**List of cognitive biases affecting judgment of global risks**

Cognitive biases might as well be called “reasoning errors,” or “biases that cause errors in reasoning”. A classic example might be generalized overconfidence, which is systematic and universal. Experts tend to be just as overconfident as laypeople if not moreso, including in their area of specialty. In the chapters ahead you will find just how many cognitive biases we have identified which may compromise reasoning with regard to global catastrophic risks, perhaps fatally so.

Cognitive biases are an innate property of the human mind. There is no “fault” for them; they are not the product of malice. They are universal, ancient, and taken for granted. Usually, they do not matter much, because being slightly mistaken in daily circumstances is rarely fatal. Natural selection has removed the worst errors. With respect to global risks, which concern extremely dangerous and poorly understood phenomena, they may very well be. So the importance that we understand the cognitive biases therein is extremely great. The typical way in which reasoning contaminated with cognitive biases is overcome on a daily basis is through trial and error, a luxury we lack with respect to extinction risks.

Even if the probabilistic contribution of several dozen cognitive biases is relatively low, in combination they may throw a likelihood estimate off completely and cause millions of dollars to be wasted on unnecessary prevention efforts. An example would be money spent on unnecessary geoengineering efforts, or money spent fighting genetically modified organisms which should go towards stopping autonomous killer robots.

To illustrate the ubiquity of cognitive biases you need only ask a few experts about their forecast for the greatest risks of the 21st century. Usually, they will fixate on one or two future technologies, subconsciously overemphasizing them relative to others, which they know little about. One person may focus on Peak Oil, another possess too much trust in the promise of renewable energy, another focus on a global pandemic, and so on. One may find the use of nuclear weapons to be extremely large, another greatly improbable. A serious researcher on the topic of global risks should know about these wild differences in position and some of the common reasoning errors which may cause them.

The analysis of possible reasoning errors with respect to global risks is a step on the way to creating a holistic methodology of global risks, and therefore, to their prevention. The purpose of this second part of the book is to put potential cognitive biases applicable to global risks in a complete and structured list. Thus first priority is given to the completeness of these lists, rather than exhaustively illustrating each individual cognitive bias. Fully explaining each cognitive bias would take its own chapter.

It is important to recall that small errors resulting from small deviations of estimates with regard to high-power, high-energy systems, especially those such as nanotechnology or Artificial Intelligence, could lead to threads which unravel and mutually reinforce, eventually culminating in disaster. Thus, those developing these systems and analyzing them must be especially careful to examine their own biases, and hold themselves to a much higher standard than someone designing a new widget.

Intellectual errors can lead to real-life catastrophes. It is easy to find examples of how erroneous reasoning of pilots led to airplane catastrophes that cost hundred of lives. For instance a pilot pulling upwards on the throttle when he thought the nose of the plane was pointed down though it was not, resulting in a stall and subsequent nosedive. In fact the majority of technological disasters are both caused by human error and preventable by proper procedures without error.

There are two main categories of error which are relevant to consideration of global risks, we'll call them “designer errors” and “pilot errors”. Designer errors are made by big groups of people for many years, whereas pilot errors are made by small groups of people or individuals in seconds or minutes while controlling critical systems in realtime.

Of course, there is the possibility that certain risks we highlight in these sections are erroneous themselves, or at least overstated. We also can be certain that the list is not complete. Therefore, any given list should be used as a launching pad for the critical analysis of reasoning on global risks, not as a conclusive tool for any definitive diagnosis.

The most dangerous illustration consists in making the assumption that errors in reasoning on global risks are insignificant, or that they could be easily found and eliminated. For instance, people might think, “Planes fly, despite all possible errors, and life in general on the Earth proceeds, so the value of these errors must be insignificant”. This projection is incorrect. Planes fly because during the course of their development, there were many tests and thousands of planes crashed. Behind each crashed test plane was someone's errors. We do not have a thousand planets to test; just one, and it has to count. The explosive combination of bio, nano, nuclear, and AI technologies are very risky, and one fatal error could end it all. The fact that the Earth is whole today gives us no reassurances for the future. There is no time in history when it was more important to be correct than now.

There is the possibility that we will discover cognitive biases, or hidden pieces of knowledge, which will completely change our reasoning on global risks and change the course of all our reasoning. For instance, there are some extremely intelligent people who see human-friendly smarter-than-human intelligence is the one best solution to address all global risks, as we explored in the last section of the book. Many of the errors we discuss, in any case, are endemic across researchers, and much insight is to be gained from attempting to debias them.

Under the term cognitive biases we mean distortions which are not only logic infringements, but also any intellectual procedures or designs which can influence final conclusions and result in an increased risk of global catastrophe.

Possible errors and cognitive biases in this section are divided into the following categories:

* + Errors specific to estimates to global risks,
  + Errors concerning any risk, with specific analysis applied to global risks,
  + Specific errors arising in discussions concerning Artificial Intelligence;
  + Specific errors arising in discussions concerning nanotechnology;

Read carefully through the lists of risks, and consider situations in which you may have encountered them outside the context of global risks.

## Chapter 2. Cognitive biases concerning global catastrophic risks

## **Chapter 2. Cognitive biases concerning global catastrophic risks**

1. **Confusion regarding the difference between catastrophes causing human extinction and catastrophies non-fatal to our species**

There is a tendency to conflate global catastrophes causing the extinction of mankind (*existential risks*) and other enormous catastrophes which could bring about major civilizational damage without being completely fatal (most conceivable nuclear war scenarios, for instance). The defining features of *existential* disaster are *irreversibility* and *totality*. If there is a disaster which wipes out 99.9% of the world's population, 7 million humans would still remain in the planet's refuges, about 400-4,000 times the number of humans that were alive at the time of the Toba catastrophe 77,000 years ago. This would be more than enough to start over, and they would have a tremendous amount of intact records, technology, and infrastructure to assist them. So, the difference between a disaster that wipes out 100% of humanity and merely 99.9% of humanity is immense. Even the difference between a disaster that wipes out 99.999% of humanity and 100% would be immense. People seldom intuitively grasp this point, and it may have to be explained several times.

**2. Underestimating non-obvious risks**

Global risks are divided into two categories: obvious and non-obvious. An obvious risk would be nuclear war, a non-obvious risk would be the destruction of an Earth by the creation of a stable strangelet during a particle accelerator experiment. Non-obvious risks may be more dangerous, because their severity and probability are unknown, and therefore suitable countermeasures are generally not taken. Some non-obvious risks are known only to a narrow circle of experts who express contradictory opinions about their severity, probability, and mechanism of emergence. A detached onlooker, such as a military general or a head of state, may be completely unable to distinguish between the expert advice given and might as well flip a coin to determine who to listen to regarding the non-obvious risks. This makes inadequate preparation for these risks highly likely, whereas well-understood risks such as nuclear war are better anticipated and ameliorated.

Making estimates based on past rates of discovery of new global risks, it seems as if the number of new risks expands exponentially over time. Therefore we can anticipate a great increase in the number of global risks in the 21st century, the nature of which may be impossible for us to guess now, and fall into the category of non-obvious risks.

Obvious risks are much more convenient to analyze. There are huge volumes of data we can use to assess these perils. The volume of this analysis can conceal the fact that there are other risks about which little is known. Their assessment may not be amenable to rigorous numerical analysis, but they are severe risks all the same (for example, risks from incorrectly programmed Artificial Intelligence).

**3. Bystander effect: global risks have less perceived national security importance**

The focus of each country is on risks to that specific country. Global risks are only considered if they are well understood, like nuclear war or global warming. As a result, there is a bystander effect1 or "Tragedy of the Commons" whereby risks that threaten us all but no particular nation individually are poorly analyzed and prepared for. This is analogous to how a collapsed person in a busy area is less likely to be helped than if they are in a less populated area.

**4. Bias connected to psychologization of the problem**

There is a social stereotype whereby those who warn of risk are considered "doomsayers," with the implication that these people are social outcasts or merely warning of risk for attention and to increase social status. This may always be the case, yet studies show that pessimistic people actually tend to estimate more accurate probabilities of events than more optimistic people. This is called depressive realism2. Only precise calculations can define the real weight of risk. Psychologizing the problem is just a way of sticking our heads in the sand. This approach to the problem will be popular among people who clearly understand social stereotypes about doomsayers but have difficulty grasping the complex scientific details surrounding global risks. It is easier to call someone a doomsayer than to understand the risk on a technical or scientific level.

**5. A false equivocation of global catastrophe with the death of all individual humans**

It is possible to imagine scenarios in which humanity survives in a literal sense but in a deeper sense, civilization comes to an end. This includes superdrug scenarios, where humanity becomes addicted to virtual reality and gradually becomes obsolete, replaced by machines that eventually turn us into detached brains or even human-like computer programs that lack phenomenological consciousness. So, we should note that a global catastrophe need not involve the sudden death or all individual humans; there are more subtle ways our story might come to an end.

**6. Perceptive stereotype of catastrophes from mass media coverage of prior risks**

Mass media creates a false image of global catastrophe that has a subconscious and profound impact on our estimates. Experience of watching television reports on catastrophes has developed the subconscious stereotype that doomsday will be shown to us on *CNN*. However, a scenario might not unfold in that way. A global event that affects everyone on the planet very quickly may not have time to be adequately covered by the media. Doomsday may not be televised.

Television also creates the perception that there will be an abundance of data regarding a disaster as it emerges, as has been the case with threats such as earthquakes and bird flu mutations. However, the amount of information available may actually be quite small in proportion to the magnitude of the risk, so detailed reporting may not be forthcoming.

**7. Bias connected with the fact that global catastrophe is by definition a unique event**

Global doom sounds fantastic. It has never happened before. If it does happen, it can only happen once. Therefore normal inductive processes of sampling are ineffective to predict it. If something is true at t=1, t=2, t=3, and so on, we can reasonably assume it will be true at t+1 (or all t). This methodology is useful during smooth conditions, but ineffective for predicting abrupt, extreme phenomena without precedent. A separate issue is that a lethal effect that kills off or disables humanity a little bit at a time may never appear to be a global catastrophe at first, but leads to human extinction when operating over a sufficient duration of time.

**8. Underestimating global risks because of emotional reluctance to consider your personal demise**

People are not keen on considering global threats because they have become accustomed to the inevitability of their personal death and have developed psychological protective mechanisms against these thoughts. Yet, amelioration of global risks demands that people of all kinds—scientists, politicians, businessmen—put aside these thoughts and devote collective effort towards considering these risks anyway, even if they are threatening on a personal and visceral level.

**9. Too many cooks in the kitchen: decision paralysis**

If there are 20 scientists who have 20 different, equally plausible-sounding accounts of global risks, decision paralysis (also called analysis paralysis or option paralysis3) may set in and there is the temptation to do nothing. A.P. Chekhov wrote, “If many remedies are prescribed for an illness, you may be certain that the illness has no cure.” If too many remedies are prescribed for global risk, it may be that we cannot come up with a “cure”. More simplistically, there may be a large pool of experts and we just listen to the wrong ones.

**10. Global risks receive less attention than small-scale risks**

Consider the risk of a nuclear explosion in Washington. This is specific, vivid, and has received top-level attention from the United States government. Consider the risk of the creation of a suite of powerful new superviruses. This has received greatly less attention, though one risk only affects one city, whereas the other could be terminal to the entire planet. The impact of permanent termination of the human species is much greater than comparatively small individual risks, and deserves correspondingly greater attention, even though local risks may be more vivid.

**11. Trading off a "slight" risk to humanity for some perceived benefit**

There may be scientists who feel that it is worth developing powerful Artificial Intelligence without adequate safeguards because there is only a “small risk” that it will get out of control and threaten the human species. Or, some other advance, in the area of weapons research or biotechnology may emit the same siren song. Alfred Nobel justified the invention of dynamite with the notion that it would end all war. Instead, dynamite greatly increased the killing power of weapons and indirectly contributed to the most lethal century in the history of mankind.

**12. Absence of clear understanding of to whom instructions on global risks are directed**

It is not clear exactly who should be heeding our advice to deal with global risks. There are many different parties and their relative ability to influence the global situation is unknown. Many parties may underestimate their own influence, or estimate it correctly but hesitate to do anything because they see it as so small. Then everyone will do nothing, the bystander effect kicks in again, and we may may the price. An important part of risk research should be identifying to whom literature on global risks should be directed. High-ranking military and intelligence personnel seem like obvious candidates, but there are many others, including rank-and-file scientists of all kinds.

**13. Poor translation between the theoretical and practical**

Global risks are theoretical events which require practical planning. There is no way to empirically test the likelihood or impact or global disasters. There is no way to test the practical probability of thermonuclear war because it hasn't happened yet, and if it did, it would be too late. Similar properties of other global risks make it difficult to connect practical planning with theoretical risk assessments.

**14. Predisposal to risk-taking and aggression**

Humans evolved to take risks in competition with one another, to play chicken. The first guy who chickens out loses the girl. This tendency becomes a problem in a world of doomsday weapons.

**15. Erroneous representation that global risks are in the far future and not relevant today**

The possible self-destruction of mankind has been a rather serious issue ever since the United States and U.S.S.R. built up their nuclear stockpiles. Though unlikely that nuclear war could wipe out humanity alone, in conjunction with biological weapons it could be possible today, in a worst-case scenario. Within 20 or 30 years, there will be many more tools for dealing death, and the risk will increase. The risk is here today; it is not futuristic.

**16. The idea that global catastrophe will be personally painless**

There is a natural tendency to think of global disaster by analogy to taking a bullet in the head: quick and painless. Yet global disaster could stretch out for decades or even centuries and provide profound misery to everyone involved. You might not be able to enjoy the benefit of dying so quickly. Even if you did, it seems irresponsible to discount global risk based on the anticipation that it would be like getting a bullet through the head. Others may not be so lucky, and humanity's entire future being lost has a profound negative value, even if doom were painless.

**17. The representation that books and articles about global risks can change a situation considerably**

Books and articles about global risk are not enough, and may create a false sense of security that knowledge is being spread. What is needed is real money and manpower being put towards creating systems and protocols to prevent global risks, systems that have barely been conceived of today. For instance, we should be acutely aware that there is no plan and no proposals to regulate the synthesis of dangerous genetic sequences, such as the Spanish flu virus. What are we going to do about it? A book about global risk alone is not sufficient to prevent these gene sequences from popping out of synthesizers and into real viruses which find their way into airports.

**18. Views that global risks are either inevitable, depend entirely on casual factors not subject to human influence, or depend on the highest-level politicians who are unreachable**

The circulation of certain ideas in a society, namely that global risks are real and it is necessary to make efforts for their prevention, can create a certain background which indirectly drives mechanisms of decision-making. Consider the process of crystallization; it requires nucleation points and certain saturation levels of a chemical in solution to get started. Unless the entire solution is supersaturated and nucleation points are provided, crystallization does not occur. Taking difficult society-wide steps like preparing for global risks is an analogous process. Every little bit of action and knowledge matters, no matter where it is distributed. Higher-ups are reluctant to take action unless they perceive there will be a certain level of understanding and approval among their subordinates. This effect becomes even more acute as one approaches the top of the pyramid. Understanding must be commonplace near the bottom and middle of the pyramid before the top does anything about it.

**19. Intuition as a source of errors in thinking about global risks**

As global risks concern events which never happened, they are fundamentally unintuitive. Intuition can be useful to come up with new hypotheses, but is less useful for more systematic analyses and nailing down probabilities. Intuition is more susceptible to subconscious biases, such as a latent unwillingness to consider unpleasant scenarios, or contrariwise, the urge to see them where they are not present. As intuition takes less mental effort than deliberative thought, there is a constant compulsion to substitute the former for the latter.

**20. Scientific research of global risks also faces a number of problems**

As mentioned before, experimentation is not a good way of establishing the truth about global risks. In connection with the impossibility of experiment, it is impossible to measure objectively what errors influence estimates of global risks. There cannot be rigorous statistics on global risks. The fundamental concept of falsifiability is also inapplicable to theories about global risks.

**21. The errors connected with unawareness of little-known logical arguments for global risk**

In the case of global risks, difficult inferential reasoning as the Doomsday Argument and observation selection effects start to operate. These are important to assessing global risk, however they are unknown to the majority of people, and a considerable share of researchers reject them due to their unintuitiveness and the conceptual challenge of understanding them.

**22. Methods applicable to management of economic and other risks are not applicable to global catastrophic risks**

Global catastrophic risks cannot be insured, and are not easily amenable to prevention within a logical economic context. They break the rules because they are universal and their economic downside, being extreme or total, is hard to quantify in conventional economic terms. Furthermore, there is a tendency for countries to fixate on economic risks relevant only to themselves, and not to the globe as a whole.

**23. Erroneous conception that global risks threaten people only while we are Earthbound, and resettlement in space will automatically solve the problem**

The scale of weapons and energies which can be created by people on the Earth grows faster than the rate of space expansion. Meaning, by the time we can expand into space, we will have created threats which can self-replicate and intelligently target human beings, namely advanced Artificial Intelligence and autonomous self-replicating nanotechnology. Information contamination, such as viruses or a computer attack, would also spread at close to the speed of light. Space is not a panacea.

**24. Scope neglect**

There is a cognitive bias known as scope neglect, meaning that to save the life of a million children, a person is only willing to pay a slightly greater amount than to save the life of one child4. Our minds are not able to scale up emotional valence or preventive motivation linearly in accordance with the risk; we discount large risks. Most money and attention consistently goes to projects which only affect a much smaller number of lives rather than all of the planet's humans.

**25. Exaggeration of prognostic values of extrapolation**

In futurism, there is an "extrapolation mania" by which futurists take present trends, extrapolate them outwards, and predict the future based on those. However, trends often change. As an example, Moore's law, the improvement in the cost-performance of computers, is already starting to level off. Our experience with futurism shows that extrapolation of curves tends to only be suited to short-term forecasts. Extrapolation poorly accounts for feedback effects between predictions and future events and the effects of fundamental innovations.

An example of an extrapolation that failed was Malthus' prediction that human beings would run out of food at some point during the 19th century. He failed to account for a number of future innovations, such as artificial fertilizer and the Green Revolution. Besides being too pessimistic, it is possible to be too optimistic. Many futurists today anticipate an era of abundance where all global risks are under control5. However, technological progress may prove to be slower than they anticipate, and global risks may remain threatening for a long time, past 2100. We ought to be wary of simple extrapolation and not take our models of the future too seriously, because models are made to be broken. Helmuth von Moltke the Elder, a famous Prussian general, said “No battle plan survives contact with the enemy.”

**26. Erroneous representation that people as a whole do not want catastrophe and a doomsday**

Humans, especially males, like to engage in thrill-seeking behavior. Humans, especially males, are also liable to fantasies of revenge and destruction. These qualities are serious liabilities in an era of global risk, where a war started anywhere in the world has the potential to explode into a conflagration between superpowers. We like to whimsically imagine that no one will press “the button,” but just because it hasn't been done before, doesn't mean that it will never happen. Heads of state and military generals are human too, and subject to errors in judgment with long-term consequences. Saying, “a doomsday will never happen, because everyone fears it too much” is wishful thinking and liable to get us into trouble.

**27. Future Shock: Cognitive biases connected with futuristic horizons**

If we transported a human being from 1850 to today, they would be bewildered—shocked—by our level of technology. Flying machines, Green Revolution, nuclear reactors, the Internet... these were not really foreseen by the people of that era. Over the last fifty years, things have changed so fast than futurist Alvin Toffler used the term “future shock” to describe common reactions to it6. Certain futuristic risks, like risks from biotechnology, nanotechnology, and Artificial Intelligence, may seem so shocking that many people have trouble taking them seriously. There has only been a gap of 13 years between the first full sequencing of the human genome and the synthesis of the first organisms with entirely artificial chromosomes. Many people are still digesting the implications. In his "Future Shock Levels" article, Eliezer Yudkowsky outlined five general levels of future shock7:

**Shock Level 0**: technology used in everyday life which everyone is familiar with, or which is so widely discussed that nearly everyone is aware of it. (Catastrophe levels: nuclear war, exhaustion of resources.)

**Shock Level 1**: the frontiers of modern technology: virtual reality, living to a hundred, etc. (Catastrophe levels: powerful biological warfare and the application of military robotics.)

**Shock Level 2**: medical immortality, interplanetary exploration, major genetic engineering. Star Trek, only moreso. Quite a few futurists anticipate we'll reach this technology level close to the end of the 21st century. (Catastrophe levels: deviation of asteroids towards the Earth, superviruses that change the behavior of people, creation of advanced artificial bacteria or nanobots immune to the body's defenses.)

**Shock Level 3**: nanotechnology, human-level AI, minor intelligence enhancement. Not necessarily technologically more difficult than Shock Level 2, but more shocking to think about. Difficult to predict the arrival of. (Catastrophe levels: grey goo, intelligence-enhanced people taking over the planet and wiping out humanity.)

**Shock Level 4**: advanced Artificial Intelligence and the Singularity (creation of greater-than-human intelligence). (Catastrophe levels:superhuman AI converting the entire planet into “computronium”.)

Future shock levels are based on the idea that people define the horizon of the possible based on what they are psychologically comfortable with, rather than objectively analyzing the technological difficulty of various propositions. For instance, we have already enhanced intelligence in mice, and we know that such tech could theoretically be applied to humans, and we could anticipate major shifts in the balance of power because GDP is known to correlate highly with the IQ of the smartest fraction in a society8, but many people would have difficulty coming to grips with such a scenario because it seems so fantastic unless one is familiar with all the points in the argument. In general, people seem to have an aversion to considering the effects of major intelligence enhancement. Nuclear war might be easier to understand than Artificial Intelligence, but it seems like the risk of human extinction in the long term rests more heavily on the latter than the former.

