.K. Treasury, 2006), influenced by environmental risks such as climate change, reports an almost 10% chance of extinction by the end of this century.

When groups estimate the risk, the answers are even more interesting. In an informal survey (Sandberg & Bostrom, 2008) at the Global Catastrophic Risk Conference, participants rated the median human extinction risk from events such as being killed by super-intelligent artificial intelligence (5%), natural pandemic (0.05%), and overall risk of extinction before 2100 (19%). In an international survey of the general public, 45% of respondents believed that humans would become extinct within 1000 years (Tonn, 2009a).

This paper synthesizes three frameworks to address the sets of conditions under which society ought to act to reduce human extinction risks as measured by the probability of extinction per year as defined at present. The first framework deals with philosophical perspectives about obligations to future

generations (see Section 2)—no regrets, maintaining options, and fairness. Acceptable thresholds of human extinction risk have been developed for each of these obligational concepts (Tonn, 2009b).

The second framework details patterns of uncertainty in exceeding the acceptable threshold of human extinction risk over time (see Section 3), which represent the upper and lower annual probabilities of human extinction. Patterns are distinguished both by the depiction of the likelihood or unlikelihood of human extinction and by the magnitude of uncertainty represented (i.e., the amount of area between the upper and lower probability curves) and how the likelihoods of extinction and magnitude of uncertainty are estimated to change over time. We also discuss the ways in which each philosophical perspective relates to uncertainty over time.

The third framework addresses six levels of actions that society should consider as the risk of human extinction becomes more dire (see Section 4):

I. do nothing;
II. minor tax incentives, deployment programs;
III. major programs (e.g., carbon tax) and major public investments;
IV. Manhattan scale projects;
V. rationing, population control, major command and control regulations; and
VI. extreme war footing, economy organized around reducing human extinction risk.

Section 5 synthesizes the three frameworks by positing the levels of action society should take given sets of patterns of upper and lower probabilities of human extinction by obligational perspective. We show that the three philosophical perspectives argue for different levels of societal action depending on the changes in risk over time and the varying magnitudes of uncertainty.

Milan Ćirković, Anders Sandberg, and Nick Bostom (Oct 2010), "Anthropic shadow: Observation selection effects and human extinction risks."

Risk Analysis 30.
http://www.nickbostrom.com/papers/anthropicshadow.pdf

Abstract

We describe a significant practical consequence of taking anthropic biases into account in deriving predictions for rare stochastic catastrophic events. The risks associated with catastrophes such as asteroidal/cometary impacts, supervolcanic episodes, and explosions of supernovae/gamma-ray bursts are based on their observed frequencies. As a result, the frequencies of catastrophes that destroy or are otherwise incompatible with the existence of observers are systematically underestimated. We describe the consequences of this anthropic bias for estimation of catastrophic risks, and suggest some directions for future work.

1. INTRODUCTION: EXISTENTIAL RISKS AND OBSERVATION SELECTION EFFECTS

Humanity faces a series of major global threats, both in the near- and in the long-term future. These are of theoretical interest to anyone who is concerned about the future of our species, but they are also of direct relevance to many practical and policy decisions we make today. General awareness of the possibility of global catastrophic events has risen recently, thanks to discoveries in geochemistry, human evolution, astrophysics, and molecular biology.(1−6) In this study, we concentrate on the subset of catastrophes called existential risks (ERs): risks where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential.(7) Examples of

potential ERs include global nuclear war, collision of Earth with a 10-km sized (or larger) asteroidal or cometary body, intentional or accidental misuse of bio- or nanotechnologies, or runaway global warming.

There are various possible taxonomies of ERs.(7) For our purposes, the most relevant division is one based on the causative agent. Thus we distinguish: (1) natural ERs (e.g., cosmic impacts, supervolcanism, nonanthropogenic climate change, supernovae, gamma-ray bursts, spontaneous decay of cosmic vacuum state); (2) anthropogenic ERs (e.g., nuclear war, biological accidents, artificial intelligence, nanotechnology risks); and (3) intermediate ERs, ones that depend on complex interactions between humanity and its environment (e.g., new diseases, runaway global warming). In what follows, we focus mainly on ERs of natural origin.(8)

Our goal in this article is to study a specific observation selection effect that influences estimation of some ER probabilities, threatening to induce an anthropic bias into the risk analysis.3 Anthropic bias can be understood as a form of sampling bias, in which the sample of observed events is not representative of the universe of all events, but only representative of the set of events compatible with the existence of suitably positioned observers. We show that some ER probabilities derived from past records are unreliable due to the presence of observation selection effects. Anthropic bias, we maintain, can lead to underestimation of the probability of a range of catastrophic events.

We first present a simple toy model of the effect in Section 2, which we generalize in Section 3. We develop the argument in more detail in Section 4, and consider its relevance to various types of global catastrophic risks in Section 5. Finally, in Section 6, we discuss how the theory of observation selection effects might generally be applied to global catastrophes.

5. WHICH ERs ARE SUBJECT TO ANTHROPIC SHADOW?

Anthropic shadow bias will downwardly influence probability estimates of hazards: (1) that could have destroyed our species or its predecessors; (2) that are sufficiently uncertain; and (3) for which frequency estimates are largely based on terrestrial records. There are many hazards satisfying these broad criteria, including:

(i) Asteroidal/cometary impacts (severity gauged by the Torin scale or the impact crater size).
(ii) Supervolcanism episodes (severity gauged by the so-called volcanic explosivity index or a similar measure).
(iii) Supernovae/gamma-ray burst explosions (severity gauged by the variations in the distance and the intrinsic power of these events).
(iv) Superstrong solar flares (severity gauged by the power of electromagnetic and corpuscular emissions).

Various hazards can be distinguished by the degree to which they satisfy these criteria. For instance, the asteroidal and cometary impact history of the solar system is, in theory, easier to obtain for the Moon, where the erosion is orders of magnitude weaker than on Earth.8 In practice, this is still not feasible for obtaining the fair sampling of the impactors because: (1) precise dating of a large set of lunar craters is beyond our present capacities9 and (2) most of the large lunar craters are known to originate in a highly special epoch of the so-called Late Heavy Bombardment,(15,16) ca. 4.0–3.8 billion years B.P., thus strongly skewing any attempt to plot the empirical distribution function of impacts for “normal” times. In practice, in the current debates about the rates of cometary and asteroidal impacts, it is the terrestrial cratering rates that are used as an argument for or against the existence of a dark impactor population,(17−21) thus offering a good case on which the anthropic model bias can, at least potentially,

be tested.10 The amount of bias of the cratering record, in principle, can be decreased through extrapolation from the smaller sizes and comparing such extrapolation with the size-frequency distribution on other solar system bodies, which could be obtained without the need for technically unfeasible measurements of the age of craters. In practice, however, not only is it unclear where the extrapolation should start—since we know little about contingencies of biological evolution leading to the emergence of observers—but the size-frequency distribution expresses only temporal averages of the relevant relationships (between velocities, angles, sizes, and consistencies of impactors vs. crater size). The loss of information in averaging is important if the impactor population may significantly vary in time.

Distribution frequencies of large cosmic explosions (supernovae and gamma-ray bursts) are also inferred—albeit much less confidently—from observations of distant regions: external galaxies similar to the Milky Way. This external evidence decreases the anthropic bias affecting probability estimates of extinction-level supernovae/gamma-ray bursts events. The degree of importance of these explosive processes for the emergence and evolution of life has been the subject of considerable research in recent decades.(22−32) Fragmentary geochemical traces of such events in the past could be found in the terrestrial record, especially ice cores.(33) The same applies to a lesser degree to giant solar flares.(34)

Supervolcanic episodes are perhaps the best example of global terrestrial catastrophes. They are interesting for two recently discovered reasons: (1) supervolcanism has been suggested as a likely causative agent that triggered the end-Permian mass extinction (251.4 ± 0.7 Myr B.P.), killing up to 96% of the terrestrial nonbacterial species.(35,36) (2) Supervolcanism is perhaps the single almost-realized existential catastrophe: the Toba supererruption (Sumatra, Indonesia, 74,000 B.P.) conceivably reduced human population to ∼1,000 individuals, nearly causing the extinction of humanity.(9,37) In that light, we would do well to consider seriously this threat, which despite well-known calamities like Santorini, Pompeii, and Tambora, has become an object of concern only recently.(38,39,3)

Other rare physical disasters might be caused by close passages of normal stars,(11) or by exotic objects, like neutron stars or black holes. If we knew nothing about astronomy, we could not accurately estimate the probability that Earth will be destroyed in a collision with a black hole tomorrow, even if we possessed complete knowledge of the Earth's history. But because we have some knowledge of the solar neighborhood in the Milky Way and the mass function of stellar objects, and because this knowledge is not based on terrestrial evidence, our estimate of these risks will not be appreciably afflicted by anthropic bias.

Unlike for some natural hazards, it is generally difficult to derive information about anthropogenic hazards through statistical analysis of deep history. One exception is the possibility of a catastrophic quantum field process, which may (speculatively) occur naturally, but may conceivably also be caused by high-energy physics experiments, such as those conducted in particle accelerators. This risk is discussed below.

Nick Bostrom, Anders Sandberg, and Tom Douglas (28 Feb 2013), "The Unilateralist's Curse: The case for a principle of conformity."

Future of Humanity Institute, Oxford University.
http://www.nickbostrom.com/papers/unilateralist.pdf

Abstract

In some situations a number of agents each have the ability to undertake an initiative that would have significant effects on the others. Suppose that each of these agents is purely motivated by an altruistic concern for the common good. We show that if each agent acts on her own personal judgment as to whether the initiative should be undertaken, then the initiative will move forward more often than is optimal. We suggest that this phenomenon, which we call the unilateralist's curse, arises in many contexts, including some that are important for public policy. To lift the curse, we propose a principle of conformity, which would discourage unilateralist action. We consider three different models for how this principle could be implemented, and respond to some objections that could be raised against it.

Excerpts

1. Introduction

Consider the following hypothetical scenarios:

1. A group of scientists working on the development of an HIV vaccine have accidentally created an airborne transmissible variant of HIV. They must decide whether to publish their discovery, knowing that it might be used to create a devastating biological weapon, but also that it could help those who hope to develop defenses against such weapons. Most members of the group think publication is too risky, but one disagrees. He mentions the discovery at a conference, and soon the details are widely known.
2. A sports team is planning a surprise birthday party for its coach. One of the players decides that it would be more fun to tell the coach in advance about the planned event. Although the other players think it would be better to keep it a surprise, the unilateralist lets word slip about the preparations underway.
3. Geoengineering techniques have developed to the point that it is possible for any of the world's twenty most technologically advanced nations to substantially reduce the earth's average temperature by emitting sulfate aerosols. Each of these nations separately considers whether to release such aerosols. Nineteen decide against, but one nation estimates that the benefits of lowering temperature would exceed the costs. It presses ahead with its sulfate aerosol program and the global average temperature drops by almost 1 degree.

It is plausible that, in each of these cases, each of a number of agents is in a position to undertake an initiative, X. Each agent decides whether or not to undertake X on the basis of her own independent judgment of the value of X, where the value of X is assumed to be independent of who undertakes X, and is supposed to be determined by the contribution of X to the common good.1 Each agent's judgment is subject to error—some agents might overestimate the value of X, others might underestimate it. If the true value of X is negative, then the larger the number of agents, the greater the chances that at least one agent will overestimate X sufficiently to make the value of X seem positive. Thus, if agents act unilaterally, the initiative is too likely to be undertaken, and if such scenarios repeat, an excessively large number of initiatives are likely to be undertaken. We shall call this phenomenon the unilateralist's curse.

Though we have chosen to introduce the unilateralist's curse with hypothetical examples, it is not merely a hypothetical problem. There are numerous historical examples, ranging from the mundane to the high-tech. Here is one:

> Until the late 1970s, the mechanism of the hydrogen bomb was one of the world's best kept scientific secrets: it is thought that only four governments were in possession of it, each having decided not to divulge it. But staff at the Progressive magazine believed that nuclear secrecy was fuelling the Cold War by enabling nuclear policy to be determined by a security elite

> without proper public scrutiny. They pieced together the mechanism of the bomb and published it in their magazine, arguing that the cost, in the form of aiding countries such as India, Pakistan and South Africa in acquiring hydrogen bombs, was outweighed by the benefits of undermining nuclear secrecy.2

It is perhaps too soon to say whether this was the wrong decision. But in other cases, it is clearer that unilateral action led to a suboptimal outcome:

> In the mid-nineteenth century there were virtually no wild rabbits in Australia, though many were in a position to introduce them. In 1859, Thomas Austin, a wealthy grazier, took it upon himself to do so. He had a dozen or two European rabbits imported from England and is reported to have said that "The introduction of a few rabbits could do little harm and might provide a touch of home, in addition to a spot of hunting."3 However, the rabbit population grew dramatically, and rabbits quickly became Australia's most reviled pests, destroying large swathes of agricultural land.4

5. Concluding thoughts

We have described a moral analogue of the winner's curse. The unilateralist's curse arises when each of a group of agents can, regardless of the opposition of others, undertake or spoil an initiative that has significant effects on others. In such cases, if each agent decides whether to undertake (or spoil) the initiative based on his own independent naive assessment of its value, there will be a group-level bias towards undertaking (spoiling) the initiative. Importantly, this effect arises even if all the agents are assumed to be motivated solely by concern for the common good.

We proposed a principle—the principle of conformity—which instructs agents faced with a unilateralist situation to reduce their likelihood of unilaterally undertaking (or spoiling) the initiative. We then outlined three models for accomplishing this. They involved, respectively, (1) sharing information and reasoning before forming one's evaluation of the initiative, (2) adjusting one's evaluation in the light of the curse, and (3) deferring to the group in making one's decision.

As we acknowledged in the previous section, there may be considerations that militate against the principle of conformity. For example, if there is already a group-level bias against unilateralism, then compliance with the principle would exacerbate this bias. However, we maintain that there is a prima facie case for complying with the principle. Moreover, since the level of bias due to such other factors towards or against unilateralism presumably varies across different contexts, it is likely that there will be some contexts in which the prima facie case for complying with the principle will be decisive. Those will be the contexts in which the group-level bias due to the unilateralist's curse is greater than the any countervailing bias against unilateralism.

It is also possible that, at least within the domain of science, the principle of conformity is more relevant today than it was in, say, Galileo's time. At that time, there was, plausibly, a strong bias against thinking and acting independently in intellectual matters, at least where this would involve diverging from the views of the Church. Since the Enlightenment, however, there may have been a significant weakening of this bias. Independence of thought and action is now more widely regarded as a virtue in scientists and other intellectuals. Honors and prizes are won based on claims to originality and precedence. There may now be no bias, or only a weak bias, against unilateralism in science. Thus, the risk posed by the unilateralist curse in scientific contexts may be greater now than ever.

To resist the unilateralists' curse one first has to become aware of when one is in a curse situation. We hope this paper will help achieve that.

Stuart Armstrong (14 Dec 2012), "Nash equilibrium of identical agents facing the Unilateralist's Curse."

Technical Report #2012-3, Future of Humanity Institute, Oxford University.
http://www.fhi.ox.ac.uk/nash-equilibrium-unilaterialists-curse.pdf

Abstract

This paper is an addendum to the `Unilateralist's Curse' paper of Nick Bostrom, Thomas Douglas and Anders Sandberg [BDS12]. It demonstrates that if there are identical agents facing a situation where any one of them can implement a policy unilaterally, then the best strategies they can implement are also Nash equilibriums. It also notes that if this Nash equilibrium involves probabilistic reactions to observations, then it is a weak Nash equilibrium and a single agent is free to change all their non-trivial probabilistic decisions, without changing the expected utility of the outcome.

Introduction

The Unilateralist's Curse paper analyses how to make decisions when there is a certain policy under consideration, and many di
erent agents who could each unilaterally implement that policy. If each agent simply followed their own estimate's of the value of that policy, we would be in a situation similar to the winner's curse in auctions: the policy would get implemented if the most optimistic agent thought it was a good idea. Thus in these situations, agents must take care to construct a decision process that counteracts this e ect and makes the agents less likely to go ahead on personal, marginally optimistic, information. The problem is isomorphic, in reverse, to policies that require unanimity: there the policy's implementation is dictated by the opinion of the most pessimistic agent.

This paper looks at a specific simplified version of this problem. It assumes that all the agents have identical preferences (they judge each outcome as equally good or equally bad), that they are equally likely to see any given piece of evidence about the value of the policy, and that they can't communicate. They will attempt to construct the best (probabilistic) strategy they can, given these constraints. Because they are identical, they will all construct the same probabilistic strategy. This paper demonstrates that if this is indeed the best strategy (or even a local maxima), then it is a Nash equilibrium: it cannot be improved by unilateral changes by a single agent.

If the strategy is probabilistic (given certain observations, the agent is neither entirely certain to implement the policy, nor entirely certain to refrain), then it is a weak Nash equilibrium – a single agent can change their strategy without making the situation worse. Indeed, a single agent can change all the non-trivial probabilities in their strategy (those neither zero nor one), without changing the expected utility at all.

Bruce Tonn and Dorian Stiefel (Oct 2013), "Evaluating methods for estimating existential risks."

Risk Analysis 33.
http://onlinelibrary.wiley.com/doi/10.1111/risa.12039/full

Abstract

Researchers and commissions contend that the risk of human extinction is high, but none of these estimates have been based upon a rigorous methodology suitable for estimating existential risks. This article evaluates several methods that could be used to estimate the probability of human extinction.

Traditional methods evaluated include: simple elicitation; whole evidence Bayesian; evidential reasoning using imprecise probabilities; and Bayesian networks. Three innovative methods are also considered: influence modeling based on environmental scans; simple elicitation using extinction scenarios as anchors; and computationally intensive possible-worlds modeling. Evaluation criteria include: level of effort required by the probability assessors; level of effort needed to implement the method; ability of each method to model the human extinction event; ability to incorporate scientific estimates of contributory events; transparency of the inputs and outputs; acceptability to the academic community (e.g., with respect to intellectual soundness, familiarity, verisimilitude); credibility and utility of the outputs of the method to the policy community; difficulty of communicating the method's processes and outputs to non-experts; and accuracy in other contexts. The article concludes by recommending that researchers assess the risks of human extinction by combining these methods.

Fig. 1. Existential risks by source and interaction with prevention/adaptation options. [see article for this useful chart]

Concluding thoughts

Estimating existential risks is daunting. This article presents seven methods for meeting this challenge. The methods range from the simplest approach, in which experts and non-experts are asked for their direct subjective probability assessments, to the most complex, the comprehensive possible-worlds modeling approach. Other methods are built upon classic probability methods, the non-classical probability methods utilized in evidential reasoning, non-probalistic methods utilized by the environmental scanning approach, and traditional scenario development. Each method has its strengths and weaknesses.

For the simpler approaches, one needs to carefully consider: Who should be the probability assessors? How do we find consensus amongst the assessors? What events and/or pieces of evidence are needed to yield defensible results? Similar concerns surround the environmental scanning and scenario approaches: How many components are needed in the former's framework? How many scenarios, written by whom, are needed to provide a sound basis for indicative reasoning?

The methods we present above are not applicable only to the study of existential risk. Indeed, they can be used to study catastrophic risks that stop short of human extinction as well as many other risk-related topics. Additionally, our focus on existential risk does not reflect our lack of concern about global catastrophic risks that could kill many millions of humans. Our focus on existential risk is motivated by our ultimate goals of estimating the risk of human extinction, comparing this risk to ethical standards (e.g., 10−20), and then, as appropriate, exploring and advocating appropriate policy responses to risk management.

Owen Cotton-Barratt and Toby Ord (9 Jan 2015), "Existential risk and existential hope: Definitions."

Technical Report #2015-1, Future of Humanity Institute, Oxford University.
http://www.fhi.ox.ac.uk/Existential-risk-and-existential-hope.pdf

Abstract
We look at the strengths and weaknesses of two existing definitions of existential risk, and suggest a new definition based on expected value. This leads to a parallel concept: 'existential hope', the chance of something extremely good happening.

4. Existential eucatastrophes and existential hope

If we enter the totalitarian regime and then manage to escape and recover, then we had an existential catastrophe which was balanced out by a subsequent gain in expected value. This kind of event gives us a concept parallel to that of an existential catastrophe:

> Definition (iv): An existential eucatastrophe2 is an event which causes there to be much more expected value after the event than before.

This concept is quite natural. We saw it in the context of escape from a regime which threatened the existence of a prosperous future. Our world has probably already seen at least one existential eucatastrophe: the origin of life. When life first arose, the expected value of the planet's future may have become much bigger. To the extent that they were not inevitable, the rise of multicellular life and intelligence may also have represented existential eucatastrophes.

In general successfully passing any 'great filter'3 is an existential eucatastrophe, since beforehand the probability of passing it is small, so the expected value is much smaller than after the filter is dealt with.

Armed with this concept, we can draw a new lesson. Just as we should strive to avoid existential catastrophes, we should also seek existential eucatastrophes.

In some ways, this isn't a new lesson at all. Under Bostrom's definition we are comparing ourselves to the most optimistic potential we could reach, so failing to achieve a eucatastrophe is itself a catastrophe. However we think more naturally in terms of events than non-events. If life fails to arise on a planet where it might have, it's much clearer to think of a failure to achieve a eucatastrophe than of an existential catastrophe stretching out over the billions of years in which life did not arise.

Just as we tend to talk about the existential risk rather than existential catastrophe, we want to be able to refer to the chance of an existential eucatastrophe; upside risk on a large scale. We could call such a chance an existential hope.

In fact, there are already people following both of the strategies this suggests. Some people are trying to identify and avert specific threats to our future – reducing existential risk. Others are trying to steer us towards a world where we are robustly well-prepared to face whatever obstacles come – they are seeking to increase existential hope.

Seth Baum (Jun 2015), "Risk and resilience for unknown, unquantifiable, systemic, and unlikely/catastrophic threats."

Environment Systems and Decisions 35.
http://link.springer.com/article/10.1007%2Fs10669-015-9551-8
http://sethbaum.com/ac/2015_RiskResilience.html

Abstract

Risk and resilience are important paradigms for analyzing and guiding decisions about uncertain threats. Resilience has sometimes been favored for threats that are unknown, unquantifiable, systemic, and unlikely/catastrophic. This paper addresses the suitability of each paradigm for such threats, finding that they are comparably suitable. Threats are rarely completely unknown or unquantifiable; what limited information is typically available enables the use of both paradigms. Either paradigm can in practice

mishandle systemic or unlikely/catastrophic threats, but this is inadequate implementation of the paradigms, not inadequacy of the paradigms themselves. Three examples are described: (a) Venice in the Black Death plague, (b) artificial intelligence (AI), and (c) extraterrestrials. The Venice example suggests effectiveness for each paradigm for certain unknown, unquantifiable, systemic, and unlikely/catastrophic threats. The AI and extraterrestrials examples suggest how increasing resilience may be less effective, and reducing threat probability may be more effective, for certain threats that are significantly unknown, unquantifiable, and unlikely/catastrophic.

Excerpts

2 Risk and resilience paradigms

The concepts of risk and resilience have each been defined in multiple ways. A prominent definition of risk comes from Kaplan and Garrick (1981), who define risk as the triplet of possible threats, the probabilities of the threats occurring, and the magnitudes of their consequences if they do occur. The risk paradigm thus involves identifying threats, analyzing their probabilities and magnitudes, and seeking means of reducing both probabilities and magnitudes. The risk paradigm sometimes also considers potential gains in addition to potential losses, but this is less common. A prominent definition of resilience comes from the National Academy of Sciences, which defines resilience as ‘‘the ability to prepare and plan for, absorb, recover from, and more successfully adapt to adverse events’’ (NRC 2012: 1). The resilience paradigm thus involves protecting systems from the impacts of threats so as to ensure that critical system functionality is preserved, even if it means adapting other system attributes to the changed circumstances brought by the impacts of the threats.

It has been claimed that the risk paradigm is poorly suited, and the resilience paradigm is well suited, for cases in which four conditions are met:

1. Threats are unknown. For example, Park et al. (2013: 359) write that ‘‘where hazards are unknown, risk analysis is impossible,’’ and additionally that ‘‘Resilience approaches… require preparing for the unexpected, whereas risk analysis proceeds from the premise that hazards are identifiable’’ (emphasis original).

2. Threat probabilities and magnitudes cannot readily be quantified. For example, Park et al. (2013: 359) write that ‘‘even when hazards can be identified, a risk-based approach emphasizes understanding of probabilities of harm that may be unknowable’’ (emphasis original). Linkov et al. (2013a: 10108) write that ‘‘resilience has a broader purview than risk and is essential when risk is incomputable.’’

3. Threats are systemic, targeting multiple specific system components and/or with significant effects on the rest of the system or other connected systems. For example, Linkov et al. (2014a: 408) write ‘‘Unlike risk-based design, which focuses on one component at a time, resilience engineering identifies critical system functionalities that are valuable to stakeholders and society.’’

4. Threats are unlikely/catastrophic. For example, Park et al. (2013: 359) write that ‘‘in some known, lowprobability, high-consequence events… the traditional risk analysis approach has been unsatisfactory.’’ Park et al. (2013: 360, Table I) further write that while risk management aims for ‘‘minimization of probability of failure, albeit with rare catastrophic consequences and long recovery times,’’ resilience aims for ‘‘minimization of consequences of failure, albeit with more frequent failures and rapid recovery times.’’

Because resilience increases the ability of systems to handle disturbances in general, it often protects systems against a range of known and unknown threats. Thus, when threats are not known or cannot readily be characterized or quantified, resilience is argued to be the more suitable paradigm. However, taking a closer look at each of these four conditions shows that the four reasons for favoring resilience are mistaken. The risk paradigm is up to the task when properly implemented, and the resilience paradigm on its own may be inadequate for guiding decisions about protecting systems.

3 Unknown threats

When a threat is completely unknown prior to its occurrence, risk analysis is indeed impossible. An analyst cannot estimate probabilities and magnitudes of something that lies completely outside her imagination. Likewise, she cannot do anything to manage this risk. But this fact does not make resilience any more suitable for such threats. When a threat is completely unknown, the resilience paradigm is as useless as the risk paradigm. A system manager cannot increase the resilience of a system to a threat without knowing something about how the threat would affect the system.

However, for something to be completely unknown, there must be literally zero available information about it. This is an extremely high standard. In practice, it is often possible to identify some information about threats that seem unknown. Such threats are not unknown—they seem unknown but are actually known to at least a minimal, nonzero extent. The nonzero information available about these threats makes it feasible to apply both the risk and resilience paradigms to the threats. For example, Joshi and Lambert (2011) use diversification for the management of unknown risks.