**28. Representation that global catastrophe will be caused by only one reason**

When people usually think about global catastrophes, or watch movies about them, there is usually only one unitary effect doing all the damage; an asteroid impact, or nuclear war. In reality, however, there is barely a limit to how many complex factors may intertwine and mutually reinfroce. There may be a nuclear war, which triggers the explosion of a Doomsday Machine network of nuclear bombs that were put in place decades ago, which is followed up by grueling biological warfare and nuclear winter. Disasters tend to come in groups, since they have the propensity to trigger one another. Few academic studies of global risks have adequately analyzed the potential of complex overlapping catastrophes.

Besides disasters consisting of discrete events triggering one another, new technologies also enable one another. For instance, today the greatest risk to humanity is nuclear warfare, but the age of nuclear danger may only be just beginning; advances in nanotechnology will make it much easier to enrich uranium cheaply. Even the smallest of rogue states might become able to mass-produce nuclear weapons for nominal cost. How dangerous would the world become then? Few attention is given to the effects of technological convergence on global risks.

**29. Underestimating systemic factors of global risk**

Systemic factors are not separate events, like the sudden appearance of a superviruses, but certain overall properties which concern an entire system. For instance, the conflict between the nature of exponentially growing modern civilization but a linearly increasing (or less) availability of material resources. The conflict is not localized in any one place, and does not depend on any one concrete resource or organization. Self-magnifying crisis situations which tend to involve a greater number of people over time do exist, but do not necessarily depend on human behavior and have no center. We are only beginning to understand the scope of these possible issues. Another example would be the Kessler syndrome, which states that as the amount of space junk increases, it will impact other space junk and eventually cause an unnavigable mess of obstacles to orbit the Earth, making space travel very difficult9. Though this is not a global risk, there may be other risks of this nature which we haven't foreseen yet.

**30. Underestimate of precritical events as elements of coming global catastrophe**

If as a result of some events, the probability of global catastrophe substantially increases (for instance, we find an extremely cheap way of enriching uranium), that event may contribute a large amount of probability mass to the likelihood of global disaster, but be underweighted in importance because it is not a disaster itself. Another example is that of a nuclear war--although a nuclear might kill “only” one billion people, the ensuing nuclear winter and breakdown of polite civilization would be much more likely to usher in human extinction over the long-term than if the war never occurred. We call these “precritical events”. More effort needs to be devoted to dangerous precritical events, which may kill no one in and of themselves, but contribute substantial probability mass to global catastrophe.

**31. Cognitive biases based on the idea: “It is too bad to be true” or “It couldn't happen to me”**

It is easy to imagine someone else's car wrecked, but much more difficult to imagine the future fragments of your own car. It is easy to imagine someone else getting shot and killed, but harder to imagine yourself being shot and bleeding out. Bad thoughts aren't fun to think about. Even discussing them is associated with low social status. Instead of dealing with challenges to our existence, it is much easier to just ignore them and pretend they will never happen.

**32. Cognitive biases based on the idea: “It is too improbable to be the truth”**

We have many historical examples of how something, that was "improbable", like powered flight or splitting the atom, suddenly became possible, and then ordinary. Moreover, it may become mortally dangerous, or responsible for millions of lost lives. It is necessary to separate “improbable” from physically impossible. If something is physically possible and can be done, people will do it, even if it sounds crazy. It is more useful to humanity to discover an “improbable” global risk than it is to build a new widget that provides a momentary distraction or slightly optimizes some industrial process.

**33. Ideas about "selective relinquishment," deliberately deciding not to develop certain technologies as a way of dealing with global risks**

In the short term, it may be possible to hold back the development of certain technologies in a certain place. But in the longer term, as scientists and military leaders in more countries encounter the prospect of a powerful technology, it will eventually be developed. There is no world government that can use force to prevent the development of a certain technology everywhere. If a technology is attractive, funds and people will gravitate towards wherever it can be suitably developed. This already happens in the realm of medical tourism and thousands of other areas, both licit and illicit.

**34. Representations that human adaptability is high and continues to grow beyond all bounds thanks to new technologies**

Human adaptability is high, but the power of attack is progressively overtaking the power of defense. There is no missile defense barrier that can resist a full-on nuclear attack. There is no universal vaccine that can immunize people to resurrected or engineered superviruses. There is no easy way to defend against a Predator drone lobbing a missile in your general direction. It is always easier to break something than to fix it.

**35. Inability of a system to simulate itself**

Though we cannot investigate global catastrophes experimentally, we can build 'best guess' probabilistic models with computers. However, there are naturally chaotic effects that cannot be predicted with typical models. A model can never be complete because it is difficult for a system to model itself, besides taking into account feedback effects between modeling outcomes and subsequent human decisions. A model can never fully model a situation because it would have to predict its own results and their downstream effects before they happen, which is impossible.

**36. Approaching life in the spirit of the proverb: “after us the deluge”**

Madame de Pompadour, a French aristocrat who was the courtesan of Louis XV, wrote “After us, the deluge. I care not what happens when I am dead and gone.” Cognitive psychology experiments on time discounting have shown that people strongly discount risks that apply only after their expected lifespan, or even earlier than that10. A "devil take all" attitude in the contexts of catastrophic risks could be rather dangerous, but some of it is inevitable. There is also the thought, "if I'm going to die, other people ought to die too," which actually enhances global risks rather than having a neutral effect.

**37. Religious outlooks and eschatological cults**

The study of global risks infringes on territory that has since time immemorial been associated with religion. This gives it an unscientific gloss. However, mankind imagines many things before they happen, and religion and myth are the primary historical engines of such imagination, so such an association is inevitable. Consider all the concepts from myth that became or were verified as real: flying machines (Vedas), weapons that destroy entire cities (the Vedas), transubstantiation (New Testament and medieval occultism, particle accelerators and nuclear reactions transform basic elements), catastrophic flooding (Old Testament, search for "Black Sea deluge hypothesis" on Google), miraculous healing (the benefits of modern medicine would be considered miraculous from the perspective of a couple centuries ago), and so on. Just because something has been mentioned in stories does not mean it can't actually happen.

**38. Uncertainty and ambiguity of novel terminology**

Describing global catastrophic risks, we may use terms that have no unequivocal interpretation, since they describe events and technologies might may not be mature yet. The use of such terms throws up red flags and may lead to misunderstanding. Yet, just because we use difficult terms does not mean that the technologies we refer to will not reach maturation and threaten our continued survival as a species.

**References**

* Darley, J. M. & Latané, B. "Bystander intervention in emergencies: Diffusion of responsibility". 1968. Journal of Personality and Social Psychology 8: 377–383.
* Michael Thomas Moore, David Fresco. "Depressive Realism: A Meta-Analytic Review". *Clinical Psychology Review* **32** (1): 496–509. 2012.
* Barry Schwartz. *The Paradox of Choice: Why More is Less*. 2004. Harper Perennial.
* Daniel Kahneman. "Evaluation by moments, past and future". In Daniel Kahneman and Amos Tversky (Eds.). Choices, Values and Frames. p. 708. 2002.
* Peter Diamandis. *Abundance: The Future Is Better Than You Think*. 2012. Free Press.
* Alvin Toffer. *Future Shock*. 1970. Random House.
* Eliezer Yudkowsky. “Future Shock Levels”. 1999. <http://sl4.org/shocklevels.html>
* La Griffe du Lion. “The Smart Fraction Theory and the Wealth of Nations”. *La Griffe du Lion*, Volume 4 Number 1, 2002. <http://www.lagriffedulion.f2s.com/sft.htm>
* Donald J. Kessler and Burton G. Cour-Palais (1978). "Collision Frequency of Artificial Satellites: The Creation of a Debris Belt". Journal of Geophysical Research 83: 2637–2646.
* Frederick, Shane; Loewenstein, George; O'Donoghue, Ted (2002). "Time Discounting and Time Preference: A Critical Review". Journal of Economic Literature 40 (2): 351–401.

## Chapter 3. How cognitive biases in general influence estimates of global risks

## **Chapter 3. How cognitive biases in general influence**

## **estimates of global risks**

**1. Overconfidence**

Overconfidence, in the formal language of heuristics and biases research, means an overinflated conception of the accuracy of one's model of the world, probability estimates, and a diminished ability to update beliefs sufficiently in response to new information1. Numerous studies in cognitive psychology have confirmed that overconfidence is a universal human bias, present even in professionals such as statisticians who have been trained to avoid it2. Overconfidence is associated with other cognitive tendencies such as egocentric biases, which have adaptive value in an evolutionary context3. An overconfident person is more likely to acquire status in a tribe and to attract high-value mates, passing on their genes and producing offspring with the same cognitive tendencies. This shows why overconfidence tends to persist in populations over time despite its downsides.

In the context of global risks, overconfidence can manifest in a number of ways. The most notable is to fixate on a preferred future, rather than considering a probability distribution over many possible futures4, and to ignore data contradicting that preferred future. The tendency to selectively interpret information to suit one's biases is called confirmation bias, which overlaps with overconfidence. One possible way of overcoming overconfidence is to solicit the opinions of others regarding possible futures and seriously consider these alternatives before building a more conclusive model. This is called “taking the outside view,” outside view being a term from the field of reference class forecasting, in which predictions are made using a reference class of roughly similar cases.

**2. Excessive attention to slow processes and underestimation of fast processes**

Slower processes are more convenient for analysis and prediction, and tend to have more data useful for interpreting them. However, dynamic systems are more likely to adapt to slow processes and collapse from fast processes, making it more important to examine fast processes in the context of global risk, including those which seem improbable at first glance. A sudden catastrophe is much more dangerous than a slow decline, as there is less time to prepare for it or intelligently respond. Many emerging technologies of the 21st century—nanotechnology, biotechnology, Artificial Intelligence—concern fast processes which are poorly understood. This does not mean that slow threats shouldn't be analyzed as well, just that we should be aware of a natural tendency to ignore fast processes and focus only on slow ones.

**3. Age features in perception of global risks**

Younger people are more oriented towards risk-taking, fun, overconfidence, neophilia, and gaining new territory, rather than on safety5. Older people may be more risk-averse than younger people6, overvalue highly certain outcomes both in risks and gains7, value safety and predictability more, and are underrepresented in groups exhibiting interest in and concern with emerging technologies, such as those under the umbrella of transhumanism8. So, different age groups have different cognitive biases particular to their generation which may confound attempts at realistically assessing global risks. Some of these biases may be advantageous, others not—the word “bias” does not always necessarily have a negative connotation, but must be evaluated in a context-dependent way.

**4. Polarization through discussion**

Discussions within groups usually lead to polarization of opinions. This has been demonstrated by hundreds of studies and is formally called “group polarization”9,10. If someone approaches a discussion with two hypotheses that he gives about equal probability, he is likely to polarize in a direction opposite to that of an opponent. One facet of this is that many people have some tendency to argue for fun, or just for the sake of arguing. As enjoyable as this may be, in practice it narrows our representation of possible futures. Bayesian rationalists, who are trying to build the most realistic possible model of the future, should approach discussions with truthseeking in mind and be wary of arbitrary polarization, such as that caused by group polarization or otherwise. In fact, true Bayesian rationalists with common priors cannot agree to disagree on matters of fact, and must converge towards consensus on probability estimates11.

**5. Skill at arguing is harmful**

In the book chapter “Cognitive biases potentially affecting judgment of global risks,” Eliezer Yudkowsky makes the point that a skillful arguer may be superior at arguing any arbitrary point, making it more likely for him to anchor on a mistaken hypothesis12. In the study of cognitive biases, this is known as anchoring. The better knowledge of cognitive biases an arguer has, the more readily he can attribute biases to his opponent, while being blind to his own biases. In this way, a smarter thinker might actually use his intelligence against himself, making it more difficult to construct an accurate model of risks. He may also fall prey to framing effects—that is, certain ways of thinking about things—and use his intelligence to reinforce that frame, rather than considering alternate framings13. For instance, someone preoccupied with the idea that advanced Artificial Intelligence would do X may never be argued out of that hypothesis and consider that AI might actually do Y instead.14

**6. Excessively conservative thinking about risk possibilities   
associated with memetic selection effects**

In the book *The Selfish Gene*, Richard Dawkins introduced the idea of ideas as replicators circulating in society as an ecosystem, analogous to genes. He used the term “memes” to refer to ideas. In order to be protected from stupid or useless ideas, human minds have a natural ideological immune system that causes us to dismiss most novel ideas we run across. In general this is useful, but in the context of global risks, which may involve blue sky concepts such as the long-term future of humanity and the power of newly emerging technologies, it can be a stumbling block. Consider times in the past where useful new ideas were rejected but could have saved many lives or much misery. The concept of utilizing nitrogen oxide for anesthesia in surgery was discovered near the end of the 18th century, but it wasn't until 1840, almost 50 years later, that it was actually applied. There was a similar story with hand-washing as a tool for hygiene; Ignaz Semmelweis pioneered its use in the 1840s, when it lowered death rates in his clinic by a factor of 3 to 4 times, yet the practice didn't catch on with skeptical doctors until decades later. Semmelweis was so enthusiastic about handwashing that authorities had him committed to a mental institution, where he soon died. In the 21st century, a time of rapid changes, it would behoove us not to make the same mistakes, as the stakes could be much higher, and the punishment for ignorance greater.

**7. The burden of proof is on a designer to prove that a system is safe,   
not on the users to prove that a catastrophe is likely**

In creating safe systems, there are two kinds of reasoning—first, trying to prove that a certain system is safe. Secondly, arguing that the system is susceptible to certain concrete dangers. These modes of reasoning are not logically equivalent. To deny the general statement that something, say the Space Shuttle, is safe only requires one counterexample. In the case of the Challenger disaster, we received a vivid display that a failed O-ring seal was enough to cause the craft to break up on reentry, causing tragic loss of life. However, refuting a particular criticism of safety—say, demonstrating that the Space Shuttle could sustain micrometeorite impacts in orbit—does not constitute a proof of the general case that the Space Shuttle is safe. There may be a thousand other threats, like the failed O-ring seal, which could cause the craft to break apart. However, our minds have a tendency to think as if a comprehensive refutation of all the threats *we can think of at the time* is sufficient to declare a piece of equipment “safe”. It may very well remain unsafe. Only through many years of repeated use and in-depth analysis can we truly verify if a system is safe with a high level of confidence. Empiricism is important here.

In analyzing the safety of a system, we must exercise vigilance to make sure that errors in reasoning have not compromised our ability to understand the full range of possible disaster scenarios, and not get too excited when we refute a few simple disaster scenarios we initially think of. In complex technical systems, there are always conditions of *maximum load* or *design failure threshold*, where if certain external or internal conditions are met, the system will become unsafe. Eventual failure is guaranteed, it's just a question of what level of environmental conditions need to be met for failure to occur. It is important to consider the full range of these possible scenarios and check our reasoning processes for signs of bias.

From a scientific point of view, it is always easier to prove that an object does exist, than to prove that an object conclusively does not exist. This effect is the same when it comes to safety—we can prove that a design is safe in general conditions, but we cannot conclusively prove that unusual conditions will not combine to threaten the safety of the design. The only way we can do that is through an extensive service lifetime, and even then, we can never be completely certain. The amount of rigor we put towards evaluating the danger of a new construct should be proportional to the damage it can cause if it undergoes critical failure or goes out of control. Large, important systems, such as nuclear reactors can do quite a lot of damage. The burden of proof is always on the designers to prove that a system is safe, not on critics to show that a system is unsafe. Steven Kaas put it pithily: “When you have eliminated the impossible, whatever remains is often more improbable than your having made a mistake in one of your impossibility proofs.” In other words, we often make basic mistakes in our reasoning, and should reevaluate disaster modes from time to time even if we think they are ruled out.

**8. Dangerous research tends to be self-preserving**

Every institution aspires to its own self-preservation, and talk of dangers can lead to the end of funding for new projects. Accordingly, scientists involved in these projects have every personal incentive to defend them and downplay the dangers. No funding, no paycheck. Unfortunately, in cases where the area of research in question threatens to annihilate mankind for eternity, this can be rather dangerous to the rest of us. For instance, one wonders why Google bought and is now pumping money into Boston Dynamics, who are primarily known for manufacturing military robotics. We thought their motto was, “Do no evil..?” It is unfortunate that the most powerful and exciting technologies tend to be the most dangerous, and have the most qualified and charismatic scientists and spokespersons to defend their continued funding. This is why every project must be subject to some degree of objective outside scrutiny. People close to potentially dangerous corporate projects—like the robotics projects at Google—have an ethical obligation to monitor what is going on and notify groups concerned with global catastrophic risk in case anything thing is amiss. The safety of the planet is more important than Google's stock price.

**9. Erroneous representation that when the   
problem arises there will be time to prepare for it**

Most serious problems arise suddenly. The more serious a problem, the greater its energy and—usually—the faster the threat develops after it initially emerges. This makes it all the more difficult to prepare for. Global catastrophes are powerful problems. Therefore they can develop too quickly to prepare for. We do not have experience which would allow us to define harbingers of global catastrophe in advance. Analogously, auto accidents occur suddenly. There may not be time to prepare after signs of the problem appear; we need to forecast it in advance and get safeguards in place. An analogous situation would be Hurricane Katrina causing water to rise above the levees in New Orleans. If New Orleans were adequately prepared in advance, it would have built levees tall enough to hold back water from even the worst foreseeable storms. But it did not, and catastrophic damage resulted. By the time the hurricane is sighted on radar, it is already too late.

**10. Specific risks are perceived as more   
dangerous than more general risks**

An experimental result in cognitive psychology is that stories with more details sound more plausible, even if its probability is lower than the general case and the entire story is made up. This is called the conjunction fallacy15. For example, “mutiny on a nuclear submarine” sounds more dangerous than “a large sea catastrophe,” though the former is a subcategory of the latter. Yudkowsky writes: “From the point of view of probability theory, adding more detail to the story makes it less likely… but in terms of human psychology, the addition of each new detail makes the story all the more credible.” Many people are not familiar with this basic logical and probabilistic truth, and focus instead on highly specific scenarios at the expense of more general concerns.

**11. Representations that thinking about global risks is pessimistic**

This is related to earlier points about overconfidence and memetic selection effects. Considering doom (or being pessimistic at all) is associated with low status, so people don't want to be associated with it or approve of it. Contrast this with the experimental result that pessimistic people tend to make more accurate probability estimates, called “depressive realism”. The 21st century is a minefield; if we're going to traverse it, we should do so cautiously. Dancing on it blindly is not optimism, but stupidity.

**12. Conspiracy theories as an obstacle for the scientific analysis of global risks**

Global risk may sometimes be associated with conspiracy theories such as the Illuminati or the idea that lizard men control the world. Of course, every effort must be made to develop the study of catastrophic global risks into a serious discipline that has no association with such nutty theorizing.

When a conspiracy theory predicts a certain risk, it tends to be highly specific: the world will end on such-and-such a date, caused by such-and-such event. In contrast, the scientific analysis of risk considers many possible risks, as probability distributions over time, including the possibility of complex overlapping factors and/or cascades. In addition, conspiracy theory risks will often come with a pre-packaged solution; the promulgator of the conspiracy theory happens to know the one solution to deal with the risk. In contrast, those engaged in the scientific study of global catastrophic risk will not necessarily claim that they have any idea how to deal with the risk. If they do have a proposed solution, it will not be entirely certain—rather, it will be a work in progress, amenable to further suggestions, elaborations, redundant safety elements and precautions.

The more poorly we predict the future, the more dangerous it is. The primary danger of the future is its unpredictability. Conspiracy theories are harmful to future prediction, not just because of their general lunacy, but because they focus attention on too narrow a set of future possibilities. Furthermore, they assume superconfidence in the prognostic abilities of their adherents. A good prediction of the future does not predict concrete facts, but describes a space of possible scenarios. On the basis of this knowledge it is possible to determine central points in this space and deploy countermeasures to deal with them.

Such "predictions" undermine trust in any sensible basis underlying them, for example that a large act of terrorism could weaken the dollar and potentially cause an economic collapse as part of a chain reaction. Conspiracy theories also tend to fall prey to the fundamental attribution error—that is, that there must be a “Them,” a deliberately malevolent actor on whom to place the blame. In reality, there is usually no such arch-villain; there may be a group, or the disaster might happen as a complete accident, or as a result of human actions which are not deliberately malevolent. Focusing on conspiracy theories, or allowing conspiracy theories to influence our thinking in any way, biases our thinking in this way.

At the same time, although the majority of conspiracy theories are false, there is always the wild chance that one of them could turn out to be true. There was a conspiracy theory by industrialists to take over the U.S. Government in a coup during 1933, the so-called Business Plot, but it was foiled when a General in on the plot decided to report it to the government. Conspiracies do exist, but conspiracy theorists tend to overestimate the degree to which groups and individuals are capable of covert coordination. Regardless, consider the saying: if you cannot catch a cat in a dark room, that does not mean that the cat is not present.

**13. Errors connected with the conflation of short-term,   
intermediate term and long-term forecasts**

A short-term forecast considers the current condition of a system, and the majority of discussions focus on that theme when considering policies for actions to actually take. Intermediate term forecasts consider further possibilities of a system and consider its current tendencies and direction rather than just its immediate state. Long-term forecasting is much more expansive and considers a variety of long-term possibilities and end states.