4 Unquantifiable threats

Some threats are known to exist but resist quantification. Their probabilities and/or their magnitudes are deemed unquantifiable. If the threat probabilities and magnitudes actually were unquantifiable, then calculating risk would be impossible, at least assuming that risk is calculated per the standard probability-times-magnitude formulation. Some treatments of risk do not require full quantification, for example the study of Karvetski and Lambert (2012) on risk analysis under deep uncertainty. But the threats are not entirely unquantifiable. Instead, they only seem unquantifiable. The situation here is much like the one about unknown threats. For something to be completely unquantifiable, there must be zero available information about what its quantity might be. As with the unknown, this is an extremely high standard, and one that often does not exist in practice even when it is believed to exist. The partial quantifiability makes it feasible to apply both the risk and resilience paradigms.

5 Systemic threats

Some threats threaten multiple system components or even multiple systems. When risk analysis and risk management only consider one component at a time, they are bound to perform poorly. When attention to resilience prompts analysts and managers to treat threats more systemically, this will often yield better results.

However, it is important to distinguish between the risk and resilience paradigms as they are sometimes practiced and the paradigms as they exist in theory. In theory, both paradigms can handle systemic threats. In some practice, they do. In other practice, they do not. In particular, some risk practice focuses narrowly on components when it should be more systemic. Linkov et al. (2014b: 379) identify this problem in an observation that ‘‘risk assessment has been primarily focusing on the physical domain of the system, while the information, cognitive, and social domains are often ignored.’’ The solution, however, is not to shift from risk to resilience, but to practice risk more systemically—for example, by risk assessment paying attention to the information, cognitive, and social domains.

Risk analysis practice is indeed often not systemic. The problem can be seen, for example, in risk analysis of global catastrophes (Baum et al. 2013). The risk paradigm often leads analysts to think in reductionist, nonsystemic terms. This tendency of risk analysis is unfortunate. To the extent that resilience prompts more systemic thinking, analysts should in many cases use the resilience paradigm. That said, systemic risk analysis and risk management is quite feasible, even if it is not always practiced. In this direction, Haimes (2009a, b) develops and advocates a systems approach to risk.

Unlikely/catastrophic threats

Some threats are unlikely to occur, but if they do occur, the consequences would be catastrophic. The issue here is similar to that for systemic threats. When risk analysis and risk management neglect these threats, they are bound to perform poorly. When attention to resilience prevents these threats from being neglected, better results will often accrue.

The issue here is likewise similar to that for systemic threats, rooted in the distinction between theory and practice. There is nothing inherent to the risk paradigm that requires neglecting unlikely/catastrophic threats. To the contrary, there is a significant literature using the risk paradigm for the analysis and management of such threats, often using the term extreme events (e.g., Bier et al. 1999; Tsang et al. 2002; Zhou et al. 2012), and there is a significant literature using the risk paradigm to argue that these are often the most important threats to address (e.g., Matheny 2007; Posner 2004). The reasoning is straightforward: If risk is calculated as probability times consequence, then low-probability risks can be very important if the probability is sufficiently high. Park et al. (2013: 359) are correct in stating that ‘‘systematic bias in risk analysis… can lead to underestimation or even ignorance of such risks’’ (emphasis added). But when risk analysis neglects these risks, it is an error of practice, not an error of theory.

It is true that risk management practice often neglects unlikely threats even if they are catastrophic. This occurs in the widespread use of de minimis thresholds in risk regulation (Adler 2007). Even when de minimis thresholds are not specified, the risk of unlikely events is often underestimated due to psychological biases (Weber 2006). A similar situation occurs in the dismissal of scientific theories that are perceived as unlikely but, if true, have catastrophic implications (C´ irkovic´ 2012). However, when risk management practice neglects unlikely/catastrophic threats, it can be corrected through better risk management practice without reference to resilience.

Meanwhile, practice of the resilience paradigm can also be accused of neglecting unlikely/catastrophic threats. Indeed, resilience research and practice has traditionally focused on local-scale threats. The highest consequence threats are the global catastrophes, which include natural threats, such as supervolcano eruptions, and human-made threats, such as nuclear war; AI and extraterrestrials can also be counted among these threats. The global catastrophes are only just beginning to be studied in resilience terms and by researchers motivated by the risk paradigm (e.g., Maher and Baum 2013; Baum and Handoh 2014). Of note is an analysis by Jebari (2014) of unknown global catastrophic risks (or existential risks, in terminology of that paper). While this analysis is not framed in terms of resilience, it is in a similar spirit.

Risk Posed by Nuclear Weapons

E.J. Konopinski, C. Marvin, and Edward Teller (1946), "Ignition of the atmosphere with nuclear bombs."

Report LA-602, Los Alamos National Laboratory.
http://www.fas.org/sgp/othergov/doe/lanl/docs1/00329010.pdf

Abstract

It is shown that, whatever the temperature to which a section of the atmosphere may be heated, no self-propagating chain of nuclear reactions is likely to be started. The energy losses to radiation always overcompensate the gains due to the reactions. This is true even with rather extravagant assumptions concerning the reactivity of the nitrogen nuclei of the air. The only disquieting feature is that the "safety factor", i.e. the ratio of losses to gains of energy, decreases rapidly with initial temperature, and descends to a value of only about 1.6 just beyond a 10 MeV temperature. It is impossible to reach such temperatures unless fission bombs or thermonuclear bombs are used which greatly exceed the bombs now under consideration. But even if bombs of the required volume (i.e. greater than 1000 cubic meters) are employed, energy transfer from electrons to light quanta by Compton scattering will provide a further safety factor and will make a chain reaction in air impossible.

Herman Kahn (20 Jan 1960), "The nature and feasibility of war and deterrence."

P-1888-RC, RAND Corporation.
http://www.rand.org/content/dam/rand/pubs/papers/2005/P1888.pdf
[Image .pdf only available.]

Summary

An evaluation of the impact of a thermonuclear war and a description of some of the risks that might cause decisionmakers to weigh the alternatives of whether or not to go to war (namely, genetic problems, postwar medical problems, and long-term recuperation). The kinds of deterrence discussed are (1) deterrence of a direct attack, (2) the use of strategic threats to deter an enemy from engaging in very provocative acts other than a direct attack on the United States, and (3) acts that are deterred because the potential aggressor is afraid that the defender or others will take limited actions, military or nonmilitary, to make the aggression unprofitable.

Page 1 footnote

This paper summarizes, sometimes rather cursorily, some of the points discussed by the author in a forthcoming book, *Thermonuclear War: Three Lectures and Several Suggestions*, to be published by the Princeton University Press lage in 1960.

Philip Quarles (26 Oct 2012), "Herman Kahn on world annihilation (Audio)."

NEH Preservation Project, WNYC.
http://www.wnyc.org/story/191163-herman-kahn

Excerpt from introductory essay

Herman Kahn addresses the members of the Overseas Press Club about "The Likelihood of Nuclear War at Some Point in the 20th Century," proclaiming the outlook is safer and calmer than five years before.

Kahn -- physicist, well-known "futurist," and partial inspiration for Stanley Kubrick's "Dr. Strangelove" -- claims that the biggest change has been in nuclear proliferation estimates. For example, China has tested a nuclear device, but India still seems to be debating whether such an escalation would be worthwhile, even though it has the means and know-how. He argues that other countries with the capability to develop atomic weapons are hesitating, or have decided against it. Kahn also appears surprised that the U.S. government is showing "a higher degree of self-control than expected," e.g., in Vietnam it appears that even if a large number of U.S. troops were in grave danger, the nuclear option would probably not be considered. In general, Kahn argues that most world leaders (particularly the Soviet Union) are recognizing that aggression does not seem to "pay," although this is only a short-term prognostication. In the next 20 years or so he imagines there may well be significant proliferation leading to a myriad of smaller wars, though these wars would not necessarily go nuclear.

In the question segment of the talk, much of the conversation focuses on Vietnam. Kahn enthusiastically endorses bombing the North and makes interesting observations of the structure and composition of the South Vietnamese army. "We can hold Vietnam," he insists but adds that this should be done primarily with Vietnamese troops.

Armed Forces Special Weapons Project, Sandia Base (1957), "Acceptable premature probabilities for nuclear weapons."

http://www-ee.stanford.edu/~hellman/resources/schlosser.pdf

Abstract

This report establishes acceptable premature probabilities for nuclear weapons exposed to the conditions experienced in stockpile-to-target sequences. Utilising data from a study by United States Continental Army Command, Office of Special 'Weapons Developments, wherein major U.S. catastrophes of the past 50 years were analysed and assigned equivalent nuclear yields, the author assumes a stockpile configuration and composition, and by straightforward mathematical methods reaches conclusions and makes recommendations on numbers to be used for future weapon systems designs. Accidents due to random component failure are assumed to be one-tenth of those attributable to human error. Values given in recent military characteristics are tabulated for comparison.

Martin E. Hellman (2014), "Comments on and analysis of 1957 Sandia Report, 'Acceptable military risks from accidental detonation of atomic weapons'."

Stanford University.
http://www-ee.stanford.edu/~hellman/resources/schlosser_meh.pdf

Conclusion

Temporarily forgetting about the sloppy work, and whether the report is correct in allowing the expected number of American deaths per year due to a nuclear failure to be comparable to the number killed in natural disasters, here is what the same basic methodology would predict is an acceptable level of risk for nuclear deterrence failing in two different modes:

A failure of nuclear deterrence which results in a nuclear terrorist incident destroying part of an American city such as New York is likely to kill on the order of 100,000 Americans. Using 500 American deaths per year from natural disasters results in a required annual risk of such a nuclear terrorist event of at most 1 in 200 or 0.5%. Equivalently, we would have to expect such an event not to happen for approximately 200 years. Contrasting this with Henry Kissinger's estimate of 10 years in the Nuclear Tipping Point documentary, or Dr. Richard Garwin's congressional testimony estimating the risk at 10-20% per year, we see that if those experts are right, we need to reduce the risk by more than a factor of 10. While both of those estimates are subjective, the come from men with significant expertise in matters of national security.

Similarly, if nuclear deterrence suffers a complete failure, resulting in an all-out nuclear exchange, and if that is assumed to kill even 100 million Americans (1,000 times as many as in the assumed nuclear terrorist incident), then the risk of that event would have to be 1-in-200,000 per year (1,000 times smaller than for the nuclear terrorist incident). Equivalently, we would have to expect such an event not to happen for approximately 200,000 years. Again, this seems at least an order of magnitude lower than any reasonable estimate of the current level of risk.

Some have argued that a full-scale nuclear war might kill all of humanity, in which case an even lower probability of failure would be required, even using the methodology of the 1957 report. Some have also argued, either on moral grounds or on the cost to unborn generations, that the cost of human extinction cannot be measured on the basis of the expected number of lives lost. Such arguments would predict an even lower acceptable probability of failure. It is instructive to note, however, that even using what some would see as the overly optimistic methodology of the report, the acceptable failure rate is much lower than most people's subjective estimate: When I ask people for an order of magnitude estimate for how long they think nuclear deterrence will work before failing and causing a full-scale nuclear war, the vast majority see 10 years as too short and 1,000 years as too long, leaving 100 years as their order of magnitude, subjective estimate.

Eric Schlosser (2013), *Command and Control: Nuclear Weapons, the Damascus Accident, and the Illusion of Safety*.

Penguin.
http://www.amazon.com/Command-Control-Damascus-Accident-Illusion/dp/1594202273
(See excerpts and author interviews below.)

Michael Mechanic (15 Sep 2013), "A sneak peek at Eric Schlosser's terrifying new book on nuclear weapons."

Mother Jones.
http://www.motherjones.com/print/232731

Excerpt

Update (1/16/2014): The Air Force announced yesterday that it had suspended and revoked the security clearances of 34 missile launch officers at the Malmstrom base in Montana after it came to light that they were cheating — or complicit in cheating — on monthly exams to ensure that they were capable of safely babysitting the nuclear warheads atop their missiles. Eleven launch officers, two of whom were

also involved in the cheating episode, were targeted in a separate investigation of illegal drug use. The revelations were just the latest fiasco in the Air Force's handling of America's nuclear arsenal, which military officials invariably insist is safe. Then again, as Schlosser reveals in his book, they've lied about that before.

ON JANUARY 23, 1961, a B-52 packing a pair of Mark 39 hydrogen bombs suffered a refueling snafu and went into an uncontrolled spin over North Carolina. In the cockpit of the rapidly disintegrating bomber was a lanyard attached to the bomb-release mechanism. Intense G-forces tugged hard at it and unleashed the nukes, which, at four megatons, were 250 times more powerful than the weapon that leveled Hiroshima. One of them "failed safe" and plummeted to the ground unarmed. The other weapon's failsafe mechanisms — the devices designed to prevent an accidental detonation — were subverted one by one, as Eric Schlosser recounts in his new book, Command and Control:

When the lanyard was pulled, the locking pins were removed from one of the bombs. The Mark 39 fell from the plane. The arming wires were yanked out, and the bomb responded as though it had been deliberately released by the crew above a target. The pulse generator activated the low-voltage thermal batteries. The drogue parachute opened, and then the main chute. The barometric switches closed. The timer ran out, activating the high-voltage thermal batteries. The bomb hit the ground, and the piezoelectric crystals inside the nose crushed. They sent a firing signal...

Unable to deny that two of its bombs had fallen from the sky — one in a swampy meadow, the other in a field near Faro, North Carolina — the Air Force insisted that there had never been any danger of a nuclear detonation. This was a lie.

Michael Mechanic (15 Sep 2013), "Eric Schlosser: If we don't slash our nukes, 'a major city is going to be destroyed'."

Mother Jones.
http://www.motherjones.com/politics/2013/09/interview-eric-schlosser-command-control-nuclear-weapons-accidents

Excerpt

ES: Throughout the '50s and '60s, it was almost boilerplate for Defense Department officials to say that during an accident there was no possibility of a nuclear detonation, while privately, at the weapons laboratories, there were physicists and engineers who were extremely worried and were well aware that we had come close to having it happen on American soil. If you look at the official list of broken arrows that the Pentagon released in the '80s, it includes 32 serious accidents involving nuclear weapons that might have threatened the public safety. The list is entirely arbitrary: Some of those accidents didn't even involve weapons that had a nuclear core, so they never could have detonated. But many, many serious accidents aren't on that list.

One document I got through a Freedom of Information Act request listed more than 1,000 weapons involved in accidents, some of them trivial and some of them not trivial. There's somebody who worked at the Pentagon who has read this book, and one of his criticisms was that I'm so hard on the Air Force — he said that there were a great number of accidents involving Army weapons that I don't write about.

You know, it's very difficult to get this information. I did the best that I could, but I have no doubt that there are other incidents and accidents that still have not been reported, so I can't blame the mainstream media so much as blame this national security apparatus. Again and again I would see by comparing documents that what was being redacted wasn't information that would threaten the national security — it was information that would be embarrassing, or put these defense bureaucracies in an unflattering light.

Eric Schlosser (Mar/Apr 2014), "Accidents will happen: An excerpt from 'Command and Control'."

Bulletin of the Atomic Scientists 70.
http://thebulletin.org/accidents-will-happen-excerpt-command-and-control

Excerpt

Two weeks after an accident that could have detonated a hydrogen bomb in Morocco, the Department of Defense and the Atomic Energy Commission issued a joint statement on weapon safety. "In reply to inquiries about hazards which may be involved in the movement of nuclear weapons," they said, "it can be stated with assurance that the possibility of an accidental nuclear explosion... is so remote as to be negligible."

Eric Schlosser (Mar/Apr 2014), "Eric Schlosser: Uncovering nuclear weapons history from the ground up."

Bulletin of the Atomic Scientists 70.
http://bos.sagepub.com/content/70/2/1.full

Abstract

In this interview, author and investigative journalist Eric Schlosser talks with the Bulletin about his recently published book Command and Control. He explains why he decided to tell the history of America's nuclear arsenal "from the bottom up," largely through interviews with ordinary people who were tasked with developing and safely deploying nuclear weapons. Schlosser describes some of the safety issues that have plagued the nuclear weapons program, and he expresses frustration that many government documents exposing these very issues — some dating as far back as the Cold War — have not yet been made public. He recommends increased spending on training and maintenance of aging nuclear weapons, but says that the main purpose of his book is not to push a particular policy but rather to encourage public debate about nuclear weapons and to raise questions about their current military purpose.

Seth D. Baum (30 Mar 2015), "Confronting the threat of nuclear winter."

Futures (in press).
http://www.sciencedirect.com/science/article/pii/S0016328715000403

Abstract

Large-scale nuclear war sends large quantities of smoke into the stratosphere, causing severe global environmental effects including surface temperature declines and increased ultraviolet radiation. The temperature decline and the full set of environmental effects are known as nuclear winter. This paper surveys the range of actions that can confront the threat of nuclear winter, both now and in the future. Nuclear winter can be confronted by reducing the probability of nuclear war, reducing the environmental severity of nuclear winter, increasing humanity's resilience to nuclear winter, and through indirect interventions that enhance these other interventions. While some people may be able to help more than others, many people—perhaps everyone across the world—can make a difference. Likewise, the different opportunities available to different people suggests personalized evaluations of nuclear winter, and of catastrophic threats more generally, instead of a one-size-fits-all approach.

Risk Posed by Environmental Catastrophes

P.A. Carpenter and P.C. Bishop (Dec 2009), “A review of previous mass extinctions and historic catastrophic events."

Futures 41.
http://www.sciencedirect.com/science/article/pii/S0016328709001037

Abstract

This paper discusses historical evidence and speculations of the causes of past prehistoric extinctions. It also describes previous catastrophic events and recent species extinctions that serve as a basis for understanding the types of interactions and interwoven events that would be necessary for future human extinction to occur.

Excerpt

4. Conclusion

The strongest likelihood for human extinction lies 5 billion years from now, if humans still exist and have not colonized other planets. That is the point at which our Sun will expand as it nears the end of its life, engulfing the inner planets — including the Earth. The next most likely possibility lies in a rare impact by an asteroid or comet of great enough size to shock the entire planet. Such an impact occurs every 100 million years or so. Beyond these two events, the likelihood of complete extinction drops greatly unless multiple points of failure occur that would clearly make the biosphere unsustainable for continued life.

Dr. Eldredge’s statement about life being resilient and always recovering after a major extinction still holds true [31]. However, if human-induced changes continue to impact biodiversity to the point that the long-term survival of ecosystems is clearly in jeopardy, then human extinction itself might not be far behind. Complete human extinction is very unlikely, but it could occur if enough events happen that cause points of failure such that human life as we know it can no longer be sustained within the biosphere. Based on the issues discussed above, we have prepared a scenario that is presented in a paper entitled The Seventh Mass Extinction: Human-Caused Events Contribute to a Fatal Consequence, which is also included in this human extinction series. In that paper, we show how such extinction might be possible given an unfortunate chain of events. It is our hope that (i) these presentations will influence leaders of society to think about the strategic consequences of decisions that have been made in the past and are yet to be made in the future and (ii) that foresight will be used in planning and analysis within a holistic strategy, ensuring that no such events ever occur.

Seth Baum, Timothy Maher, Jr., and Jacob Haqq-Misra (2013), "Double catastrophe: Intermittent stratospheric geoengineering induced by societal collapse."

Environment, Systems and Decisions 33.
http://sethbaum.com/ac/2013_DoubleCatastrophe.pdf

Abstract

Perceived failure to reduce greenhouse gas emissions has prompted interest in avoiding the harms of climate change via geoengineering, that is, the intentional manipulation of Earth system processes.

Perhaps, the most promising geoengineering technique is stratospheric aerosol injection (SAI), which reflects incoming solar radiation, thereby lowering surface temperatures. This paper analyzes a scenario in which SAI brings great harm on its own. The scenario is based on the issue of SAI intermittency, in which aerosol injection is halted, sending temperatures rapidly back toward where they would have been without SAI. The rapid temperature increase could be quite damaging, which in turn creates a strong incentive to avoid intermittency. In the scenario, a catastrophic societal collapse eliminates society's ability to continue SAI, despite the incentive. The collapse could be caused by a pandemic, nuclear war, or other global catastrophe. The ensuing intermittency hits a population that is already vulnerable from the initial collapse, making for a double catastrophe. While the outcomes of the double catastrophe are difficult to predict, plausible worst-case scenarios include human extinction. The decision to implement SAI is found to depend on whether global catastrophe is more likely from double catastrophe or from climate change alone. The SAI double catastrophe scenario also strengthens arguments for greenhouse gas emissions reductions and against SAI, as well as for building communities that could be self-sufficient during global catastrophes. Finally, the paper demonstrates the value of integrative, systems-based global catastrophic risk analysis.

Non-Technical Summary

Background: Global Catastrophic Risk Systems Analysis

Global catastrophic risks are risks of events that would significantly harm or even destroy humanity at the global scale, such as climate change, nuclear war, and pandemics. To date, most research on global catastrophes analyzes one risk at a time. A better approach uses systems analysis to capture the many important interactions between risks. This paper analyzes a global catastrophe scenario involving climate change, geoengineering, and another catastrophe. We call the scenario "double catastrophe".

Climate Change & Stratospheric Geoengineering

The rising temperatures of global climate change pose great risks to humanity and ecosystems. Climate change can be slowed by reducing emissions of greenhouse gases like carbon dioxide and methane. But humanity has been struggling to reduce emissions. One alternative is geoengineering, the intentional manipulation of Earth systems. The most promising geoengineering option may be stratospheric geoengineering, in which aerosol particles are put into the stratosphere. The particles block sunlight, lowering temperatures on Earth's surface.

Intermittency & Double Catastrophe

One problem with stratospheric geoengineering, known as intermittency, is that the particles must be continuously replaced in the stratosphere. If they're not, then in a few years they fall out, and temperatures rapidly rise back to where they would have been without the geoengineering. The rapid temperature increase would be very damaging to society. Because of this, society is unlikely to let intermittency occur - unless some other catastrophe occurs, knocking out society's ability to continue the geoengineering. Then, the rapid temperature increase hits a population already vulnerable from the initial catastrophe. This double catastrophe could be a major global catastrophe.

Implications For Decision Making

Because of how damaging global catastrophes would be to human civilization, decision making is often oriented towards minimizing the risk of global catastrophe. Stratospheric geoengineering can prevent global catastrophe from climate change alone, but it can also lead to global catastrophe from the double catastrophe scenario. If global catastrophe is more likely from climate change alone, then society should

decide to implement stratospheric geoengineering. Otherwise, society is better off without stratospheric geoengineering. This assumes (among other things) that the goal should be minimizing global catastrophic risk and that stratospheric geoengineering is the best form of geoengineering.

Amy Donovan and Clive Oppenheimer (May 2014), "Extreme volcanism; Disaster risks and societal implications."

In *Extreme Natural Hazards, Disaster Risks and Societal Implications*, ed. Alik Ismail-Zadeh, et al. Cambridge Books Online.
http://ebooks.cambridge.org.ezproxy.cul.columbia.edu/ebook.jsf?bid=CBO9781139523905

Conclusion

Large magnitude volcanic eruptions have occurred in the past with significant impacts on climate, societies, and landscapes. Such eruptions will occur in the future, and the probability of a magnitude 7 (or even 8) eruption occurring in the next century is finite (Oppenheimer, 2011). We have identified and discussed three potential scenarios for such eruptions, and have indicated a number of global hotspots for high impacts. A key challenge for scientists is intensive data gathering and collation concerning the many unstudied volcanic eruptions around the world in addition to impact assessment and disaster planning.

This brings us to the crux of the challenge of volcano catastrophe risk management: what international institutional models and organisations are relevant to addressing the threat? In fact, there are numerous existing platforms to consider and prior experience on which to draw. They include UNESCO's Intergovernmental Oceanographic Commission (International Oceanographic Commission / UNESCO, 2009), which promotes (among other things) regional tsunami warning networks, development of national disaster plans, community awareness programmes, and evacuation drills. Another is the World Meteorological Organization (WMO, also an agency of the United Nations). Reading its 'vision and mission' statement it is easy to replace 'meteorological' by 'volcanological', and instructive, therefore, to reproduce it here (http://www.wmo.int/pages/about/mission_en.html):

The vision of WMO is to provide world leadership in expertise and international cooperation in weather, climate, hydrology and water resources and related environmental issues and thereby contribute to the safety and well-being of people throughout the world and to the economic benefit of all nations.

We might borrow further from the mission statement of the WMO, substituting 'meteorology' with 'volcanology', so as to propose the following remit, of a global volcanological organisation, which would be to:

Facilitate worldwide cooperation in the establishment of networks of stations for the making of volcanological observations as well as hydrological and other geophysical observations related to volcanology, and to promote the establishment and maintenance of centres charged with the provision of volcanological and related services;

Promote the establishment and maintenance of systems for the rapid exchange of volcanological and related information; Promote standardization of volcanological and related observations and to ensure the uniform publication of observations and statistics;

Further the application of volcanology to aviation, shipping, water problems, agriculture and other human activities;

Promote activities in operational volcanology and to further close cooperation between Meteorological and Hydrological Services;

Encourage research and training in volcanology and, as appropriate, in related fields, and to assist in coordinating the international aspects of such research and training.

...Many communities today tolerate proximity to geophysical hazards: the perceived benefits outweigh often incalculable longer-term risks, and uprooting an individual family let alone a whole community or megacity is traumatic. Thus, the many co-locations of people, flood plains, fault lines, and/or volcanoes (Tokyo, Los Angeles, Naples, Istanbul, and Port-au-Prince) pose considerable challenges. As the global human population heads towards ten billion, an increasing proportion of whom will likely live in poverty in cities and near coasts, it is undeniable that future large and very large volcanic eruptions pose vital management challenges for national governments and the global community. A major problem in risk management is that most extreme possible scenarios become a kind of science fiction. On the other hand, considering only the 'worst' probable eruptions, earthquakes, tsunami, and so on may fail to prepare us for more frequent events.

David C. Denkenberger and Joshua M. Pearce (29 Nov 2014), "Feeding everyone: Solving the food crisis in event of global catastrophes that kill crops or obscure the sun."

Futures (in press).
http://www.sciencedirect.com/science/article/pii/S0016328714001931

Abstract

Mass human starvation is currently likely if global agricultural production is dramatically reduced for several years following a global catastrophe, e.g. super volcanic eruption, asteroid or comet impact, nuclear winter, abrupt climate change, super weed, extirpating crop pathogen, super bacterium, or super crop pest. This study summarizes the severity and probabilities of such scenarios, and provides an order of magnitude technical analysis comparing caloric requirements of all humans for 5 years with conversion of existing vegetation and fossil fuels to edible food. Here we present mechanisms for global-scale conversion including natural gas-digesting bacteria, extracting food from leaves, and conversion of fiber by enzymes, mushroom or bacteria growth, or a two-step process involving partial decomposition of fiber by fungi and/or bacteria and feeding them to animals such as beetles, ruminants (cattle, sheep, etc.), rats and chickens. We perform an analysis to determine the ramp rates for each option and the results show that careful planning and global cooperation could maintain humanity and the bulk of biodiversity.