Consider we have a ship with gunpowder on which sailors go and smoke. It is in short term possible to argue that if sailors smoke in a certain place, as far away as possible from the gunpowder, an explosion will not happen. But in the intermediate term, it is more important to consider general statistics, such as the quantity of gunpowder and smoking sailors which define the probability of explosion, because sooner or later a smoking sailor will appear in the wrong place. In the long term, there is an essentially unsafe situation, and an explosion is bound to occur. The same holds with the threat of nuclear war. When we discuss its probability over the next two months, the current concrete behavior of world powers matters. When we consider the next five years, we should take to the account the overall quantity of nuclear powers and missiles, and not focus too much on current events, which may change quickly. When we speak about the prospect of nuclear war over the timescale of decades, an even more fundamental variable comes into play, which is the overall technological difficulty of enriching uranium or producing plutonium. Different logical frames are useful for best considering different time frames.

Thus in different areas of knowledge the appropriate time scale of a forecast may differ. As interesting as the relationship between Obama and Putin may be for current world affairs, it is not likely to be relevant to the long-term unfolding of nuclear war risk. Depending on the industry or area under consideration, what is considered a “short-term forecast” or a “long-term forecast” may vary. For example, in the field of coal output, 25 years is a short-term forecast. In the field of microprocessor fabrication, it may be as short as 4 months.

**14. Fear**

Fear evolved in human beings in response to concrete stimuli in concrete situations. Our visceral emotion of fear is not attuned to deal with remote, general, or abstract risks. In fact, when it comes to remote risks, fear works against us, because it motivates us not to consider them, as the prospect is vaguely scary, but not terrifying enough to make us actually care. An example would be when a man refuses to get a medical analysis of a bump on his prostate because he is afraid that something malignant will be found.

**15. Underestimating importance of remote events (temporal discounting)**

It is a natural tendency of human reasoning to assign lesser importance to events which are distant in space and/or time. This is called discounting, and quantified with a variable known as the *discount rate*. It obviously makes sense in an evolutionary context, but in the modern age, the usefulness of our ancestral intuitions is beginning to waver16. A hundred thousand years ago, in the environment in which humans evolved, there was no such thing as a nuclear missile, or a drone that can travel around the world and hit you with a bomb in a few hours. Today, global risks might originate from distant lands and decades in the future, but we might need to begin preparing for them now. We can hardly do that if we engage in hyperbolic discounting, that is, completely ignore risks outside of a certain time or space window, say, 5 years in the future. We might rationalize ignoring such risks by saying there is nothing we can do about them, but there most certainly is something we can do about them. Look at the grassroots effort to deal with global warming; this is an example of a risk that is being prepared for far in advance.

**16. Effect of displacement of attention.**

The more someone gives attention to one global catastrophe, the less he gives to another and as a result his knowledge will become specialized. Therefore supervaluation of any one global risk is conducive to the underestimation of others and may be harmful. The negative effect may be ameliorated if the thinker cooperates with other specialists and gives due respect to their area of global risk expertise. We do see this effect in the global catastrophic risk community today, though there are certain gaping holes in activity and knowledge, such as with respect to the issue of nanotechnological arms races. In the current global catastrophic risk analysis community, the primary focus is on Artificial Intelligence. While Artificial Intelligence does indeed seem to be a major risk, there may be other land mines on the road to it that need taking care of for us to even approach advanced AI.

**17. The Internet as a source of possible errors**

The Internet naturally promotes a certain kind of bias; mostly for the sensational. Search engines like Google even optimize their returned results based on your prior searches, showing you what they think you want to see. This can make it difficult to branch out from a certain niche, and exacerbates confirmation bias, the reception of data that confirms what we already think. In addition, there is always a lot of low-quality noise associated with any concept or idea. Even quality journals like *Nature* cannot necessary be trusted, as peer review is fraught with all kinds of bias—for sensational results, for results that operate within a certain scientific paradigm, results that adhere to the framing of a dominant scientist, and so on. Older scientists receive all the grant money, meaning they dictate the flavor of much of contemporary research. Max Planck's old saying comes to mind: “Science advances one funeral at a time.” Of course, the wider amount of content on the Internet means that if there is good content, and if someone is diligent about searching for it, it will eventually be found. The Internet also improves the speed of research, allowing a researcher to cycle through poor research more quickly and cheaply than may otherwise have been possible.

**18. Beliefs**

Strong social or religious beliefs, especially religious beliefs, can powerfully influence or bias estimates of global risk. Many Christians believe that the course of history is directed by God, and that he would never allow humanity to be exterminated for an arbitrary reason. It would be disappointing if we were then wiped out because not enough Christians cared about global risks to participate in doing anything about them.

Aside from Christianity, there are also pseudo-religious beliefs associated with progressivism and liberalism that bias risk estimates. Questioning global warming, or the severity of global warming, is often considered heresy according to the dominant progressive paradigm. Emails from the ClimateGate controversy, of which there were two rounds of emails, made it very clear that one prominent group of climate scientists were only looking for the “right” conclusions, scientific objectivity be damned. Another example is that those with engrained liberal beliefs might find it personally difficult to consider the effects of mass intelligence enhancement creating humans with superior capabilities, because of distaste for the idea of innate inequality.

In his book on global risks *Our Final Hour*, Sir Martin Rees writes that in Reagan's environmental administration, the religious fundamentalist James Watt, Secretary of the Interior, believed that the arrival of the Apocalypse might be accelerated by destruction of the environment17. There are many other examples in this vein, too many to count.

**19. Congenital fears**

Many people have congenital fears of snakes, heights, waters, fear of impending collision, illnesses, and so on. Many of these fears are likely human universal. It is not so difficult to assume that they might overestimate the seriousness of events reminding them of these fears, or to underestimate those which are unlike them. Post-traumatic stress syndrome, such as an earlier illness, may also influence how one looks at certain risks. Most people have a preferred life narrative that is absent of global catastrophe (don't we all?), so they may irrationally underestimate its objective likelihood.

**20. Shooting the messenger**

Discussion of risks can trigger discontent. This discontent may be directed towards the bearer of the message rather than addressing the threat itself.

**21. Difficulty in delimitation of own knowledge**

I do not know what I do not know. It is tempting to subconsciously think that we personally know everything that is important to know, even in outline. This leads to a false sensation of omniscience, conducive to intellectual blindness and an unwillingness to accept new data. Albert Camus said that the only system of thought that is “faithful to the origins” is one which “notes its limits”. We should always be keeping an eye out for unknown unknowns, and realize that they must exist, even if we cannot currently imagine what they are.

**22. Humor**

It is possible to misinterpret a genuine threat as a joke, or interpret a joke as a genuine threat. Reagan once joked to technicians, prior to a televised speech, *“My fellow Americans, I'm pleased to tell you today that I've signed legislation that will outlaw Russia forever. We begin bombing in five minutes.”* This comment was never broadcast, but it did leak later. If it were leaked early or accidentally taken seriously, who knows what kind of mayhem it may have caused. At the very least, it would have heightened tensions. Similarly, senator John McCain “jokingly” sung the tune, *“Bomb, Bomb Iran,”* in response to an audience question at a campaign stop during the 2008 Presidential elections.

**23. Panic**

A hyperactive reaction to stress leads to erroneous and dangerous actions. For example, a man may jump out of a window during a fire although the fire has not reached him, causing his premature death. It is obvious that panic influences and thoughts and actions of people in a stress condition. For example, famous engineer Barnes Wallis was described as a religious man and a pacifist during peacetime, but during World War II developed a plan of using “bouncing bombs” to destroy dams in Germany to flood the Ruhr valley18. This is an example of how panic and acute threats change normal modes of behavior. Panic can be long-term, not just acute. Short-term panic is also very dangerous, as a situation of global risk may develop very quickly, in hours or even minutes, and calm strategic decisions will need to be made in that time.

**24. Drowsiness and other mundane human failings**

According to one account, Napoleon lost at Waterloo because had a chill19. How is it reasonable to expect that the President of the United States would make the best possible decision if he is abruptly awoken in the middle of the night? Add in the basic inability of human beings to precisely follow instructions, and the finiteness of instructions it is possible to execute, and you have plenty more obvious limitations which come into play in a crisis. Someone might be cranky that their girlfriend dumped them, they have loose stools, or they just haven't had enough to eat lately and their blood sugar is low. Even a trained soldier might experience a slight lapse in consciousness for no reason at all. Given how easy it is for a head of state to make a call and initiate nuclear war, it is hard to overestimate how mundane or seemingly stupid the reason for starting a conflict or missing a crucial safety detail may be.

**25. Propensity of people to struggle with dangers which are in the past**

After the tsunami of 2004, Indonesians and other southeast Asians began to build many systems of prevention and warning for tsunamis. However, the next major disaster in the area might not be a tsunami, it could be something else. In this fashion, people may be preparing for a disaster which already happened and is not a threat, neglecting a future disaster.

**26. Weariness from catastrophe expectation**

If you live in a city which is constantly being bombed, you may eventually stop caring, even if there is a constant risk you will be blown to bits. During the London Blitz of World War II or the Siege of Leningrad, many citizens went about their business normally. This effect has been called crisis fatigue. After September 11th, many skyscrapers around the world were put on alert, in expectation of further attacks, but none occurred, and security went back down again. Since the periodicity of major disasters might consist of many years, people may become complacent even as the objective probability of such an event gradually increases. The probability of a large earthquake in California continues to increase, but many of the buildings in San Francisco were built in a hurry after the last major earthquake of 1906, and are not at all earthquake-safe or earthquake-ready, meaning the effects of another major earthquake could lead to even more loss of life than in 1906. These buildings house at least 7 percent of current residents, likely more20. The weariness of catastrophe expectation is expressed by the loss of sensitivity of a society to warnings.

**27. An expert’s estimates which are not based on strict calculations  
cannot serve as a measure of real probability**

Unlike in the stock markets, where the average estimate of the best experts is used as a forecast of market behavior, we cannot select our experts and average them based on their track record of predicting human extinction, because there is no track record of such an event. If it had happened, we would all be dead, and quite incapable of predicting anything.

Slovic, Fischhoff, and Lichtenstein (1982, 472)21, as cited in Yudkowsky (2008)22 observed:

*A particularly pernicious aspect of heuristics is that people typically have great confidence in judgments based upon them. In another followup to the study on causes of death, people were asked to indicate the odds that they were correct in choosing the more frequent of two lethal events (Fischoff, Slovic, and Lichtenstein, 1977) In Experiment 1, subjects were reasonably well calibrated when they gave odds of 1:1, 1.5:1, 2:1, and 3:1. That is, their percentage of correct answers was close to the appropriate percentage correct, given those odds. However, as odds increased from 3:1 to 100:1, there was little or no increase in accuracy. Only 73% of the answers assigned odds of 100:1 were correct (instead of 99.1%). Accuracy “jumped” to 81% at 1000:1 and to 87% at 10,000:1. For answers assigned odds of 1,000,000:1 or greater, accuracy was 90%; the appropriate degree of confidence would have been odds of 9:1. . . . In summary, subjects were frequently wrong at even the highest odds levels. Moreover, they gave many extreme odds responses. More than half of their judgments were greater than 50:1. Almost one-fourth were greater than 100:1. 30% of the respondents in Experiment 1 gave odds greater than 50:1 to the incorrect assertion that homicides are more frequent than suicides.*

The point of this quote is to illustrate that experts are consistently overconfident, often ridiculously so. From Parkin's *Management Decisions for Engineers*23:

*Generally, people have a displaced confidence in their judgment. When asked general knowledge or probability questions, experimental subjects performed worse than they thought they had (Slovic et al., 1982). Calibration experiments that test the match between confidence and accuracy of judgment, demonstrate that those without training and feedback perform badly. Lichtenstein et al. (1982) found that from 15,000 judgments, when subjects were 98% sure that an interval contained the right answer they were wrong 32% of the time. Even experts are prone to some overconfidence. Hynes and Vanmarke (1976) asked seven geotechnical gurus to estimate the height of a trial embankment (and their 50% confidence limits), that would cause a slip fracture in the clay bed. Two overestimated the height and five underestimated. None of them got it within their 50% confidence limits. The point estimates were not grossly wrong but all the experts underestimated the potential for error.*

Simply put, “experts” are often wrong. Sometimes their performance is equal to random chance, or to that of a person pulled off the street. Statistical prediction rules often outperform experts24. This creates trouble for us when we rely on experts to evaluate the probability and nature of catastrophic global risks.

**28. Ignoring a risk because of its insignificance according to an expert**

This ties into the above point. If an expert thinks a risk is insignificant, he may be wrong. It is also necessary for the expert to quantify the precise degree of insignificance they are talking about, which they often refuse to do or are unable to do. For instance, say that we determine that the probability that a certain particle accelerator experiment destroys the planet is one in a million. That sounds low, but what if they are running the same experiment a hundred times a day? After a year, the probability of doom is already 1/300. After 500 or so years, the probability of doom approaches unity. So, “one in a million” may actually be a quite significant risk if the risk is repetitive enough.

Aside from having a proper understanding of insignificance, we also ought to keep in context the relationship between an estimate of insignificance and the chance that the expert making the prediction is in error. Say an expert says that an event only has a one in a billion chance, but there is a 50% probability that they are completely wrong. In that case, the real insignificance might be just one in two, or one in twenty. The probability that an expert is just plainly wrong often throws off the estimate, unless there is a rigorous statistical or empirical basis to confirm the estimate. Even then, the empirical data may lie, or there may be an error in the statistical calculations, or a mistaken prior.

**29. Underestimating or overestimating our ability to resist global risks**

If we underestimate our ability to resist global risks, we might fail to undertake actions which could rescue us. If we overestimate our abilities to resist it, it could lead us to excessive complacency. We need to find a balance.

**30. Stockholm syndrome**

Most humans have a relationship to death similar to Stockholm syndrome; that is, similar to the relationship between hostages who become attached to their kidnappers. The hostage feels helpless and controlled by the kidnapper, and, perhaps as a survival mechanism, begins to fall for them. This can proceed to the point of the hostage being willing to risk their life for the kidnapper. The same thing applies to humans and our relationship to death. We feel it is inevitable we'll die one day, and even begin to acquire an aesthetic love for it. Just because we will die one day doesn't mean that we shouldn't attempt to ameliorate global catastrophic risks.

**31. Behind errors of the operator there is an improper preparation**

Behind the concrete errors of pilots, operators, dispatchers, and politicians, there is often a conceptual error in operator training or a warning flag that something was wrong with the pilot or operator in advance which goes ignored. In March 2014, a cheating scandal among proficiency tests at Maelstrom Air Force Base, Montana, for nuclear force officers led to the firing of nine missile wing commanders and the resignation of the commander of the 341st missile wing25. This is an example of setting standards and sticking to them. If a nuclear force officer cannot pass a proficiency test without cheating, what business do they have controlling devices which could cause the end of the world as we know it? The more important the technology, the more rigorous and failproof the evaluation and competency testing processes need to be. Scientific analysis of global risks and the promulgation of such knowledge should be considered part of “proper preparation” for mankind as a whole.

**32. A group of people can make worse decisions than each person separately**

Depending on the form of organization of a group, it can promote or interfere with the development of intelligent decisions. A good example might be a scientific research institute, a bad example would be a mob or a country in the middle of civil war. The influence of a crowd can bring the thinking level down to the lowest common denominator26. The “wisdom of crowds” is often better suited to estimating the number of gumballs in a large jar than making highly complex, technical decisions. That is why most “expert surveys” are restricted to a relatively small number of experts. The majority of people do not have knowledge to make complex decisions, and should not be asked to. Until there is a uniform decision making and threat evaluation process for global risk, we are probably in a sub-par situation.

**33. Working memory limits**

A person can only focus on a few things at a time. Will A attack B? Maybe yes, or maybe no, but even that framing alone leaves out important details, a sort of attention shade. One human or even an organization cannot capture all aspects of world problems and perfectly arrange them by order of their degree of danger and priority. This is why computer models can be helpful, because they process more simultaneous details than a human can. Of course, computer models can fail, and need to augment human reasoning, not replace it.

**34. Futurology is split across different disciplines   
as though the underlying processes occur independently**

There are several variants or genres of thinking about the future, and they have the propensity not to be interconnected in the intellectual world or in thoughtspace very much, as if these futurist domains were in entirely different worlds.

* Forecasts around the theme of accelerating change, NBIC convergence, and the Singularity. Supercomputers, biotechnologies, and nanotechnology.
* Forecasts of system crises in economy, geopolitics and warfare. This tends to be a different crowd than the Singularity crowd, though there is some overlap.
* Forecasts in the spirit of traditional futurology, such as demographics, resource limitations, global warming, *et cetera*.
* Special type of forecasts for big catastrophes: asteroids, supervolcanoes, coronal mass ejections from the Sun, the Earth's magnetic field flipping, and so on.

To accurately predict the future, we must reconcile data from each of these domains and take a holistic view.

**35. A situation when a bigger problem follows  
a smaller one, but we are incapable of noticing it**

There is a Russian proverb, “trouble does not come alone”. The American version is “When it rains, it pours”. Global catastrophe could occur as a result of a chain of progressively worse events, however we might get distracted by the first disasters in the chain and fail to prepare for the larger ones. The reasons may be:

* Our attention at the first moment of failure is distracted and we make a critical error. For instance, a driver almost rear-ends the driver in front of him, then decides to quickly go around him without thinking, causing him to get hit by a car in the other lane. Or, a man carrying out a robbery decides to shoot at a policeman who comes to arrest him, and is gunned down in return. Or, something falls off a cliff, a person goes to look for it, and ends up falling off himself. The possibilities here are quite expansive.
* Misunderstanding that the first failure creates a complex chain of causes and effects which causes the person or civilization under threat to inadequately respond to subsequent threats. The first disaster weakens the organism and it becomes susceptible to further disasters or maladies. For example, flu can lead to pneumonia, or nuclear war could lead to nuclear winter.
* Euphoria from overcoming the first catastrophe causes the group to lose prudence. For instance, someone who suffers an accident and is in the hospital begins to recover somewhat, and decides to leave the hospital prematurely, leading to inadequate healing and permanent injury.

**36. Selectivity of attention**

Often, when people are looking for certain weaknesses, for instance in the economy, they may tend to overfocus on one issue, like subprime mortgages. This causes a certain selectivity of attention, where there is then a tendency see everything through a lens pertaining to one issue, rather than the bigger picture.

This can lead to a vicious cycle of selective accumulation of information (confirmation bias) about only one aspect of instability in the system, ignoring the reasons for its basic stability, or other risks connected with the system. Overestimating the magnitude or importance of certain risks can then cause a society to become complacent with a certain expert or set of experts, confounding future preparation efforts. For instance, science fiction films that focus on robotic takeovers tend to emphasize unrealistic scenarios, such as robots with anthropomorphic psychology, and cause desensitization of the public at large to the very real risk of Artificial Intelligence in the longer term. Another example: in Thailand in 2004, when the Indian Ocean tsunami hit, the Warning Service decided not to inform the public, assuming it was a less severe event than it actually was, for fear of scaring tourists. Unfortunately, this cost many lives.

**37. Subconscious desire for catastrophe**

Similar to Stockholm syndrome, this risk consists of the aspiration of a catastrophic risk expert for his forecasts to be proven correct. It may push him to exaggerate harbingers of a coming catastrophe, or to tolerate those events which may lead to catastrophe. People may also want catastrophes from boredom or due to the masochistic mechanism of “negative pleasure”.

**38. Use of risk warnings to attract attention or social status**

This type of behavior may be called “Scaramella syndrome” after the Italian security professional (born 1970) Mario Scaramella. Quoting Wikipedia's entry on him:

*While working for the Intelligence and Mitrokhin Dossier Investigative Commission at the Italian Parliament, Scaramella claimed a Ukrainian ex-KGB officer living in* [*Naples*](https://en.wikipedia.org/wiki/Naples)*, Alexander Talik, conspired with three other Ukrainians officers to assassinate Senator Guzzanti. The Ukrainians were arrested and special weapons including granades were confiscated, but Talik claimed that Scaramella had used intelligence to overestimate the story of the assassination attempt, which brought the calumny charge on him. Talik also claimed that rocket propelled grenades sent to him in Italy had in fact been sent by Scaramella himself as an undercover agent.*

Sometimes, an expert will make up a risk in his mind because he knows that society or the mass media will sharply react to it, and gain him attention. This is a problem because some of the worst risks may not be amenable to sensationalism or suited to media attention. Also, it may cause inappropriate desensitization to serious risks in a society, because of their association with publicity stunts.

**39. Use of the theme of global risks as a plot for entertaining movies**

There are many dumb movies about global risks. This causes us to associate them with whimsy, entertainment, or frivolity. This is a problem.

**40. Generalizing from fictional evidence**

In the ancestral environment, where our bodies and brains evolved and were formed, there was no such thing as movies. If you saw something happening, it was real. So our brains are not well suited to telling the difference between movies and reality. We think about movies subconsciously, or even consciously, as if they actually happened, though they are completely made up. Unfortunately, the scientific understanding level of a Hollywood screenwriter is usually somewhere between that of a 7th grader and an 8th grader. The tendency to recall movies and books as if they were actual events is called generalizing from fictional evidence27.

The most audacious example are movies about Artificial Intelligence, which postulate that AIs would have human-like, or anthropomorphic thinking, such as clannishness, anthropomorphic rebellion, or a desire for revenge. An AI, being a machine, would not have any of these animalistic tendencies unless they were explicitly programmed into it, which they likely would not be. Thus, the greatest danger to humanity is from AI that is indifferent or insufficiently benevolent to us, and emphatically not AI that has a specific grudge or malevolence against us28. This crucial point makes all the difference in the world in terms of how we will design an AI to optimize for safety.