Seth Baum (2015), "Winter-safe deterrence: The risk of nuclear winter and its challenge to deterrence."

Contemporary Security Policy 36.
http://www.tandfonline.com/doi/abs/10.1080/13523260.2015.1012346

Abstract

A new line of nuclear winter research shows that even small, regional nuclear wars could have catastrophic global consequences. However, major disarmament to avoid nuclear winter goes against the reasons nuclear weapon states have for keeping their weapons in the first place, in particular deterrence. To reconcile these conflicting aims, this paper develops the concept of winter-safe deterrence, defined as military force capable of meeting the deterrence goals of today's nuclear weapon states without risking catastrophic nuclear winter. This paper analyses nuclear winter risk, finding a winter-safe limit of about 50 nuclear weapons total worldwide. This paper then evaluates a variety of candidate weapons for winter-safe deterrence. Non-contagious biological weapons (such as anthrax or ricin), neutron bombs detonated at altitude, and nuclear electromagnetic weapons show the most promise. Each weapon has downsides, and the paper's analysis is only tentative, but winter-safe deterrence does appear both feasible and desirable given the urgency of nuclear winter risk.

Hsi-Hua Huang, et al. (15 May 2015), "The Yellowstone magmatic system from the mantle plume to the upper crust."

Science 348.
http://www.sciencemag.org/content/348/6236/773.abstract
https://www.iris.edu/hq/files/workshops/2015/06/earthscope_national_meeting_2015/abstracts/earthscope2015_abstract.pdf

Abstract

The Yellowstone supervolcano is one of the largest active continental silicic volcanic fields in the world. An understanding of its properties is key to enhancing our knowledge of volcanic mechanisms and corresponding risk. Using a joint local and teleseismic earthquake P-wave seismic inversion, we revealed a basaltic lower-crustal magma body that provides a magmatic link between the Yellowstone mantle plume and the previously imaged upper-crustal magma reservoir. This lower-crustal magma body has a volume of 46,000 cubic kilometers, ~4.5 times that of the upper-crustal magma reservoir, and contains a melt fraction of ~2%. These estimates are critical to understanding the evolution of bimodal basaltic-rhyolitic volcanism, explaining the magnitude of CO2 discharge, and constraining dynamic models of the magmatic system for volcanic hazard assessment.

Karim Jebari (Jun 2015), "Existential risks: Exploring a robust risk reduction strategy."

Science and Engineering Ethics 21.
http://link.springer.com.ezproxy.cul.columbia.edu/article/10.1007%2Fs11948-014-9559-3

Abstract

A small but growing number of studies have aimed to understand, assess and reduce existential risks, or risks that threaten the continued existence of mankind. However, most attention has been focused on known and tangible risks. This paper proposes a heuristic for reducing the risk of black swan extinction events. These events are, as the name suggests, stochastic and unforeseen when they happen. Decision theory based on a fixed model of possible outcomes cannot properly deal with this kind of event. Neither can probabilistic risk analysis. This paper will argue that the approach that is referred to as engineering safety could be applied to reducing the risk from black swan extinction events. It will also propose a conceptual sketch of how such a strategy may be implemented: isolated, self-sufficient, and continuously manned underground refuges. Some characteristics of such refuges are also described, in particular the psychosocial aspects. Furthermore, it is argued that this implementation of the engineering safety strategy safety barriers would be effective and plausible and could reduce the risk of an extinction event in a wide range of possible (known and unknown) scenarios. Considering the staggering opportunity cost of an existential catastrophe, such strategies ought to be explored more vigorously.

Conclusion

The notion of black swan extinction events present us with a daunting task. How to even start thinking about risks that are unknown? The stakes are further raised when considering that, on a large number of normative theories, an existential catastrophe implies a staggering loss of value. Thus, it is unwise to ignore the risk such an event represents. In engineering safety, a number of heuristics and strategies are device to prevent a catastrophic failure in a large number of possible scenarios. These strategies could be employed in thinking about how to reduce the risk of a black swan extinction event. Safety barriers are an instance of such a strategy. These could be actual physical barriers in some systems, or subsystems that prevent catastrophic failure by compartmentalization and physical separation. This article has discussed an example implementation of this strategy: isolated, continuously manned and self-sufficient underground refuges that could protect a large enough number of people to ensure the continued existence of mankind. While building such a "doomsday shelter" is less glamorous than colonizing the Moon, it may give us much more risk reduction for the money invested. The conceptual sketch of the project in this paper should be further developed in an interdisciplinary research project, which could benefit from the extensive literature on isolated, self-containing habitats. Architecture, engineering, social psychology and decision theory would probably be needed to fully assess the costs, and social and technological challenges.

Risk Posed by SETI

David Brin (1983), "The 'great silence': The controversy concerning extraterrestrial intelligent life."
Quarterly Journal of the Royal Astronomical Society 24.
http://articles.adsabs.harvard.edu/cgi-bin/nph-iarticle_query?1983QJRAS..24..283B&defaultprint=YES&filetype=.pdf

Abstract

Recent discussions concerning the likelihood of encountering extraterrestrial technological civilizations have run into an apparent paradox. If, as many now contend, interstellar exploration and settlement is possible at non-relativistic speeds, then reasonable calculations suggest that space-faring species, or their machine surrogates, should pervade the Galaxy. The apparent absence of evidence for extraterrestrial civilizations, herein called 'the Great Silence' places severe burdens on present models. Many of the current difficulties are due to inadequate exploration of the parameters of the problem. A review of the topic shows that present approaches may be simplistic.

Stuart Armstrong and Anders Sandberg (12 Mar 2013), "Eternity in six hours: Intergalactic spreading of intelligent life and sharpening the Fermi paradox."
Acta Astronautica.
http://www.fhi.ox.ac.uk/intergalactic-spreading.pdf

Abstract

The Fermi paradox is the discrepancy between the strong likelihood of alien intelligent life emerging (under a wide variety of assumptions), and the absence of any visible evidence for such emergence. In this paper, we extend the Fermi paradox to not only life in this galaxy, but to other galaxies as well. We do this by demonstrating that traveling between galaxies -- indeed even launching a colonisation project for the entire reachable universe -- is a relatively simple task for a star-spanning civilization, requiring modest amounts of energy and resources. We start by demonstrating that humanity itself could likely accomplish such a colonisation project in the foreseeable future, should we want to, and then demonstrate that there are millions of galaxies that could have reached us by now, using similar methods. This results in a considerable sharpening of the Fermi paradox.

Risk Posed by Synthetic Biology

Jonathan Tucker and Raymond Zilinskas (Spring 2006), "The promise and perils of synthetic biology."
The New Atlantis.
http://www.grid.unep.ch/FP2011/step1/pdf/028_syntheticBiology_references.pdf/028_Tucker_2006.pdf

Excerpt

[W]e can use history as a guide—particularly the history of recombinant DNA technology—to discern three main areas of risk in synthetic biology. First, synthetic microorganisms might escape from a research laboratory or containment facility, proliferate out of control, and cause environmental damage or threaten public health. Second, a synthetic microorganism developed for some applied purpose might cause harmful side effects after being deliberately released into the open environment. Third, outlaw states, terrorist organizations, or individuals might exploit synthetic biology for hostile or malicious purposes.

The Risk of Accidental Release

…[I]t would be prudent to adopt the "precautionary principle" and treat synthetic microorganisms as dangerous until proven harmless. According to this approach, all organisms containing assemblies of BioBricks would have to be studied under a high level of biocontainment (Biosafety Level 3 or even 4) until their safety could be demonstrated in a definitive manner. As George Church argued in Nature in 2005, "Learning from gene therapy, we should imagine worst-case scenarios and protect against them. For example, full physical isolation and confined lab experiments on human and agricultural pathogens should continue until we have data on a greater number of potential consequences—ecological and medical—of engineering such systems."

The Risk of Testing in the Open Environment

…Theoretically, three types of negative effects could result from releasing a synthetic microorganism into the environment. First, the organism could disrupt local biota or fauna through competition or infection that, in the worst case, could lead to the extinction of one or more wild species. Second, once a synthetic organism has successfully colonized a locale, it might become endemic and thus impossible to eliminate. Third, the synthetic organism might damage or disrupt some aspect of the habitat into which it was introduced, upsetting the natural balance and leading to the degradation or destruction of the local environment.

The Risk of Deliberate Misuse

…[T]wo possible scenarios for the deliberate misuse of synthetic biology provide some grounds for concern. The first involves a "lone operator," such as a highly trained molecular biologist who develops an obsessive grudge against certain individuals or groups (or society as a whole). If Theodore Kaczynski, the "Unabomber," had been a microbiologist instead of a mathematician, he might have fit this profile; perhaps the perpetrator of the 2001 anthrax-letter attacks does fit it. So-called "lone wolf" terrorists

have proven very innovative and difficult to locate; if armed with a weapon of mass destruction, such a lone operator could cause as much damage as an organized group.

The second scenario of concern is that of a "biohacker," an individual who does not necessarily have malicious intent but seeks to create bioengineered organisms out of curiosity or to demonstrate his technical prowess—a common motivation of many designers of computer viruses. The reagents and tools used in synthetic biology will eventually be converted into commercial kits, making it easier for biohackers to acquire them. Moreover, as synthetic-biology training becomes increasingly available to students at the college and possibly even high-school levels, a "hacker culture" may emerge, increasing the risk of reckless or malevolent experimentation.

Emily Singer (30 May 2006), "The dangers of synthetic biology."

MIT Technology Review.
http://www.technologyreview.com/news/405882/the-dangers-of-synthetic-biology/

Excerpt

TR: What issues are you most worried about today?

David Baltimore: The real danger today is from organisms that already exist. The idea of synthesizing something worse than that, of taking bits of Ebola and other viruses to create something more deadly, underertimates how hard it is to survive in the natural world.

Adapting to the human lifestyle is very complicated, so I would guess that we would fail if we tried to engineer a dangerous organism. Ebola, for example, is very pathogenic. It infects families and health workers, but it never spreads widely because it is too lethal – it isn't in the community long enough to spread. Bird flu is not likely to spread widely until it mutates to become less pathogenic.

TR: Among existing organisms, what has the biggest potential for harm?

DB: I think viruses are the major focus of concern. They are relatively simple to make and control and some are quite lethal. Smallpox, for example, is very potent, and we are not protected against it. The smallpox sequence is published, so you could recover it by synthesis if you had the lab facilities to do that. But getting the pieces of DNA to make smallpox is not a backyard experiment. You need a large lab with significant biosafety precautions. I don't see this as something that would happen clandestinely in the U.S., but a well-funded lab outside of this country could do something quite nefarious.

Jonathan Tucker (5 Aug 2011), "Could terrorists exploit synthetic biology?"

The New Atlantis
http://www.thenewatlantis.com/publications/could-terrorists-exploit-synthetic-biology
http://www.thenewatlantis.com/docLib/20110805_TNA31Tucker.pdf

How Great Are the Risks?

In addition to the potential benefits of de-skilling and open access, a number of commentators have warned that the democratization of synthetic biology could give rise to new safety and security risks. One concern is that substantially expanding the pool of individuals with access to synthetic-biology techniques would inevitably increase the likelihood of accidents, creating unprecedented hazards for the environment and public health.[32] Even Dyson's generally upbeat article acknowledges that the recreational use of synthetic biology "will be messy and possibly dangerous" and that "rules and regulations will be needed to make sure that our kids do not endanger themselves and others."

Beyond the possible safety risks, Mukunda, Oye, and Mohr warn that the de-skilling of synthetic biology would make this powerful technology accessible to individuals and groups who would use it deliberately to cause harm. "Synthetic biology," they write, "includes, as a principal part of its agenda, a sustained, well-funded assault on the necessity of tacit knowledge in bioengineering and thus on one of the most important current barriers to the production of biological weapons."[33] Drawing on the precedent of "black-hatted" computer hackers, who create software viruses, worms, and other malware for criminal purposes, for espionage, or simply to demonstrate their technical prowess, some have predicted the emergence of "bio-hackers" who engage in reckless or malicious experiments with synthetic organisms in basement laboratories.[34] Such nightmare scenarios are probably exaggerated, however, because the effective use of synthetic biology techniques relies on socio-technical resources that are not generally available to hobbyists. According to Andrew Ellington, a biochemistry professor at the University of Texas, "There is no 'Radio Shack' for DNA parts, and even if there were, the infrastructure required to manipulate those parts is non-trivial for all but the richest amateur scientist."[35]

Indeed, when assessing the risk of misuse, it is important to distinguish among potential actors that differ greatly in financial assets and technical capabilities — from states with advanced bio-warfare programs, to terrorist organizations of varying size and sophistication, to individuals motivated by ideology or personal grievance. The study of past state-level bio-warfare programs, such as those of the Soviet Union and Iraq, has also shown that the acquisition of biological weapons requires an interdisciplinary team of scientists and engineers who have expertise and tacit knowledge in fields such as microbiology, aerobiology, formulation, and delivery.[36] States are generally more capable of organizing and sustaining such teams than are non-state actors.

Conceivably, the obstacles posed by the need for personal and communal tacit knowledge might diminish if a terrorist group managed to recruit a group of scientists with the required types of expertise, and either bribed or coerced them into developing biological weapons. But Vogel and Ben Ouagrham-Gormley counter this argument by noting that even in the unlikely event that terrorists could recruit such a scientific A-team, its members would still face the challenge of adapting the technology to a local context.[37] Dysfunctional group dynamics, such as a refusal by some team members to work together, would also create obstacles to interdisciplinary collaboration in areas requiring communal tacit knowledge.

Taking such factors into account, Michael Levi of the Council on Foreign Relations has questioned the ability of terrorists to construct an improvised nuclear device from stolen fissile materials. He notes that the process of building a functional weapon would involve a complex series of technical steps, all of which the terrorists would have to perform correctly in order to succeed.[38] The same is true of assessing bioterrorism risk: one must examine not only the likelihood of various enabling conditions, but also the probability that all of the steps in the weapon development process will be carried out successfully.

Finally, problem-solving is crucial to the mastery of any complex technology. Biotechnologists must be creative and persistent to overcome the technical difficulties that inevitably arise during the development of a new process. Thus, a key variable affecting the risk that terrorists could exploit synthetic biology for harmful purposes would be their ability to perform multiple iterations of a technique until they get it right, a requirement that presupposes a stable working environment and ample time for experimentation. Such amenities would probably be lacking, however, for individuals working in a covert hideaway or conducting illicit activities (such as the synthesis and weaponization of a deadly virus) in an otherwise legitimate laboratory.

Seth Baum and Grant Wilson (2013), "The ethics of global catastrophic risk from dual-use bioengineering."

Ethics in Biology, Engineering and Medicine 4.
http://sethbaum.com/ac/2013_BioengineeringGCR.pdf

Abstract

Global catastrophic risks (GCRs) are risks of events that could significantly harm or even destroy civilization at the global scale. GCR raises a number of profound ethical issues, with a range of ethical theories suggesting that GCR reduction should be society's top priority. This paper discusses GCR ethics in the context of dual-use bioengineering: bioengineering that can cause either benefit or harm, including increases and decreases in GCR. Advances in bioengineering offer great promise, but also introduce new perils. Key ethical questions include what phenomena hold intrinsic value and how the phenomena are valued across space and time. Another key question is how decisions about bioengineering risks should be made. The global scope of bioengineering and GCR suggests a role for international law. Bioengineering does not fall neatly within existing international regimes such as the Convention on Biological Diversity, Cartagena Protocol, and Biological Weapons Convention. An international regime with comprehensive coverage of bioengineering would help address dual-use bioengineering as it relates to GCR.

Non-Technical Summary

Background: Dual-Use Bioengineering and Global Catastrophic Risk

Dual-use technologies are technologies that can be used in both beneficial and harmful ways. Some technologies produced through biological engineering (bioengineering) are dual-use. Of all the possible harms from dual-use technologies, global catastrophic risk is a significant concern. Global catastrophic risks (GCRs) are risks of events that could significantly harm or even destroy civilization at the global

scale. This paper discusses ethical issues raised by those bioengineered technologies that pose a GCR. The paper also explores how international law can reduce GCR from dual-use bioengineering.

Ethics of Global Catastrophic Risk

Different ethical views will reach different conclusions about the importance of GCR. Some views consider GCR to be very important; others do not. Certain views could even consider global catastrophe to be a good thing. For example, if the main priority is to reduce suffering, then global catastrophes that kill many people would be good because those people would no longer suffer. But ethical views that let the good in life outweigh the bad generally conclude that global catastrophes are bad. Views that value all people equally consider GCR to be especially bad because global catastrophes involve so many people. Indeed, these egalitarian views often find that reducing the risk of global catastrophe is a top priority for humanity.

Benefits and Risks of Bioengineering

Bioengineering has already led to major benefits in fields like medicine, leading to breakthroughs like a vaccine for human papillomavirus and gene therapy techniques to treat diseases, and agriculture, where genetically engineered crops increase yields and possess favorable traits such as drought and herbicide resistance. But dual-use bioengineering has also been used to create weapons and other threats. Some bioengineering technologies also have the potential to influence GCR, either by reducing the risk, increasing it, or both. For example, while a genetically engineered virus could help create vaccines to prevent a pandemic, such a virus could also unintentionally escape from the lab. Or, bioterrorists could use the virus as a blueprint to create an extremely deadly biological weapon.

International Regulation of GCRs from Bioengineering

GCR arising from bioengineering has an inherently international scope. A global catastrophe from bioengineering could impact the entire planet, and bioengineering research and development can be done anywhere in the world. Therefore, international law is an appropriate tool to regulate bioengineering. While some aspects of bioengineering already fall under existing international treaties, these treaties do not sufficiently curtail the GCRs that arise from dual-use bioengineering. One solution is to create a new international legal regime that either regulates bioengineering alone or both bioengineering and other emerging technologies. Other international law options include nonbinding international norms (‘soft law’) and the establishment of an organization dedicated to mitigating GCR from emerging technologies.

Gigi Gronvall (Feb 2015), "Mitigating the risks of synthetic biology."

Council on Foreign Relations.
http://www.cfr.org/health/mitigating-risks-synthetic-biology/p36097

Preventive Options

Complete prevention of the accidental or deliberate misuse of synthetic biology is not possible. However, the level of risk can be reduced and preparations can be made to diminish the consequences.

Regarding accident prevention, laboratory safety can be improved and better enforced, and standards can be promulgated internationally. There is now adequate guidance for laboratories to develop oversight systems to catch and contain accidents, but not all research institutions adhere to such guidance, require adequate training, or have sufficient resources to dedicate to biosafety. There is also great variability from one research institution to another, even within a nation. Implementing nationwide biosafety norms could begin to address this deficit. For example, it would be helpful to know that GOF research is performed with sufficient safety systems in place, including national standards for equipment maintenance, worker safety training, health monitoring, surveillance, and other measures to help keep researchers and the public safe. It would be ideal to promote such safety standards for all laboratory work to protect workers everywhere. However, confining norms to research with the potential for international consequences, more often performed in donor nations, would allow routine medical work and public health diagnoses to continue even in less-resourced environments. Without nationwide standards for biosafety, organizations will remain reluctant to commit the resources required to achieve high levels of biosafety.

In the last decade, amateur biology, or DIY bio, has gained popularity. Although DIY bio may not require specific regulation now, the situation is likely to evolve, so mechanisms to improve the safety and knowledge of the DIY bio community's activities may help to prevent either an accident or misuse of DIY laboratories to develop a weapon. There is an "ask a biosafety expert" program, funded by the Alfred P. Sloan Foundation, for DIY bio enthusiasts to tap volunteers from the American Biological Safety Association to provide advice, and there are various codes of conduct. For the iGEM competition, teams have to complete safety forms for their projects and biosecurity experts are involved in the judging. For the last five years, a successful Federal Bureau of Investigation (FBI) outreach program has been working with DIY bio practitioners to raise their awareness of potential misuse and to give them points of contact to report suspicious behavior. These mechanisms can forestall an incident and generate law enforcement awareness of potential malign actors.

Another preventive measure to consider is a review of regulations to determine if specific synthetic biology applications have adequate oversight from the appropriate federal agency. There appear to be gaps and a need for updating; a 2013 fundraising campaign on Kickstarter caused consternation by producing glowing plants and distributing seeds to more than eight thousand supporters. The mechanisms used to produce the plants, distribute them, and plant them do not violate any current rules or regulations; however, allowing glowing plants to be introduced into the environment without regulatory review struck many scientists as inappropriate. Congressional action will be necessary to expand and update regulatory authority for the USDA and the Environmental Protection Agency (EPA) to appropriately regulate modern methods to alter plant genomes.5 9

There are already mechanisms in place that can be expanded to cover the potential use of synthetic biology for weapons. Many gene synthesis companies adhere to a code of conduct and screen customer orders for pathogen matches, according to U.S. Department of Health and Human Services guidance. There are no options for the United States to impose regulations on international companies that do not screen orders, but actions can be taken to encourage other nations to issue similar guidance, promote industrywide screening standards, and champion a common code of conduct. The United States also participates in international agreements that prohibit biological weapons development and use, especially the Biological Weapons Convention (BWC). Critics point out that the treaty does not have a verification mechanism and that it has been violated several times in the past, most notably by the Soviet Union. However, the dual-use challenges of biology make a verification mechanism unfeasible. The treaty has effectively reinforced the norm against biological weapons and served as a vehicle to discuss other issues, such as the potential for biological accidents or the misuse of synthetic biology. UN

Resolution 1540, another legally binding mechanism, requires nations to have and enforce measures against the proliferation of nuclear, chemical, and biological weapons. The U.S. Cooperative Biological Engagement Program assists partner nations in fulfilling their Resolution 1540 obligations and boosting their public health infrastructure to mitigate biological incidents.

Risk Posed by AI

Anders Sandberg and Nick Bostrom (2008), "Whole brain emulation: A roadmap."
Future of Humanity Institute, Oxford University.
http://www.fhi.ox.ac.uk/brain-emulation-roadmap-report.pdf
[Probably much more detail than needed just yet.]

Daniel Dewey (2014), "Long-term strategies for ending existential risk from fast takeoff."
Future of Humanity Institute, Oxford University.
http://www.danieldewey.net/fast-takeoff-strategies.pdf

Abstract

If, at some point in the future, each AI development project carries some amount of existential risk from fast takeoff, our chances of survival will decay exponentially until the period of risk is ended. In this paper, I review strategies for ending the risk period. It seems that effective strategies will need to be resilient to government involvement (nationalized projects, regulation, or restriction), will need to account for the additional difficulty of solving some form of the control problem beyond the mere development of AI, and will need to deal with the possibility that many projects will be unable or unwilling to make the investments required to robustly solve the control problem. Strategies to end the risk period could take advantage of the capabilities provided by powerful AI, or of the incentives and abilities governments will have to mitigate fast takeoff risk. Based on these considerations, I find that four classes of strategy – international coordination, sovereign AI, AI empowered project, or other decisive technological advantage – could plausibly end the period of risk.

Introduction

It has been argued that after some level of artificial intelligence capability is reached, an AI might be able to improve very quickly, and could gain great enough cognitive capability to become the dominant power on Earth.1 Call this “fast takeoff”. In this paper, I assume that fast takeoff will be possible at some point in the future, and try to clarify the resulting strategic situation.

Most work on existential risk2 from long-term AI capabilities has focused on the problem of designing an AI that would remain safe even if it were to undergo a fast takeoff. Bostrom calls this the control problem.3

Imagine an optimistic future: the control problem has been solved, and a prudent, conscientious project has used the solution to safely develop human-level or even superintelligent AI. The AI race has been won, and the control problem solved in time to keep this project from causing harm. Has the danger now passed?

Solving the control problem leaves a major issue: other projects are probably developing AI, each carrying the potential for an existential disaster, and not all of those projects will be as safe as the first one. Some additional strategy is needed to end the period of existential risk (x-risk) from fast takeoff. Furthermore, strategies we could take are probably not equally likely to succeed; maximizing the chances of a positive outcome will require us to choose well among them.

The need for a long-term strategy is not a new insight (see, for example, Muehlhauser and Bostrom, "Why we need friendly AI", and Yudkowsky, "Artificial intelligence as a positive and negative factor in global risk"), but I have not found an overview of strategies for ending AI x-risk, nor much in the way of comparing their strengths and weaknesses (Sotala and Yampolskiy, Responses to catastrophic AGI risk: A survey comes closest). In this paper, I attempt such an overview. After introducing the exponential decay model of fast takeoff x-risk (§2) and reviewing what seem to be the most relevant considerations (§3), I find that plausible strategies fall into four categories (§4):

1. International coordination
2. Sovereign AI
3. AI-empowered project
4. Other decisive technological advantage

Implementing one of these strategies may be the best thing one could do to reduce overall existential risk from fast takeoff – in fact, if the considerations underlying my analysis are correct, then it seems plausible that existential risk from fast takeoff cannot be mitigated significantly without using one of these strategies. Based on this analysis, projects aiming to reduce fast takeoff x-risk should be aiming to eventually implement one of these strategies, or to enable future projects to implement one of them.

Nick Bostrom (2014), *Superintelligence*.

Oxford University Press.
https://global.oup.com/academic/product/superintelligence-9780199678112

Excerpt (from Chapter 9: The Control Problem)

A quick synopsis might be called for before we close this chapter. We distinguished two broad classes of methods for dealing with the agency problem at the heart of AI safety: capability control and motivation selection. Table 10 gives a summary.

Table 10 Control methods

Capability control

- Boxing methods: The system is confined in such a way that it can affect the external world only through some restricted, pre-approved channel. Encompasses physical and informational containment methods.
- Incentive: The system is placed within an environment that provides appropriate incentives. This could involve social integration into a world of similarly powerful entities. Another variation is the use of methods (cryptographic) reward tokens. "Anthropic capture" is also a very important possibility but one that involves esoteric considerations.
- Stunting: Constraints are imposed on the cognitive capabilities of the system or its ability to affect key internal processes.
- Tripwires: Diagnostic tests are performed on the system (possibly without its knowledge) and a mechanism shuts down the system if dangerous activity is detected.

Motivation selection

- Direct specification: The system is endowed with some directly specified motivation system, which might be consequentialist or involve following a set of rules.
- Domesticity: A motivation system is designed to severely limit the scope of the agent's ambitions and activities.
- Indirect normativity: Indirect normativity could involve rule-based or consequentialist principles, but is distinguished by its reliance on an indirect approach to specifying the rules that are to be followed or the values that are to be pursued.
- Augmentation: One starts with a system that already has substantially human or benevolent motivations, and enhances its cognitive capacities to make it superintelligent.