Another issue with fiction, previously discussed, is that futuristic stories tend to make the future similar to today, but with just a few added details. For instance, in *Total Recall*, the technology was very similar to that of the year when the movie was made (1990), except there was interplanetary travel and slightly more advanced computer. In *Back to the Future* (1985), the main differences of the future appeared to be hoverboards and flying cars. In the real future, many details will simultaneously be different, not just a few.

Yet another problem unique to fiction is that forces that clash tend to be equally balanced. If *Star Wars* were real, the Empire would just use the Death Star to blow the entire Rebel fleet out of the sky. If *Terminator* were real, the assassin robot would just snipe the protagonist from a mile away, without ever being seen. If *The Matrix Reloaded* were real, the AI would just destroy the subterranean human city of Zion with nuclear weapons. In reality, extreme power asymmetries and “unfair” match-ups happen all the time. In 1518-1520, about 90-100 Spanish cavalry and 900-1,300 infantry were able to conquer and subjugate the Aztec civilization, an empire of millions of people.

**41. Privacy as a source of errors in management of risks**

Research conducted in private can't be examined by outside auditors or receive external feedback. As a result, it can contain more errors than more open sources. Contrariwise, open source data might be of poor quality because more unskilled or stupid people have the opportunity to edit or contribute to it29. When disasters or catastrophes are kept secret, as may have been the case with early Soviet space missions, we lose valuable feedback that might be used to prevent future disasters.

**42. Excessive intellectual criticism or skepticism**

Safety is often threatened by improbable coincidences or circumstances. Therefore, a “long tail” of strange ideas can be useful. Before narrowing down a narrow range of failure scenarios, it is helpful to brainstorm as many ideas as possible, including the most weird. The problem, however, is that it is much easier to criticize something than it is to come up with a solid risk, which may cause analysts to dismiss crazy-sounding ideas prematurely.

**43. The false belief that it is possible to prove safety conclusively**

The *Titanic* was “proven” to be safe, and called “unsinkable”. We all know how that turned out. There is no such thing as something perfectly proven to be safe. There is always some wild combination of circumstances that will cause something to break. A black hole may fly into the solar system and completely swallow the Earth, destroying everything on the planet, for instance. You can never be completely sure what will happen. The only way the safety of something can be proven to a high degree is observing many instances of it in action over a long period of time. For instance, aircraft are generally highly reliable, since the chance of dying in a plane crash is far less than the chance of dying in a car ride. The Space Shuttle was not very reliable, since two of the six orbiters exploded in flight, a 33 percent critical failure rate. Keep in mind that the Space Shuttle was highly over-engineered, and ran millions of line of code which had to be vigorously verified, but two of them still experienced catastrophic failures.

**44. Underestimate of the human factor**

Somewhere between 50 and 80 percent of catastrophes occur because of errors by operators, pilots or other people exercising direct administration of the system30. Other catastrophic human errors happen during maintenance service, preflight preparation or design errors. Even a super-reliable system can be put into a critical condition by the right sequence of commands. We should never underestimate the power of human stupidity and error. If someone can break something by accident, he will. This applies even to the best trained military officers.

**45. The false idea that it is possible to create a faultless system**

It is not possible to create a faultless system unless it is extremely simple. Any system with any kind of complexity will be put together through the efforts of thousands of people, people who will occasionally make mistakes. The wrong confluence of mistakes will produce a catastrophic outcome. Even if the design seems perfect, a disaster may be caused because the design is not followed to the letter, or followed to the letter but not in the right spirit.

**46. Statistics as a source of possible errors**

In the nature of statistics there is the possibility of errors, distortions, and false interpretations. This may derive from sampling, unsuitable framing, different methods of calculation or chosen formulae, rounding errors, interpretation of the received results, and cognitive biases connected to the visual interpretation of numbers and charts.

**47. Availability bias**

Certain glitzy facts are more accessible to our minds than others. For instance, everybody knows where the first nuclear bomb was used in warfare—Hiroshima, Japan. This attack killed about 120,000 people. But, do you know where the Spanish flu of 1918, which killed a *hundred million* people, originated? (According to most analyses, it was on March 8th, 1918, in Haskell county, Kansas.) There is a large literature on availability bias, which people looking into global risks would be well-advised to be at least superficially familiar with31.

**48. Analysis of global risks and making futurist forecasts are not identical**

A futurist forecast often contains concrete data about a time and place (though it would be more accurate if it were a probability distribution over times and places). Yet, it is extremely rare that such specific predictions are on-point. Moreover, futurist forecasts and the analysis of global risks demand different attitudes. In making futurist forecasts, prognosticators often make pie-in-the-sky predictions, throwing their hat into the ring to see if they get lucky or even famous for correct predictions. Analysis of global risk, however, requires more caution and care than this approach.

**49. Hindsight bias**

After something important happens, it feels like we knew it would all along. This is hindsight bias, another cognitive bias with an extensive literature32. The phrase “I knew it all along,” when the speaker actually didn't, exemplifies this way of thinking. Unless there is a written record or recording with the exact text of a prediction, it is difficult to verify how correct it is or how likely it was to be made by luck. Concerning global risks, by their very nature we cannot have hindsight knowledge of them, therefore we are stuck with having to predict things in advance the very first time. We may overconfidently predict global risks by analogizing them to simpler risks for which we have the benefit of hindsight.

**50. False positives**

The dollar weakening by several percent may be inappropriately taken as a sign that the dollar is imminently going to crash, and some number of pundits may point to it as an indicator of such, though ultimately it does not occur. This would be an example of false positive indicators of risk, which can undermine belief in the possibility of accurately forecasting and preparing for genuine disasters.

**51. An overly simplistic explanation is the most prominent**

Popular science writers often simplify complex scientific issues. The quality standard for articles in *Popular Science*, for instance, is guaranteed to be inferior to that of the best scientific journals. Understanding a domain of risk, for instance nanotechnological arms races, may require years of analysis. If a version of this risk does reach the public (such as the “grey goo” scenario) it may be a highly simplified and fantastic form of the true risk, and motivate improper preparations33. Another variant is that a minor mishap occurs and many years of analysis are needed to find its exact cause, delaying efforts to deal with future dangers from the same source. In certain conditions, such a backlog of knowledge acquisition could lead to a critical failure.

**52. Misuse of apocalyptic scenarios to draw attention and financing to projects**

This is essentially a variant on the Scaramella scenario described earlier, but applied to projects. Obviously, using apocalyptic scenarios as a foil may draw attention and financing to certain projects. Nearly every project which is genuinely, authentically, and truthfully trying to mitigate some kind of global catastrophic risk has fallen under this accusation. It is important to realize that there are people who are really motivated by lowering global risk, however, and there are probably easier ways of striking it rich than using an apocalypse scare. This applies doubly so for those of a scientific bent, who will be surrounded by secular intellectuals who are innately very skeptical of apocalypse claims because they pattern-match to religious apocalypse claims.

**53. Aspiration of people to establish a certain risk level acceptable to them**

Everybody has different intuitions of what constitutes an acceptable risk. Someone may choose a safer car so that they can drive it a bit more dangerously. Some people may consider walking in a bad neighborhood at night an acceptable risk, others may dare it once in their life. Consider a sport or activity which has a risk of 1 in 100,000 of causing death for each time that it is practiced, say, skydiving. For most adventure-seeking human beings, this is considered an acceptable risk. When it comes to global risk, however, a 1 in 100,000 chance of doom per event would be unacceptable. If the event were a daily occurrence, over the course of a hundred years, the probability of extinction would approach 30 percent. That would be an unacceptable level of risk. So, our natural intuitions about “acceptable risk” may be too carefree when it comes to mitigating global risks.

**54. “Overconfidence of the young professional”**

Talented young men and women who are professionals in a certain area, especially a dangerous area like racecar driving or base jumping, eventually get to a level of skill where they may acquire a false sense of invulnerability and become overconfident. Due to this very human factor of overconfidence, they run into catastrophes. This model can be used to consider mankind as a whole in relation to super-technologies such as nuclear, biological, and nanotechnologies. We may get so drunk on our own technological mastery that we fail to take basic safeguards and annihilate ourselves through carelessness.

**55. Sensation of invulnerability through survival**

The natural overconfidence of the young professional is aggravated from the observation selection effect which consists of the experience that, for example, at war after a certain duration without wounds, soldiers start to feel “invulnerability,” and will take increasingly risky maneuvers34. The same could occur with civilization—the longer period of time we go without nuclear war, the more complacent we will be about it, and the more brinksmanship we will be willing to undergo because it hasn't happened yet. This is distinguished from the prior point in that the salient factor is length of time of survival rather than skill level.

**56. Dunning-Kruger effect—overconfidence in one's professional skills**

According to the Dunning-Kruger effect, which has been extensively studied, those who are less skilled in a certain area are more likely to judge that they are competent35. Dunning and Kruger proposed the following, that incompetent people will:

* tend to overestimate their own level of skill;
* fail to recognize genuine skill in others;
* fail to recognize the extremity of their inadequacy;
* recognize and acknowledge their own previous lack of skill, *if* they are exposed to training for that skill.

Various studies have confirmed all these hypotheses. In considering global risks, which span many spheres of knowledge—from biology to astrophysics to psychology to public policy—trying to create an adequate picture of the situation, any expert will be compelled to venture outside his limits of knowledge. As it is pleasant to feel knowledgeable, people may test the boundaries of their propensity to exaggerate their own abilities. They may become overconfident and stop consulting experts about vital issues. The stereotype of a “savior of the world,” a single hero who is capable of anything without effort, may possess them. This effect may discourage other researchers from participating or even create a hole in our knowledge of global risk if that one researcher subsequently turns out to have been systematically wrong. One example which comes to mind is the Japanese composer Mamoru Samuragochi, who was known as the “Beethoven of Japan” due to his deafness and prodigal classical compositions. However, in early 2014 it was revealed that he had a ghostwriter for all his compositions and was in fact not deaf36. This caused quite a stir when many planned performances of his music were suddenly canceled.

**57. The error connected with concentrating on prevention of small catastrophes instead of prevention of the greatest possible catastrophe**

In Yellowstone Park, wildfires were prevented effectively for many years. This prevention was so effective that it allowed a buildup of dry woody material, which culminated in a trio of catastrophic blazes in the summer of 1988, which required 9,000 firefighters and $120 million ($240 million as of 2014) to contain37. Yudkowsky (2008) has a similar example related to flooding:

Burton, Kates, and White (1978) report that when dams and levees are built, they reduce the frequency of floods, and thus apparently create a false sense of security, leading to reduced precautions. While building dams decreases the frequency of floods, damage per flood is so much greater afterward that the average yearly damage increases.

Another example is the weakening of the average immune system today due to insufficient exposure to pathogens. Our sanitized lives are devoid of pathogens, which could make us highly vulnerable to genetically engineered viruses of the 21st century. Analogously, American Indians were highly susceptible to European pathogens, which were incubated in the filthy gutters and alleyways of overcrowded European cities. Some Atlantic coast tribes lost 90% of their adult members to disease shortly after the arrival of the Europeans.38

**58. Weariness of researchers**

The enthusiasm of people moves in waves. Someone who sends out a certain warning or bulletin may grow weary of sending out the same message after a few years. This may cause others to think that the risk has passed, although it persists. This is what has happened with regard to the risk of nanotechnological arms race since the Center for Responsible Nanotechnology (CRN) lapsed into inactivity around the year 2010. The researchers who work on global risk topics may never receive gratitude or see any immediate benefits of their work. Only in the movies does the “savior of the world” get the gratitude of mankind. Recall that Churchill lost reelection right after the war despite fervently believing that he deserved reelection. To avoid the effect of burnout, during WWI the American fleet had a constant rotation of their personnel, with one group at war, the other ashore. Furthermore, people rarely become heroes for successful preventive measures. There was a major wave of research into global catastrophic risks between the years 2001 and 2010, but activity seriously slowed down from 2011 onwards.

**59. Fear of loss of the social status by researchers**

Social status is a basic good that people in society seek. In our society there are a number of themes of interest which are perceived to be the symptom of a certain kind of inferiority. People who are interested in these areas are automatically perceived as second-grade, mad, clowns, and marginal (and could be squeezed out of social niches where they reside). Other researchers will avoid contact with such people and avoid reading their papers. When such people lose status, they also lose the ability to inform the thoughts of officials in power. The study of global catastrophic risk sometimes falls into a category of status-lowering activity, but not always. Sir Martin Rees, former Astronomer Royal and President of the Royal Society (2005-2010) has recently taken a serious interest in global risk, and due to his extremely eminent stature, has only suffered from a minor drop in status, if any. Yet, there is still not enough of a critical mass of researchers in the area of global catastrophic risks to give it robust credibility.

**60. The quantity of the attention which society can give to risks is limited**

Of course, society has limited attention to give to global risks. It may focus on a few risks at the expense of others, depriving lesser-known risk mitigation efforts of crucial resources they need to deal with threats. Also, there will be many people who are calmed by taking simple actions such as filling up their car with ethanol instead of gasoline. This action then satisfies them and makes them consider it unnecessary to pursue more comprehensive or serious risk mitigation. They might think “my contribution doesn't really matter, I'm just one person.” Of course, if everyone believes this, as most people do, that pretty much guarantees that nothing useful will be done. This is the bystander effect39.

**61. Neglect of economic risks**

Expressions such as “money is only pieces of paper”, or “bank accounts are only bits in computers” can be a reflection of the widespread opinion that the economy is not so important, as, say, war or natural disasters. However, the economy is the material embodiment of the structure of most human activity. To understand the role of the economy, it is important to note that the Great Depression of 1929 arguably caused more personal misery for the United States than World War II. It's also important to look at the crash of the USSR and the resulting economic troubles, as billions of dollars worth of state assets were seized by monopolists for pennies on the dollar. This crisis occurred because of structural-economic reasons, not any external threat. The same factor can occur in large extinctions; the large sauropods were in decline prior to the arrival of the fatal asteroid, as a result of complex ecological factors and changing patterns of competition between species40.

All disasters have an economic cost. Even small terrorist attacks can have economic effects hundreds of times larger than the initial damage itself. The September 11th terrorist attacks did at least $100 billion in damage to the American economy41, or closer to $6 trillion if you view the Afghanistan and Iraq wars as direct responses to that one terrorist action, which they certainly appear to have been42. A purely economic action such as a decrease in interest rates by the Fed might lead to a tremendous amount of economic damage by causing a bubble in the real estate market. In 2001, a mere seven letters laced with anthrax was capable of causing $320 million in cleanup costs alone43. On a dollar-per-pound basis, that is quite a sum.

Even small failures can lead to a huge damage and loss of stability of economy, and an economic crash would make the whole system less steady and more vulnerable to even larger catastrophes. It could lead to positive feedback that is a self-amplifying catastrophic process. During the process of economic globalization, a possibility of global systemic crisis continues to increase. It is difficult for some to believe that many of the world's most powerful nations would collapse because some large banks go bankrupt, but it is a definite possibility44.

**62. The errors connected with overestimating, underestimating, or   
failing to appreciate the moral condition of a society and its elite**

One popular account of the decline of the Roman empire is moral decadence and cultural decay45. This may have been a degradation of its elite, insofar as governors and leaders of all levels operated exclusively in pursuit of their personal short-term interests, that is, foolishly and selfishly. Generally speaking, people who pursue the long-term interests of a society unselfishly do more to help it flourish. A term from economics to describe this is *time preference*: a *high time preference* means a desire for immediate consumption a *low time preference* refers to saving and planning for the future46. Civilization is fundamentally based on low time preference47. Other metaphor is the comparison of “moral spirit,” for example, in armies—with the ability to molecules in some substance to turn into a uniform crystal (this theme is explored in detail by Lev Tolstoy in War and Peace). If the crystal is broken, the society collapses into feudalism and localism.

Today, both bad and good can be done by coordinated groups operating over long time scales, or subtle processes of decay or growth over the same. Two large societies may conflict and mutually influence one another, such as the Occident and the Orient, or NATO vs. Russia. Ideals such as the tension between authoritarianism and disorganization, rules and liberty, democracy and tradition, may expose large rifts that cause mass terrorism or even civil war. Even a moral paragon might unleash a powerful weapon or process by mistake, while an immoral man may be impotent by virtue of the fact that he is always drunk, or occupied with petty theft, and never becomes a genuine risk to the survival of humanity, never getting hold of powerful technologies.

**63. Popularity bias**

This is similar to availability bias, but the concept has no literature, and is original to this work. The easier an idea is to popularize or transform into bright propaganda, the more attention it will receive. It is easier to advertise a threat from global warming than from an Artificial Intelligence because the latter is difficult to portray and less visceral. People have to be involved in the process of spreading an idea among the masses, and that leads to an identification with the idea, and the aspiration to make it easier and more accessible. This also means that complex risks will tend to get watered down or excessively simplified as they become better known, which can cause sophisticated people to dismiss them unnecessarily, since the nuances of the concept are not spread to a mass audience. So, the drive for popularity and its results has many complex impacts on the spread of an idea.

**64. Propensity of people to offer "simple" and "obvious" decisions   
in difficult situations—not having thought them though**

We all know this happens. It is followed up by a persistence, defending the decision through argument, and resistance to considering other options. H.L. Mencken said, “For every complex problem there is an answer that is clear, simple, and wrong.” Yudkowsky writes in detail about the importance of a time interval between the moment of appearance of a question and the moment in which a human being makes a definitive choice in favor of an answer is the interval in which any real thinking happens, and it may be quite short, even a few seconds. Norman R.F. Maier wrote, “Do not propose any solutions until the problem has been discussed as thoroughly as possible without suggesting any.” It is psychologically difficult for someone to change their mind once they have proposed a solution and begun to take a liking to it, partially because in every human society, spending too much time considering solutions is seen as weakness. Once someone is seen as advocating a solution publicly, it becomes a subject of dispute that they get emotionally attached to, which represents them, and they feel the need to defend it, either consciously or subconsciously.

**65. Error connected with incorrect correlation of force and safety**

Emotionally, there may be a tendency to associate strong technologies as good, and weaker technologies as bad. Maybe because on a daily basis the strongest technologies we are exposed to are generally good. However, it actually follows that the stronger a tool, the more capable it is of influencing the world, and (usually) the more destructive it can theoretically be. Seemingly safe technologies, such as air travel, can be turned to destructive ends by the use of bombers. The destructive variants of many of the common technologies we use are hidden on military bases or in other places. An analysis based on incomplete information of technology is inclined to interpret a technology emotionally or whimsically, softening perceived risk.

**66. Premature investments**

If a large quantity of funds and efforts are put towards a project prematurely, such as electric cars or Artificial Intelligence, and it does not bear fruit, it can put a field on ice for a matter of decades, even after it becomes economically feasible. If people were informed in 1900 that nuclear weapons would be developed in 1941, and would threaten the safety of the world by the 1950s, they would likely spend tens or hundreds of millions building bomb shelters, trying to develop quick flying machines, anti-aircraft batteries, or aerospace technology to ameliorate the anticipated strategic risk. In all likelihood, this would cause expenditure fatigue, so that by the time nuclear weapons were actually developed, there would be a reluctance to invest in dealing with them that there would not have otherwise been.

Humans, and humanity, do not have the greatest attention span or planning capability, even when the world hangs in the balance. We need quick results and gratification to move forward on projects. According to some recently released information, in the 80s, the USSR got wind of an unmanned aerial vehicle, or drone, built by the United States, and spent a great deal of money and military research trying to come up with their own version, to no avail48. As a result of that program, by the time drones actually became cheaper and reliable, in the 00s, Russian military leaders were already exhausted with the idea, and lagged behind in their development accordingly. Timing can make a crucial psychological and economic difference which makes the difference between success and failure for a given technology or safeguard.

**67. Planning fallacy and optimism bias**

Yudkowsky writes49:

*When asked for their “most probable” case, people tend to envision everything going exactly as planned, with no unexpected delays or unforeseen catastrophes: the same vision as their “best case.” Reality, it turns out, usually delivers results somewhat worse than the “worst case.”*

In large projects, “Cost overruns and benefit shortfalls of 50 percent are common; cost overruns above 100 percent are not uncommon”50. Same with respect to the time it takes to write books, complete papers, and so on. Our optimism tends to consider the best case the most probable, possibly because it is the easiest to imagine. But, as Murphy law goes, if something can go wrong, it will, and when many things in a row go wrong, the worst-case scenario turns out to in fact be worse than anyone imagined it could be.

**68. Bystander effect**

Previously mentioned, the bystander effect refers to the fact that people are less likely to do anything if they think others will do it. A man lying on the ground in a sorry state is less likely to be helped by a crowd than if someone came upon him while walking through the woods. We have a tendency to avoid personal responsibility for events if possible, and if we are not specifically called out, will avoid contributing. This condition arises subconsciously, as simply as a reflex. Global risks conjure up the ultimate bystander effect, as they effect the whole planet, but so few do anything about them. Nick Bostrom points out that there are more academic research papers published on *Star Trek* or the reproductive habits of dung beetles than there are on global catastrophic risks.

**69. Need for closure**

People have a need for closure. As soon as there is a disturbing open question, we have the desire to immediately find a solution for it and put it behind us. We prefer a fast decision whose correctness is uncertain to a long and grueling search for a complete solution which may appear to be endless. Although we do not have infinite time to come up with answers, it is advisable that we think well before coming to any conclusions, probably a bit longer and more thoroughly than we would prefer to.