Excerpt (from Conclusion)

To reduce the risks of the machine intelligence revolution, we will propose two objectives that appear to best meet all those desiderata: strategic analysis and capacity-building. We can be relatively confident about the sign of these parameters —more strategic insight and more capacity being better. Furthermore, the parameters are elastic: a small extra investment can make a relatively large difference. Gaining insight and capacity is also urgent because early boosts to these parameters may compound, making subsequent efforts more effective. In addition to these two broad objectives, we will point to a few other potentially worthwhile aims for initiatives.

Peter Eckersley and Anders Sandberg (2013), "Is brain emulation dangerous?"

Journal of Artificial General Intelligence 4.
http://www.degruyter.com/view/j/jagi.2013.4.issue-3/jagi-2013-0011/jagi-2013-0011.xml

Abstract

Brain emulation is a hypothetical but extremely transformative technology which has a non-zero chance of appearing during the next century. This paper investigates whether such a technology would also have any predictable characteristics that give it a chance of being catastrophically dangerous, and whether there are any policy levers which might be used to make it safer.

We conclude that the riskiness of brain emulation probably depends on the order of the preceding research trajectory. Broadly speaking, it appears safer for brain emulation to happen sooner, because slower CPUs would make the technology's impact more gradual. It may also be safer if brains are scanned before they are fully understood from a neuroscience perspective, thereby increasing the initial population of emulations, although this prediction is weaker and more scenario-dependent.

The risks posed by brain emulation also seem strongly connected to questions about the balance of power between attackers and defenders in computer security contests. If economic property rights in CPU cycles1 are essentially enforceable, emulation appears to be comparatively safe; if CPU cycles are ultimately easy to steal, the appearance of brain emulation is more likely to be a destabilizing development for human geopolitics.

Furthermore, if the computers used to run emulations can be kept secure, then it appears that making brain emulation technologies open would make them safer. If, however, computer insecurity is deep and unavoidable, openness may actually be more dangerous. We point to some arguments that suggest the former may be true, tentatively implying that it would be good policy to work towards brain emulation using open scientific methodology and free/open source software codebases.

Future of Life Institute (12 Jan 2015), "Research priorities for robust and beneficial artificial intelligence: An open letter."

http://futureoflife.org/AI/open_letter

Full text

Artificial intelligence (AI) research has explored a variety of problems and approaches since its inception, but for the last 20 years or so has been focused on the problems surrounding the construction of intelligent agents - systems that perceive and act in some environment. In this context, "intelligence" is related to statistical and economic notions of rationality - colloquially, the ability to make good decisions, plans, or inferences. The adoption of probabilistic and decision-theoretic representations and statistical learning methods has led to a large degree of integration and cross-fertilization among AI, machine learning, statistics, control theory, neuroscience, and other fields. The establishment of shared theoretical frameworks, combined with the availability of data and processing power, has yielded remarkable successes in various component tasks such as speech recognition, image classification, autonomous vehicles, machine translation, legged locomotion, and question-answering systems.

As capabilities in these areas and others cross the threshold from laboratory research to economically valuable technologies, a virtuous cycle takes hold whereby even small improvements in performance are worth large sums of money, prompting greater investments in research. There is now a broad consensus that AI research is progressing steadily, and that its impact on society is likely to increase. The potential benefits are huge, since everything that civilization has to offer is a product of human intelligence; we cannot predict what we might achieve when this intelligence is magnified by the tools AI may provide, but the eradication of disease and poverty are not unfathomable. Because of the great potential of AI, it is important to research how to reap its benefits while avoiding potential pitfalls.

The progress in AI research makes it timely to focus research not only on making AI more capable, but also on maximizing the societal benefit of AI. Such considerations motivated the AAAI 2008-09 Presidential Panel on Long-Term AI Futures and other projects on AI impacts, and constitute a significant expansion of the field of AI itself, which up to now has focused largely on techniques that are neutral with respect to purpose. We recommend expanded research aimed at ensuring that increasingly capable AI systems are robust and beneficial: our AI systems must do what we want them to do. The attached research priorities document gives many examples of such research directions that can help maximize the societal benefit of AI. This research is by necessity interdisciplinary, because it involves both society and AI. It ranges from economics, law and philosophy to computer security, formal methods and, of course, various branches of AI itself.

In summary, we believe that research on how to make AI systems robust and beneficial is both important and timely, and that there are concrete research directions that can be pursued today.

[signed by over 1,000 researchers]

Future of Life Institute (23 Jan 2015), "Research priorities for robust and beneficial artificial intelligence."
http://futureoflife.org/static/data/documents/research_priorities.pdf

Executive Summary

Success in the quest for artificial intelligence has the potential to bring unprecedented benefits to humanity, and it is therefore worthwhile to research how to maximize these benefits while avoiding potential pitfalls. This document gives numerous examples (which should by no means be construed as an exhaustive list) of such worthwhile research aimed at ensuring that AI remains robust and beneficial.

2.2 Law and Ethics Research

The development of systems that embody significant amounts of intelligence and autonomy leads to important legal and ethical questions whose answers impact both producers and consumers of AI technology. These questions span law, public policy, professional ethics, and philosophical ethics, and will require expertise from computer scientists, legal experts, political scientists, and ethicists. For example:

1. Liability and law for autonomous vehicles: If self-driving cars cut the roughly 40,000 annual US traffic fatalities in half, the car makers might get not 20,000 thank-you notes, but 20,000 lawsuits. In what legal framework can the safety benefits of autonomous vehicles such as drone aircraft and selfdriving cars best be realized [88]? Should legal questions about AI be handled by existing (softwareand internet-focused) “cyberlaw”, or should they be treated separately [14]? In both military and commercial applications, governments will need to decide how best to bring the relevant expertise to bear; for example, a panel or committee of professionals and academics could be created, and Calo has proposed the creation of a Federal Robotics Commission [15].

2. Machine ethics: How should an autonomous vehicle trade off, say, a small probability of injury to a human against the near-certainty of a large material cost? How should lawyers, ethicists, and policymakers engage the public on these issues? Should such trade-offs be the subject of national standards?

3. Autonomous weapons: Can lethal autonomous weapons be made to comply with humanitarian law [18]? If, as some organizations have suggested, autonomous weapons should be banned [23, 85], is it possible to develop a precise definition of autonomy for this purpose, and can such a ban practically be enforced? If it is permissible or legal to use lethal autonomous weapons, how should these weapons be integrated into the existing command-and-control structure so that responsibility and liability be distributed, what technical realities and forecasts should inform these questions, and how should “meaningful human control” over weapons be defined [66, 65, 3]? Are autonomous weapons likely to reduce political aversion to conflict, or perhaps result in “accidental” battles or wars [6]? Finally, how can transparency and public discourse best be encouraged on these issues?

4. Privacy: How should the ability of AI systems to interpret the data obtained from surveillance cameras, phone lines, emails, etc., interact with the right to privacy? How will privacy risks interact with cybersecurity and cyberwarfare [73]? Our ability to take full advantage of the synergy between AI and big data will depend in part on our ability to manage and preserve privacy [48, 1].

5. Professional ethics: What role should computer scientists play in the law and ethics of AI development and use? Past and current projects to explore these questions include the AAAI 2008–09 Presidential

Panel on Long-Term AI Futures [43], the EPSRC Principles of Robotics [8], and recently-announced programs such as Stanford's One-Hundred Year Study of AI and the AAAI committee on AI impact and ethical issues (chaired by Rossi and Chernova).

2.3 Computer Science Research for Robust AI

As autonomous systems become more prevalent in society, it becomes increasingly important that they robustly behave as intended. The development of autonomous vehicles, autonomous trading systems, autonomous weapons, etc. has therefore stoked interest in high-assurance systems where strong robustness guarantees can be made; Weld and Etzioni have argued that "society will reject autonomous agents unless we have some credible means of making them safe" [91]. Different ways in which an AI system may fail to perform as desired correspond to different areas of robustness research:

1. Verification: how to prove that a system satisfies certain desired formal properties. ("Did I build the system right?")

2. Validity: how to ensure that a system that meets its formal requirements does not have unwanted behaviors and consequences. ("Did I build the right system?")

3. Security: how to prevent intentional manipulation by unauthorized parties.

4. Control: how to enable meaningful human control over an AI system after it begins to operate. ("OK, I built the system wrong, can I fix it?")

Tom Dietterich and Eric Horvitz (23 Jan 2015), "Benefits and risks of artificial intelligence." *Medium.*

https://medium.com/@tdietterich/benefits-and-risks-of-artificial-intelligence-460d288cccf3

Excerpt

One set of risks stems from programming errors in AI software. We are all familiar with errors in ordinary software. For example, apps on our smartphones sometimes crash. Major software projects, such as HealthCare.Gov, are sometimes riddled with bugs. Moving beyond nuisances and delays, some software errors have been linked to extremely costly outcomes and deaths. The study of the "verification" of the behavior of software systems is challenging and critical, and much progress has been made. However, the growing complexity of AI systems and their enlistment in high-stakes roles, such as controlling automobiles, surgical robots, and weapons systems, means that we must redouble our efforts in software quality.

There is reason for optimism. Many non-AI software systems have been developed and validated to achieve high degrees of quality assurance. For example, the software in autopilot systems and spacecraft systems is carefully tested and validated. Similar practices must be developed and applied to AI systems. One technical challenge is to guarantee that systems built automatically via statistical "machine learning" methods behave properly. Another challenge is to ensure good behavior when an AI system encounters unforeseen situations. Our automated vehicles, home robots, and intelligent cloud services must perform well even when they receive surprising or confusing inputs.

A second set of risks is cyberattacks: criminals and adversaries are continually attacking our computers with viruses and other forms of malware. AI algorithms are no different from other software in terms of their vulnerability to cyberattack. But because AI algorithms are being asked to make high-stakes decisions, such as driving cars and controlling robots, the impact of successful cyberattacks on AI systems could be much more devastating than attacks in the past. US Government funding agencies and corporations are supporting a wide range of cybersecurity research projects, and artificial intelligence techniques in themselves will provide novel methods for detecting and defending against cyberattacks. Before we put AI algorithms in control of high-stakes decisions, we must be much more confident that these systems can survive large scale cyberattacks.

A third set of risks echo the tale of the Sorcerer's Apprentice. Suppose we tell a self-driving car to "get us to the airport as quickly as possible!" Would the autonomous driving system put the pedal to the metal and drive at 300 mph while running over pedestrians? Troubling scenarios of this form have appeared recently in the press. Other fears center on the prospect of out-of-control superintelligences that threaten the survival of humanity. All of these examples refer to cases where humans have failed to correctly instruct the AI algorithm in how it should behave.

This is not a new problem. An important aspect of any AI system that interacts with people is that it must reason about what people intend rather than carrying out commands in a literal manner. An AI system should not only act on a set of rules that it is instructed to obey?—?it must also analyze and understand whether the behavior that a human is requesting is likely to be judged as "normal" or "reasonable" by most people. It should also be continuously monitoring itself to detect abnormal internal behaviors, which might signal bugs, cyberattacks, or failures in its understanding of its actions. In addition to relying on internal mechanisms to ensure proper behavior, AI systems need to have the capability?—?and responsibility?—?of working with people to obtain feedback and guidance. They must know when to stop and "ask for directions"?—?and always be open for feedback.

Some of the most exciting opportunities ahead for AI bring together the complementary talents of people and computing systems. AI-enabled devices are allowing the blind to see, the deaf to hear, and the disabled and elderly to walk, run, and even dance. People working together with the Foldit online game were able to discover the structure of the virus that causes AIDS in only three weeks, a feat that neither people nor computers working alone could come close to matching. Other studies have shown how the massive space of galaxies can be explored hand-in-hand by people and machines, where the tireless AI astronomer understands when it needs to occasionally reach out and tap the expertise of human astronomers.

In reality, creating real-time control systems where control needs to shift rapidly and fluidly between people and AI algorithms is difficult. Some airline accidents occurred when pilots took over from the autopilots. The problem is that unless the human operator has been paying very close attention, he or she will lack a detailed understanding of the current situation.

Future of Life Institute (19 Feb 2015), "A survey of research questions for robust and beneficial AI."
http://futureoflife.org/static/data/documents/research_survey.pdf

Contents

Stuart Dredge (29 Jan 2015), "Artificial intelligence will become strong enough to be a concern, says Bill Gates."

Guardian.
http://www.theguardian.com/technology/2015/jan/29/artificial-intelligence-strong-concern-bill-gates

Excerpt

“I am in the camp that is concerned about super intelligence. First the machines will do a lot of jobs for us and not be super-intelligent. That should be positive if we manage it well,” wrote Gates.

“A few decades after that though the intelligence is strong enough to be a concern. I agree with Elon Musk and some others on this and don’t understand why some people are not concerned.”

Musk spoke out in October 2014 during an interview at the AeroAstro Centennial Symposium, telling students that the technology industry should be thinking hard about how it approaches AI advances in the future.

“I think we should be very careful about artificial intelligence. If I had to guess at what our biggest existential threat is, it’s probably that. So we need to be very careful,” said Musk. “I’m increasingly inclined to think that there should be some regulatory oversight, maybe at the national and international level, just to make sure that we don’t do something very foolish.”

In a December interview, Professor Hawking went further. “The primitive forms of artificial intelligence we already have, have proved very useful. But I think the development of full artificial intelligence could spell the end of the human race.”

Tom Simonite (7 Apr 2015), "AI doomsayer says his ideas are catching on (Interview with Nick Bostrom)."

MIT Technology Review.
http://www.technologyreview.com/news/536381/ai-doomsayer-says-his-ideas-are-catching-on

Excerpt

What kind of research is possible on something so far from being real today?

What’s been produced up to date is a clearer understanding of what the problem is and some concepts that can be used to think about these things. These may not look like much on paper, but before, it wasn’t possible to go to the next stage, which is developing a technical research agenda.

Can you give an example of a technical project that might be on that?

For example, could you design an AI motivation system [so] that the AI doesn’t resist the programmer coming in to change its goal? There is a whole set of things that could be practically useful, like boxing methods—tools that can contain an AI before it is ready to be released.

Anthony Kosner (20 Apr 2015), "What really scares tech leaders about artificial intelligence."

Forbes.
http://www.forbes.com/sites/anthonykosner/2015/04/20/what-really-scares-tech-leaders-about-artificial-intelligence

Excerpt

My own research into this fear has led me to conclude that it is about something far more mundane and predictable: regulation. The concern most central to a select sampling of the language of tech leaders who have weighed in on AI recently is the need for regulation to avoid unintended consequences. Musk said as much right after he identified AI as “our biggest existential threat.” “There should be some regulatory oversight, maybe at the national and international level,” he warned, “just to make sure that we don’t do something very foolish.” Compared to regulation, though, fear of an existential threat to humanity is the least correlated concept on this list. See my forthcoming companion post, “Graph Theory Helps To Decode The AI Fears Of Tech Leaders,” in which I explain my methodology and give credit to its source (hat tip to Dan Shipper). For the present post, I will explore the implications of my findings.

...If machine intelligence is essentially predictive of and responsive to external data streams, what might there be to regulate? The answer, at this point, is not the AIs themselves but the uses to which humans put data. This is correct but an understandable pain point for tech companies who complain that regulation inhibits innovation. The intersection of this Venn diagram occurs where these innovations clearly benefit the most humans.

I don’t think Musk or the rest are irresponsibly pointing to the threat of AI, but this gesture has served to take attention away from the effect that humans at tech companies are having on the rest of the planet’s humans. As I have commented about biotechnology, the best way to lower the public fear about new technologies is to demonstrate clear benefits for a lot of people. The unintended side effect of fomenting fear of AI is that governments may react to the pressures of a sensationalized public with regulations that prove counterproductive. In many ways, I think the Future of Life Institute, to which Elon Musk has donated $10 million to research the threats of AI, is an attempt by technologists and scientists to regulate themselves before governments do it for them, and not as well.

Nick Bilton (20 May 2015), "Ava of 'Ex Machina' is just sci-fi (for now)."

New York Times.
http://www.nytimes.com/2015/05/21/style/ava-of-ex-machina-is-just-sci-fi-for-now.html

Full text

Are technology companies running too fast into the future and creating things that could potentially wreak havoc on humankind?

That question has been swirling around in my head ever since I saw the enthralling science-fiction film “Ex Machina.”

The movie offers a clever version of the robots versus humans narrative. But what makes “Ex Machina” different from the usual special-effects blockbuster is the ethical questions it poses.

Foremost among them is something that most techies don’t seem to want to answer: Who is making sure that all of this innovation does not go drastically wrong?

In the film, advances in artificial intelligence take place in a secret laboratory beyond the reach of governments and concerned citizens. (The robot's name is Ava.) That is not unlike how most innovations occur in real life today.

Alex Garland, the writer and director of "Ex Machina," said in a phone interview last week: "I have no idea if technology companies are doing anything wrong or not, but they are so powerful, and the work they are doing has such potential for seismic human change of how we live, they have to have oversight.

"If you've got corporations that are investigating areas that can change fundamental things about the way we live, someone needs to be looking at them."

While Mr. Garland's film is focused on A.I., his concern about unchecked innovations could apply to all kinds of disciplines, including bioengineering, smart homes, self-driving cars and medical nanobots, to name a few. And while these breakthroughs are intended to help humanity, they could backfire without the proper oversight.

This fear isn't just confined to science-fiction filmmakers, or people who wear tinfoil hats. In recent years, experts in robotics, cosmology and artificial intelligence have set out to tackle the issue of oversight, holding symposiums and creating research organizations.

Elon Musk, founder of Tesla, recently donated $10 million to the Future of Life Institute, an organization that seeks to "mitigate existential risks facing humanity" from "human-level artificial intelligence."

The Lifeboat Foundation is a nonprofit that tries to help humanity combat the "existential risks" of genetic engineering, nanotechnology and the so-called singularity, which refers to the hypothetical moment when artificial intelligence surpasses the human intellect.

And in 2012, philosophers and scientists at Cambridge University formed the Center for Study of Existential Risk, with the goal to ensure "that our own species has a long-term future."

Sir Martin Rees, an emeritus professor of cosmology and astrophysics at Cambridge, who helped start the research center, said that what makes the existential risk today so much greater is the ease with which a single person or company can cause catastrophic harm.

"Unlike the past, the empowerment of individuals is much greater," Mr. Rees said. "You can't make a clandestine H-bomb today, but you can make a clandestine biological virus or a clandestine computer virus."

Mr. Rees said that his biggest worry is not robots or A.I., but biological agents. He cited research done by scientists at the University of Wisconsin, who created a bird flu virus that can be transmitted to people through the air. (Scientists later played down the danger.)

It's not hard to imagine other potential doomsday outcomes. Last month, plant geneticists at the University of Minnesota created a DNA-engineered potato that doesn't accumulate sugars, so it can sit on a shelf for years without rotting. It's unclear how consuming that potato may affect the human body.

Scientists are experimenting with altering the human immune system to fight certain viruses. But yet we don't know if this will create super viruses.

Adding to the concern is the lack of oversight, so that private companies and researchers are basically policing themselves. For example, there is no government body that oversees the development of A.I., so Google created its own ethics committee, conveniently made up of A.I. experts.

But the real-world implications of technological breakthroughs are often not apparent to those entrenched in those fields, said Ronald C. Arkin, a robotics expert and professor at the Institute for Robotics and Intelligent Machines at Georgia Tech. Mr. Arkin, who has designed software for battlefield robots under contract with the Army, said that it wasn't until he saw his robots in the field that some risks became apparent.

"Seeing the robots move out of our lab and into the real world gave me some pause," he said, noting that he saw robots that were becoming "killing machines fully capable of taking human life, perhaps indiscriminately."

The main characters in "Ex Machina" come to this realization as well, but do so too late. Toward the end of the film, the character Nathan Bateman, a genius programmer, realized that he may have done just what he set out to do.

Nathan, drunk, mutters: "The good deeds a man has done before defend him." The line is a reference to what J. Robert Oppenheimer, the father of the atomic bomb, said after witnessing the explosion of the first such bomb, Trinity.

"I remembered the line from the Hindu scripture, the Bhagavad Gita," Oppenheimer said, before uttering the now famous quote. "Now I am become Death, the destroyer of worlds."

Stuart Russell, et al. (27 May 2015), "Robotics: Ethics of artificial intelligence."

Nature.
http://www.nature.com/news/robotics-ethics-of-artificial-intelligence-1.17611
http://www.nature.com/polopoly_fs/1.17611!/menu/main/topColumns/topLeftColumn/pdf/521415a.pdf

Excerpt from Stuart Russell: Take a stand on AI weapons

LAWS [lethal autonomous weapons systems] could violate fundamental principles of human dignity by allowing machines to choose whom to kill — for example, they might be tasked to eliminate anyone exhibiting 'threatening behaviour'. The potential for LAWS technologies to bleed over into peacetime policing functions is evident to human-rights organizations and drone manufacturers.

In my view, the overriding concern should be the probable endpoint of this technological trajectory. The capabilities of autonomous weapons will be limited more by the laws of physics — for example, by constraints on range, speed and payload — than by any deficiencies in the AI systems that control them. For instance, as flying robots become smaller, their manoeuvrability increases and their ability to be targeted decreases. They have a shorter range, yet they must be large enough to carry a lethal payload — perhaps a one-gram shaped charge to puncture the human cranium. Despite the limits imposed by physics, one can expect platforms deployed in the millions, the agility and lethality of which will leave humans utterly defenceless. This is not a desirable future.

Other essays:

- Sabine Hauert: Shape the debate, don't shy from it
- Russ Altman: Distribute AI benefits fairly
- Manuela Veloso: Embrace a robot–human world

Nature (28 May 2015), " Interview with Stuart Russell (Audio)."
http://www.nature.com/polopoly_fs/7.26593!/audiofile/nature-2015-05-28-automated-AI.mp3

Amir Mizroch (8 Jun 2015), "Google on artificial-intelligence panic: Get a grip."
Wall Street Journal Blogs.
http://blogs.wsj.com/digits/2015/06/08/google-on-artificial-intelligence-panic-get-a-grip

Excerpt

Google's artificial-intelligence researchers believe there are more urgent matters than the potential destruction of humanity at the hands of superintelligent machines, and that anyone talking about how AI will destroy us all is being "preposterous."

"Whether it's Terminator coming to blow us up or mad scientists looking to create quite perverted women robots, this narrative has somehow managed to dominate the entire landscape, which we find really quite remarkable," said Mustafa Suleyman, the head of applied AI at Google DeepMind, the London-based AI company he co-founded and which Google bought last year for about $400 million.

"The narrative has shifted from 'Isn't it terrible that AI has been such a failure?' to 'Isn't it terrible that AI has been such a success?' " he said. Suleyman was speaking at a machine learning event in London last Friday.

...Just over this past year, figures such as astrophysicist Stephen Hawking, Microsoft MSFT +1.24%'s Bill Gates, and Tesla's Elon Musk—an early investor in DeepMind—have voiced concern over AI's potential to harm humanity.

"On existential risk, our perspective is that it's become a real distraction from the core ethics and safety issues, and it's completely overshadowed the debate," Suleyman said. "The way we think about AI is that it's going to be a hugely powerful tool that we control and that we direct, whose capabilities we limit, just as you do with any other tool that we have in the world around us, whether they're washing machines or tractors. We're building them to empower humanity and not to destroy us."

Stuart Russell and John Bohannon (17 Jul 2015), "Artificial intelligence. Fears of an AI pioneer."
Science 349.
http://www.ncbi.nlm.nih.gov/pubmed/26185241
http://www.sciencemag.org/content/349/6245/252.long

Excerpt

Q: What do you see as a likely path from AI to disaster?

A: The basic scenario is explicit or implicit value misalignment: AI systems [that are] given objectives that don't take into account all the elements that humans care about. The routes could be varied and complex—corporations seeking a supertechnological advantage, countries trying to build [AI systems] before their enemies, or a slow-boiled frog kind of evolution leading to dependency and enfeeblement not unlike E. M. Forster's The Machine Stops.

Sebastian Anthony (27 Jul 2015), "Musk, Hawking, Wozniak call for ban on autonomous weapons and military AI."

Ars Technica UK.
http://arstechnica.co.uk/gadgets/2015/07/musk-hawking-wozniak-call-for-ban-on-autonomous-weaponry-and-military-ai

Excerpt

A very large number of scientific and technological luminaries have signed an open letter calling for the world's governments to ban the development of "offensive autonomous weapons" to prevent a "military AI arms race."

The letter, which will be presented at the International Joint Conferences on Artificial Intelligence (IJCAI) in Buenos Aires tomorrow, is signed by Stephen Hawking, Elon Musk, Noam Chomsky, the Woz, and dozens of other AI and robotics researchers.

For the most part, the letter is concerned with dumb robots and vehicles being turned into smart autonomous weapons. Cruise missiles and remotely piloted drones are okay, according to the letter, because "humans make all targeting decisions." The development of fully autonomous weapons that can fight and kill without human intervention should be nipped in the bud, however.

Here's one of the main arguments from the letter:

> The key question for humanity today is whether to start a global AI arms race or to prevent it from starting. If any major military power pushes ahead with AI weapon development, a global arms race is virtually inevitable, and the endpoint of this technological trajectory is obvious: autonomous weapons will become the Kalashnikovs of tomorrow.

Later, the letter draws a strong parallel between autonomous weapons and chemical/biological warfare:

> Just as most chemists and biologists have no interest in building chemical or biological weapons, most AI researchers have no interest in building AI weapons — and do not want others to tarnish their field by doing so, potentially creating a major public backlash against AI that curtails its future societal benefits.

The letter is being presented at IJCAI by the Future of Life Institute. It isn't entirely clear who the letter is addressed to, other than the academics and researchers who will be attending the conferences. Perhaps

it's just intended to generally raise awareness of the issue, so that we don't turn a blind eye to any autonomous weapons research being carried out by major military powers.

Elon Musk and Stephen Hawking have both previously warned of the dangers of advanced AI. Musk said that AI is “potentially more dangerous than nukes,” while Hawking was far more optimistic, merely saying that AI is "our biggest existential threat."

The main issue with AI in general, and autonomous weapons in specific, is that they are transformational, sea-change technologies. Once we create an advanced AI, or a weapons system that can decide for itself who to attack, there's no turning back. We can't put gunpowder or nuclear weapons back in the bag, and autonomous weaponry would be no different.

Edward Moore Geist (30 Jul 2015), "Is artificial intelligence really an existential threat to humanity?"