**70. Influence of authority and the social pressure of a group**

The famous Milgram experiment showed what evil everyday people can do when they are ordered to. In the experiment, “examinees” were hooked up to an electric current in a sealed room (actually, conspirators connected to no real current whatsoever) while on the other side of glass, the “participants” had access to a dial which allowed them to control the flow to electricity to the victim. 66% of the participants increased the voltage all the way to 400 volts—a mortal dose—when they were ordered to by a researcher, even when the “victim” begged them to stop. In this experiment, factors such as authority, the remoteness of the victim and the influence of similar behavior by being in the same room with other people in the same role all combined together to cause people to take actions which were ostensibly horrible. The same factors apply to us when we estimate the risk connected with some future factor or technology. The potential victims, even if our future selves are among them, are far away from us in time and space that if a strong authority expresses favor in the dangerous technology, and we are surrounded and influenced by a group of people apt to do the same thing, all of these factors will strongly influence our choice.

**71. Conflict between general research and applied research**

Ray Kurzweil points to a phenomenon he calls “engineer's pessimism,” that engineers working on a difficult problem will overestimate its difficulty because they are so closely immersed in details51. Similarly, in nanotechnology there is a split between theoretical researchers focused on long-term goals, like Rob Freitas and Eric Drexler, and more applied researchers, like Nadrian Seeman and the late Richard Smalley. This led to a bitter rivalry between Drexler and Smalley until Smalley's death in 2005. Generalists sometimes accuse engineers of focusing overmuch on what has already been done rather than the space of what's possible, or not taking the long-term view, or considering developing new basic capabilities, while engineers accuse the generalists of being overoptimistic or pie-in-the-sky. There is a grain of truth to both charges.

**72. Mind projection fallacy**

The mind projection fallacy is when we unconsciously attribute to subjects properties which only exist in our representations of them. The concept originates with E.T. Jaynes, a physicist and Bayesian philosopher with a highly sophisticated and nuanced understanding of probability theory. He used the phrase to argue against the Copenhagen interpretation of quantum mechanics52:

*[I]n studying probability theory, it was vaguely troubling to see reference to "gaussian random variables", or "stochastic processes", or "stationary time series", or "disorder", as if the property of being gaussian, random, stochastic, stationary, or disorderly is a real property, like the property of possessing mass or length, existing in Nature. Indeed, some seek to develop statistical tests to determine the presence of these properties in their data...*

*Once one has grasped the idea, one sees the Mind Projection Fallacy everywhere; what we have been taught as deep wisdom, is stripped of its pretensions and seen to be instead a foolish non sequitur. The error occurs in two complementary forms, which we might indicate thus: (A) (My own imagination) → (Real property of Nature), [or] (B) (My own ignorance) → (Nature is indeterminate)*

Yudkowsky (2008) uses the term to refer to the way people are prone to think about Artificial Intelligence. They take a disposition, say, “nice,” which may be their own or what they hope AI will be, and they project that on to every possible consideration of advanced, agent-like Artificial Intelligence that they can come up with. They are engaging in projecting rather than considering the full expanse of possibilities. This is bound to occur when people are considering complex new technologies, particularly Artificial Intelligence. Another pernicious aspect of the mind projection fallacy concerns the property of ignorance—taking our own ignorance and projecting it onto an external object, as if the object were itself inherently mysterious. That is impossible, however. The mystery is a property of our mind, not the object itself.

**73. Confusion between objective and subjective threat**

People who end up taking actions that are risky to the survival of the human species may be fair, noble, beautiful people who are not personally malicious to us. They might just not know what they are getting into, for instance developing advanced Artificial Intelligence without adequate safeguards. This highlights the difference between objective and subjective threat. If someone is playing a zero-sum game with you, say competing in a business field, he may be your “enemy” or a threat to you objectively, but bear no personal malice towards you or your friends. Conversely, someone with a personal grudge who is clearly out to get you is your subjective enemy. With regard to global risks, it is important to remember that those taking actions dangerous to humanity may bear no personal malice in any way, may be competing casually, or attacking a third party nowhere near you, yet their actions could threaten you in the long run regardless. This becomes even harder to grasp when it comes to the economy. A Federal Reserve chairman who prints too much money could cause the economy to collapse through runaway inflation and currency devaluation, even if they are ostensibly doing it to improve the economy and many leading economists support them.

**74. Predictions or dreams of catastrophe, caused by envy**

A vivid example of this is a phenomena which occurred on Internet forums in the early 90s in Russia. People who were offended by the disintegration of the USSR began to dream of a similar crash of the United States, pouring over data and news stories to discover signs of this process. When the dreamed-for schadenfreude fails to pan out, this can influence interpretations of data.

**75. Fear of loss of identity**

Systems can be resistant to change, because change can compromise core and essential identity. This is one of the reasons for the struggles against globalization and immigration. Someone can prefer death to identity loss. That is, he might prefer global catastrophe to a transformation of the world in which he lives.

**76. Clear catastrophe can be more attractive than an uncertain future**

As bizarre as it may seem, global catastrophe is easier to imagine than an uncertain future, and may be more intuitively acceptable for that reason. Uncertainty can cause fear and intimidation, whereas certain doom offers a sort of closure.

**77. Incorrect application of Occam's razor**

Occam's razor is the scientific heuristic that “entities must not be multiplied beyond necessity”. However, this is just a guideline, and is subjective. In the hands of a clumsy operator, Occam's razor can simply be used to exclude ideas that are too complicated to understand, but nonetheless valid failure modes. As mentioned earlier, catastrophes tend to involve the confluence of improbable scenarios, which breach a hole in safety mechanisms. So, overlapping and combinatorial failure sequences ought to be considered. Occam's razor may be suitable when it comes to deriving naturalistic explanations for complex natural phenomena, but isn't as useful when it comes to considering complex failure modes or scenarios or global risk.

**78. The upper limit of possible catastrophe is   
formed on the basis of past experience**

Previously we mentioned the anecdote about how a river was dammed and the frequency of floods decreased, but their magnitude increased, leading to greater overall construction. One factor was a false sense of security created by the dam, which caused building closer to the river. This tends to be a general feature of dams and embankments in that they create a false feeling of safety. The river and dam example is also useful for considering the fact that our imagination of the upper limit of a possible catastrophe is formed on the basis of our past experience. We do not account for once-in-a-hundred-year events, because we haven't lived through them. Few large structures on the Hayward fault in the San Francisco Bay Area are built to cope with a once-in-a-hundred-year earthquake, though one will eventually occur sooner or later. We need to prepare for disasters which are significantly larger than anything we have seen.

**79. Underestimating the fragility of complex systems**

A person can be quickly killed by a small incision if it punctures a vital organ. A tree can be killed by removing a ring of bark, which prevents fluids from the roots from reaching the leaves, a technique called girdling. Every complex system has a weak point, sometimes several weak points. Our power grid is exactly the same way; an overload of current in one area can fry transformers in a long series, potentially shutting down power in large areas53. In a disaster situation this could lead to widespread looting, as during the New York City Blackout of 1977. Many do not appreciate how many weak points our complex society has.

There is an empirical generalization that technological systems decrease in proportion to the fourth degree of energy density. This empirical generalization (exact value varies depending on different factors) can be derived by comparing the reliability of planes and rockets54. A similar empirical generalization holds for statistics of deadly car crashes in relation to speed55. It is necessary to observe that the installed power per employee of mankind is constantly growing56.

**80. Ambiguity and a polysemy of any statement as a source of a possible error**

From the point of view of the authors of the regulation instructions for the Chernobyl nuclear reactor, personnel had broken their requirements, whereas from the point of view of the personnel, they operated precisely according to its requirements57. Regulations required for operators to “muffle the reactor,” from the point of view of the authors this was to be done immediately, but from the point of view of the operators is was to be gradual. Another example is when there is an automatic rescue system in a plane as well as a manual rescue system, and if they are both executed simultaneously, they run into each other and lead to catastrophe (nose-dive and crash of Aeroflot Flight 953 in 1994 in Siberia). It is difficult to reach an unequivocal understanding of terms in cases where we do not have experimental experience, as in global catastrophes.

**81. Refusal to consider a certain scenario because of its "incredibility"**

As mentioned before, the majority of catastrophes happen as a result of improbable coincidence of circumstances. The destruction of the HMS Titanic was connected with an incredible combination of no less than 24 unfortunate and totally avoidable circumstances58.

**82. Transition from deliberate deceit to self-deception**

Conscious deceit for the purpose of gaining a certain benefit—in our context, the concealment of risks—can imperceptibly take the form of self-hypnosis. This sort of self-deception can be much steadier than illusion or inadvertent error. Another version of self-hyponosis is simple procrastination—a command to think to yourself, “I will think of this tomorrow”, but tomorrow never comes.

**83. Overestimate of own possibilities in general and survival rate in particular**

Quoting Nick Bostrom's article on existential risks59:

*The empirical data on risk-estimation biases is ambiguous. It has been argued that we suffer from various systematic biases when estimating our own prospects or risks in general. Some data suggest that humans tend to overestimate their own personal abilities and prospects. About three quarters of all motorists think they are safer drivers than the typical driver. Bias seems to be present even among highly educated people. According to one survey, almost half of all sociologists believed that they would become one of the top ten in their field, and 94% of sociologists thought they were better at their jobs than their average colleagues. It has also been shown that depressives have a more accurate self-perception than normals except regarding the hopelessness of their situation. Most people seem to think that they themselves are less likely to fall victims to common risks than other people. It is widely believed that the public tends to overestimate the probability of highly publicized risks (such as plane crashes, murders, food poisonings etc.), and a recent study shows the public overestimating a large range of commonplace health risks to themselves. Another recent study, however, suggests that available data are consistent with the assumption that the public rationally estimates risk (although with a slight truncation bias due to cognitive costs of keeping in mind exact information).*

**84. Aspiration to the wonderful future, masking perception of risks**

This phenomenon can be seen in revolutionaries. The experience of the French Revolution showed us that revolution leads to civil war, dictatorship, and external wars, however the Russian Revolution at the beginning of the 20th century took dangerous actions based on the same idealism. If someone is fanatical about achieving a certain goal, they will ignore risks on the way to that goal, no matter how great they may be. In this sense, many modern transhumanists and technologists are exactly the same, in that they see a glorious future, and will do anything to reach for it, ignoring risks in the process. They do not realistically anticipate new weapons, and new applications of those weapons. They just barge on ahead.

**85. Filters between information reception and management**

Any complex, breakable system will have safeguards and detection apparatus to listen to signs of trouble. Value of information is defined by its novelty and the ability of the whole system to react to it. There are several filters in the way between reception of possible danger signs and management with the authority to act on it.

The first filter is what the monitoring system is designed to detect. This will be based on past events, and may not pick up sudden or unanticipated changes. The second filter is psychological, which consists of an aversion by technicians and management to information owing to its strategic novelty or ambiguity. The third filter has to do with limitations inherent in any hierarchical system; an officer that receives certain information may lack sufficient power to officially recognize the urgency of the situation or compel a superior to do anything about it. The fourth filter is filter has to do with a connection between the warning signal and what superiors are psychologically capable of recognizing as danger; it must bear sufficient similarity to past signals or training on danger signals to be recognized as worth taking action on.

**86. Curiosity can be stronger than fear of death**

Any information on global risks is useful. For example, if we run a certain dangerous experiment and survive, we learn that that kind of experiment is safe. This game has a limit though, where we eventually run an experiment that blows up in our faces. People risk their lives for the sake of knowledge or experiences. Alfred Nobel's (inventor of dynamite) experiments accidentally caused the death of five workshop assistants, including his brother Emil. Various scientists studying diseases have deliberately infected themselves with the pathogen of study to observe its effects firsthand. In 1993 in Russia, Boris Yeltsin put down a coup attempt where his political enemies had occupied the Parliament building, and some innocents who had crowded around the building out of curiosity were shot. We can be certain that there will be some people, with the power to destroy the world in their hands, who will be extremely curious about what would happen if they unleashed this power. People will agree to run dangerous experiments for the sake of curiosity.

**87. Systematic regulatory failure**

A global catastrophe, or any technogenic failure, may not be a result of any one fatal error, but the culmination of ten insignificant errors. Due to limitations in space and knowledge, for it is necessary to avoid too many regulations or instructions on trifles. Yet, small things—a smoldering cigarette butt, an open tank porthole, an improper start-up procedure, a restart returning a system to default settings—all these things can set the stage for a disaster. In disasters, there is rarely one “main thing” that causes critical failure, but a series of things. Running a system “by the book,” may not help, because the book may not have an entry for the mistake you have made.

**88. Scapegoating**

It is nearly always easier to make someone take the fall and move on after a disaster than go through the trouble of forming a commission of inquiry and figuring out what actually happened. Even after a commission of inquiry has been formed and come to conclusions, open questions may remain. Depending on how corruption or busy the administrators of the system or country in question are, they may avoid seeking out the real sources of the problem, setting the stage for it to happen again—only worse next time.

**89. Minimum perceived risk**

Aside from a ceiling on maximum risk set by past experience, people also have a floor of minimum perceived risk based on the minimum probabilities that a human being can intuitively care about. According to cognitive psychology experiments, this probability is about 0.01%, or one in 10,00060. If an experiment or procedure has a one in 10,000 chance of going awry, even if the consequences would be catastrophic, there is a strong tendency to ignore it. This can be a problem if the risk is a daily event, imperceptible on a day-by-day basis, but over a period of years or decades, a disaster is all but assured.

**90. Influence of emotional reactions to catastrophe**

It is known the emergencies or catastrophes provoke a certain sequence of psychological reactions, each of which compromises the objectivity of decision-making and action-taking. The book *Psychogenesis in Extreme Conditions* says, “psychological reactions to catastrophe are subdivided into four phases: heroism, a honeymoon phase, disappointment, and recovery”61. In addition, there is often a phase of panic or paralysis during the first moments of a catastrophe, which can precede the heroism phase. Each of these stages creates its own kind of bias

**91. Problems with selection of experts**

On each separate question related to global risk—biotechnology, nanotechnology, AI, nuclear war, and so on—we are compelled to rely on the opinions of the most competent experts in those areas, so it is necessary for us to have effective methods of selecting which experts are most trustworthy. The first criteria is usually the quality and quantity of their publications—citation index, publication ranking, recommendations from other scientists, web traffic from reputable sources, and so on.

Secondly, we can evaluate experts by their track record of predicting the future. An expert on technology who does not make future predictions, even if qualified predictions made only a year or so in advance is probably not a real expert. If their predictions fail, they may have a poor understanding of the subject area. For instance, nanotechnologists who predicted in the 1990s that a molecular assembler would be built “around 2016” have been proven to be mistaken, and have to own up to that before they can be taken seriously.

A third strategy is to simply not trust any expert, and to always recheck everyone's calculations, either from first principles or based on comparisons to other expert claims. Lastly, it is possible to select people based their views pertaining to theories relevant to predicting the future of technology—whether they have an interest or belief in the technological Singularity, or Hubbert's peak oil theory, a neoliberal model of the economy, or whatever. It is possible to say “an expert should not have any concrete beliefs,” but this is false. Anyone who has been thinking about the future of technology must eventually make ideological commitments to certain patterns or systems, even if they are just their own, or it shows that they have not thought about the future in detail. An “expert” who never goes out on a limb in prediction is indistinguishable from a non-expert or random guesser.

**92. Fault and responsibility as factors in the prevention of risks**

It is possible to over-focus on the attribution of guilt in a catastrophe, thereby under-focusing on the systemic and accidental factors that led to the disaster or potential disaster at hand. Furthermore, when it comes to global catastrophes, there may be no time to punish the guilty, so their guilt is not as relevant as it would otherwise be. Still, this doesn't mean that it isn't worthwhile to select competent, loyal, and predictable personnel to manage dangerous systems, just that a military-style focus on guilt and responsibility may not be as relevant to global risks as we are prone to thinking.

**93. Underestimating inertia as a factor of stability**

Besides general concerns about complexity, competence, safety-checking feedback mechanisms, and hundreds of other factors that keep a system safe and steady, it is possible to use Gott's formula (see chapter on indirect estimation of risks) to estimate the future expected lifespan of a system, based on how long it has already existed. For instance, there has not been a major meteor impact in at least 11,000 years, so it is not likely that one will occur tomorrow. Or, the pyramids at Giza have been in place for more than 4,500 years, meaning it is not likely they will topple or erode away in the next 100 years. Accordingly, this lets us take into account the inherent inertia of systems as a factor of judging their stability. Even if a system seems to have certain fragility, if it has been around for a long time, it could very well be more stable than we think. Correspondingly, if a system is extremely new and untested, although we consider it foolproof, we cannot confidently assume that it will not collapse at some point, since it has not built up a track record of stability.

**94. Bias caused by differences in outlook**

This error consists in us underestimating or overestimating the validity of statements based on the outlook of the person which produces them. For instance, nearly all discussions of global catastrophic risk in an academic setting have been undertaken by people with a certain common scientific, cultural, and historical outlook which is so obvious to us that it seems transparent and imperceptible. However, it is possible that a representative of another culture and religion will make a relevant point, which, due to biases in favor of our own outlook, we are oblivious to.

**95. Search terms**

A scientific concept may have different systems of terminology associated with it, such that an internet or journal search for a given term misses important results associated with other terms that the searcher did not know about. Furthermore, there may be no links or mutual citations whatsoever between these fields. They may be in direct competition and therefore as silent as possible about one another. Thus, care should be taken to ensure that “comprehensive” surveys of an academic field on global risk are truly comprehensive.

**96. Errors connected with the conscious and unconscious unwillingness of people to recognize the fault and scale of catastrophe**

When a catastrophe is in the middle of happening, there may be denial about its scale, leading to further problems. At Chernobyl, the organizer of reactor testing, Anatoly Diatlov, believed that it was not a reactor, but a cooling tank which exploded, and he continued to submit commands to a nonexistent reactor. This kind of overoptimism can also operate forward in time, compelling us not to make adequate preparations for a probable catastrophe, or denying its scale.

**97. Egocentrism**

Mundane egocentrism may cause people to overestimate the influence they can have over a certain situation, indirectly making them powerless to influence it. Or, egocentric cowardice might cause someone to attempt to escape in the middle of a disaster, saving his own skin, instead of taking necessary actions which may save millions of people. In general, egocentrism is one of the most prevalent human biases, and will affect every decision and action we take. One way of avoiding such effects, which may not be malicious in any way, is to take what economist Robin Hanson calls “the outside view,” that is, the view of a detached and objective observer. Role-playing such an observer can help people or groups overcome the “inside view,” which can have biasing effects.

**98. Excessive focus on needing a villain**

In the ancestral environment of mankind, the external threat we were exposed to most frequently would be rival human tribes with malignant intent. We lacked knowledge of how to combat disease aside from basic purity and domestic cleanliness intuitions. Our instincts switch into high gear when there is a clear enemy threatening us. Less so when the risk is subtle or complex, as may often be the case with global risks, where there may not even be a nameable enemy. Even if there is an enemy, the risk may derive from a particular technology used by many people simultaneously, all over the globe making it hard to point the finger, or some systemic effect that does not involve deliberate human intent.

**99. Dependence of reaction on speed of change of size**

This one is the inverse of #2, “excessive attention to slow processes and underestimation of fast processes”. A frog may be boiled to death in a pot, since if the heat is only turned up a little bit at a time, it does not notice. The same effect must have happened with the deforestation of Easter Island, which involved a slow reduction of trees, over a long enough number of generations, that there was never a point of alarm until it was too late.