The Bulletin of the Atomic Scientists.
http://thebulletin.org/artificial-intelligence-really-existential-threat-humanity8577

Conclusion

For all its entertainment value as a philosophical exercise, Bostrom’s concept of superintelligence is mostly a distraction from the very real ethical and policy challenges posed by ongoing advances in artificial intelligence. Although it has failed so far to realize the dream of intelligent machines, artificial intelligence has been one of the greatest intellectual adventures of the last 60 years. In their quest to understand minds by trying to build them, artificial intelligence researchers have learned a tremendous amount about what intelligence is not. Unfortunately, one of their major findings is that humans resort to fallible heuristics to address many problems because even the most powerful physically attainable computers could not solve them in a reasonable amount of time. As the authors of a 1993 textbook about problem-solving programs noted, “intelligence is possible because Nature is kind,” but “the ubiquity of exponential problems makes it seem that Nature is not overly generous.” As a consequence, both the peril and the promise of artificial intelligence have been greatly exaggerated.

But if artificial intelligence might not be tantamount to “summoning the demon” (as Elon Musk colorfully described it), AI-enhanced technologies might still be extremely dangerous due to their potential for amplifying human stupidity. The AIs of the foreseeable future need not think or create to sow mass unemployment, or enable new weapons technologies that undermine precarious strategic balances. Nor does artificial intelligence need to be smarter than humans to threaten our survival—all it needs to do is make the technologies behind familiar 20th-century existential threats faster, cheaper, and more deadly.

Risk Posed by Nanotechnology

Eric Drexler (1986), Engines of creation: The coming era of nanotechnology.
http://www.nanowerk.com/nanotechnology/reports/reportpdf/report47.pdf

Excerpt

THE THREAT FROM THE MACHINES

In Chapter 4, I described some of what replicating assemblers will do for us if we handle them properly. Powered by fuels or sunlight, they will be able to make almost anything (including more of themselves) from common materials.

Living organisms are also powered by fuels or sunlight, and also make more of themselves from ordinary materials. But unlike assembler-based systems, they cannot make "almost anything".

Genetic evolution has limited life to a system based on DNA, RNA, and ribosomes, but memetic evolution will bring life-like machines based on nanocomputers and assemblers.I have already described how assembler-built molecular machines will differ from the ribosome-built machinery of life. Assemblers will be able to build all that ribosomes can, and more; assembler-based replicators will therefore be able to do all that life can, and more. From an evolutionary point of view, this poses an obvious threat to otters, people, cacti, and ferns - to the rich fabric of the biosphere and all that we prize.

The early transistorized computers soon beat the most advanced vacuum-tube computers because they were based on superior devices. For the same reason, early assemblerbased replicators could beat the most advanced modern organisms. "Plants" with "leaves" no more efficient than today's solar cells could out-compete real plants, crowding the biosphere with an inedible foliage. Tough, omnivorous "bacteria" could out-compete real bacteria: they could spread like blowing pollen, replicate swiftly, and reduce the biosphere to dust in a matter of days. Dangerous replicators could easily be too tough, small, and rapidly spreading to stop - at least if we made no preparation. We have trouble enough controlling viruses and fruit flies.

Among the cognoscenti of nanotechnology, this threat has become known as the "gray goo problem." Though masses of uncontrolled replicators need not be gray or gooey, the term "gray goo" emphasizes that replicators able to obliterate life might be less inspiring than a single species of crabgrass. They might be "superior" in an evolutionary sense, but this need not make them valuable. We have evolved to love a world rich in living things, ideas, and diversity, so there is no reason to value gray goo merely because it could spread. Indeed, if we prevent it we will thereby prove our evolutionary superiority.

The gray goo threat makes one thing perfectly clear: we cannot afford certain kinds of accidents with replicating assemblers.

Chris Phoenix and Eric Drexler (9 Jun 2004), "Safe exponential manufacturing."

Nanotechnology 15.
http://rachel.org/lib/safe_exponential_manufacturing.040601.pdf

Abstract

In 1959, Richard Feynman pointed out that nanometre-scale machines could be built and operated, and that the precision inherent in molecular construction would make it easy to build multiple identical copies. This raised the possibility of exponential manufacturing, in which production systems could rapidly and cheaply increase their productive capacity, which in turn suggested the possibility of destructive runaway self-replication. Early proposals for artificial nanomachinery focused on small self-replicating machines, discussing their potential productivity and their potential destructiveness if abused. In the light of controversy regarding scenarios based on runaway replication (so-called 'grey goo'), a review of current thinking regarding nanotechnology-based manufacturing is in order. Nanotechnology-based fabrication can be thoroughly non-biological and inherently safe: such systems need have no ability to move about, use natural resources, or undergo incremental mutation. Moreover, self-replication is unnecessary: the development and use of highly productive systems of nanomachinery (nanofactories) need not involve the construction of autonomous self-replicating nanomachines. Accordingly, the construction of anything resembling a dangerous self-replicating nanomachine can and should be prohibited. Although advanced nanotechnologies could (with great difficulty and little incentive) be used to build such devices, other concerns present greater problems. Since weapon systems will be both easier to build and more likely to draw investment, the potential for dangerous systems is best considered in the context of military competition and arms control.

7. Conclusion

Early proposals for manufacturing systems based on molecular nanotechnology included devices that had some similarity to runaway self-replicating machines, in that they were, at least, self-replicating. It has since become clear that all risk of accidental runaway replication can be avoided, since efficient manufacturing systems can be designed, built, and used without ever making a device with the complex additional capabilities that a hypothetical 'grey goo robot' would require. However, this does not mean that molecular nanotechnology is without risks. Problems including weapon systems, radical shifts of economic and political power, and aggregate environmental risks from novel products and largescale production will require close attention and careful policymaking.

The Royal Society (29 Jul 2004), "Possible adverse health, environmental and safety impacts."

In *Nanoscience and nanotechnologies: opportunities and uncertainties*.
http://www.nanotec.org.uk/report/chapter5.pdf

5.1 Introduction

1 In Chapters 3 and 4 we have outlined the ways in which researchers and industry hope to exploit the unique properties of nanomaterials and the processes of nanomanufacturing for medical applications and to deliver environmental benefits. Current medical applications of nanotechnologies include anti-

microbial wound dressings, and it is anticipated that future applications will include more durable and better prosthetics and new drug delivery mechanisms. Current research into applications of nanotechnology includes efforts to reduce the amount of solvents and other harmful chemicals in manufacturing, to improve energy efficiency and energy storage capabilities, and to remove persistent pollutants from soil and water supplies, all of which offer hope of benefiting the environment and increasing sustainability. In section 4.5 we highlighted the need to incorporate a life cycle assessment approach into the research and development of products and processes arising from nanotechnologies to ensure that they do not result in a net increase in resource use. In this chapter we consider potential adverse health, environmental and safety impacts of nanotechnologies.

2 Whereas the potential health and environmental benefits of nanotechnologies have been welcomed, concerns have been expressed that the very properties that are being exploited by researchers and industry (such as high surface reactivity and ability to cross cell membranes) might have negative health and environmental impacts and, particularly, that they might result in greater toxicity. The public who participated in the market research that we commissioned expressed worries about possible long-term side effects associated with medical applications and whether nanomaterials would be biodegradable. Analogies were made with plastics, which were once hailed as 'the future' but which have proved to have accompanying adverse effects on individuals and the environment (BMRB 2004).

3 Almost all the concerns expressed to us, in evidence and during our workshop on health and environmental impacts of nanotechnologies, related to the potential impacts of manufactured nanoparticles and nanotubes (in a free rather than fixed form) on the health and safety of humans, non-human biota and ecosystems. The fact that nanoparticles are on the same scale as cellular components and larger proteins has led to the suggestion that they might evade the natural defences of humans and other species and damage cells. It is important to set these concerns in context by noting that humans have always been exposed to some types of nanoparticles arising from natural sources such as atmospheric photochemistry and forest fires, and exposures to millions of pollutant nanoparticles per breath have been commonplace since the first use of fire.

4 Manufactured nanoparticles and nanotubes are important because they are among the first nanoscale technologies used in consumer products, but as Table 4.1 makes clear, the production rates of these materials is only a small fraction of the predicted potential for nanotechnologies. The IT industry also uses nanotechnologies, both in techniques used and the minimum feature size of devices; however, in contrast to manufactured nanoparticles and nanotubes, it does not present any unique hazards. There is an important distinction between applications that use nanoscalar active areas on larger objects (for example, nanometre-scale junction regions in transistors, which form part of a millimetre-sized chip and are therefore fixed), and chemicals or pharmaceuticals in which the nanometrescale 'active area' is a discrete nanoparticle or nanotube. Although a computer chip with 100 million nanostructures presents a potential hazard for manufacture, disposal or recycling, these issues are related to the bulk materials, which make up the chips (for example, gallium), rather than to the nanostructures within them. Although nanoscience and nanotechnologies may involve individual scientists and other workers using or being exposed to a range of chemical reagents and physical processes that could imply harm to their health, such exposures to substances and materials other than nanoparticles are covered by existing understanding and regulation. They are not considered further in this report except in that they may be in the form of discrete particles incorporated into materials in the nanometre size range.

Robert Freitas, Jr. (23 Jan 2006), "Molecular manufacturing: Too dangerous to allow?"

Nanotechnology Perceptions 2.
http://www.rfreitas.com/Nano/MMDangerous.pdf

Excerpt

One common argument against pursuing a molecular assembler or nanofactory design effort is that the end results are too dangerous. According to this argument, any research into molecular manufacturing (MM) should be blocked because this technology might be used to build systems that could cause extraordinary damage. The kinds of concerns that nanoweapons systems might create have been discussed elsewhere, in both the nonfictional and fictional literature. Perhaps the earliest-recognized and best-known danger of molecular manufacturing is the risk that self-replicating nanorobots capable of functioning autonomously in the natural environment could quickly convert that natural environment (e.g., 'biomass') into replicas of themselves (e.g., 'nanomass') on a global basis, a scenario often referred to as the 'gray goo problem' but more accurately termed 'global ecophagy'.4 As Drexler first warned in Engines of Creation in 1986:8

'Plants' with 'leaves' no more efficient than today's solar cells could out-compete real plants, crowding the biosphere with an inedible foliage. Tough omnivorous "bacteria" could out-compete real bacteria: They could spread like blowing pollen, replicate swiftly, and reduce the biosphere to dust in a matter of days. Dangerous replicators could easily be too tough, small, and rapidly spreading to stop—at least if we make no preparation....We cannot afford certain kinds of accidents with replicating assemblers.

...Attempts to block or 'relinquish' molecular manufacturing research will make the world a more, not less, dangerous place. This paradoxical conclusion is founded on two premises. First, attempts to block the research will fail. Second, such attempts will preferentially block or slow the development of defensive measures by responsible groups. One of the clear conclusions reached by Freitas4 was that effective countermeasures against self-replicating systems should be feasible, but will require significant effort to develop and deploy. (Nanotechnology critic Bill Joy, responding to this author, complained in late 2000 that any nanoshield defense to protect against global ecophagy "appears to be so outlandishly dangerous that I can't imagine we would attempt to deploy it.") But blocking the development of defensive systems would simply insure that offensive systems, once deployed, would achieve their intended objective in the absence of effective countermeasures.

...We can reasonably conclude that blocking the development of defensive systems would be an extraordinarily bad idea. Actively encouraging rapid development of defensive systems by responsible groups while simultaneously slowing or hindering development and deployment by less responsible groups ('nations of concern') would seem to be a more attractive strategy, and is supported by the Foresight Guidelines. As even nanotechnology critic Bill Joy finally admitted in late 2003: "These technologies won't stop themselves, so we need to do whatever we can to give the good guys a head start."

Center for Responsible Nanotechnology (Feb 2008), "Dangers of molecular manufacturing."

http://www.crnano.org/dangers.htm

Excerpt

Molecular manufacturing raises the possibility of horrifically effective weapons. As an example, the smallest insect is about 200 microns; this creates a plausible size estimate for a nanotech-built antipersonnel weapon capable of seeking and injecting toxin into unprotected humans. The human lethal dose of botulism toxin is about 100 nanograms, or about 1/100 the volume of the weapon. As many as 50 billion toxin-carrying devices—theoretically enough to kill every human on earth—could be packed into a single suitcase. Guns of all sizes would be far more powerful, and their bullets could be self-guided. Aerospace hardware would be far lighter and higher performance; built with minimal or no metal, it would be much harder to spot on radar. Embedded computers would allow remote activation of any weapon, and more compact power handling would allow greatly improved robotics. These ideas barely scratch the surface of what's possible.

An important question is whether nanotech weapons would be stabilizing or destabilizing. Nuclear weapons, for example, perhaps can be credited with preventing major wars since their invention. However, nanotech weapons are not very similar to nuclear weapons. Nuclear stability stems from at least four factors. The most obvious is the massive destructiveness of all-out nuclear war. All-out nanotech war is probably equivalent in the short term, but nuclear weapons also have a high long-term cost of use (fallout, contamination) that would be much lower with nanotech weapons. Nuclear weapons cause indiscriminate destruction; nanotech weapons could be targeted. Nuclear weapons require massive research effort and industrial development, which can be tracked far more easily than nanotech weapons development; nanotech weapons can be developed much more rapidly due to faster, cheaper prototyping. Finally, nuclear weapons cannot easily be delivered in advance of being used; the opposite is true of nanotech. Greater uncertainty of the capabilities of the adversary, less response time to an attack, and better targeted destruction of an enemy's visible resources during an attack all make nanotech arms races less stable. Also, unless nanotech is tightly controlled, the number of nanotech nations in the world could be much higher than the number of nuclear nations, increasing the chance of a regional conflict blowing up.

Steffen Foss Hansen, et al. (20 Jul 2008), "Late lessons from early warnings for nanotechnology," Nature Nanotechnology 3.

http://www.nature.com/nnano/journal/v3/n8/pdf/nnano.2008.198.pdf

Abstract

A new technology will only be successful if those promoting it can show that it is safe, but history is littered with examples of promising technologies that never fulfilled their true potential and/or caused untold damage because early warnings about safety problems were ignored. The nanotechnology community stands to benefit by learning lessons from this history.

Box 1: The 12 lessons outlined by the EFA [European Environment Agency]

1. Acknowledge and respond to ignorance, uncertainty and risk in technology appraisal.
2. Provide long-term environmental and health monitoring and research into early warnings.

3. Identify and work to reduce scientific 'blind spots' and knowledge gaps.
4. Identify and reduce interdisciplinary obstacles to learning.
5. Account for real-world conditions in regulatory appraisal.
6. Systematically scrutinize claimed benefits and risks.
7. Evaluate alternative options for meeting needs, and promote robust, diverse and adaptable technologies.
8. Ensure use of 'lay' knowledge, as well as specialist expertise.
9. Account fully for the assumptions and values of different social groups.
10. Maintain regulatory independence of interested parties while retaining an inclusive approach to information and opinion gathering.
11. Identify and reduce institutional obstacles to learning and action.
12. Avoid 'paralysis by analysis' by acting to reduce potential harm when there are reasonable grounds for concern.

Chris Toumey (8 Oct 2012), "Lessons from before and after nanotech."

Nature Nanotechnology 7.
http://www.nature.com/nnano/journal/v7/n10/full/nnano.2012.173.html

Excerpt

Yes, new technologies deserve to be examined for the ways they will touch our lives, including their economic, environmental and existential consequences. So be it. At the same time, however, an organization discredits its own response to a new technology when it shows us that its response has little connection to reality. A moratorium on synthetic biology is as unrealistic as a moratorium on nanotechnology.

In one sense, synthetic biology is not as problematic as nanotechnology because it embodies fewer scientific disciplines and subdisciplines than nanotech. SynBio, as some call it, can be described in terms of genetics, genetic engineering, information technology and cell biology. This is much more intellectually compact than nanotech, which embraces molecular biology, microelectronics, catalytic chemistry, atomic physics, materials science, targeted drug delivery, electron microscopy, scanning probe microscopy and many more fields. And yet synthetic biology is more than enough of a challenge for legislators and regulators because centred around it is a new culture named do-it-yourself biology, which exists parallel to the labs of universities, corporations and government agencies like the US National Institutes of Health. The basic operations of synthetic biology are remarkably accessible to the non-scientists and would-be scientists who populate DIYbio, as this movement is called. DIYbio clubs can legally purchase chemically synthesized segments of a genome from suppliers as easily as you or I order clothes or gadgets. This new variation of citizen science is not about to be suppressed by any moratorium.

Linda F. Hogle (3 Jan 2013), "Concepts of risk in nanomedicine research."
Journal of Law, Medicine & Ethics 40.
http://onlinelibrary.wiley.com/doi/10.1111/j.1748-720X.2012.00709.x/abstract

Abstract

Risk takes center stage in ethical debates over nanomedical technologies. Yet concepts of risk may hold different meanings, and they are embedded within particular political, economic, and social contexts. This article discusses framings of risk in debates over medical innovations such as nanomedicine, and draws attention to organizational and institutional forms of risk which are less visible in bioethical policy debates. While significant, possibly unique risks may exist in specific nano-based products, risk may also arise from the very processes and procedures that develop, test, and oversee any novel technology. This supports recommendations to coordinate efforts through an interagency Working Group and a Secretary-level Advisory Committee to provide flexibility and sensitivity to emerging issues of concern.

Excerpt

Some scientists, ethicists, and policymakers have been extremely vocal about the need to revisit risk review for nanotechnologies, citing multiple causes for concern. Others suggest that existing systems are sufficient and claim that current procedures already recognize and deal with dangers so that adding new layers of review will only slow the introduction of potentially enormously useful products.58 Indecision about the best course of action will likely result in no increased oversight. However, to create new bodies for review or to institute new guidelines, procedures, and tests requires the political will to do so. In an era of pressure to see returns from national investment in research, this may not be likely. Nanotechnology is also situated in a historical moment in which broader issues of expertise, evidence, and capacity are being called into question. Calls for regulatory reform seek to reduce redundancies and make agencies more efficient, which may reduce some of the systematic error as described by Vaughn above, but could also potentially strip them of resources needed to analyze risk in novel products. Some would argue that agencies such as the FDA already lack the resources and expertise necessary to deal with the variety of novel entities submitted for review. Here is where the recommendation to create an inter-agency working group is most compelling: critical information may be available, but is simply not accessed effectively or equally across regulatory agencies and other relevant organizations, due to infrastructure weaknesses or some of the institutional or procedural issues as described above. An inter-agency working group could bridge gaps in knowledge and types of expertise, be better able to conduct more comprehensive and interdisciplinary analyses, and more readily flag potential problems.

Novel products are also entering review in the context of evidence-based medicine policies, which have a higher bar for demonstrating effectiveness and may affect risk-benefit analyses accordingly. Expertise at the local level of IRBs is at issue as well: can they be expected to bear most of the burden of recognizing relevant risks, and is there enough consistency among IRBs to be confident that human subject protections are commensurable across locations?

Perhaps what we should ask is not what is different about nanomedicine that might trigger new or different oversight mechanisms, but what (if anything) is changing in human subjects protection. Explorations into different ways of conducting pre-clinical trials, including the suggestion of introducing bioinformatics and predictive algorithms or cell-based, in vitro preclinicals with the addition of visualization techniques during and post-administration of nanomedicines may change analyses, or at

least the way IRBs and agencies review risk data. Because of the sophistication of some technologies being tried in humans, IRBs have begun focusing more on the complex technical aspects of the science rather than the bigger clinical picture for trial subjects. If true, what is the impact on the well-being of patients? Jonathan Kimmelman argues that the purview of IRBs should be much broader than assessing risk alone, and that the context in which trials are conducted should be a primary consideration.

Medical risk assessments performed by bioethicists, quantitative risk experts, and regulatory and other oversight authorities have not sufficiently considered the broader landscape of risk, including business and market risk assessments made by those translating concepts into products. Decisions made about novel products from this perspective — from clinical trial to market entry — are based on different assumptions and priorities than those used by bioethicists and regulators, yet there is a distinct interaction between the two decision-making processes.

One way to deal with risk might be to incentivize trial sponsors themselves to become more reflective about risk and risk practices. Stimulated by financial collapses as much as technological and natural disasters, many organizations have become aware of how greatly high-visibility disasters might affect the welfare of the organization for the near and long term. Certainly this has proven to be the case in some areas of medicine, such as gene transfer research.

Risk Posed by Computerization of Public Infrastructure and Financial Markets

Munther Dahleh (2012), "The future power grid: Resilience and systemic risk (Powerpoint),"
2012 Cyber Physical Systems Virtual Organization PI Meeting.
http://cps-vo.org/file/7827/download/29453

Daron Acemoglu, Aso Ozdaglar, and Alireza Tahbaz-Salehi (30 Jun 2013), "The network origins of large economic downturns."
Working Paper 13-16, Department of Economics, Massachusetts Institute of Technology.
http://www.dklevine.com/archive/refs4786969000000000944.pdf

Abstract

This paper shows that large economic downturns may result from the propagation of microeconomic shocks over the input-output linkages across different firms or sectors within the economy. Building on the framework of Acemoglu et al. (2012), we argue that the economy's input-output structure can fundamentally reshape the distribution of aggregate output, increasing the likelihood of large downturns from infinitesimal to substantial. More specifically, we show that an economy with non-trivial intersectoral input-output linkages that is subject to thin-tailed productivity shocks may exhibit deep recessions as frequently as economies that are subject to heavy-tailed shocks. Moreover, we show that in the presence of input-output linkages, aggregate volatility is not necessarily a sufficient statistic for the likelihood of large downturns. Rather, depending on the shape of the distribution of the idiosyncratic shocks, different features of the economy's input-output network may be of first-order importance. Finally, our results establish that the effects of the economy's input-output structure and the nature of the idiosyncratic firm-level shocks on aggregate output are not separable, in the sense that the likelihood of large economic downturns is determined by the interplay between the two.

Daron Acemoglu, Aso Ozdaglar, and Alireza Tahbaz-Salehi (Feb 2015), "Systemic risk and stability in financial networks.'
American Economic Review 105.
http://economics.mit.edu/files/10433

Abstract

This paper argues that the extent of financial contagion exhibits a form of phase transition: as long as the magnitude of negative shocks affecting financial institutions are sufficiently small, a more densely connected financial network (corresponding to a more diversified pattern of interbank liabilities) enhances financial stability. However, beyond a certain point, dense interconnections serve as a mechanism for the propagation of shocks, leading to a more fragile financial system. Our results thus highlight that the same factors that contribute to resilience under certain conditions may function as significant sources of systemic risk under others.

Matteo Chinazzi and Giorgio Fagiolo (3 Jun 2015), "Systemic risk, contagion, and financial networks: A survey."

LEM Working Paper Series, No. 2013/08, Laboratory of Economics and Management (LEM), Sant'Anna School of Advanced Studies; Social Sciences Research Network.
http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2243504
http://www.econstor.eu/bitstream/10419/89353/1/740929119.pdf (Original version)

Abstract

The recent crisis has highlighted the crucial role that existing linkages among banks and financial institutions plays in channeling and amplifying shocks hitting the system. The structure and evolution of such web of linkages can be fruitfully characterized using concepts borrowed from the theory of (complex) networks. This paper critically surveys recent theoretical work that exploits this concept to explain the sources of contagion and systemic risk in financial markets. We taxonomize existing contributions according to the impact of network connectivity, bank heterogeneity, existing uncertainty in financial markets, portfolio composition of the banks. We end with a discussion of the most important challenges faced by theoretical network-based models of systemic risk. These include a better understanding of the causal links between network structure and the likelihood of systemic risk and increasingly using the empirical knowledge about real-world financial-network structures to calibrate theoretical models.

Risk Construction, Perception & Response

Richard Posner (2006), "Efficient responses to catastrophic risk."
Chicago Journal of International Law 6.
http://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=2895&context=journal_articles

Introduction

The Indian Ocean tsunami of December 2004 has focused attention on a type of disaster to which policymakers pay too little attention-a disaster that has a very low or unknown probability of occurring but that if it does occur creates enormous losses. Great as the death toll, physical and emotional suffering of survivors, and property damage caused by the tsunami were, even greater losses could be inflicted by other disasters of low (but not negligible) or unknown probability. The asteroid that exploded above Siberia in 1908 with the force of a hydrogen bomb might have killed millions of people had it exploded above a major city. Yet that asteroid was only about two hundred feet in diameter, and a much larger one (among the thousands of dangerously large asteroids in orbits that intersect the earth's orbit) could strike the earth and cause the total extinction of the human race through a combination of shock waves, fire, tsunamis, and blockage of sunlight wherever it struck. Other catastrophic risks, besides earthquakes such as the one that caused the recent tsunami, include natural epidemics (the 1918-1919 Spanish influenza epidemic killed between twenty and forty million people), nuclear or biological attacks by terrorists, certain types of lab accidents (including one discussed later in this Article), and abrupt global warming. The probability of catastrophes resulting, whether or not intentionally, from human activity appears to be increasing because of the rapidity and direction of technological advances. It is natural to suppose that the prediction, assessment,

Excerpt

Why, then, were such measures not taken in anticipation of a tsunami on the scale that occurred? Tsunamis are a common consequence of earthquakes, which are themselves common, and tsunamis can have causes besides earthquakes-a major asteroid strike in an ocean would create a tsunami that would dwarf the one in the Indian Ocean. Again, economics can yield some useful insights.

First, although a once-in-a-century event is as likely to occur at the beginning of the century as at any other time, it is much less likely to occur at some time in the first decade of the century than at some time in the last nine decades of the century. (The point is simply that the probability is greater the longer the interval being considered; one is more likely to catch a cold in the next year than in the next forty-eight hours.) Politicians with limited terms of office, and thus foreshortened political horizons, are likely to discount low risk disaster possibilities steeply since the risk of harm to their careers from failing to take precautionary measures is truncated.

Second, to the extent that effective precautions require governmental action, the fact that government is a centralized system of control makes it difficult for officials to respond to the full spectrum of possible risks against which cost-justified measures might be taken. Given the variety of matters to which they must attend, officials are likely to have a high threshold of attention below which risks are simply ignored.

Third, where risks are regional or global rather than local, many national governments, especially in poorer and smaller countries, may drag their heels in the hope of taking a free ride on richer and larger

countries. Knowing this, the latter countries may be reluctant to take precautionary measures that would reward and thus encourage free riding.

Fourth, countries are often poor because they are run by weak, inefficient, or corrupt governments, and such governments may disable poor nations from taking cost-justified precautions.

And fifth, the positive correlation of per capita income with value of life suggests that it is quite rational for even a well-governed poor country to devote proportionately less resources to averting calamities than a rich country.

Conclusion

To conclude, catastrophic risks-in the sense of low-probability events that if they occur will inflict catastrophic harm-are, despite their low probability, well worth the careful attention of policymakers. There are, however, a variety of psychological and political obstacles to such attention. In addition, there is a sense that the uncertainties surrounding catastrophic risks are so great as to make such risks analytically intractable. My purpose in this Article has been to contest that sense. There are a variety of useful analytical techniques for dealing with catastrophic risks; greater use of those techniques would enable a rational response to those risks.

Eliezer Yudkowsky (2008), "Cognitive biases potentially affecting judgment of global risks."

in *Global Catastrophic Risks*, ed. Nick Bostrom and Milan M. Ćirković, Oxford University Press.
https://intelligence.org/files/CognitiveBiases.pdf

Introduction

All else being equal, not many people would prefer to destroy the world. Even faceless corporations, meddling governments, reckless scientists, and other agents of doom require a world in which to achieve their goals of profit, order, tenure, or other villainies. If our extinction proceeds slowly enough to allow a moment of horrified realization, the doers of the deed will likely be quite taken aback on realizing that they have actually destroyed the world. Therefore I suggest that if the Earth is destroyed, it will probably be by mistake.

The systematic experimental study of reproducible errors of human reasoning, and what these errors reveal about underlying mental processes, is known as the heuristics and biases program in cognitive psychology. This program has made discoveries highly relevant to assessors of global catastrophic risks. Suppose you're worried about the risk of Substance P, an explosive of planet-wrecking potency which will detonate if exposed to a strong radio signal. Luckily there's a famous expert who discovered Substance P, spent the last thirty years working with it, and knows it better than anyone else in the world. You call up the expert and ask how strong the radio signal has to be. The expert replies that the critical threshold is probably around 4,000 terawatts. "Probably?" you query. "Can you give me a 98% confidence interval?" "Sure," replies the expert. "I'm 99% confident that the critical threshold is above 500 terawatts, and 99% confident that the threshold is below 80,000 terawatts." "What about 10 terawatts?" you ask. "Impossible," replies the expert.

The above methodology for expert elicitation looks perfectly reasonable, the sort of thing any competent practitioner might do when faced with such a problem. Indeed, this methodology was used in the Reactor Safety Study (U.S. NRC 1975), now widely regarded as the first major attempt at

probabilistic risk assessment. But the student of heuristics and biases will recognize at least two major mistakes in the method—not logical flaws, but conditions extremely susceptible to human error.

The heuristics and biases program has uncovered results that may startle and dismay the unaccustomed scholar. Some readers, first encountering the experimental results cited here, may sit up and say: “Is that really an experimental result? Are people really such poor guessers? Maybe the experiment was poorly designed, and the result would go away with such-and-such manipulation.” Lacking the space for exposition, I can only plead with the reader to consult the primary literature. The obvious manipulations have already been tried, and the results found to be robust.

Christopher Niemiec, et al. (Aug 2010), "Being present in the face of existential threat: The role of trait mindfulness in reducing defensive responses to mortality salience."

Journal of Personality and Social Psychology 99.
http://www.ncbi.nlm.nih.gov/pubmed/20658848
[Potentially useful in discussions of the influence of individual psychology on public policy re “existential threat,” such as policymakers’ feelings about mortality.]

Abstract

Terror management theory posits that people tend to respond defensively to reminders of death, including worldview defense, self-esteem striving, and suppression of death thoughts. Seven experiments examined whether trait mindfulness – a disposition characterized by receptive attention to present experience – reduced defensive responses to mortality salience (MS). Under MS, less mindful individuals showed higher worldview defense (Studies 1-3) and self-esteem striving (Study 5), yet more mindful individuals did not defend a constellation of values theoretically associated with mindfulness (Study 4). To explain these findings through proximal defense processes, Study 6 showed that more mindful individuals wrote about their death for a longer period of time, which partially mediated the inverse association between trait mindfulness and worldview defense. Study 7 demonstrated that trait mindfulness predicted less suppression of death thoughts immediately following MS. The discussion highlights the relevance of mindfulness to theories that emphasize the nature of conscious processing in understanding responses to threat.

Lisa Keränen (Oct 2011), "Concocting viral apocalypse: Catastrophic risk and the production of bio(in)security."

Western Journal of Communication 75.
http://www.tandfonline.com/doi/abs/10.1080/10570314.2011.614507

Abstract

The post-9/11 era featured an unprecedented expansion of global biodefense initiatives. This essay chronicles the rise of biodefense by tracking biological risk construction across political, scientific, and cultural rhetoric from the late 1990s to the present. It maintains that the production of bio(in)security entails two interlocking rhetorical operations—framing biological threats as catastrophic risk and enlisting the specter of viral apocalypse—that license technological solutions to imagined

vulnerabilities. The essay concludes by considering the implications of such rhetoric for public health and national security.

Conclusion: A Brave New World of Biorisks

The dominant critical read of the U.S.'s post-9=11 biodefense bonanza is that it represents a dangerous extension of the War on Terror into a technoscientific front that strips funding from crucial areas such as malaria, tuberculosis, and HIV=AIDS (Goldstein, 2003; Klotz & Sylvester, 2009). Supporters counter that because it extends public health response capacity, biodefense could potentially counter a host of naturally occurring outbreaks and lead to new medical advances (Palmquist, 2008).

Whether or not either or both of these claims bears out upon empirical scrutiny, this paper locates the biodefense buildup in a widespread vision of bio(in)security collectively produced through representations of catastrophic viral apocalypse that, in turn, licenses a proliferation of biological weapons agents in the name of biodefense. Indeed, a collection of experts from security circles, the pharmaceutical industry, the scientific community, citizen advocacy groups, international policy circles, and even Hollywood have—across a variety of political, technical, and cultural fronts—pushed the guiding notion of biological vulnerability that may in fact be promulgating bio(in)security in order to justify and perpetuate its existence. In short, while these elite decision-makers do not control the endless loop of Hollywood imagery and simulated confabulations that lodge the germ threat so firmly in the American psyche, they do confront such visions of viral apocalypse through a series of technological fixes that make germ work routine, and which sustain biodefense writ large. The rhetoric of biological threats as catastrophic risk that emerged out of the mid–late 1990s and intensified after the post-9=11 anthrax mailings thus signifies a reconfiguration of anxieties about emerging infectious disease to the realm of national security, encouraging a robust "biodefense." As necessary as protections from epidemic may be, this development nevertheless raises questions about the interlacing of national security and public health. It also raises questions about which health risks merit large-scale economic and cultural outlays. For instance, while acknowledged acts of bioterrorism killed fewer than 10 people in the last 100 years, cell phone–related distractions are responsible for 2,600 annual deaths and 333,000 accidents with moderate to severe injuries (Richtel, 2009). Routine medical errors kill tens of thousands of citizens each year, food-borne pathogens cause more than 76 million illnesses each year in the United States with 5,000 deaths (Institute of Medicine, 2009; Mead et al., 1999), while cancer and heart disease kill more than a million (Goldstein, 2003). Yet, concerns about bioterrorism and possible pandemic—more than the more mundane and regularly occurring killers—prompt large-scale funding and action; this imbalance is fueled, in part, through viral apocalyptic imaginations.

This essay represents but a beginning inspection of how naturally occurring germs and newly created biological agents are rising in prominence and symbolic power. Future investigations of the rhetorical constitution, deployment, and operation of perceived biological threats are needed. For instance, much work remains to account for the evolution of viral apocalypse as a rhetorical form that cuts across political, technical, and cultural domains. The visual imaginary of viral apocalypse in particular deserves scholarly scrutiny, as does the technical and public framing of biological risks across multiple time periods and contexts. Additionally, scholars should explore the meanings and consequences of the rhetoric of "public health security." Indeed, the implications of biodefensive activities for research ethics, genetic manipulation, health and safety, and global transparency and international relations remain to be seen (and operate often under the radar), but deserve intense discussion and scrutiny from scholars and broader global community. These are but a few of the projects that scholars in communication and rhetoric can undertake to help explain how—and with what effect—biorisks are being generated, understood, and activated in public and private life.

Citing Mitchell Dean, Scott (2006) maintains that scholars should analyze how changing conceptions of risk "become latched onto different political programmes and social imaginaries that invest them with a specific ethos" (Dean as cited in Scott, 2006, p. 120). Scott concludes his essay concerning 9=11, BigPharma, and bioterrorism with the "hope that others will join me in exploring rhetoric's interdependent and relative roles in the construction, functions, and effects of risk across global socio-political contexts" (p. 138). By supplying a preliminary rhetorical history and example of biocriticism (Kera¨nen, 2011a), this essay has attempted precisely that task. Harkening back to Hay and Andrejevic's (2006) notion of "homeland insecurities," it contributes a biological component to critical homeland security research and ends with an invitation for others to contribute to this emerging vein of scholarship. If, indeed, biological threats are multiplying both symbolically and materially, then rhetoricians and critical communication scholars can at the very least play a more significant part in explaining how biodefense might be reproducing the very bio(in)security that gives it meaning and power, hence generating a brave new world wherein biological weapons agents are normalized and awaiting further action.

Howard Kunreuther and Geoffrey Heal (Jun 2012), "Managing catastrophic risk."

NBER Working Paper No. 18136, National Bureau of Economic Research.
http://www.nber.org/papers/w18136

Abstract

A principal reason that losses from catastrophic risks have been increasing over time is that more individuals and firms are locating in harm's way while not taking appropriate protective measures. Several behavioural biases lead decision-makers not to invest in adaptation measures until after it is too late. In an interdependent world with no intervention by the public sector, it may be economically rational for those at risk not to invest in protective measures. Risk management strategies that involve private-public partnerships that address these issues may help in reducing future catastrophic losses. These may include multi-year insurance contracts, well-enforced regulations, third-party inspections, and alternative risk transfer instruments such as catastrophe bonds.

2. Why Decision-Makers Do Not Invest in Protective Measures

Budgeting Heuristics

The simplest explanation as to why decision-makers may fail to invest in protection is affordability. If the decision-maker has limited capital on hand, there may be little point in undertaking a benefit-cost analysis of whether to incur the upfront cost of investing in protection….

Safety-first Behavior

Decision-makers may utilize a simplified decision rule that determines whether to invest in protective measures only if the probability of the event (p) is above their threshold level of concern (p*). If the decision-makers perceives p < p*, then they will not undertake any protection. If, on the other hand, p >p* then they will want to invest in protection….

Under-weighing the Future

There is extensive experimental evidence revealing that human temporal discounting tends to be hyperbolic: temporally distant events are disproportionately discounted relative to immediate ones. As

an example, people are willing to pay more to have the timing of the receipt of a cash prize accelerated from tomorrow to today, than from the day after tomorrow to tomorrow (in both cases a one-day difference) (Loewenstein and Prelec, 1992)….

Myopic Behavior

An extreme form of hyperbolic discounting is when the decision-maker considers only the expected benefits from the protective measure over the next year or two, rather than over the life of the protective measure. Elected officials are likely to view the decision by reflecting on how their specific decisions will affect their chances of re-election. If the perceived expected benefits from the measure achieved before their next re-election campaign are less than the costs of protection, they will very likely oppose the expenditure. They will prefer to allocate funds where they can see an immediate return. The fact that protective measures yield positive returns only when a disaster occurs makes it even more difficult to justify these measures. This reluctance to incur upfront costs that do not yield immediate benefits highlights a NIMTOF (Not in My Term of Office) behavior….

Procrastination

The tendency to shy away from undertaking investments that abstractly seem worthwhile is exacerbated if individuals have the ability to postpone investments—something that is almost always the case with respect to protection. A community might recognize the need to invest in irrigation measures to reduce the consequences of a disaster but may still fail to act….

Underestimation of Risk

Another factor that has been shown to suppress investments in protection is underestimation of the likelihood of a hazard—formally, under-estimation of p in (1). For one thing, decisions about protection are rarely based on formal beliefs about probabilities. Magat, Viscusi and Huber (1987) and Camerer and Kunreuther (1989), for example, provide considerable empirical evidence that individuals do not seek out information on probabilities in making their decisions. In a study by Huber, Wider and Huber (1997), only 22 percent of subjects sought out probability information when evaluating risk managerial decisions. When asked to justify their decisions on purchasing warranties for products that may need repair, consumers rarely use probability as a rationale for purchasing this protection (Hogarth and Kunreuther, 1995)….

Howard Kunreuther, Paul Slovic, and Kimberly Giusti Olson (Aug 2014), "Fast and slow thinking in the face of catastrophic risk."

Working Paper No. 2014-06, Risk Management and Decision Processes Center, The Wharton School, University of Pennsylvania.
http://opim.wharton.upenn.edu/risk/library/WP201406_Fast-and-Slow-Thinking-in-Catastrophes-HK-PS-KGO.pdf

Abstract

Studies of behavior in the face of natural disasters and mass atrocities provide common lessons about managing catastrophic threats. We cannot assume that the massive destructiveness of an event will lead us to appreciate and appropriately respond to the risk. The potential consequences, whether in billions of dollars or millions of endangered lives, often fail to convey the emotional meaning necessary to

motivate effective protective actions. Rather than trusting our desensitized feelings as our moral compass, we must employ slow and careful thinking, coupled with short-term incentives, to create policies, procedures, laws, and institutions that will nudge or even require us to behave in ways that accord with our considered values for protecting human lives and property.

Intuitive and Deliberative Thinking

A large body of cognitive psychology and empirical research during the past 30 years has revealed that individuals, small groups, and organizations often make decisions by employing a blend of intuitive and deliberative thought processes. In his thought-provoking book, Thinking, Fast and Slow, Nobel Laureate Daniel Kahneman characterized these two modes of thinking as System 1 and System 2, building on a large body of cognitive psychology and behavioral decision research as depicted in Table 1.

System 1 thinking tends to be fast and effortless. System 2 thinking requires more time and attention. Intuitive thinking works well for routine decisions but can be problematic for low-probability high-consequence events when there is limited opportunity to learn from personal experience and when the consequences are likely to occur far into the future. With respect to floods and hurricanes, individuals are likely to discount the impact climate change will have on sea level rise and future damage. With respect to genocide, the atrocities involve large numbers of faceless people in distant places, whose actual or predicted deaths fail to spark the feelings and emotions necessary to motivate protective action.

This paper documents how fast intuitive thinking causes underreaction to natural disasters and mass atrocities. Informed by these descriptive insights, we then argue for prescriptive measures that reflect trade-offs informed by slow and deliberative thinking.

From Natural Disasters to Mass Atrocities: Common Lessons

Natural disasters and mass atrocities, though different in many respects, share similar features for understanding behavior and its implications for preventing catastrophic losses. For both these risks, intuition may mislead us. Specifically, we cannot assume that the massive destructiveness of a future event will lead us to appreciate and appropriately respond to the threat. The low probability of natural disasters at any one place or time leads us to treat the event as below our threshold level of concern much as the geographical remoteness of atrocities does. The potential consequences, whether in billions of dollars or millions of endangered lives, fail to convey the emotional meaning necessary to motivate effective protective actions.

Fortunately, the modern brain is able to deliberate, to think slowly and analytically when appropriately motivated, and thus recognize the enormity of these catastrophic events and the importance of taking immediate steps to prevent or reduce the consequences of future catastrophes. Rather than trusting our desensitized feelings or simplistic heuristics, we must employ slow and careful thinking coupled with short-term incentives to create policies, procedures, laws, and institutions that will nudge or even require us to behave in ways that accord with our considered values for protecting human lives and property.

Grant Wilson (2013), "Minimizing global catastrophic and existential risks from emerging technologies through international law."

Virginia Environmental Law Journal 31.
http://lib.law.virginia.edu/lawjournals/sites/lawjournals/files/3.%20Wilson%20-%20Emerging%20Technologies.pdf

Abstract

Mankind is rapidly developing "emerging technologies" in the fields of bioengineering, nanotechnology, and artificial intelligence that have the potential to solve humanity's biggest problems, such as curing all disease, extending human life, or mitigating massive environmental problems like climate change. However, if these emerging technologies are misused or have an unintended negative effect, the consequences could be enormous, potentially resulting in serious, global damage to humans (known as "global catastrophic harm") or severe, permanent damage to the Earth—including, possibly, human extinction (known as "existential harm"). The chances of a global catastrophic risk or existential risk actually materializing are relatively low, but mankind should be careful when a losing gamble means massive human death and irreversible harm to the planet. While international law has become an important source of global regulation for other global risks like climate change and biodiversity loss, emerging technologies do not fall neatly within existing international regimes, and thus any country is more or less free to develop these potentially dangerous technologies without practical safeguards that would curtail the risk of a catastrophic event. In light of these problems, this paper serves to discuss the risks associated with bioengineering, nanotechnology, and artificial intelligence; review the potential of existing international law to regulate these emerging technologies; and propose an international regulatory regime that would put the international world in charge of ensuring that low-probability, high-risk disasters never materialize.

Conclusion

A series of fantastical scientific breakthroughs are leading toward or, in some instances, have already created technologies that question basic premises of life: that man cannot create life, that humans are the ultimate intelligent being, or that we are limited by the basic building blocks we find on Earth. Nanotechnology, bioengineering, and AI offer great benefits to society, but they also have the potential to cause global catastrophic or even existential harm to humans. While bioengineering has caused a revolution in crop production, genetically engineered viruses have the potential to cause global devastation if accidentally or purposefully released. Nanotechnology has yielded stronger and lighter materials, yet nanomaterials also pose unknown human and animal health effects, and weapons developed from advanced nanotechnology could be far more destructive and concealable than nuclear bombs. And while AI could innovate every technology on the planet, a superintelligent machine could outcompete humans or be programmed to act maliciously.

While the chances of massive destruction from these technologies are not high, states should still act quickly to create a flexible, binding international treaty that limits GCRs/ERs arising from emerging technologies to a degree that society deems acceptable. As this paper demonstrates, emerging technologies do not fall squarely within current international law, and allowing a small group of self-interested scientists to regulate themselves is undesirable when a single misstep could result in global catastrophic or existential harm. Instead, the international community, with the guidance of a body of experts representing a wide range of interests and strong considerations of the precautionary principle, should develop a binding framework to regulate emerging technologies at the international level. Furthermore, because emerging technologies will likely affect the entire world, society should help

determine which risks they are willing to take and what moral, ethical, and other beliefs should influence an international regulatory regime. If the international community successfully concludes a treaty on GCRs/ERs from emerging technologies, then perhaps society can thrive in an age of technological innovation without suffering from the associated risks.

World Economic Forum (12 Jan 2015), "Global Risks 2015."

http://reports.weforum.org/global-risks-2015
[Includes results of WEF's *Risk Perception Survey*, in charts]

Box 1.1: The evolution of the risks of highest impact/likelihood

As the report's 10th anniversary approaches, the evolution of the perceived top five global risks can be viewed in terms of impact and likelihood as documented in the Global Risks reports from 2007 to 2015. As Table 1.1.1 shows, economic risks largely dominated from 2007 to 2014, with the risk of an asset-price collapse heading the list in the run-up to the financial crisis, giving way to concerns about the more immediate but slow-burning consequences of constrained fiscal finances, a major systemic financial failure in the immediate post-crisis years, and income disparity. This year features a radical departure from the past decade; for the first time in the report's history, economic risks feature only marginally in the top five. In the 25th year after the fall of the Berlin Wall, geopolitical risks are back on the agenda. The dispute over Crimea in March 2014 serves as a forceful reminder of the consequences of interstate conflicts with regional consequences that seemed long forgotten and unfathomable, as further explored in this report. Similarly, together with other events in 2014, such as the prominent rise of the Islamic State, it has brought state collapse and the failure of national governance back into public consciousness. At the same time, health-related risks, such as pandemics – last considered impactful in 2008 – have made it back into the unglamorous top, following the unprecedented spread of Ebola.

On a higher level, Table 1.1.1 also indicates a shift over past years away from economic risks in general to environmental risks – ranging from climate change to water crises. While this highlights a recognition of the importance of these slow-burning issues, strikingly little progress has been made to address them in light of their far-reaching and detrimental consequences for this and future generations.

Table 1.1.1: The Evolving Risks Landscape (2007-2015) [see paper for table]

Ian Martin and Robert Pindyck (Jun 2014; revised Apr 2015), "Averting catastrophes: The strange economics of Scylla and Charybdis."

NBER Working Paper No. 20215, National Bureau of Economic Research.
http://www.nber.org/papers/w20215
http://personal.lse.ac.uk/martiniw/AvoidCatastrophesApril2015.pdf

Abstract

Faced with numerous potential catastrophes—nuclear and bioterrorism, "megaviruses," climate change, and others—which should society attempt to avert? A policy to avert one catastrophe considered in isolation might be evaluated in cost-benefit terms. But because society faces multiple catastrophes, simple cost-benefit analysis fails: Even if the benefit of averting each one exceeds the cost, we should

not necessarily avert them all. We explore the policy interdependence of catastrophic events, and develop a rule for determining which catastrophes should be averted and which should not.

Excerpt

For instance, one apparently sensible response to the non-marginal nature of large catastrophes is to decide which is the most serious catastrophe, avert that, and then decide whether to avert other catastrophes. This approach is intuitive and plausible—and wrong. We illustrate this in an example with three potential catastrophes. The first has a benefit w1 much greater than the cost t1, and the other two have benefits greater than the costs, but not that much greater. Naive reasoning suggests we should proceed sequentially: eliminate the first catastrophe and then decide whether to eliminate the other two, but we show that such reasoning is flawed. If only one of the three were to be eliminated, we should indeed choose the first; and we would do even better by eliminating all three. But we would do best of all by eliminating the second and third and not the first: the presence of the second and third catastrophes makes it suboptimal to eliminate the first.

In the next section we use two very simple examples to illustrate the general interdependence of large projects, and show why, if faced with two potential catastrophes, it might not be optimal to avert both, even if the benefit of averting each exceeds the cost. In Section 2 we introduce our framework of analysis by first focusing on the WTP to avert a single type of catastrophe (e.g., nuclear terrorism) considered in isolation. We use a constant relative risk aversion (CRRA) utility function to measure the welfare accruing from a consumption stream, and we assume that the catastrophe arrives as a Poisson event with known mean arrival rate; thus catastrophes occur repeatedly and are homogeneous in time. Each time a catastrophe occurs, consumption is reduced by a random fraction.4 These simplifying assumptions make our model tractable, because they imply that the WTP to avoid a given type of catastrophe is constant over time.

This tractability is critical when, in Section 3, we allow for multiple types of catastrophes. Each type has its own mean arrival rate and impact distribution. We find the WTP to eliminate a single type of catastrophe and show how it depends on the existence of other types, and we also find the WTP to eliminate several types at once. We show that the presence of multiple catastrophes may make it less desirable to try to mitigate some catastrophes for which action would appear desirable, considered in isolation. Next, given information on the cost of eliminating (or reducing the likelihood of) each type of catastrophe, we show how to find the welfare-maximizing combination of projects that should be undertaken.

Section 4 presents some extensions. First, we show that our framework allows for the partial alleviation of catastrophes, i.e., for policies that reduce the likelihood of catastrophes occurring rather than eliminating them completely. The paper's central intuitions apply even if we can choose the amount by which we reduce the arrival rate of each catastrophe optimally. Second, our framework easily handles catastrophes that are directly related to one another: for example, averting nuclear terrorism might also help avert bioterrorism. Third, our results also apply to bonanzas, that is, to projects such as blue-sky research that increase the probability of events that raise consumption (as opposed to decreasing the probability of events that lower consumption).

Conclusions

How should economists evaluate projects or policies to avert major catastrophes? We have shown that if society faces more than just one catastrophe (which it surely does), conventional cost-benefit analysis breaks down; if applied to each catastrophe in isolation, it can lead to policies that are far from optimal. The reason is that the costs and benefits of averting a catastrophe are not marginal, in that they have

significant impacts on total consumption. This creates an interdependence among the projects that must be taken into account when formulating policy. In fact, as we demonstrated in Example 1, cost-benefit analysis can even fail when applied to small catastrophes if they have a non-marginal aggregate impact.

Using WTP to measure benefits and a permanent tax on consumption as the measure of cost (both a percentage of consumption), we derived a decision rule (Result 2) to determine the optimal set of catastrophes that should be averted. And we have shown that this decision rule can yield “strange” results. For instance, as we demonstrated in Example 3, although naive reasoning would suggest using a sequential decision rule (e.g., avert the catastrophe with the largest benefit/cost ratio, then decide on the one with the next-largest ratio, etc.), such a rule is not optimal. In general, in fact, there is no simple decision rule. Instead, determining the optimal policy requires evaluating the objective function (16) of Result 2 for every possible combination of catastrophes. In a strong sense, then, the policy implications of different catastrophe types are inextricably intertwined.

Given that the complete elimination of some catastrophes may be impossible or prohibitively expensive, a more realistic alternative may be to reduce the likelihood that the catastrophe will occur, i.e., reduce the Poisson arrival rate ?. We have shown how that alternative can easily be handled in our framework. In the previous section we examined the costs and benefits of completely averting seven catastrophes, but we could have just as easily considered projects to reduce the likelihood of each, and given the amounts of reduction and corresponding costs, determined the optimal set of projects to be undertaken.

The theory we have presented is quite clear. (We hope most readers will agree.) But there remain important challenges when applying it as a tool for government policy, as should be evident from Section 5. First, one must identify all of the relevant potential catastrophes; we identified seven, but there might be others. Second, for each potential catastrophe, one must estimate the mean arrival rate ?i , and the probability distribution for the impact fi . Finally, one must estimate the cost of averting or alleviating the catastrophe, which we expressed as a permanent tax on consumption at the rate ti . As we explained, for some catastrophes (floods, storms and earthquakes), a relatively large amount of data are available. But for others (nuclear and bio-terrorism, or a mega-virus), estimates of ?i , ßi and ti are likely to be subjective and perhaps speculative. On the other hand, one can use our framework to determine optimal policies for ranges of parameter values, and thereby determine which parameters are particularly critical, and should be the focus of research.

The Concept of “Catastrophic Terrorism”

Ashton Carter, John Deutch, and Philip Zelikow (Nov/Dec 1998), "Catastrophic terrorism: Tackling the new danger."
Foreign Affairs.
http://www.foreignaffairs.com/articles/54602/ashton-b-carter-john-deutch-and-philip-zelikow/catastrophic-terrorism-tackling-the-new-danger

Introduction

Terrorism is not a new phenomenon. But today's terrorists, be they international cults like Aum Shinrikyo or individual nihilists like the Unabomber, act on a greater variety of motives than ever before. More ominously, terrorists may gain access to weapons of mass destruction, including nuclear devices, germ dispensers, poison gas weapons, and even computer viruses. Also new is the world's dependence on a nearly invisible and fragile network for distributing energy and information. Long part of the Hollywood and Tom Clancy repertory of nightmarish scenarios, catastrophic terrorism has moved from far-fetched horror to a contingency that could happen next month. Although the United States still takes conventional terrorism seriously, as demonstrated by the response to the attacks on its embassies in Kenya and Tanzania in August, it is not yet prepared for the new threat of catastrophic terrorism.