**References**

* Marc Alpert, Howard Raiffa. “A progress report on the training of probability assessors”. In Daniel Kahneman, Paul Slovic, Amos Tversky. *Judgment under uncertainty: Heuristics and biases*. 1982. Cambridge University Press. pp. 294–305.
* Ulrich Hoffrage. “Overconfidence”. In Rüdiger Pohl. *Cognitive Illusions: a handbook on fallacies and biases in thinking, judgement and memory*. 2004. Psychology Press.
* Dominic D. P. Johnson & James H. Fowler. “The Evolution of Overconfidence”. *Nature* **477**, 317—320. 15 September 2011.
* Steve Rayhawk et. al. “Changing the Frame of AI Futurism: From Storytelling to Heavy-Tailed, High-Dimensional Probability Distributions”. 2009. Paper presented at the 7th European Conference on Computing and Philosophy (ECAP), Bellaterra, Spain, July 2–4.
* Hamish A. Deery. “Hazard and Risk Perception among Young Novice Drivers”. *Journal of Safety Research*, volume 30, issue 4, Winter 1999, pp. 225–236.
* Steven M. Albert, John Duffy. “Differences in risk aversion between young and older adults”. *Neuroscience and Neuroeconomics*, vol 2012:1, pp. 3-9. February 2012.
* Mara Mather et al. “Risk preferences and aging: The “Certainty Effect” in older adults’ decision making”. *Psychology and Aging*, Vol 27(4), Dec 2012, 801-816.
* Hank Pellessier. “Transhumanists: Who Are They? What Do They Want, Believe, and Predict? (Terasem Survey, Part 5).” Institute for Ethics and Emerging Technologies. Sep. 9, 2012. <http://ieet.org/index.php/IEET/more/pellissier20120909>
* Myers, D.G.; H. Lamm (1975). "The polarizing effect of group discussion". *American Scientist* 63 (3): 297–303.
* Isenberg, D.J. “Group Polarization: A Critical Review and Meta-Analysis”. J*ournal of Personality and Social Psychology* 50 (6): 1141–1151. 1986.
* Robin Hanson. “For Bayesian Wannabes, Are Disagreements Not About Information?” 2002. <http://hanson.gmu.edu/disagree.pdf>
* Eliezer Yudkowsky. “Cognitive Biases Potentially Aﬀecting Judgment of Global Risks.” In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan M. Ćirković, 91–119. 2008. New York: Oxford University Press.
* Daniel Kahneman and Amos Tversky. *Choices, Values, and Frames*. 2000. Cambridge University Press.
* Eliezer Yudkowsky. *Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures*. 2001. The Singularity Institute, San Francisco, CA, June 15
* Amos Tversky and Daniel Kahneman.. "Extension versus intuitive reasoning: The conjunction fallacy in probability judgment". *Psychological Review* 90 (4): 293–315. October 1983.
* Frederick, Shane; Loewenstein, George; O'Donoghue, Ted (2002). “Time Discounting and Time Preference: A Critical Review”. *Journal of Economic Literature* 40 (2): 351–401.
* Martin Rees. *Our Final Hour: A Scientist's Warning: How Terror, Error, and Environmental Disaster Threaten Humankind's Future In This Century—On Earth and Beyond*. 2003. Basic Books.
* Stephen Flower. *A Hell of a Bomb: How the Bombs of Barnes Wallis Helped Win the Second World War*. 2002. Tempus Publishers Limited.
* Napoleon's chill.
* Noah Arroyo. “Potentially Earthquake-Unsafe Residential Buildings — a (Very Rough) List”. *Public Press*. Jan 14, 2013. <http://sfpublicpress.org/news/2013-1/potentially-earthquake-unsafe-residential-buildings-a-very-rough-list>
* Paul Slovic, Baruch Fischhoff, Sarah Lichtenstein. “Facts Versus Fears: Understanding Perceived Risk”. In *Judgment under uncertainty: Heuristics and biases*. 1982. Cambridge University Press. pp. 463-492.
* Yudkowsky 2008.
* James Parkin. *Management Decisions for Engineers*. 1996. Thomas Telford Ltd.
* William M. Grove and Paul E. Meehl. “Comparative Efficiency of Informal (Subjective, Impressionistic) and Formal (Mechanical, Algorithmic) Prediction Procedures: The Clinical–Statistical Controversy.” *Psychology, Public Policy, and Law*, 1996, 2, 293–323.
* Helene Cooper. “Air Force Fires 9 Officers in Scandal Over Cheating on Proficiency Tests”. *The New York Times*. March 27, 2014.
* Jan Lorenz et al. “How social inﬂuence can undermine the wisdom of crowd effect”. *Proceedings of the National Academy of Sciences of the United States of America*. May 31, 2011 vol. 108 no. 22.
* Eliezer Yudkowsky. “"The Logical Fallacy of Generalization from Fictional Evidence”. *Less Wrong*. October 16, 2007.
* Stephen Omohundro. “The Basic AI Drives”. *Proceedings of the First AGI Conference*, Volume 171, Frontiers in Artificial Intelligence and Applications, edited by P. Wang, B. Goertzel, and S. Franklin, February 2008, IOS Press.
* Andrew Keen. *The Cult of the Amateur: How Today's Internet Is Killing Our Culture*. 2007. Currency.
* “between 50 and 80 percent of catastrophes occur because of errors by operators, pilots or other people exercising direct administration of the system”
* Amos Tversky, Daniel Kahneman. “Availability: A heuristic for judging frequency and probability”. *Cognitive Psychology* 5 (1): 207–233, 1973.
* Roese, N. J., & Vohs, K. D. “Hindsight bias”. *Perspectives on Psychological Science*, 7, 411-426, 2012.
* Chris Phoenix and Eric Drexler, “Safe Exponential Manufacturing”. *Nanotechnology* 15 (August 2004) 869-872.
* soldiers start to feel “invulnerability,” and will take increasingly risky maneuvers
* Justin Kruger, David Dunning. “Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments”. 1999. *Journal of Personality and Social Psychology* 77 (6): 1121–34.
* David Ng. “Mamoru Samuragochi, disgraced Japanese composer, apologizes for lies”. *Los Angeles Times*. March 7, 2014.
* National Park Service. “The Yellowstone Fires of 1988”. 2008.

<http://www.nps.gov/yell/naturescience/upload/firesupplement.pdf>

* Some Atlantic coast tribes lost 90% of their adult members to disease shortly after the arrival of the Europeans.
* Darley, J. M. & Latané, B. (1968). "Bystander intervention in emergencies: Diffusion of responsibility". *Journal of Personality and Social Psycholog*y 8: 377–383.
* Paul M. Barrett and Paul Upchurch. “Sauropodomorph Diversity through Time”. In *The Sauropods: Evolution and Paleobiology*, edited by Kristina Curry Rogers and Jeffrey A. Wilson, pp. 125-150 (double check). 2005. University of California Press.
* Gail Makinen. "The Economic Effects of 9/11: A Retrospective Assessment". Congressional Research Service. pp. CRS–4. September 27, 2002.
* Sabir Shah. “US Wars in Afghanistan, Iraq to Cost $6 trillion”. Global Research. September 20, 2013. (is there a better source for this?)
* Ketra Schmitt and Nicholas A. Zacchia. “Total Decontamination Cost of the Anthrax Letter Attacks”. *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science* Volume 10, Number 1, 2012.
* It is difficult for some to believe that many of the world's most powerful nations would collapse because some large banks go bankrupt, but it is a definite possibility.
* One popular account of the decline of the Roman empire is moral decadence and cultural decay.
* A term from economics to describe this is *time preference*: a *high time preference* means a desire for immediate consumption a *low time preference* refers to saving and planning for the future.
* Civilization is fundamentally based on low time preference.
* According to some recently released information, in the 80s, the USSR got wind of an unmanned aerial vehicle, or drone, built by the United States, and spent a great deal of money and military research trying to come up with their own version, to no avail.
* Yudkowsky 2008.
* Bent Flyvbjerg, Massimo Garbuio, and Dan Lovallo. “Delusion and Deception in Large Infrastructure Projects: Two Models for Explaining and Preventing Executive Disaster”. *California Management Review*, vol. 51, no. 2, Winter 2009, pp. 170-193.
* Ray Kurzweil. *The Singularity is Near: When Humans Transcend Biology*. 2005. Viking.
* E.T. Jaynes. *Probability Theory: the Logic of Science*. 2003. Cambridge University Press.
* Congressional EMP Commission. *Report of the Commission to Assess the Threat to the United States from Electromagnetic Pulse (EMP) Attack*. April 2008.
* This empirical generalization (exact value varies depending on different factors) can be derived by comparing the reliability of planes and rockets.
* A similar empirical generalization holds for statistics of deadly car crashes in relation to speed.
* It is necessary to observe that the installed power per employee of mankind is constantly growing.
* From the point of view of the authors of the regulation instructions for the Chernobyl nuclear reactor, personnel had broken their requirements, whereas from the point of view of the personnel, they operated precisely according to its requirement.
* The destruction of the HMS Titanic was connected with an incredible combination of no less than 24 unfortunate and totally avoidable circumstances.
* Nick Bostrom. “Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards”. *Journal of Evolution and Technology*, Vol. 9, No. 1, 2002.
* people also have a floor of minimum perceived risk based on the minimum probabilities that a human being can intuitively care about. According to cognitive psychology experiments, this probability is about 0.01%, or one in 10,000
* The book *Psychogenesis in Extreme Conditions* says, “psychological reactions to catastrophe are subdivided into four phases: heroism, a honeymoon phase, disappointment, and recovery”

## Chapter 4. The universal logical errors, able to be shown in reasoning on global risks

1. Mess between probability, as a measure of variability of object, and confidence degree, as a measure of the information on object

The first concerns likelihood process, for example, to radioactive disintegration, and the second to unknown process - for example, to card guessing. However global risks concern the phenomena, estimating which we are compelled to state likelihood judgments about processes, which simultaneously both likelihood, and unknown to humans. Here we start to speak about degree of confidence of this or that probability. In this case the probability and confidence degree are multiplied.

2. Substitution of the analysis of possibilities by the analysis of the purposes

For example, reasoning like «terrorists never will want to use the bacteriological weapon because it will put strike and on what they are interesting». The structure of the purposes can be is very complex or to comprise errors simply.

3. The incorrect use of inductive logic of a following kind: if something did not occur for a long time, it will not occur the same long period.

This statement works only in the event that we unitary observed something during the casual moment of time and probability is in that case described by formula of Gott. It gives chance of end for any event in 50 percent in an interval from 1/3T to 3T, where T - age of object at the moment of its casual detection. However if we very long observe a certain process it evidently, comes nearer to the end. For example, if we take the casual human it, most likely, will be middle age. However if to take the casual human and then to observe very long him, we will unequivocally receive the very old man who can die at any moment. (See my article «Natural catastrophes and Anthropic principle» for more details.)

4. The thinking caused by desire something to prove

Depending on that human wishes to prove, it will select those or other arguments, is frequent unconsciously. Other name for this model - "rationalization", that is selection of pseudo-rational arguments under certain initially irrational statement.

5. The logic error arising at attempts to prove what it is necessary to do, proceeding only from the description of the facts

If in the first and second premise of inference there are only facts then in the conclusion there can be only facts. Any reasoning on the purposes should lean against certain representations about the values, set is axiomatic. However it means randomness of such purposes, and their understanding can differ at different researchers of global risks that can conduct to different definitions of catastrophe and representations that will be from it an exit. Besides, any system of axioms allows to formulate indemonstrable statements (as has shown Godel in the theorem of incompleteness), and concerning obligations it is easy to be convinced of it: almost any system of base values easily allows to create in itself contradictions that is the basic maintenance of many literary works where the character has to choice between, let us assume, love to his family and to the native land (should make that else is called the existential choice). It is not known, whether the consistent system of values as it will look is possible in general, and whether will be applicable in practice. However work on consistent system of values is important, as it will need to be enclosed in the future computers possessing artificial intellect.

6. The errors connected with substitution of the analysis of risks by the analysis of those commercial motives who speaks about them

It is possible to argue as follows: if human investigates risks free of charge he is unemployed and marginal nut, but if he wishes to receive for it money, he parasitizes on public fears, and if it is his direct official duties, to trust him is impossible because he is the agent of the state and washes brains of the population. From here it is visible, that the direct communication between money and the analysis of risks is not present, though in some cases it is possible. The explanation through simplification is called «reductionism» and allows to explain everything.

7. Use of so-called «authoritative knowledge»

«The authoritative knowledge» was the basic source of data on the world in the Middle Ages when the truth was searched in Aristotle's works; have then invented an empirical method. References to opinions of great people should not form the sufficient basis to recognize something safe. Only regularly repeated calculations can specify in it.

8. Wrong application of idea that the theory should be considered as true, only if it is proved

If to consider scientific method, as a way of reception of the most authentic knowledge this methodology is true. However from the point of view of safety maintenance the opposite approach is necessary: a certain assumption should be considered dangerous until it is not denied. For example, the new model of the plane is considered dangerous, until then will not be proved yet, by theoretical calculations and test flights in all modes, that it is safe; the same principle underlies clinical testing of new medicines. Not clearly the same as to apply a principle falsification concerning theories about those or other global catastrophes.

9. Perception of the new information through a prism of the old

In the course of perception human only a part of the information a beret from an external world, and the rest completes on the basis of the memory, expectations and associations. Alas, the same is true and for texts, including on global risks. Reading to the review of different people of the same text, it is not difficult to be convinced, that they have apprehended it absolutely differently. Hardly it is connected by that one people were essentially cleverer than others - more likely, that they applied different filters of perception. Moreover, if human has started to adhere to a certain point of view he subscribes for those editions and chooses those articles which confirm it. Thus, as this illusion is created, than the statistics on the data confirming his point of view, grows. It even more strengthens both its filter, and its confidence of these data.

10. An error in a choice of a neutral position

Each human understands in due course, that he is not quite objective, and his point of view has some tendentiousness. To compensate this deviation, he can choose a certain neutral source of the information. The error consists in that the people adhering to opposite sights, will choose different neutral points, each of which will be closer to a position of the one who has chosen it. We described a similar error above when discussed results of experiences in which examinees have been warned about a possible error and did the amendment on it - and, nevertheless, all the same underestimated. Possibly, it was necessary to give the amendment not only to key parameter, but also to the amendment.

11. Confidence as a source of errors

The more human doubts his point of view, the more is often he changes it under the influence of the new facts, and the it is more than chances, that it will get to more authentic knowledge. If human is too assured of the opinion, it is difficult to him to change it. If it is too changeable, it does not come nearer to true, and goes on a circle.

12. Use completely the erroneous logic

Alas, the situation when human in the reasoning makes mistakes «in each line» is possible. In this case he could not find errors even if he would like. It can be or one repeating regular error, or such density of different errors which does impossible a faultless reasoning. Even I now do not know for certain, whether I do any regular logic errors at the moment. It can occur more often, than we think - the analysis of scientific texts has shown, that usually people use the reduced conclusions and heuristics receptions - and do not realize it.

13. Pre-science and pseudo-science mixture

While the hypothesis is in process of a formulation, it yet has not acquired all scientific apparatus and is, more likely, a product of brain storm on a certain theme, probably, carried out collectively by an exchange of opinions in printing editions. And during this moment it is a pre-science - however it is aimed at becoming a science part, that is to pass corresponding selection and to be accepted or rejected. The pseudo science can simulate all attributes of scientific character - ranks, references, a mathematical apparatus, - nevertheless, its purpose - not search of authentic knowledge, but visibility of reliability. All statements about global risks are hypotheses which we almost never can check up. However we should not reject them on early phases of maturing. In other words, the phase of brain storm and a phase of critical elimination should not mix up - though both should be present.

14. The error connected with wrong definition of the status of « universalis »

The reality problem universalis (that is generalizations) was the basic in medieval philosophy, and it consisted in a question, what objects actually really exists. Whether there are, for example, birds in general, or there are only separate copies of birds, and all kinds, sorts and families of birds - no more than a conditional invention of human reason? One of possible answers is that objectively there is our ability to distinguish birds and not-birds. Moreover, each bird too possesses this quality, and owing to it universalis exist objectively. In reasoning on risks the ambiguity of universalis creeps as follows: properties of one object are transferred on a certain class as though this class was object. Then there are reasoning like «America wants …» or «it is peculiar to Russian …» whereas behind these concepts there is not a uniform object, but the set, which exact definition depends on the observer. Any discussions about the politician are poisoned by such shift. Arguing on an artificial intellect is easy to make such mistake as it is not clear, whether there is a speech about one device or about a class of objects.

15. Statements about possibility of something and about impossibility are not equal

The statement about impossibility is much stronger, for enough one object concerns all set of potential objects, and for the validity of the statement about possibility. Therefore statements about impossibility something are false much more often. Assuming any event or coincidence of circumstances impossible, we cause a damage of our safety. In certain circumstances probably all. Thus any discussions about the future catastrophes is always discussions about possibilities. As Artuhov said: “I am very skeptical man. And if someone said me – it is impossible, I ask to prove it.”

16. Evidence as a source of errors

The correct conclusion always leans on two premise, two true judgments. However the analysis of texts shows, that people very seldom use the full form of conclusions, and instead use reduced where only one premise obviously is called, and another is meant by default. Are held back usually evidence - the judgments, seeming so true and doubtless, that there is need no them to sound. Moreover, it is frequent they are so obvious, that are not realized It is clear, that such state of affairs is the reason of numerous errors because evidence is not necessarily validity, and that is obvious to one, is not obvious to another.

17. Under**estimate** of own inaccuracy

As well as any human, I inclined to be mistaken, that is connected as with the basic unreliability of a human brain connected with the likelihood nature of its work, and with incompleteness of my knowledge of the world and skills of elimination of errors. I can know nothing on 100 % because reliability of my brain is not equal 100 %. I can check up reliability, having solved a series of logic problems of average complexities, and then having counted quantity of errors. However usually it is not happened, and own inaccuracy is estimated intuitively. Precisely also human usually does not measure a characteristic inaccuracy of the judgments about the future though it probably to make experimentally: for example, to write the forecast of public life for year or five years and then to compare.

18. The error connected with representation that each event has one reason

Actually:

There are absolutely casual events.

Each event has many reasons (the glass has fallen because it was put on the edge, because it is made of glass, because force of gravitation is great, because the floor was firm, because the cat disobedient, because it should happen sooner or later).

Each reason has own reasons therefore we have dispersing in tree of the reasons. Human mind is incapable entirely this tree of the reasons to capture and is compelled to simplify. But the concept of "reason" is necessary in a society because it is connected with fault, punishment and a free will. That is here under "causal" acceptance by the free made human of the decision on crime fulfillment means. There is no need to speak about that, how many here the unevident moments. (The basic question: Who is guilty?)

And in techniques designing: where it is important to find a cause of catastrophe. That is that it is possible to eliminate - so failures like that will not happen anymore. (The basic question: What to do?)

The concept the reason less all is applicable to the analysis of the difficult unique phenomena, such as human behavior and history. The example to that is weight of the confused discussions about those reasons or other historical events. For this reason reasoning in a sort «the reason of global catastrophe will be Х» - to put it mildly, are imperfect.

19. Necessity of a choice on the basis of belief

If the head of state receives some conclusions contradicting each other about safety, he makes a choice between them, simply trusting in one of them - for the reasons which have been not connected with the logic. Here too it is possible to recollect the term «an existential choice» when human should make a choice in a non-formalizable situation. For example, between love and a obligation.

20. Effect of first and last read book

The order of receipt of the information influences its estimate, and are allocated first and last source. It is one of forms of the inaccuracy connected with availability of the information.

21. Exaggeration of a role of computer modeling

Two most elaborated models are meteorology and nuclear explosions modeling. Both are made on a huge actual material, with the account of hundreds tests which made amendments to forecasts, and both regularly gave errors. Even the most exact model remains a model. Therefore we cannot strongly rely on computer modeling of unique events to what global catastrophe concerns.

22. The proof by analogy as a source of possible errors

The main idea is not only that there cannot be analogies to the unique event, which has never happened - to irreversible global catastrophe, but also that we do not know how to draw such analogies. In any case, analogy can only illustrate. Possibly, it is useful to accept analogies when they speak about a reality of a certain threat, but not when - about safety.

**23. The error connected with discrepancy of extrapolation exponential probability function by means of the linear**

Probability function of destruction of a civilization - if to consider it process smooth in sense of probability, that is, of course, incorrect - it is possible to assimilate functions of disintegration of radioactive atom which, as is known, is described by exponent. For example, if the probability of  destruction of a civilization during the XXI century is equal 50 % as it is assumed by sir Martin Rees in the book «Our final hour» in 200 years, the chance of the survival of the civilization will be 25 %, and through one thousand years - only 0.1 % - at uniform preservation of the same tendencies. From here it is visible, that it is incorrect to conclude, that if chances of the survival within a millennium makes 0.1 % for one century it will be in only ten times more, that is 1 %. The same error in less obvious kind arises, if we need to extrapolate the same 50 % of a survival within 100 years on annual probability of  destruction. Linear approximation would give 0.5 % for a year. However the exact value calculated under formula , makes approximately 0,7 %, that is in 1,4 times above, than intuitive linear approximation gives.

24. The St.-Petersburg paradox

This paradox has the direct relation to global catastrophes as shows that infinitely big damage from the extremely rare events has bigger weight, than all other events, however psychologically people are not ready to apprehend it. G.G. Malinetsky so describes this paradox in the book «Risk. Sustainable development. Synergetrics": "We will consider the following game. The coin is thrown until the heads for the first time will not drop out. If it was required *n* throws the prize will make 2*n* units. That is prizes 2,4,8, … 2*n* will occur to probability 1/2,1/4,1/8, … 1/2*n*. The expected prize in this game is infinite:

.

It is asked, how much a man is ready to pay for the right to enter into such game. The paradox consists that the majority of people is ready to pay for this right no more than 100, and sometimes only 20 units».

25. Distinction between danger and risk

The risk is created by accepted decisions, and dangers by circumstances. As the basic source of risk of global catastrophes are new technologies, decisions on their development and application define it. However if technologies develop spontaneously and unconsciously they become similar to natural dangers.

26. The error connected by that if probability of some events is not computable, it is believed to be zero

Whereas the principle of precaution would demand that we attributed to such events 100 percentage probability. However it would lead to absurd conclusions in the spirit of: the probability of disembarkation of aliens is unknown tomorrow, therefore we should prepare for it how if it was equal to 100 percent. In this case it is possible to use indirect ways of an estimate of probability, for example, formula of Gott.

27. Omission of that safety of system is defined by its weakest link

If in a house there are three parallel doors, one of which is locked by three locks, the second - two, and the third - one the house is locked on one lock. As do not strengthen two strongest doors, it will change nothing.

28. Denial of hypotheses without consideration

To reject a certain hypothesis, it should be considered in the beginning. But it is frequent that this sequence is broken. People refuse to consider those or other improbable assumptions because they reject them. However reliably to reject a certain assumption is possible, only if it having considered carefully, and for this purpose it is necessary to accept it at least for some time seriously.

**29.Non-computability**

Variety of essentially important processes for us is so combined what to predict them it is impossible, as they are incomputable. Un-computability can have the different reasons.