American military superiority on the conventional battlefield pushes its adversaries toward unconventional alternatives. The United States has already destroyed one facility in Sudan in its attempt to target chemical weapons. Russia, storehouse of tens of thousands of weapons and material to make tens of thousands more, may be descending into turmoil. Meanwhile, the combination of new technology and lethal force has made biological weapons at least as deadly as chemical and nuclear alternatives. Technology is more accessible, and society is more vulnerable. Elaborate international networks have developed among organized criminals, drug traffickers, arms dealers, and money launderers, creating an infrastructure for catastrophic terrorism around the world.

The bombings in East Africa killed hundreds. A successful attack with weapons of mass destruction could certainly take thousands, or tens of thousands, of lives. If the device that exploded in 1993 under the World Trade Center had been nuclear, or had effectively dispersed a deadly pathogen, the resulting horror and chaos would have exceeded our ability to describe it. Such an act of catastrophic terrorism would be a watershed event in American history. It could involve loss of life and property unprecedented in peacetime and undermine America's fundamental sense of security, as did the Soviet atomic bomb test in 1949. Like Pearl Harbor, this event would divide our past and future into a before and after. The United States might respond with draconian measures, scaling back civil liberties, allowing wider surveillance of citizens, detention of suspects, and use of deadly force. More violence could follow, either further terrorist attacks or U.S. counterattacks. Belatedly, Americans would judge their leaders negligent for not addressing terrorism more urgently.

The danger of weapons of mass destruction being used against America and its allies is greater now than at any time since the Cuban missile crisis of 1962. It is a national security problem that deserves the kind of attention the Defense Department devotes to threats of military nuclear attack or regional aggression. The first obstacle to imagination is resignation. The prospects may seem so dreadful that some officials despair of doing anything useful. Some are fatalistic, as if contemplating the possibility of a supernova. Many thinkers reacted the same way at the dawn of the nuclear age, expecting doom to

strike at any hour and disavowing any further interest in deterrence as a hopeless venture. But as with nuclear deterrence, the good news is that more can be done.

Stephen Fidler (2002), "Catastrophic terrorism."

in *Catastrophic terrorism*, Report of the Meeting organised by the Centre of International Studies University of Cambridge November 18 –19, 2002.
http://www.nti.org/media/pdfs/catastrophic_terrorism.pdf?_=1316466791

Excerpt

The nature of the threat

THE SEPTEMBER 11 attacks were so devastating in part because modern societies are so highly tuned and depend so much on the smooth functioning of many separate parts. It is not just that we are dependent on technologies that most of us do not understand, but we are dependent on their working almost optimally. Indeed, it does not take a catastrophic terrorist strike to bring modern economies, or important parts of them, to a halt. Just-in-time stocking patterns, for example, have helped to increase productivity, but when they break down – as they did during the UK fuel strike of 2000 – the impact is huge. “We are creating systems and structures that are vulnerable to fairly simple acts,” said Lord Wilson, now Master of Emmanuel College at Cambridge University. It raises the question of whether and how we can make our societies more robust.

In this kind of environment, making predictions is hazardous. Even more than is usual, very different outcomes can depend on chance events: the discovery of smuggled fissile material, the capture of a terrorist before he boards an aircraft or as he crosses a border. It places a premium on good intelligence, but as the volume of raw intelligence data grows rapidly, it increases the chance that important information is overlooked.

Extrapolating from the past is unlikely therefore to provide much help in projecting the future: there are too many “wild cards” in the pack. “I don’t think this is straight line territory. This is going around corners,” Wilson said.

Given our highly-strung economies, it may be impossible to focus in advance on all the methods that terrorists could use to cause a catastrophe. The willingness of religious zealots to commit suicide is clearly an effective delivery mechanism. Jet fuel has already been turned into a means of mass destruction, and airliners could generate even greater terror if they were to be guided into nuclear installations.

Yet, there is evidence of Al Qaeda’s intention to acquire weapons of mass destruction – chemical, biological and nuclear weapons, and, for the purposes of this discussion, radiological or “dirty” bombs. Though by no means the only one, this is a cause of real concern.

The most likely source of such weapons, the raw material and the expertise to make them are the states of the former Soviet Union, where security at many important sites is questionable. Cooperative threat reduction programmes have been directed at addressing the risks they pose, by the financing of proper security for dangerous materials and legitimate programmes for scientists to reduce the temptation to sell their services to rogue actors. Originally US-financed, these initiatives now have the financial support of the Group of Eight countries and have been broadened to include other states that present

proliferation risks. Unfortunately, the US Congress is showing signs of reluctance to continue funding this programme.

Each type of weapon, in fact, poses a different risk, both in terms of the ease with which technologies could be spread to terrorists and of the potential destruction should they be used. At the bottom of the scale of destruction, the deployment of chemical and radiological devices is unlikely to be able to kill more than a few hundred people, though both represent terrorist weapons par excellence in the sense they would provoke widespread fear and panic if used in an urban environment. It is not easy, though, to develop a weapon that would effectively disperse a chemical agent in a city. Less complicated may be a radiological bomb, in which conventional explosive is packed around low-grade radioactive material, of which supplies are relatively plentiful. Experts are divided on how effectively such a device would disperse radioactivity, but it is conceivable that one could render several city blocks unlivable or deny access to symbolic buildings or monuments for generations.

More worrying still is the potential for biological attack, a risk that unfortunately seems likely to grow over time into the biggest threat. The techniques used in making biological weapons are spreading as the bio-technology industry grows around the world. The manufacture of bio-weapons, unlike nuclear weapons, can be relatively simple and easy to hide. Biological attacks could have a wide range of effects, depending in part on whether the agent is infectious. The distribution of anthrax through the mail in the US in late 2001 showed how easily a relatively unsophisticated delivery mechanism could generate panic, but it resulted in only five deaths and 23 suspected cases.

The spreading of smallpox into a largely unvaccinated population would be a different matter, infecting perhaps 80 per cent of those who came into contact with it and killing a third of them. However, since the eradication of smallpox in its natural state in 1979, the distribution of this agent has been officially limited to certain laboratories in the US and Russia. Only if it has leaked from those laboratories or if other governments kept secret reserves of smallpox could it get into the hands of terrorists. Any smallpox attack would also risk backfiring on the organisation responsible, since it would be impossible to target. Because of its extreme contagion, the disease would probably cause as much devastation in poor and Muslim countries as in rich ones, and probably more. Nonetheless, Western governments are preparing for this remote eventuality by buying smallpox vaccines.

Perhaps of greatest concern is the possibility that some kind of nuclear weapon could fall into the hands of terrorists. The difficulties of manufacturing fissile material are such that they can only be made undetected by states, or under state auspices. Yet the protection afforded weapons-grade nuclear materials, particularly in the states of the former Soviet Union, leaves much to be desired. Though strategic nuclear weapons are relatively secure, tactical nuclear weapons have never been accounted for. Fissile material outside weapons is a source of even more concern: Russia is said to have produced 1,200 tonnes of highly-enriched uranium and 200 tonnes of plutonium. In 2000, it was estimated that just 40 per cent of this material in Russia had been secured. Fifty-eight countries around the world have research reactors. In many of them, criminal syndicates operate.

Once highly-enriched uranium has been obtained – making a plutonium bomb is more difficult – constructing a crude gun-type nuclear device would be relatively simple, particularly if the services of an out-of-work nuclear scientist could be called upon. But it is not the threat of one nuclear bomb that is the greatest concern, though that prospect is horrific enough in terms of loss of life, it is the prospect that a successful nuclear attack could be followed by blackmail and the threat that others would be detonated. The probability of a successful nuclear attack by terrorists is low, and it is not even certain that the presumption the group would use nuclear weapons if they had them is correct. Since 1945, possession of nuclear weapons has been useful to states, but the states that have them have not found

it in their interests to use them. Yet, given the risk associated with any event is equivalent to the product of the probability of its occurring and its consequences (Risk = Probability x Consequences), it is an eventuality policy makers have to take seriously.

James Fearon (9 Oct 2003), "Catastrophic terrorism and civil liberties in the short and long run."

Symposium on Constitutions, Democracy, and the Rule of Law, Columbia University, October 17, 2003.
https://web.stanford.edu/group/fearon-research/cgi-bin/wordpress/wp-content/uploads/2013/10/Catastrophic-terrorism-and-civil-liberties-in-the-short-and-long-run.pdf

Excerpt

Three models for responding to the threat of catastrophic terrorism

So let’s visit again this scary future world in which individuals or small groups can fairly easily acquire the means to kill thousands. Civil liberties in the sense we have known them are inconceivable in such a world. It would be simply insane if, in this world, government did not have the power to undertake secret investigations of individuals and groups that give off warning signs, if government did not have the power to collect information about what individuals were doing on the basis of mere suspicions or indicators that correlate with a disposition to undertake terrorism. I think there could also be a strong case, in such a world, for extending the requirement of security clearances to people who acquire forms of knowledge that could easily be used for mass destruction, particularly in biochemistry and nuclear engineering. There is no denying that such a system could have major costs in terms of freedom of inquiry and the social benefits this freedom brings.

Two common analogies people use to think about the threat of terrorism suggest that such restrictive and anti-liberal measures would not be necessary. But both analogies are highly misleading applied to the problem of catastrophic terrorism as it is likely to develop in the coming decades.

The first analogy might be called the war model, and is constantly invoked by President Bush and other politicians in the phrase “the war on terrorism.” As many have pointed out, in a conventional war you know who the enemy is and where to find his forces. The central problem in war is to figure out when and how best to attack. This is the enemy is and where they are, before they strike. Once you know these things, “attacking” is relatively easy because of the huge disparity in power between the state and terrorists groups (once you know who and where they are).

The second model might be called the crime model, which wants to think about terrorism as a conventional problem of crime that our current criminal justice system can and should deal with. With ordinary crime, the idea is that the police investigate crimes after they take place, producing a high enough probability of apprehension to deter prospective criminals, so keeping the crime rate at tolerably low levels.2

Ordinary crime is indeed more analogous to terrorism than is interstate war. For example, the enemy in a war can be defeated, whereas both crime and terrorism are more like “technological” problems that can be reduced but never completely eliminated. All in all, however, crime is also misleading as a template for thinking about terrorism.

In the first place, with catastrophic terrorism we are talking about crimes so enormous and socially devastating – a nuclear bomb in New York City, for example – that apprehending the criminal after the

fact is cold comfort. The costs of such an attack are just too high. Almost all of our effort must go into preventing it from happening at all, rather than tracking down the culprits ex post. Liberals (and I count myself among them) are just sticking our heads in the sand if we go around exclaiming that “We must not compromise civil liberties!” without ever engaging the problem posed by the necessity of preventing this sort of crime (catastrophic terrorism) ex ante rather than ex post.

The second problem with the crime model is that if terrorists are willing to die to commit these crimes, or if an individual or group can reasonably hope to strike and escape detection or punishment, then deterrence does not operate. Of course, we should spend serious resources on tracking down the culprits after the fact, to get whatever value we can out of deterrence and the disruption of future operations. But we can’t count on deterrence to prevent catastrophic terrorism, the way that it makes sense, arguably, to depend on mutual deterrence in relations between states with nuclear weapons.

A lawyer might point out that our criminal law actually does have resources to investigate and prosecute crimes that have not yet been committed, under the heading of “conspiracy.” 3 This is true, and this is the part of the law that needs to debated,developed, and used as we increasingly face the long-run threat of catastrophic terrorism. It is also true, however, that defining the standards for investigating and prosecuting “conspiracy” are problematic on civil liberties grounds. This is precisely the part of the law that has been most abused in past U.S. episodes of fear of invisible attackers (as in the persecution of alleged Communist Party members after both world wars). What constitute effective procedures for investigating possible conspiracies to commit terrorist acts that are maximally consistent with the Bill of Rights is really the crux of the civil liberties problem we now face. More on this below.

There is a third way of thinking about terrorism and how to respond to it, which I will call the counterinsurgency model. If we define terrorism as violent attacks on noncombatants intended to coerce an opponent or influence third parties, then the vast majority of the world’s terrorism takes place not in the U.S., in developed countries, or even in Israel, but rather in what are mainly very poor countries wracked by civil war. Since the end of World War II, long-running civil wars have become remarkably common, and they have been concentrated in the world’s poorest countries.4 With few exceptions, these have been rural guerrilla wars characterized by the techniques of insurgency and counterinsurgency. In guerrilla war, civilians are routinely targeted by both rebels and government forces. Especially in poor countries, both sides tend to use collective punishment. Government forces massacre whole villages in attempts to “drain the sea” (in which the guerrilla “fish” swim) and to dissuade other villages from giving passive or active support to rebels. Likewise, rebel forces often massacre villages whose members are suspected of helping the government; they do this largely to influence other villages.5 Terrorist attacks in Northern Ireland, Basque Country, or Israel have killed dozens in the worst cases. Terrorist attacks in civil wars in Angola, Algeria, Sudan, Indonesia (East Timor and Aceh), Peru, Sri Lanka, and Sierra Leone – to name just a few poor-country civil wars – have killed hundreds and sometimes thousands.

The problem we face in responding to the threat of catastrophic terrorism is structurally similar to the problem faced by a state trying to run a counterinsurgency. How to distinguish the active rebels from the rest of the population in which they are hiding? If you arrest or kill the wrong people, or if you abuse the civil liberties of many to get very few, you may actually worsen things by increasing support for the guerrillas.6 But if you fail to prevent guerrilla attacks and assassinations before they happen (deterrence being problematic because of the low odds of ex post apprehension and the high motivation of the attackers), then you lose control of territory and tax revenue, while also suffering the political consequences of severe public displeasure.

How have states fared in confronting this dilemma? For the most part, I would say “terribly.” The problem of counterinsurgency is extremely difficult, if not intractable. In poor countries with administratively weak, badly financed states, the state frequently adopts scorched-earth tactics that end up being counterproductive and at any rate morally horrific. In wealthier or more democratic states, such as Britain in Northern Ireland and in the Malayan Emergency, Spain in Pais Vasco, Israel in the Occupied Territories, or India in the Punjab (“Khalistan”), mass killing has largely been avoided but civil liberties greatly abused, often in quite bloody ways.

Based on what I have read about insurgency and counterinsurgency since the end of World War II, I believe that the following two, slightly contradictory propositions are both true:

(A) It is highly unlikely that a state can defeat or minimize an insurgency without committing significant abuses of civil liberties and human rights. These will include legal changes that give the state powers of detention and investigation that go well beyond what is necessary to counter ordinary crime. Counterinsurgency always appears to be a messy business. Most likely this reflects the nature of the problem it confronts, as opposed to resulting solely from government stupidity.

(B) At the same time, abuses of civil liberties and human rights by state forces can hurt rather than help counterinsurgency, by increasing support for the rebels while undermining support for the government.

In sum, I see two arguments in favor of the view that in the long run, civil liberties as we have know them will have to be compromised, or significantly altered in their implementation, in the face of the growing threat of catastrophic terrorism. The first argument is basically from common sense: As technological developments make it increasingly easy for small groups or individuals to cause incredible devastation, it becomes more likely that some individual or group will try, even if the vast majority views government policies as largely benign. The only feasible way to counter this threat is by increasing the ability of government to monitor and investigate individual behavior, however distasteful this may be.7

The second argument is from the experience of insurgency and counterinsurgency, which I have argued provides the closest analogy for thinking about the strategic problem posed by catastrophic terrorism. Quite possibly all counterinsurgency efforts in the past 50 years have depended on major changes in laws that worked against civil liberties and often led to human rights abuses. While these changes have often been pushed to the point of being dysfunctional, successes in defeating guerrilla rebels have often clearly depended on the state’s ability to detain and investigate suspects who had not yet, or not clearly, committed any crime.

Hamid Mohtadi and Antu Murshid (Mar 2009), "The risk of catastrophic terrorism: an extreme value approach."

MPRA Paper No. 25738, Munich Personal RePEc Archive.
http://mpra.ub.uni-muenchen.de/25738/1/MPRA_paper_25738.pdf

Abstract

This paper models the stochastic behavior of large-scale terrorism using extreme value methods. We utilize a unique dataset composed of roughly 26,000 observations. These data provide a rich description of domestic and international terrorism between 1968 and 2006. Currently, a credible worst-case scenario would involve losses of about 5000 to 10,000 lives. Also, the return time for events of such magnitude is shortening every year. Today, the primary threat is from conventional weapons, rather

than from chemical, biological and/or radionuclear weapons. However, pronounced tails in the distribution of these incidents suggest that this threat cannot be dismissed.

Conclusions

In this paper, we analyzed the risks of catastrophic terrorism using a unique dataset gathered from the internet and various other primary sources. Our results suggest that currently, a credible worst-case scenario is one that involves the loss of between 5000 to 10,000 lives on a single day. However, this conclusion is sensitive to the form of terrorism. The threat of CBRn-terrorism for instance is very different from that posed by conventional attacks. Our analysis reveals that CBRn terrorism is more likely to cause injuries, as opposed to loss of life. Although by this metric, the risk can be significant.

In interpreting our results, it is important to recognize that risks are continually evolving: the distribution underlying catastrophic terrorism is unstable. Over the last forty years this instability has manifested itself in two ways. First, the right tail of the distribution has got heavier. This has been accompanied by an increase in positive skewness, i.e. a redistribution of the probability mass into a higher range of values. Second, the scale of the distribution has increased dramatically.

These developments are consistent with an overall pattern of change beginning in the late 1970s, with the emergence of radical terrorist organizations, and continuing through to present day. It seems that earlier models of terrorism, which focused on maximizing disruption, have given way to new forms of terrorism in which the metric for success is the number of fatalities. Yet, there should be no presumption that this new paradigm represents the future of terrorism. If, for instance, the social and political causes for the revival of Islamic fundamentalism were to erode, probability laws governing the distribution of terrorism today will be of little significance for understanding future risks. It is critical therefore, to identify potential determinants of the distribution of large-scale terrorist attacks. However this is not simply to establish links between future risks and specific future outcomes. At issue is also the accuracy of current forecasts. These are affected by our ability to disentangle that variation in our data, which is due to structural breaks in the distribution, from that, which is due to the distribution itself.

Since the risks associated with catastrophic terrorism are in continual flux, risk assessments must be part of an ongoing effort. In assessing these risks it is important that we take a pragmatic approach which weighs model forecasts against data from other sources relevant for the future of terrorism risk.

Chin-Kuei Tsui (Jan 2015), "Framing the threat of catastrophic terrorism: Genealogy, discourse and President Clinton’s counterterrorism approach."

International Politics 52.
http://www.palgrave-journals.com/ip/journal/v52/n1/full/ip201436a.html

Abstract

A frequent argument in the literature on the US-led war on terror is that the war and its public discourse originated with the George W. Bush administration. This article seeks to explore the political discourse of terrorism and counterterrorism practices during the Clinton administration in order to challenge this perspective. By examining US administration discourses of terrorism, this article demonstrates deep continuities in counterterrorism approaches from Ronald Reagan to Bill Clinton, through to George W. Bush. The research suggests that, based on Reagan’s initial ‘war on terrorism’ discourse, Clinton articulated the notion of ‘catastrophic terrorism’ or ‘new terrorism’, which became a formative

conception for the United States and its allies in the post-Cold War era. Clinton’s counterterrorism discourse then provided an important rhetorical foundation for President Bush to respond to the 2001 terrorist attacks. In other words, far from being a radical break, Bush’s ‘war on terror’ represents a continuation of established counter-terrorist understanding and practice.

Introduction

The central core of terrorism and counterterrorism discourse is the interpretation of threat, danger and uncertainty. Political elites also emphasize, and frequently claim, that terrorist violence is sudden, dramatic and threatening, thus requiring urgent action. However, some would question whether the threat posed by terrorism really is as dangerous as officials assert. It is argued that the danger and threat stressed by politicians is not actually an objective condition; instead, it is defined, articulated and socially constructed by authorized actors (Campbell, 1998, pp. 1–2). Specifically, danger and threat are not things that exist independently; rather, they become ‘reality’ by the way in which people analyze them and consider them to be urgent and imminent. In other words, our perception of threats, crises and risks is introduced through a series of interpretations, and as a result, is largely a product of social construction. In Foucault’s (1980, 2002) terms, the interpretation of terrorist threat constitutes the knowledge of terrorism and sustains a counterterrorism ‘regime of truth’1 that defines what can be meaningfully said and discussed about the subject. With regard to the political function of threat and danger, some scholars (Freedman, 2004; Robin, 2004; Jackson, 2008a) have argued that the political interpretation of threat, danger and war serves a number of political purposes, in particular, ‘selling’ specific foreign or domestic policies to public audiences. In efforts to prepare public opinion for extraordinary exertions and potential sacrifice, a political and social ‘reality’ of threat and danger is necessary.

The concept of threat and danger is established on the basis of the human emotion of fear (Booth and Wheeler, 2008, p. 62). Undoubtedly, emotions inform our attitudes and strategies, and tell us how to react and face the situation we are experiencing. The emotional reaction we experience as fear is caused by a sense of danger. That danger, through discursive interpretation, threatens to harm things we value, such as freedom from pain or freedom from loss of some sort. Importantly, the fear produced by threat and danger not only affects individuals’ responses to the surroundings they face; it also guides the subsequent actions and behaviors of actors in the political arena. For example, during the Cold War, US foreign and security policies were based on the scenario of nuclear devastation, and the world order was perceived to be established on the so-called ‘balance of terror’. Similarly, in the post-Cold War period, the threat and danger posed by terrorism has become a dominant framework for foreign and security policies. Through a series of discursive processes of interpretation, so-called ‘catastrophic terrorism’ (Carter and Perry, 1999, pp. 149–150) or ‘super-terrorism’ (see Sprinzak, 1998) has become a ‘reality’ that threatens the values of US society.

This article argues that counterterrorism policy was not initially the primary preoccupation of the Clinton administration; yet, through the discursive construction of the notion of ‘new terrorism’,2 terrorism came to be viewed as one of the most pressing challenges to US national security. With the occurrence of several major terrorist attacks on the United States and its allies, including the 1993 World Trade Center bombing, the 1995 Oklahoma City bombing and the 1995 Tokyo sarin gas attack, terrorism came to be seen as a serious threat to the United States, and counterterrorism was defined as one of the main tasks of US military forces in the post-Cold War period. By articulating the extraordinary threat of catastrophic terrorism, the ‘reality’ of new terrorism was accepted and shared by the key figures of Clinton’s administration and by most US citizens.

This article is divided into four sections. The first section briefly introduces the methodological approach and the texts examined in this research. The second section discusses the historical meaning of the word ‘terrorism’, and the invention of ‘terrorism’ in the US political arena. Through the method of genealogy, the article examines how the understanding of ‘terrorism’ and ‘terrorist’ has shifted historically and culturally. The third section focuses on Clinton’s discursive construction of terrorism. An analysis of the ‘intertextuality’3 of Clinton’s ‘new terrorism’ discourse demonstrates that the Reagan administration actually provided the initial framework for the Clinton administration to construct its new terrorism discourse. Through various discursive practices, so-called catastrophic terrorism involving inherent features of boundlessness, weapons of mass destruction and rogue states became a political ‘reality’ widely known by US citizens. Finally, some of the broader social effects and political consequences of Clinton’s terrorism discourse, such as anti-terrorism initiatives, and a military approach to address the threat of terrorism are discussed in the fourth section. The conclusion discusses the main continuities in US counterterrorism from Reagan to Clinton and Bush, and explores some of the implications of the main findings.

Robert Callahan (7 May 2015), "Terrorism blown out of proportion? Daniel Benjamin assesses the threat."

Chicago Policy Review.
http://chicagopolicyreview.org/2015/05/07/terrorism-blown-out-of-proportion-daniel-benjamin-assesses-the-threat

Summary

Daniel Benjamin, former advisor to both President Bill Clinton and Secretary Hillary Clinton on counterterrorism discusses progress in counterterrorism efforts and the changing nature of terrorist threats, from major networks like al Qaeda to “self-starter” terrorists today

Excerpt

How have ideas about terrorism changed since you began working in the field of counterterrorism?

I’ve co-written a book and many articles on how terrorism went from being essentially a third tier foreign policy and security concern to being a first tier one. And that largely has to do with the rise of catastrophic terrorism, in particular driven by religiously imbued motivation. When I started as Director for Counterterrorism on the [National Security Council] staff, our number one concern was state-sponsored terrorism, particularly from Iran. In February of 1998, right around the time I was starting, [Osama] bin Laden’s most famous fatwah was published and soon we were receiving intelligence about potential WMD [Weapons of Mass Destruction] development by al Qaeda which of course changed things considerably. I and Steve Simon, my coauthor, wrote about this long before 9/11, warned that this was coming, and then of course it happened. I think that that has been a key development. The rise of al Qaeda, 9/11, the embrace of violence, and ultimately the use of catastrophic terrorism.

What are we missing from media coverage of terror?

I’m not sure it’s confined to media coverage. We certainly suffer from all sorts of unsophisticated thinking about the risks, consequences, the relative threat of terror and the changing nature of the threat. For example, one of the arguments that I’ve made often is that in spite of the rise of ISIS, we are considerably safer than we were 10 or 12 years ago because al Qaeda, the group that could carry out

long distance covert operations and can cause catastrophic terror, has been much diminished by U.S. counterterrorism.

I also think there’s an element of the discourse that ignores all the strides forward that have been made in counterterrorism and Homeland Security. My own view is that if you look at the polls in where terrorism rates now as a national concern, it’s out of line with the actual threat at the moment. It remains a very serious concern, but I don’t think that, for example, it’s a bigger deal than Russia trying to rewrite the rules of the international system or the rise of China, and that is the way it’s often covered.

The Concept of "Existential Cyber Attack"

Defense Science Board (Jan 2013), "Resilient military systems and the advanced cyber threat."
Office of the Under Secretary of Defense for Acquisition, Technology and Logistics.
http://www.acq.osd.mil/dsb/reports/ResilientMilitarySystems.CyberThreat.pdf

Excerpts

Report Terminology

To discuss the cyber threat and potential responses in more detail, it is important to establish some common language. For purpose of this report, Cyber is broadly used to address the components and systems that provide all digital information, including weapons/battle management systems, IT systems, hardware, processors, and software operating systems and applications, both standalone and embedded. Resilience is defined as the ability to provide acceptable operations despite disruption: natural or man-made, inadvertent or deliberate. Existential Cyber Attack is defined as an attack that is capable of causing sufficient wide scale damage for the government potentially to lose control of the country, including loss or damage to significant portions of military and critical infrastructure: power generation, communications, fuel and transportation, emergency services, financial services, etc.

…The Task Force developed a threat hierarchy to describe capabilities of potential attackers, organized by level of skills and breadth of available resources (See Figure ES.1).