* It can be connected with incomprehensibility of process (for example, Technological Singularity, or, for example, how the theorem Fermat is incomprehensible for a dog), that is connected with basic qualitative limitation of a human brain. (Such is our situation with a prediction of behavior of Superintelligence in the form of AI.)
* It can be connected with quantum processes which do possible only a likelihood prediction, that is indeterministic systems (weather forecast, a brain).
* It can be connected with supercomplexity of systems in which force each new factor completely changes our representation about a definitive outcome. That concern: models of global warming, nuclear winter, global economy, model of exhaustion of resources. Four last fields of knowledge unite that everyone describes the unique event, which never was in history that is advancing model.
*  Noncomputability can be connected that the meant volume of calculations though is final, but it is so great, that any conceivable computer cannot execute it during Universe existence (such un-computability it is used in cryptography). This kind of un-computability can be shown in the form of chaotic determined system.
* Noncomputability is connected also by the fact that though to us the correct theory can be known (along with many other things), we cannot know, which theory is correct. That is the theory, besides correctness, should be easily demonstrable for everyone, and it not one too, in conditions when experimental check is impossible. One way to check theories is to test them by so called prediction markets, where price of oil for example reflects measure of confidence in Peak Oil theory. However besides theory market price influence many other factors: gamble, emotions or not market nature of the object. (It is senseless to be insured against global catastrophe as there nobody to pay off, and owing to it it is possible to tell, that its insurance price is equal to zero.)
* One more kind noncomputability is connected with possibility of realization of self-coming true or self-denying forecasts which do system essentially non-stable and unpredictable.
* Un-computability, connected with the *self-sampling assumption* - see about it N. Bostrom’s book. The essence of this assumption consists that in some situation I should consider myself as the casual representative from some set of people. For example, considering myself as usual human, I can conclude, that I with probability in 1/12 had chances to be born in September. Or with probability, let us assume, 1 to 1000 I could be born the dwarf. It sometimes allows to do predictions on the future: namely, if in Russia is 100 billionaires chances, that I will become the billionaire, make one to 1,5 million, in the assumption, that this proportion will remain. To incomputability it results, when I try to apply the assumption of own site to own knowledge. For example, if I know, that only 10 % of futurologists give correct predictions, I should conclude, that with chances of 90 % any my predictions are wrong. The majority of people do not notice it as for the account of superconfidence and the raised estimate consider themselves not as one of representatives of set but as "elite" of this set, the possessing raised ability to predictions. It is especially shown in gambling and game in the market where people do not follow obvious thought: «the majority of people loses in a roulette, hence, I, most likely, will lose».
* The similar form of incomputability is connected with an information neutrality of the market. (What is told further is considerable simplification of the theory of the market and problems of information value of indicators given to it. However more detailed consideration does not remove the named problem but only complicates it, creating one more level incomputability - namely the impossibility for the usual man to seize all completeness of knowledge connected with the theory of predictions, and also uncertainty of what of theories of predictions is true. See about information value of the market so-called «*no trade theorem*».) The ideal market is in balance in which half of players considers, that the goods will rise in price, and half - what to become cheaper. In other words, to win in a game with the zero sum from the majority of people, a man should be only cleverer or more informed, than they. However the majority of people are not cleverer, than all people, by definition, though are not capable to realize it because of psychological bias. For example, the price for oil is at such level that does not give obvious acknowledgment to the assumption of inevitability of the crisis connected with exhaustion of oil, the assumption of limitlessness of oil stocks. As a result the rational player does not receive any information on to which scenario to prepare. The same situation concerns disputes: If a certain human has chosen to prove the point of view opposite to yours, and you of anything do not know about his intelligence, erudition and information sources (and you agree that you are ordinary man, but not special), and also about the objective rating, that is chances 50 on 50, that he is right, instead of you. As objectively to measure own the intelligence and awareness is extremely difficultly because of desire to overestimate them, it is necessary to consider their being in the middle of spectrum.
* As in a modern society operate mechanisms of transformation of any future parameters in market indexes (for example, trade in quotas under the Kiotsky report on emissions of carbonic gas or the rate on elections, war, futures for weather etc) it brings an additional element of basic unpredictability in all kinds of activity. Owing to such trade we cannot learn for certain, whether there will be a global warming, exhaustion of oil, and how much is real threat of a bird flu.
* One more reason incomputability is privacy. If we try to consider this privacy through different «conspiracy theories» in the spirit of Simmons book «Twilight in desert» about overestimate of stocks of the Saudi oil, we receive dispersing space of interpretations. (That is, unlike a usual case when accuracy raises with number of the measurements, here each new fact only increases split between opposite interpretations.) Any man on the Earth does not possess all completeness of the classified information, as different organizations has different secrets.
* Market-based mechanisms to encourage people to lie about the quality of their products and on projections of their firms in order to get more profits for the time of their work. A clear example of this, we see the consequences of the so-called «revolution of managers», when managers changed owners in the directorate of firms in 1970e years. As a result, they became more interested in the short term profit for the duration of their work in the company without addressing the risks to companies outside this time limit.

The psychological aspect of this problem consists that people argue as if no incomputability exists. In other words, it is possible to find out some opinions and reasoning about the future in which its basic and multilateral unpredictability is not considered at all, no less than limitation of human ability to argue authentically.

# Chapter 5. Specific errors arising in discussions about danger of uncontrollable development of Artificial Intelligence

Chapter 5. Specific errors arising in discussions about danger of uncontrolled development of an artificial intelligence

Artificial Intelligence (AI) is an area of global risk particularly fraught with biases. Thinking about Artificial Intelligence evidently pushes certain buttons in the human mind which lead to exceptionally poor reasoning, especially among those new to considering AI risk, but also among those actually working in the field. In this chapter we attempt to outline some of the most common errors.

1. Belief in predetermined outcomes or quick fixes

Experts in Artificial Intelligence often state why they think advanced AI will be safe, but give mutually exclusive answers that cannot all possibly be true. Some of these experts must be wrong, it's just a question of who. In 2008, one of the authors of this volume, Alexei Turchin, ran an Internet survey among developers of Artificial Intelligence on the theme of guaranteed safety in AI, and received the following answers, with an approximately equivalent number of people expressing each view. Various AI experts or commentators who hold the stated views are cited, although few of them participated in the survey. Many of the citations are not from academic sources, though the views below are common among academics. *“AI is safe, because...”*

1. Because true AI is impossible1. (Roger Penrose: *“algorithmic computers are doomed to subservience.”*)
2. Because AI can only solve narrow problems, such as image recognition1. (Penrose: "Mathematical truth is not something that we can ascertain merely by use of an algorithm”)
3. Because the Three Laws of Robotics (or an updated variant) will solve any AI safety issues2. (David Woods and Robin Murphy.)
4. Because I know how to make AI safe3. (Mark Waser.)
5. Because AI will possess superhuman wisdom by definition, which will cause it to be benevolent or otherwise non-harmful4.
6. Because AI will not need anything from humans, allowing us to peacefully co-exist with them5. (In the cited article, the editor or writer misinterprets the quoted experts by titling the article “AI uprising: humans will be outsourced, not obliterated.”)
7. Because AI will be trapped in computers, and if anything goes wrong, it will be possible to simply pull the plug.6
8. Because AI cannot have free will.
9. AI will be dangerous, but we'll manage somehow.
10. AI is dangerous, and we are all doomed.
11. AI will destroy mankind, and we should welcome it, as AI is a higher stage of evolution7.

All of the above cannot be true simultaneously. Some—indeed, most—must be mistaken.

2. The idea that it is possible to create a faultless system   
by repeatedly checking the code

Checks bring some number of new errors, and owing to it at certain level the number of errors is stabilized. It is true and about systems definition of objectives which are laws, for example. It is not necessary to count that we can create the arch corrected behavior for AI, not containing errors.

**3. Errors in the critique of Artificial Intelligence by Roger Penrose**

In his book *The Emperor’s New Mind*, physicist and philosopher Roger Penrose asserts that strong AI is not possible through algorithmic design methods, because in human brain there are not computable quantum processes which are necessary for creative thinking and consciousness. On this basis, it is sometimes affirmed by others (but not Penrose himself) that dangerous AI is impossible or very distant in the future. However, this conclusion is flawed for the following reasons:

* 1. Penrose himself admits that strong (i.e., conscious) AI is possible. According to a review of his book by Robin Hanson8, *“Penrose grants that we may be able to artificially construct conscious intelligence, and “such objects could succeed in actually superceding human beings.”* For Penrose, the key ingredient is consciousness, which he considers to be non-algorithmic. But, he does not rule out non-algorithmic approaches to consciousness in intelligent machines. So, the notion that Penrose is predicting that strong AI will never happen does not hold water. It is merely the standard paradigm of computer scientists that Penrose is criticizing, not the feasibility of AI in general.
  2. Penrose has argued that the neurons in the brain work in some way based on the quantum decoherence of particles in neurons, but physicist Max Tegmark showed that the relevant timescales of dynamical activity in neuron firings are roughly ten billion times longer than quantum decoherence events, making a connection between the two extremely unlikely9. Penrose's ideas on quantum involvement in consciousness have been met with near-universal dismissal among experts in philosophy, computer science, physics, and robotics.10
  3. Artificial Intelligence can be dangerous without possessing phenomenal consciousness of the type that Penrose emphasizes. In an overview of AI-related risks, Luke Muehlhauser and Anna Salamon write11, *“a machine need not be conscious to intelligently reshape the world according to its preferences, as demonstrated by goal-directed “narrow AI” programs such as the leading chess-playing programs.”* In a paper on nanotechnology and national security, Mark Gubrud writes12, *“By advanced artificial general intelligence, I mean AI systems that rival or surpass the human brain in complexity and speed, that can acquire, manipulate and reason with general knowledge, and that are usable in essentially any phase of industrial or military operations where a human intelligence would otherwise be needed. Such systems may be modeled on the human brain, but they do not necessarily have to be, and they do not have to be "conscious" or possess any other competence that is not strictly relevant to their application. What matters is that such systems can be used to replace human brains in tasks ranging from organizing and running a mine or a factory to piloting an airplane, analyzing intelligence data or planning a battle.”*

4. The notion that Asimov's Three Laws of Robotics address AI risk

This was mentioned in the first bullet point for this chapter, but is worth devoting its own section to, since the misconception is so universal. Asimov's laws of robotics are the following:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given to it by human beings, except where such orders would conflict with the First Law.
3. A robot may not injure its own kind and defend its own kind unless it is interfering with the first or second rule.

Many different authors and experts in Artificial Intelligence who have thought seriously about the problem of AI safety have written at length regarding why Asimov's Laws are insufficient. There are three key points: first, Asimov's laws are too simple, secondly, they are negative injunctions rather than positive injunctions, and therefore woefully underspecified, thirdly, language is inherently ambiguous and the definition of “harm” is not exhaustive.

For instance, consider that humans could be best prevented from coming to harm by sealing every human being in a diamond bubble and hooking up them to virtual reality, *a la* The Matrix. Asimov's laws greatly underestimate the power of AIs to reshape the world in the long term. An Artificial Intelligence that can self-replicate and mine resources to mass-produce robotics could quickly come to dominate the planet. So, a positive agenda for AI is necessary to set the course of its future, not merely a list of what AI *cannot* do.

The notion of “obeying humans” is ambiguous. What if orders from humans conflict? Do thoughts count as orders? An AI could eventually gain the capability to read the thoughts of humans directly with non-invasive scanning or other monitoring techniques, do we want an AI carrying out the most fleeting wishes of every human being on Earth, even if they contradict each other from one moment to the next?

All words are ambiguous. Even if Asimov's laws were to serve as a template, or inspiration for a set of benevolent AI motivations, they would have to be fleshed out in absolutely exhaustive detail, and many important theoretical questions would need to be solved before they could be implemented. Speaking as if the problem is solved by the existence of Asimov's laws is analogous to drawing a picture of a rocket with a crayon and saying you've designed a spaceship that can go to Mars. A few sentences are not sufficient for an AI theorist to develop a well-grounded goal system.

Asimov's laws of robotics are so simplistic that they are a tautology. An AI will be safe because it will not allow humans to come to harm. Well, naturally. How do we define harm in terms of computer code, in a way that remains somewhat consistent over time as the AI's knowledge base and even fundamental ontology changes? It turns out that these problems are so complicated that we might as well throw out Asimov's laws altogether and start from scratch, beginning with the question of what positive drives—not just forbidden actions—should direct advanced AI.

5. Erroneous idea that software progress is absent

It is easier to measure progress in computing hardware than software, making it easier to claim that software progress is slow or stalling. Katja Grace, visiting fellow for the Machine Intelligence Research Institute, describes “algorithmic progress in six domains” in her paper of that title13. In the abstract, Grace writes, “gains from algorithmic progress have been roughly fifty to one hundred percent as large as those from hardware progress.” Considering that hardware cost effectiveness doubles roughly every 18 months (Moore's law), this is a fairly quick pace.

The six areas that Grace focused on in her paper were the Boolean satisfiabiality problem, game playing (chess and Go), physics simulations, operations research, and machine learning.

Though Grace demonstrated software gains in each domain similar to the hardware gains, the clearest area of improvement was in physics simulations, specifically in simulations of magnetic confinement fusion. The “effective speed” of simulations of micro-turbulence and global magnetohydrodynamics (the motion of a fluid under a magnetic field) clearly improves exponentially, even when hardware improvements are factored out. Features such as improved electron models and linear solvers are cited as major contributors to the software speedup.

In linear optimization for operations research, there are similar signs of major speedups due to improvements in sofware. Grace cites Martin Grötschel, who claims an average 2-times speedup for a production-planning problem for every year between 1988 and 2003. Robert Bixby claims a 3,300-times software-derived speedup in a linear programming problem during an unspecific period leading up to 2004. The hardware-derived improvement in the same period was about 1,600, which corresponds to improvement over the course of about 15 years. Since hardware performance doubles about every 18 months, we can calculate a software performance doubling time of roughly 16 months, for this specific problem.

None of this means that there aren't many domains in which software progress is stagnant, or that areas such as natural language processing aren't susceptible to diminishing returns. However, it appears that software performance doubling is a fairly common trend across a variety of fields and it seems likely that these doublings are having a direct impact on various subproblems of Artificial Intelligence.

6. The erroneous idea that advanced AI is considered uninteresting   
or intractable and no one is focusing on it

To address this point requires making a distinction between narrow AI and general AI. Narrow AI is artificial intelligence designed to focus on certain specific problems, like driving a car or playing chess. General AI is artificial intelligence designed to work as general problem-solvers. Naturally, the latter is much more difficult than the former. This doesn't mean that effort isn't being placed towards general AI. Google and many smaller organizations have stated publicly that they are working on it.

Google bought the artificial intelligence company DeepMind, of which an investor said14, *“*If anyone builds something remotely resembling artificial general intelligence, this will be the team.” Google's research director Peter Norvig said “less than 50 percent but certainly more than 5 percent” of the world’s leading experts on machine learning work at Google, and this statement was made prior to Google's acquisition of DeepMind. Ray Kurzweil, probably the world's leading public advocate of general artifical intelligence, was hired by Google to “work on new projects involving machine learning and language processing” in December 2012, to great fanfare15. Peter Norvig has twice given talks at the Singularity Summit, a conference closely associated with the community around general artificial intelligence, and has also participated at the Conference on Artificial General Intelligence in Mountain View.

Another major company working on general AI is Grok, which changed its name from Numenta in 2013. Grok's website states that its mission is to be a “catalyst for machine intelligence,” and to “accelerate the creation of machines that learn”. Grok is led by Jeff Hawkins, the founder of Palm.

A spinoff company from Grok is Vicarious, with similar goals and technology. Their website states, “we're building software that thinks and learns like a human”. Vicarious was founded by Dileep George, a protege of Hawkins, and has raised more than $40 million from big names like Mark Zuckerberg, Peter Thiel, and Elon Musk. Instead of using the Deep Learning approach common to many software development companies, Vicarious' AI design is directly inspired by the mammalian brain and the hierarchical structure of temporal memory.

Though there have always been “AI winters,” that is, periods of diminished activity, the field arguably emerged from any haitus in the late 2000s. Since then, the advances have been quick in coming, and many of the most interesting companies have been bought by Google. Much work remains to be done, but progress is steady.

**7. Anthropomorphism**

The stumbling block of anthropomorphism, that is, “human-shaped” thinking has been emphasized by many writers in AI risk16. Having evolved psychologies that are adapted to interact with other humans, human is what we naturally imagine other agents as being. An Artificial Intelligence may have instincts and drives nothing like ours—no jealousy, no primal confidence, no love, no hate, no boredom, no social self-consciousness—close to nothing that we associate with “agent-ness”. An AI might have no inner subjective life whatsoever. Our intuitions about what a highly intelligent agent “would” do or “ought” to do are plausibly profoundly mistaken when it comes to AI.

What can we be confident that an advanced AI would do, if not act in human-like ways? Certain extremely basic things; pursue goals (since goal-driven behavior is the foundation of cognition), preserve its own survival to an extent (or it would not be very effective at pursuing its goals), expand its resources (since resources are useful for accomplishing nearly any goal), and implement efficiency (achieve more goals with fewer resources). These are the “basic AI drives” outlined by Steve Omohundro in his paper of the same name17. Since an AI could continue to build copies of itself or agents to do its bidding, it need not experience satiation like human beings do. It could very well have an infinite “hunger”, and convert all materials on the planet to robots to be its servants, or computers to do its thinking. Any solid material can be used to build a computing device. Since humans have solid parts, our material would also be useful in building computers. In this fashion, an AI could threaten us even if it were not explicitly malevolent. Eliezer Yudkowsky put it this way: *“The AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else.”*

An AI would only have human social instincts if the computational structure of these neural objects were fully understood and programmed in with painstaking detail. Subsequently, these features would need to be retained over many rounds of the AI engaging in self-improvement and self-modification. For all of them to be retained, or even programmed in the first place, seems unlikely16. An AI would not need the full suite of human instincts to be economically useful, generally helpful, and non-harmful to humans.

It is tempting to attribute human-like features to AIs, especially in the domain of tribal warfare and resentment. However, AIs could be programmed to be willing slaves. There is no universal law that all possible agents need resent control. That is only a feature of evolved organisms, which exists for obvious adaptive reasons. Biologically evolved organisms have observer-centric goal systems, but minds in general, such as artificial intelligence, would not have them unless they are explicitly programmed that way.

8. Pulling the plug

An advanced AI will be able to access the Internet and port itself onto other machines, so it will not be possible to pull the plug. Even today, “pulling the plug” rarely solves anything.

9. Erroneous representation that, even having extended over the Internet, AI cannot influence an external world in any way

An advanced Artificial Intelligence could accomplish a great deal even if it lacked a robotic or biological body. It could split itself into thousands of software agents, putting them to work making money over the Internet simultaneously. It could pose as human beings remotely and hire hitmen or mercenaries to eliminate its enemies. It could start mass-producing robots to fulfill its goals. It could even begin researching nanotechnology or other superior manufacturing methods to allow it to build robots and weapons much stronger and faster than anything today. Just because an AI starts off within a computer does not mean it does not pose a threat.

10. Erroneous representation that AI cannot have its own desires,   
and therefore could never do harm to humans

An AI would need some goals to get started, even if they were only implicit rather than explicit. Goals means physical motion, and physical motion can kill. An AI would not necessarily have “common sense” morality like human beings do, and even if it could be programmed in, it is possible that some other goal could override it. For this reason, carefully programming a self-improving AI's goal system to be stable under self-modification is of utmost importance. We need to account for scenarios where AIs are able to massively self-replicate and overwhelm the planet, where the slightest error in programming could potentially cause our demise. Certainly we can anticipate self-correcting goal systems of a sort, but it is as best to be as conservative as possible. There will come a point where Artificial Intelligence surpasses humankind as the most powerful intelligence on Earth, and when that happens, we want to be certain that AI will improve our lives.

11. Erroneous representation that AI will master space, leaving Earth to humans

The Earth is a planet rich in resources. An Artificial Intelligence looking to pursue its goals will not just skip it, blast off into space, and leave Earth untouched. It would be far more useful to first utilize all the Earth's resources, transform it into computers or paperclips, then follow that up with the solar system and the galaxy. We will not be spared from a human-indifferent AI just because it overlooks what is right under its nose. That is wishful thinking.

12. Erroneous idea that greater intelligence inevitably   
leads to non-threatening supergoal X

Intelligence is a tool which can be directed to the achievement of any purpose. *Homo sapiens* uses our intelligence to pursue general goals, the outlines of which are programmed into our brains from birth. This includes finding food and water, being entertained, making friends, fighting off competitors, acquiring a mate, and so on. For the sake of these goals, ships are launched, theorems are proven, and plots are hatched. The presence of intelligence does not entail any unequivocal purpose. Different humans have slightly different goals, and animals have more different goals still. Artificial Intelligence would have an even higher possible level of goal variation, and might pursue goals considered quite outlandish to us, such as converting all matter in the solar system into paperclips, if that's how they were programmed.

The idea that a sufficiently wise AI will cooperate with everyone, or be kind to everyone, or some other kind of optimistic trait, is simply wishful thinking and projection16. We want the AI to be cooperative, or we feel that wiser humans tend to be cooperative, so we think advanced AI will be. We neglect our complex evolutionary history and limitations which make cooperation ecologically rational *for Homo sapiens, in the context of our innate neural architectures*. An AI would not need to cooperate to achieve its goals; it could simply wipe out every human being on the planet and do exactly as it wants without any fuss.