- Tiers I and II attackers primarily exploit known vulnerabilities
- Tiers III and IV attackers are better funded and have a level of expertise and sophistication sufficient to discover new vulnerabilities in systems and to exploit them
- Tiers V and VI attackers can invest large amounts of money (billions) and time (years) to actually create vulnerabilities in systems, including systems that are otherwise strongly protected.

…The impact of a destructive cyber attack on the civilian population would be even greater with no electricity, money, communications, TV, radio, or fuel (electrically pumped). In a short time, food and medicine distribution systems would be ineffective; transportation would fail or become so chaotic as to be useless. Law enforcement, medical staff, and emergency personnel capabilities could be expected to be barely functional in the short term and dysfunctional over sustained periods. If the attack's effects were reversible, damage could be limited to an impact equivalent to a power outage lasting a few days. If an attack's effects cause physical damage to control systems, pumps, engines, generators, controllers, etc., the unavailability of parts and manufacturing capacity could mean months to years are required to rebuild and reestablish basic infrastructure operation.

The DoD should expect cyber attacks to be part of all conflicts in the future, and should not expect competitors to play by our version of the rules, but instead apply their rules (e.g. using surrogates for exploitation and offense operations, sharing IP with local industries for economic gain, etc.).

Based upon the societal dependence on these systems, and the interdependence of the various services and capabilities, the Task Force believes that the integrated impact of a cyber attack has the potential of existential consequence. While the manifestation of a nuclear and cyber attack are very different, in the end, the existential impact to the United States is the same.

Recommendations

An overview of the Task Force’s recommendations is included in this executive summary. Recommendation details, including proposed organizational assignments and due dates, are described further in the main body of the report.

1. Protect the Nuclear Strike as a Deterrent (for existing nuclear armed states and existential cyber attack).

 Secretary of Defense (SECDEF) assign United States Strategic Command (USSTRATCOM) the task to ensure the availability of Nuclear Command, Control and Communications (C3) and the Triad delivery platforms in the face of a full-spectrum Tier V-VI attack – including cyber (supply chain, insiders, communications, etc.).

Our nuclear deterrent is regularly evaluated for reliability and readiness. However most of the systems have not been assessed (end-to-end) against a Tier V-VI cyber attack to understand possible weak spots. A 2007 Air Force study addressed portions of this issue for the ICBM leg of the U.S. triad but was still not a complete assessment against a high-tier threat.7

The Task Force believes that our capacity for deterrence will remain viable into the foreseeable future, only because cyber practitioners that pose Tier V-VI level threats are limited to a few state actors who have much to hold at risk, combined with confidence in our ability to attribute an existential level attack.

Richard Clarke and Steven Andreasen (14 Jun 2013), "Cyberwar’s threat does not justify a new policy of nuclear deterrence."

Washington Post.
http://www.washingtonpost.com/opinions/cyberwars-threat-does-not-justify-a-new-policy-of-nuclear-deterrence/2013/06/14/91c01bb6-d50e-11e2-a73e-826d299ff459_story.html

Full text

President Obama is expected to unveil a new nuclear policy initiative this week in Berlin. Whether he can make good on his first-term commitments to end outdated Cold War nuclear policies may depend on a firm presidential directive to the Pentagon rejecting any new missions for nuclear weapons — in particular, their use in response to cyberattacks.

The Pentagon’s Defense Science Board concluded this year that China and Russia could develop capabilities to launch an “existential cyber attack” against the United States — that is, an attack causing sufficient damage that our government would lose control of the country. “While the manifestation of a nuclear and cyber attack are very different,” the board concluded, “in the end, the existential impact to the United States is the same.”

Because it will be impossible to fully defend our systems against existential cyberthreats, the board argued, the United States must be prepared to threaten the use of nuclear weapons to deter cyberattacks. In other words: I’ll see your cyberwar and raise you a nuclear response.

Some would argue that Obama made clear in his 2010 Nuclear Posture Review that the United States has adopted the objective of making deterrence of nuclear attacks the “sole purpose” of our nuclear weapons. Well, the board effectively reviewed the fine print and concluded that the Nuclear Posture Review was “essentially silent” on the relationship between U.S. nuclear weapons and cyberthreats, so connecting the two “is not precluded in the stated policy.”

As the board noted, cyberattacks can occur very quickly and without warning, requiring rapid decision-making by those responsible for protecting our country. Integrating the nuclear threat into the equation means making clear to any potential adversary that the United States is prepared to use nuclear weapons very early in response to a major cyberattack — and is maintaining nuclear forces on “prompt launch” status to do so.

Russia and China would certainly take note — and presumably follow suit. Moreover, if the United States, Russia and China adopted policies threatening an early nuclear response to cyber-attacks, more countries would surely take the same approach.

It’s hard to see how this cyber-nuclear action-reaction dynamic would improve U.S. or global security. It’s more likely to lead to a new focus by Pentagon planners on generating an expanding list of cyber-related targets and the operational deployment of nuclear forces to strike those targets in minutes.

Against that backdrop, maintaining momentum toward reducing the role of nuclear weapons in the United States’ national security strategy (and that of other nations) — a general policy course pursued by the past five presidents — would become far more difficult. Further reductions in nuclear forces and changes in “hair-trigger” postures, designed to lessen the risk of an accidental or unauthorized nuclear launch, would also probably stall.

Fortunately, Obama has both the authority and the opportunity to make clear that he meant what he said when he laid out his nuclear policy in Prague in 2009. For decades, presidential decision directives have made clear the purpose of nuclear weapons in U.S. national security strategy and provided broad guidance for military planners who prepare the operations and targeting plans for our nuclear forces. An update to existing presidential guidance is one of the homework items tasked by the 2010 Nuclear Posture Review.

Cyberthreats are very real, and there is much we need to do to defend our military and critical civilian infrastructure against what former defense secretary Leon E. Panetta referred to as a “cyber Pearl Harbor” — including enhancing the ability to take action, when directed by the president, against those who would attack us. We also need more diplomacy such as that practiced by Obama with his Chinese counterpart, Xi Jinping, at their recent summit. Multinational cooperation centers could ultimately lead to shared approaches to cybersecurity, including agreements related to limiting cyberwar.

U.S. cyber-vulnerabilities are serious, but equating the impact of nuclear war and cyberwar to justify a new nuclear deterrence policy and excessive Cold War-era nuclear capabilities goes too far. It diminishes the unique threat of national devastation and global extinction that nuclear weapons represent, undermines the credibility of nuclear deterrence by threatening use for lesser contingencies and reduces the urgency for focused action to lessen nuclear dangers. Excessive rhetoric on the threat of cyberwar from the United States and blurring the distinction between cyber and nuclear attacks just makes progress toward cyber-peace more difficult.

With a stroke of his pen and his speech in Berlin, Obama can keep the United States from uploading the cyber-nuclear link.

Paul Davis (Jun 2014), "Deterrence, influence, cyber attack, and cyberwar."

WR-1049, RAND National Security Research Division, RAND Corporation.
http://www.rand.org/content/dam/rand/pubs/working_papers/WR1000/WR1049/RAND_WR1049.pdf

Excerpt

A recent discussion touching on deterrence is a study of improving resilience against the advanced cyber threat (Defense Science Board, 2013). Unlike most of the open literature, this is based on extensive conversation with the information industry and ample access to classified information. A theme in the report is the need to emphasize resilience because fully successful defense is implausible.

In domains where "real" empirical data is lacking, war gaming, red teaming, and related methods have long revealed serious problems that otherwise would have been missed or sloughed off. It is therefore of particular interest to read in the public DSB study that

DoD red teams, using cyber attack tools which can be downloaded from the internet are very successful at defeating our systems.

In charactering the degree of disruption, the report says (p. 5)

Typically, the disruption is so great, that the exercise must be essentially reset without the cyber intrusion to allow enough operational capability to proceed.

The DSB's description of cyberattack potential is no less alarming than that of Clarke and Knake (see, e.g., pp. 5). As the most tangible measure of the study's concern, the report recommends a deterrent based on a full range of response mechanisms, to include nuclear responses. Their first recommendation is

Protect the Nuclear Strike as a Deterrent (for existing nuclear armed states and existential cyber attack).

The term "existential" is important. Only in extreme circumstances might a cyberattack be arguably in the realm of existential. However, the study team saw such an attack as plausible.

The conclusion is controversial. Richard Clarke argues against blurring the distinction between cyber and nuclear threats, believing that such blurring will make cyberpeace even more difficult to attain (Clarke and Andrasen, 2013). Others, like Eldbridge Colby, disagree, arguing that the linkage—even if tentative—would be to encourage stability rather than a notion that cyberwar is a "Wild West" arena where rules are lax or nonexistent (Colby, 2013).

Kamal T. Jabbour and E. Paul Rattazzi provides another review of cyberdeterrence issues, pointing out the same problems mentioned above, concluding that what is needed are new domain-specific approaches to deterrence, including technological feasible ways to strengthen the infrastructure (Jabbour and Ratazzi, 2013). Another paper in the same volume draws on deterrence doctrine (USSTRATCOM, 2006) to describe what the authors see as a necessary "operationally responsive cyberspace" (Beeker, Mills, Grimaila, and Haas, 2013, p.35), saying that its realization not only prepares the United States to operate under duress, but sends a strong deterrence message to potential adversaries that the nation aims to deny the benefit derived from an adversary's cyberspace attacks.

Pew Research Center (29 Oct 2014), "Cyber attacks likely to increase." http://www.pewinternet.org/files/2014/10/PI_FutureofCyberattacks_102914_pdf.pdf

Summary

The Internet has become so integral to economic and national life that government, business, and individual users are targets for ever-more frequent and threatening attacks.

In the 10 years since the Pew Research Center and Elon University’s Imagining the Internet Center first asked experts about the future of cyber attacks in 2004 a lot has happened:

- Some suspect the Russian government of attacking or encouraging organized crime assaults on official websites in the nation of Georgia during military struggles in 2008 that resulted in a Russian invasion of Georgia.
- In 2009-2010, suspicions arose that a sophisticated government-created computer worm called “Stuxnet” was loosed in order to disable Iranian nuclear plant centrifuges that could be used for making weapons-grade enriched uranium. Unnamed sources and speculators argued that the governments of the United States and Israel might have designed and spread the worm.
- The American Defense Department has created a Cyber Command structure that builds Internet-enabled defensive and offensive cyber strategies as an integral part of war planning and war making.
- In May, five Chinese military officials were indicted in Western Pennsylvania for computer hacking, espionage and other offenses that were aimed at six US victims, including nuclear power plants, metals and solar products industries. The indictment comes after several years of revelations that Chinese military and other agents have broken into computers at major US corporations and media companies in a bid to steal trade secrets and learn what stories journalists were working on.
- In October, Russian hackers were purportedly discovered to be exploiting a flaw in Microsoft Windows to spy on NATO, the Ukrainian government, and Western businesses.
- The respected Ponemon Institute reported in September that 43% of firms in the United States had experienced a data breach in the past year. Retail breaches, in particular, had grown in size in virulence in the previous year. One of the most chilling breaches was discovered in July at JPMorgan Chase & Co., where information from 76 million households and 7 million small businesses was compromised. Obama Administration
- officials have wondered if the breach was in retaliation by the Putin regime in Russia over events in Ukraine.
- Among the types of exploits of individuals in evidence today are stolen national ID numbers, pilfered passwords and payment information, erased online identities, espionage tools that record all online conversations and keystrokes, and even hacks of driverless cars.
- Days before this report was published, Apple’s iCloud cloud-based data storage system was the target of a so-called “man-in-the-middle” attack in China that was aimed at stealing users’ passwords and spying on their account activities. Some activists and security experts said they suspected the Chinese government had mounted the attack, perhaps because the iPhone 6 had just become available in the country. Others thought the attack was not sophisticated enough to have been government-initiated.
- The threat of cyber attacks on government agencies, businesses, non-profits, and individual users is so pervasive and worrisome that this month (October 2014) is National Cyber Security Awareness Month.

To explore the future of cyber attacks we canvassed thousands of experts and Internet builders to share their predictions. We call this a canvassing because it is not a representative, randomized survey. Its findings emerge from an “opt in” invitation to experts, many of whom play active roles in Internet evolution as technology builders, researchers, managers, policymakers, marketers, and analysts. We also invited comments from those who have made insightful predictions to our previous queries about the future of the Internet. (For more details, please see the section “About this Canvassing.”)

Overall, 1,642 respondents weighed in on the following question:

Major cyber attacks: By 2025, will a major cyber attack have caused widespread harm to a nation’s security and capacity to defend itself and its people? (By “widespread harm,” we mean significant loss of life or property losses/damage/theft at the levels of tens of billions of dollars.)

Please elaborate on your answer. (Begin with your name if you are willing to have your comments attributed to you.) Explain what vulnerabilities nations have to their sovereignty in the coming decade and whether major economic enterprises can or cannot thwart determined opponents. Or explain why you think the level of threat has been hyped and/or why you believe attacks can be successfully thwarted.

Some 61% of these respondents said “yes” that a major attack causing widespread harm would occur by 2025 and 39% said “no.”

The Concept of “Moral Enhancement”

Ingmar Persson and Julian Savulescu (2012), *Unfit for the Future: The Need for Moral Enhancement*.
Oxford University Press.
http://ukcatalogue.oup.com/product/9780199653645.do

Blurb

Unfit for the Future argues that the future of our species depends on our urgently finding ways to bring about radical enhancement of the moral aspects of our own human nature. We have rewritten our own moral agenda by the drastic changes we have made to the conditions of life on earth. Advances in technology enable us to exercise an influence that extends all over the world and far into the future. But our moral psychology lags behind and leaves us ill equipped to deal with the challenges we now face. We need to change human moral motivation so that we pay more heed not merely to the global community, but to the interests of future generations. It is unlikely that traditional methods such as moral education or social reform alone can bring this about swiftly enough to avert looming disaster, which would undermine the conditions for worthwhile life on earth forever. Persson and Savulescu maintain that it is likely that we need to explore the use of new technologies of biomedicine to change the bases of human moral motivation. They argue that there are in principle no philosophical or moral objections to such moral bioenhancement. Unfit for the Future? challenges us to rethink our attitudes to our own human nature, before it is too late.

Ingmar Persson and Julian Savulescu (2013), "Summary of *Unfit for the Future*."
Journal of Medical Ethics (Early content).
http://jme.bmj.com/content/early/2013/12/13/medethics-2013-101323.full

Introduction

The argument of Unfit for the Future can be summed up as follows. It is easier for us to harm each other than it is for us to benefit each other. For example, it is easier for us to kill than to save life. As the progress of scientific technology has increased our powers of action, our capacity to harm has reached the point at which it is possible for us to undermine worthwhile life on Earth forever. This could be done by the use of weapons of mass destruction or by causing catastrophic climatic or other environmental changes. One central, neglected problem is that a significant cause of these problems is human behaviour, caused by limitations in our psychology as moral agents. Our moral psychology has been adapted to life in small, close-knit societies with primitive technology, in which human beings have lived for virtually all of their history. This is reflected in the fact that we are psychologically myopic, that is, disposed to care more about what happens in the near future to ourselves and some individuals who are near and dear to us. We are also incapable of responding adequately to the suffering of larger collectives. Due to the fact that it is easier to harm, we tend to have a moral reluctance to harm that is stronger than our disposition to benefit, but like the latter, it is largely confined to an ‘in-group’. Such a limited moral psychology is an ineffective brake on misuse of technology when modern weapon

technology enables us to create weapons to kill large numbers at long distances. To some extent, we have undergone moral improvement in the course of history by means of traditional moral education. But to cope with the moral problems created by the advance of scientific technology, it seems that we would have to change radically in a short time. Therefore, it is imperative that we investigate the possibility of moral enhancement by means of genetic and biomedical techniques, as well as conventional political and social reform. We need advanced technology for the foreseeable future to provide a huge, and increasing, human population on Earth with a decent standard of life.

The argument might be said to consist of four main claims:

- It is easier to harm us than to benefit us...
- Due to the progress of scientific technology, we are now in a position to cause ultimate harm, that is, to forever make worthwhile life on this planet impossible...
- Since our moral dispositions are designed for life in small communities with limited technology, there is considerable risk that we shall cause ultimate harm...
- We need to consider moral enhancement if possible by biomedical means, alongside traditional means to minimise the risk of us causing ultimate harm with the advanced technology that we need to give a huge human population good lives...

Ingmar Persson and Julian Savulescu (2014), "Unfit for the future? Human nature, scientific progress, and the need for moral enhancement."

in *Enhancing Human Capacities*, ed. Julian Savulescu, Ruud ter Meulen and Guy Kahane, Wiley.
http://onlinelibrary.wiley.com/doi/10.1002/9781444393552.ch35/summary

Summary

This chapter identifies the problems created by the misfit between a limited human moral nature and globalized, highly advanced technology. It highlights the several ways of addressing the potential catastrophic consequences of this mismatch. The chapter discusses the development of a globally responsible liberalism, with the restriction of traditional liberal neutrality, inculcation of values and “moral education” to achieve restraint, promote cooperation, respect for equality, and other values now necessary for our survival as a global community. It also discusses some consequences of the intuitive endorsement of the act-omission doctrine. Environmental problems probably provide the most worrying example of the limitations of our cooperative dispositions in the contemporary world. In the world of today, when scientific progress has vastly increased our powers of action, societies need to inculcate norms that are conducive to the good of the world community of which these societies are an integral part.

Information Risks

Nick Bostrom (Aug 2011), "Information hazards: A typology of potential harms from knowledge."
Review of Contemporary Philosophy 10.
http://www.fhi.ox.ac.uk/information-hazards.pdf
[Potentially useful in discussions of the risks entailed in publishing information about dual-use research.]

Abstract

Information hazards are risks that arise from the dissemination or the potential dissemination of true information that may cause harm or enable some agent to cause harm. Such hazards are often subtler than direct physical threats, and, as a consequence, are easily overlooked. They can, however, be important. This paper surveys the terrain and proposes a taxonomy.

1. *Basic definition*

Information hazard: A risk that arises from the dissemination or the potential dissemination of (true) information that may cause harm or enable some agent to cause harm.

2. *Six information transfer modes*

Data hazard: Specific data, such as the genetic sequence of a lethal pathogen or a blueprint for making a thermonuclear weapon, if disseminated, create risk.

Idea hazard: A general idea, if disseminated, creates a risk, even without a data-rich detailed specification.

Attention hazard: The mere drawing of attention to some particularly potent or relevant ideas or data increases risk, even when these ideas or data are already “known”.

Template hazard: The presentation of a template enables distinctive modes of information transfer and thereby creates risk.

Signaling hazard: Verbal and non-verbal actions can indirectly transmit information about some hidden quality of the sender, and such social signaling creates risk.

Evocation hazard: There can be a risk that the particular mode of presentation used to convey some content can activate undesirable mental states and processes.

3. *Adversarial risks*

Enemy hazard: By obtaining information our enemy or potential enemy becomes stronger and this increases the threat he poses to us.

Competiveness hazard: There is a risk that, by obtaining information, some competitor of ours will become stronger, thereby weakening our competitive position.

Intellectual property hazard: A faces the risk that some other firm B will obtain A’s intellectual property, thereby weakening A’s competitive position.

Commitment hazard: There is a risk that the obtainment of some information will weaken one’s ability credibly to commit to some course of action.

Knowing-too-much hazard: Our possessing some information makes us a potential target or object of dislike.

4. Risks to social organization and markets

Norm hazard: Some social norms depend on a coordination of beliefs or expectations among many subjects; and a risk is posed by information that could disrupt these expectations for the worse.

Information asymmetry hazard: When one party to a transaction has the potential to gain information that the others lack, a market failure can result.

Unveiling hazard: The functioning of some markets, and the support for some social policies, depends on the existence of a shared "veil of ignorance"; and the lifting of which veil can undermine those markets and policies.

Recognition hazard: Some social fiction depends on some shared knowledge not becoming common knowledge or not being publicly acknowledged; but public release of information could ruin the pretense.

5. Risks of irrationality and error

Ideological hazard: An idea might, by entering into an ecology populated by other ideas, interact in ways which, in the context of extant institutional and social structures, produce a harmful outcome, even in the absence of any intention to harm.

Distraction and temptation hazards: Information can harm us by distracting us or presenting us with temptation.

Role model hazard: We can be corrupted and deformed by exposure to bad role models.

Biasing hazard: When we are biased, we can be led further away from the truth by exposure to information that triggers or amplifies our biases.

De-biasing hazard: When our biases have individual or social benefits, harm could result from information that erodes these biases.

Neuropsychological hazard: Information might have negative effects on our psyches because of the particular ways in which our brains are structured, effects that would not arise in more "idealized" cognitive architectures.

Information-burying hazard: Irrelevant information can make relevant information harder to find, thereby increasing search costs for agents with limited computational resources.

6. Risks to valuable states and activities

Psychological reaction hazard: Information can reduce well-being by causing sadness, disappointment, or some other psychological effect in the receiver.

Belief-constituted value hazard: If some component of well-being depends constitutively on epistemic or attentional states, then information that alters those states might thereby directly impact well-being.

Disappointment hazard: Our emotional well-being can be adversely affected by the receipt of bad news.

Spoiler hazard: Fun that depends on ignorance and suspense is at risk of being destroyed by premature disclosure of truth.

Mindset hazard: Our basic attitude or mindset might change in undesirable ways as a consequence of exposure to information of certain kinds.

Embarrassment hazard: We may suffer psychological distress or reputational damage as a result of embarrassing facts about ourselves being disclosed.

7. Risks from information technology systems

Information system hazard: The behavior of some (non-human) information system can be adversely affected by some informational inputs or system interactions.

Information infrastructure failure hazard: There is a risk that some information system will malfunction, either accidentally or as result of cyber attack; and as a consequence, the owners or users of the system may be inconvenienced, or third parties whose welfare depends on the system may be harmed, or the malfunction might propagate through some dependent network, causing a wider disturbance.

Information infrastructure misuse hazard: There is a risk that some information system, while functioning according to specifications, will service some harmful purpose and will facilitate the achievement of said purpose by providing useful information infrastructure.

Robot hazard: There are risks that derive substantially from the physical capabilities of a robotic system.

Artificial intelligence hazard: There could be computer-related risks in which the threat would derive primarily from the cognitive sophistication of the program rather than the specific properties of any actuators to which the system initially has access.

Development hazard: Progress in some field of knowledge can lead to enhanced technological, organizational, or economic capabilities, which can produce negative consequences (independently of any particular extant competitive context).

Academic Centers Studying Existential Risk

Centre for the Study of Existential Risk

University of Cambridge.
http://cser.org

Leadership

- Huw Price, Bertrand Russell Professor of Philosophy, Cambridge
- Martin Rees, Emeritus Professor of Cosmology & Astrophysics, Cambridge
- Jaan Tallinn, Co-founder of Skype
- Seán Ó hÉigeartaigh, CSER Project Manager

Emerging risks to humanity's future

Modern science is well-acquainted with the idea of natural risks, such as asteroid impacts or extreme volcanic events, that might threaten our species as a whole. It is also a familiar idea that we ourselves may threaten our own existence, as a consequence of our technology and science. Such home-grown "existential risk" – the threat of global nuclear war, and of possible extreme effects of anthropogenic climate change – has been with us for several decades.

However, it is a comparatively new idea that developing technologies might lead – perhaps accidentally, and perhaps very rapidly, once a certain point is reached – to direct, extinction-level threats to our species. Such concerns have been expressed about artificial intelligence (AI), biotechnology, and nanotechnology, for example.

Technology and uncertainty

The common factor in such concerns is that the new capabilities of such technologies might provide direct and relatively short-term control over circumstances essential to our survival, and either place that control in dangerously few human hands, or take it out of our sphere of influence altogether, so that we cannot protect ourselves. The grounds for such concerns are presently difficult to assess. Relatively little work has been done on such problems, and experts in the fields in question often disagree. These uncertainties are themselves a ground for concern, given how much is at stake.

Investigation and mitigation

The Centre for the Study of Existential Risk is premised on the view that the task of investigating and mitigating such home-grown existential risks is a pressing and enduring responsibility for the scientific community; a task whose urgency and importance may be expected only to increase, as technology continues to develop. Yet there is at present little coherent sense of what this task amounts to – little sense of the necessary shape and components of practical and theoretical science of existential risk. Our aim is to construct and conceptualise this new science, and to begin developing a protocol for the investigation and mitigation of technologically-driven existential risk.

Future of Humanity Institute

Oxford University.
http://www.fhi.ox.ac.uk

Leadership and selected associates

- Nick Bostrom, Director
- Stuart Armstrong, James Martin Research Fellow
- Nick Beckstead, Research Fellow
- Daniel Dewey, Alexander Tamas Research Fellow
- Toby Ord, James Martin Research Fellow
- Seán Ó hÉigeartaigh, James Martin Academic Project Manager
- Milan Cirkovic, Research Associate
- Owen Cotton-Barratt, FHI-CEA Collaboration Fellow

Mission

The Future of Humanity Institute is a leading research centre looking at big-picture questions for human civilization. The last few centuries have seen tremendous change, and this century might transform the human condition in even more fundamental ways. Using the tools of mathematics, philosophy, and science, we explore the risks and opportunities that will arise from technological change, weigh ethical dilemmas, and evaluate global priorities. Our goal is to clarify the choices that will shape humanity's long-term future.

Global Catastrophic Risk Institute.

http://gcrinstitute.org

Leadership

- Seth Baum
- Tony Barrett
- Grant Wilson

About

GCRI's mission is to develop the best ways to confront humanity's gravest threats.

Global catastrophic risk (GCR) is the risk of events large enough to significantly harm or even destroy human civilization at the global scale. Major GCRs include global warming, nuclear war, pandemics, and emerging technologies such as artificial intelligence and biotechnology. Many organizations work on GCR, but they mainly look at just one risk or a few risks at a time. They do excellent work, but they don't answer the big questions about GCR. That's where we come in.

The Global Catastrophic Risk Institute (GCRI) is a nonprofit, nonpartisan think tank. GCRI was founded in 2011 by Seth Baum and Tony Barrett. GCRI studies the full range of GCRs and GCR topics in order to answer the big questions: Which risks should society be most worried about? How do the different risks

affect each other? And above all, what are the best ways to reduce the risk? These questions guide our work, and they are at the heart of our flagship GCR Integrated Assessment project.

Machine Intelligence Research Institute
https://intelligence.org

Leadership

- Luke Muehlhauser
- Eliezer Yudkowsky
- Benja Fallenstein
- Nate Soares

Mission Statement

MIRI's mission is to ensure that the creation of smarter-than-human intelligence has a positive impact. We aim to make intelligent machines behave as we intend even in the absence of immediate human supervision. Much of our current research deals with reflection, an AI's ability to reason about its own behavior in a principled rather than ad-hoc way.