Artificial Intelligence has a tendency to be a blank canvas for us to project our fears and hopes onto. If we would appreciate a benevolent father figure, that's what we imagine a superintelligent AI as. If we wish for a loving mother figure, we imagine that. If we desire a passive background “operating system” for the solar system, that's what we envision. The thing is, the structural and goals of a superintelligence do not depend so much on what we envision or prefer, but more on what goals we program into an AI and how these modify or change as the AI becomes trillions of times larger and more powerful. Conceivably all of the outcomes outlined above are possible, it's just a question of what sort of programming the AI receives when it is just a small seed.

13. Underestimating the performance of computers

Most people who casually say that Artificial Intelligence is not likely to be built for 70 or more years generally cannot say what the expert estimates of the human brain's computing power is nor how much computing power the best computers have today. Many of them do not even recognize that the brain itself is essentially a deterministic computer, with the neurons serving both as processors and memory units. Among the cognitive science community, this isn't even up for debate—the brain is a biological computer, full stop. The brain may not have a standard von Neumann architecture, but not that does not mean that it isn't a computer. A computer with enough computing power will certainly be able to simulate the brain and display intelligence itself. This means that even if we never figure out how intelligence works, we will eventually be able to create Artificial Intelligence by copying the brain's structure into a dynamic computer program. A report from the University of Oxford estimates that this will happen with a median estimate of 208018. That sets a rough upper bound for the likely creation of Artificial Intelligence.   
 How can one purport to estimate when AI could be created without knowing such simple numbers as those mentioned above? It doesn't make any sense, but alleged experts in philosophy and futurism do so all the time. The quantity of the first estiamte—the computing power of the human brain—varies from roughly 10^14 operations per second, based on Moravec's estimate19, to 10^19 operations per second, an upper bound estimated by a team at Oxford focusing on whole brain emulation17. Put another way, it ranges from a hundred exaflops (trillion floating operations per second) to 10,000 petaflops (quadrillion floating operations per second). The world's most powerful solitary supercomputer as of this writing (May 2014) is the *Tianhe-2* supercomputer in China, which has a peak performance of slightly less than 55 petaflops. So, according to some estimates, our best computers are already more than fast enough to run software with as much computing capacity as the human brain. Accordingly, most researchers in the field of artificial general intelligence do not consider hardware the salient issue, but software. One obstacle is that Moore's law (the periodic doubling of computer performance) seems to already be dead or dying, so improvements in cost effectiveness of computers may not be forthcoming. If so, this could be a major long-term barrier to the creation of Artificial Intelligence.

The issue of Moore's law slowing is ameliorated somewhat by the fact that few researchers anticipate it would take the full computing power of the human brain to create an Artificial Intelligence. It is likely that it will be many times less, due to software optimization.

14. Underestimating how much we know about intelligence and the brain

A great many scientists, philosophers, and academics without firsthand knowledge of cognitive science or detailed familiarity with some of its subfields tend to radically underestimate just how much we know about the human brain and its function. Although we do not have a wiring diagram of the brain, we understand its gross structure consists of macrocolumns and microcolumns, divided into 52 functional areas. Google Scholar brings up about 275,000 articles on neuroanatomy and 35,000 articles on “functional neuroanatomy.” The MIT Encyclopedia of Cognitive Science (MITECS), a “brief” overview reference work on the brain, is 1096 pages. We have brain-computer interfaces so precise that we can measure the electrical activity in someone's visual cortex and use it to build a fuzzy video of what that person is seeing20. We are so close to artificial telepathy that the Pentagon is spending millions of dollars trying to seal the deal21.   
 Many algorithmic details about speech, reading, attention, executive control, scene scanning, happiness, pleasure, pain, sadness, motor processing, scent, and so on are known. We don't know enough to simulate these functions in a computer, but enough that MIT scientists have created a wiring diagram of the cochlea has been ported into a software program used to process sounds22. What details we do know about the brain, especially sensory perception, are sufficient to constrain many aspects of our software if we did want to emulate the human brain. It is still the early days of functional understanding of the brain's details, but we know far more than nothing, and progress is accelerating thanks to new tools and modern computers. Tools have been developed to connect neurons directly with light pulses, a field known as optogenetics. This has been used this to discover where exactly in the brain memories are stored23, though we have yet to decode them.   
 Before claiming “we know close to nothing about the brain,” skeptics should spend an hour or two looking through a large cognitive science textbook like MITECS, and see what we do know. They will discover that it's quite a lot.

15. The erroneous representation that because humans can do X ,   
AIs can never do X and consequently AI does not present any threat

Humans can perform many tasks that AIs currently cannot, and we sometimes like to get excited about that fact, boasting that because an AI cannot enjoy a sunset, strong AI is therefore extremely distant, and never worth worrying about. This is incorrect. Computer vision can already process a sunset in extreme detail, and it is only a matter of time before we create machines with the knowledge and perception to appreciate the nuances of a sunset in an ideographic way. Many of the faculties we care about the most are social, reinforced by highly niche-adapted neural modules. It may be difficult to appreciate an AI that gradually increases in its social skills until it quickly equals and then surpasses us. For a great deal of time it may remain in the “Uncanny Valley” until it becomes transhuman.

A general AI may be designing its own Iron Man androids but still lack the social skills to shake someone's hand without looking like a machine. Operating systems will still have glitches by the time AI is developed. Competence in the domain of general intelligence, which is what makes AI a threat, may be achieved without an AI reaching proficiency in any of the domains that human beings traditionally associate with success, like social dominance. An AI may have highly advanced narrow skills—planning, military skills, manufacturing, communicating in parallel with thousands of people, an exploding bank account, drones, and so on—but be a really odd fellow to sit down to tea with. So, just because an AI can't compose a concerto you like, or make your mother laugh, doesn't mean that it can't kill you and everyone you care about.

16. Erroneous conception that AI is impossible because   
AI thinks algorithmically and humans non-algorithmically

In *The Emperor's New Mind*, Roger Penrose spent a good deal of pages explaining why he thinks consciousness is non-algorithmic. David Chalmers has raised the question of the “hard problem” of consciousness in a similar context24. Though most people in cognitive science subscribe to casual functionalism, the notion that the mind is what the brain does, and thus that consciousness is merely the result of neuron firings, we should always be willing to consider that consciousness is in some way non-algorithmic.

Thankfully, we could overcome that obstacle by building non-algorithmic AI. If AI proves to be impossible through conventional algorithmic means, we can always try out new things, like growing neurons on a chip. An Artificial Intelligence could consist of a brain in a vat surrounded by a supercomputer. That would hardly make it less threatening.

17. Erroneous representation that AI will be roughly human-equivalent

Quoting Eliezer Yudkowsky25:

*Many have speculated whether the development of human-equivalent AI, however and whenever it occurs, will be shortly followed by the development of transhuman AI (Moravec 1988; Vinge 1993; Minsky 1994; Kurzweil 1999; Hofstadter 2000; McAuliffe 2001). Once AI exists it can develop in a number of different ways; for an AI to develop to the point of human-equivalence and then remain at the point of human-equivalence for an extended period would require that all liberties be simultaneously blocked at exactly the level which happens to be occupied by Homo sapiens sapiens. This is too much coincidence. Again, we observe Homo sapiens sapiens intelligence in our vicinity, not because Homo sapiens sapiens represents a basic limit, but because Homo sapiens sapiens is the very first hominid subspecies to cross the minimum line that permits the development of evolutionary psychologists.*

*Even if this were not the case—if, for example, we were now looking back on an unusually long period of stagnation for Homo sapiens—it would still be an unlicensed conclusion that the fundamental design bounds which hold for evolution acting on neurons would hold for programmers acting on transistors. Given the different design methods and different hardware, it would again be too much of a coincidence.*

*This holds doubly true for seed AI. The behavior of a strongly self-improving process (a mind with access to its own source code) is not the same as the behavior of a weakly self-improving process (evolution improving humans, humans improving knowledge). The ladder question for recursive self-improvement—whether climbing one rung yields a vantage point from which enough opportunities are visible that they suffice to reach the next rung—means that effects need not be proportional to causes. The question is not how much of an effect any given improvement has, but rather how much of an effect the improvement plus further triggered improvements and their triggered improvements have. It is literally a domino effect—the universal metaphor for small causes with disproportionate results. Our instincts for system behaviors may be enough to give us an intuitive feel for the results of any single improvement, but in this case we are asking not about the fall of a single domino, but rather about how the dominos are arranged. We are asking whether the tipping of one domino is likely to result in an isolated fall, two isolated falls, a small handful of toppled dominos, or whether it will knock over the entire chain.*

We have included a long quote from Yudkowsky because this point is rather important and may be difficult to grasp. It is common to think of human-equivalent AI as reasonable, but transhuman or superhuman AI as beyond the pale. This makes no sense, however, because humanity does not occupy a priviliged section in cognitive statespace, we are just one rung on a long ladder. The fact that we are the first truly intelligent species on the planet actually suggests that we are close to the *bottom* of possible general intelligences. A fairly limited set of genetic mutations and brain volume changes caused the upgrade from chimps to humans, and, being conservative, we should consider the possibility that another similar improvement will lead to beings qualtitatively smarter than ourselves as we are above chimpanzees. “Human-equivalent” AI is a misnomer, like a “human-equivalent” locomotive. If we figure out AI at all, it will be possible to throw the entire planet's computing resources and talent at it (including the AI's own talent) which all but ensures it will not remain “human-equivalent” for long, but will quickly become superintelligent.

18. Derailing discussion with talk about AI rights

Whether or not all or some AIs will be considered persons worthy of being granted rights is an interesting question, but besides the point when it comes to global risk. Advanced AI will eventually become a major possible threat and boon to humanity, whether or not we grant them “rights” or consider them equal persons. Similarly, AIs will not despair if we do not treat them as persons, as they will not have a human-like ego, by default. The notion that such an ego would arise organically, without deliberate programming, derives from a misunderstanding of how the brain works and the complicated evolutionary history that gave rise to its features14. Similar to how a human-like ego will not arise organically in an AI unless it is programmed in, a fear of snakes will not arise in an AI, even though all mammals have this fear. Fear of snakes and a human-like ego are similar in that they are complex functional adaptations crafted by evolution over the course of millions of years to solve concrete ecological challenges. Artificial Intelligence is not likely to be evolved, and even if it were, its selection pressures would be vastly different than our own. Genetic algorithms have only been used to solve narrow optimization problems and are not a viable candidate for the construction of general Artificial Intelligence. In conclusion, worrying about tribalistic AI rebellion or assertion of rights is just another form of anthropomorphism. If we choose to create AIs with human-like egos and grant them rights, that will be very interesting, but it is likely that many egoless AIs will be created first. Scenarios of anthropomorphic AIs are featured in fiction because they are easier for the audience to understand.

19. Erroneous representation that the risk is assauged by there being multiple AIs

It may be pointed out that there might not be just one powerful, superintelligent, human-threatening AI, but many. This does not make the risk any less. A multiplicity of AIs does not mean that they will be on par or capable of checking one another. If there is a multiplicity of AIs, as there is certain to eventually be, we had better be sure that they all have robust moralities, in addition to overarching safeguards. Multiple AIs exacerbates the risk, not ameliorates it.

20. Sidetrack of debating the definition of intelligence

Defining intelligence is not easy. Ben Goertzel has proposed the definition “the ability to achieve complex goals in complex environments using limited computational resources,” which is serviceable, but still vague. AI researcher Shane Legg has compiled a list of 52 definitions of intelligence26. Every armchair philosopher in the world loves to pontificate on the topic. Instead of discussing “intelligence,” we should ask, “can this system improve itself open-endedly and threaten the planet, or not?” That is all that really matters. Engaging in philosophical tail-chasing regarding the definition of intelligence is fun, but not useful in appraising concrete risk. An AI might become extremely powerful and threatening even if it does not meet a textbook, anthropocentric definition of intelligence. To separate the idea of powerful AI from controversies about the definition of intelligence, Yudkowsky has proposed calling a superintelligent AI a “Really Powerful Optimization Process”27.

21. Erroneous unequivocal identification of AI with a discrete object

It is tempting to think of an intelligence as a distinct object, as a human being is. However, a powerful AI would be more like a ghost, or a god, distributed across many different systems at once. It would be very difficult to defeat it as it could always hack into unguarded machines and upload its code to them.

22. Erroneous idea that AI can be boxed

As a thought experiment, some have proposed the idea of putting an AI in a box with no way of accessing the Internet, only releasing it if it proves trustworthy. The problem with this is that the AI can lie, or promise people things to let it out.

23. Objection that because AI has previously failed, it is guaranteed to fail

Just because general AI has not been previously created does not mean it will never be created. This objection has begun to fade recently, as the field of AI has made a comeback during the 2010s.

24. Erroneous idea that giving AI a simple command will work

This objection relates to the earlier items on anthropomorphism and Asimov's laws. Giving an AI a command like “love all people,” “cause no one any harm,” or “obey only me” is likely to work only if the AI already has an extremely sophisticated goal system that predisposes it to listen to commands, fully understand them, and follow through on them only in ways which its master intends. The problem is setting this goal system up in the first place, not just uttering a command. An AI is not a human—it would not have the “common sense” we have about a millon different issues which makes it behave in ways which are intuitive to us. For instance, in response to the command, “love all people,” it might decide to hook people up to machines that directly stimulate their pleasure center all day. Giving an AI the common sense to know why this would be bad is the challenge, and the challenge is huge. The proposal of “just tell the AI what to do” does not diminish the size of the technical challenge.

25. Objection in the spirit of “when we have AI, then we'll worry about safety”

It's best to start worrying about safety now. Safety considerations are likely to impact fundamental considerations about how an AI is constructed, and will only be integrated into the architecture if they are considered well ahead of time. The Machine Intelligence Research Institute refers to this as “Friendly AI theory”.

**26. Difference of ability and intention**

Yudkowsky calls this the “Giant Cheesecake Fallacy”--the idea that if an AI has the ability to build a gigantic, 30-foot tall cheesecake, it will surely do so. Just because an agent has the ability to do something, does not mean it will do it. Ability and intention have to align for an agent to take action.

**Citations**

1. Roger Penrose. *The Emperor's New Mind*. Oxford University Press, 1999.
2. Jeremy Hsu for Space.com. "Science Fiction's Robotics Laws Need Reality Check”. <http://www.space.com/7148-science-fictions-robotics-laws-reality-check.html> Retrieved 4/21/2014.
3. Mark Waser. Rational Universal Benevolence: Simpler, Safer, and Wiser than “Friendly AI. *Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3-6, 2011. Proceedings*. Springer Berlin Heidelberg, 2011.
4. Mark Waser. “Wisdom Does Imply Benevolence.” Becoming Gaia blog. <http://becominggaia.wordpress.com/papers/wisdom-does-imply-benevolence/> Retrieved 4/21/2014.
5. Mark Piesing for Wired.co.uk. “AI uprising: humans will be outsourced, not obliterated.” <http://www.wired.co.uk/news/archive/2012-05/17/the-dangers-of-an-ai-smarter-than-us> Retrieved 4/21/2014
6. Chris Willis for Android World. “Are super-intelligent machines a danger to humanity?” <http://www.androidworld.com/prod90.htm> Retrived 4/21/2014
7. Hugo de Garis. *The Artilect War*. Etc Publications, 2005.
8. Robin Hanson. Has Penrose Disproved AI? Foresight Update No 12, pp. 4,5, 1991, publication of the Foresight Institute.
9. Max Tegmark. “The importance of quantum decoherence in brain processes”. <http://arxiv.org/abs/quant-ph/9907009>. 1999.
10. Victor Stenger. “The Myth of Quantum Consciousness”. *The Humanist*. 53 No 3 (May—June 1992). pp. 13-15.
11. Luke Muehlhauser and Anna Salamon. Intelligence Explosion: Evidence and Import. *Singularity Hypothesis: A Scientific and Philosophical Assessment*. Springer, 2012.
12. Mark Gubrud. “Nanotechnology and International Security”. Draft paper for the Fifth Foresight Conference on Molecular Nanotechnology, 1997. <http://www.foresight.org/Conferences/MNT05/Papers/Gubrud/>
13. Katja Grace. “Algorithmic Progress in Six Domains”. Technical report 2013-3. Berkeley, CA: Machine Intelligence Research Institute, 2013.
14. Dan Rowinski for ReadWriteWeb. “Google's Game of Moneyball in the Age of Artificial Intelligence”. January 29, 2014. <http://readwrite.com/2014/01/29/google-artificial-intelligence-robots-cognitive-computing-moneyball>
15. KurzweilAI.net. “Kurzweil joins Google to work on new projects involving machine learning and language processing”. December 14, 2012. <http://www.kurzweilai.net/kurzweil-joins-google-to-work-on-new-projects-involving-machine-learning-and-language-processing>
16. Eliezer Yudkowsky. 2001. C*reating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures*. The Singularity Institute, San Francisco, CA.
17. Stephen Omohundro. “The Basic AI Drives”. November 30, 2007. Self-Aware Systems, Palo Alto, California.
18. Anders Sandberg and Nick Bostrom. *Whole Brain Emulation: A Roadmap, Technical Report #2008‐3*, Future of Humanity Institute, Oxford University. [www.fhi.ox.ac.uk/reports/2008](http://www.fhi.ox.ac.uk/reports/2008)‐3.pdf
19. Hans Moravec. “When will computing hardware match the human brain?” *The Journal of Evolution and Technology*, vol 1. Dec. 1997. <http://www.transhumanist.com/volume1/moravec.htm>
20. UC Berkeley News Center. “Scientists use brain imaging to reveal the movies in our mind.” September 22, 2011. <https://newscenter.berkeley.edu/2011/09/22/brain-movies/>
21. Eko Armunanto. “Artificial telepathy to create Pentagon’s telepathic soldiers”. May 10, 2013. <http://digitaljournal.com/article/349839>
22. Lloyd Watts, Richard F. Lyon, Carver Mead. "A Bidirectional Analog VLSI Cochlear Model”. 1991. California Institute of Technology. <http://www.lloydwatts.com/WattsBidirectional1991.pdf>
23. Sebastian Anthony. “MIT discovers the location of memories: Individual neurons”. Extreme Tech. <http://www.extremetech.com/extreme/123485-mit-discovers-the-location-of-memories-individual-neurons>
24. David Chalmers. “Facing Up to the Problem of Consciousness”. *Journal of Consciousness Studies* 2(3):200-19, 1995.
25. Eliezer Yudkowsky. 2007. “Levels of Organization in General Intelligence.” In Artificial General Intelligence, edited by Ben Goertzel and Cassio Pennachin, 389–501. Cognitive Technologies. Berlin: Springer. Doi:10.1007/ 978-3-540-68677-4\_12.
26. Shane Legg. “A Collection of Definitions of Intelligence”. IDSIA Technical Report 07-07. June 15th, 2007. <http://arxiv.org/pdf/0706.3639.pdf>

Eliezer. Yudkowsky. *Coherent Extrapolated Volition*. 2004. The Singularity Institute, San Francisco, CA

Chapter 6. The specific reasoning errors applicable   
to risks from use of nanotechnology

Like Artificial Intelligence and many other global risks, there are numerous reasoning errors people often make with respect to nanotechnology. Most risks from nanotechnology concern risks from molecular assemblers, nanofactories, and nanorobots; speculative future technologies whose properties are poorly understood. However, this has not stopped arms control experts and risk analysts such as Jurgen Altmann1 and Mark Gubrud2 from examining the field. Issues around nanotechnology and molecular manufacturing can seem intimidating from the perspective of newcomers to the study of global risk, but it is important to recognize that there is a consensus view which is fairly well developed, and can serve as a starting point for further discussion and risk analysis. A list of common reasoning errors around nanotechnology risk follows.

**1. Erroneous idea that molecular robotics are impossible**

There are thousands if not millions of examples of molecular robotics within the human body. Scientists have created DNA nanodevices with a special coating that can avoid the safeguards of the mouse immune system3. There are self-assembling nanodevices that move and change shape on demand4. There are molecular robotics on nano-assembly lines that can be programmed to 'manufacture' eight different molecular products5. DNA nanorobots have been injected into live cockroaches and programmed to perform complex tasks6. Molecular manufacturing cannot be said to be impossible because it already exists. A more relevant question is how long it will take to mature, which is unknown.

**2. Erroneous idea that grey goo is a greater risk than nano arms-races**

In 1986 a landmark book introducing many of the fundamental concepts of nanotechnology, *Engines of Creation*, by Eric Drexler, was published7. The book mentioned out-of-control, biomass-consuming self-replicating nanotechnology, or “grey goo,” as an aside, and the concept was latched onto by the media. In a 2004 paper, “Safe Exponential Manufacturing,” Eric Drexler and co-author Chris Phoenix emphasized that grey goo is not the greatest threat from nanotechnology8, saying, *“Fictional pictures of MNT commonly assume that pulling molecules apart would be as easy as putting them together—that assemblers could swarm over a pile of grass clippings and convert it directly into a personal spaceship. This is not the case.”* They highlighted non-replicating weapons as a larger risk, writing, *“The authors do not mean to imply that advanced mechanochemical manufacturing will create no risks. On the contrary, the technology introduces several problems more severe than runaway replicators. One of the most serious risks comes from non-replicating weapons.”* Even today, many people with a casual, non-academic understanding of nanotechnology continue to perpetuate the grey goo risk, not realizing that the “father of nanotechnology” himself, Eric Drexler, has repeatedly emphasized that it is not the primary risk.

**3. Erroneous idea that nanotechnology is connected only with materials science, photolithography, chemisty, and nanotubes**

In the early 2000s, a federal government program called the National Nanotechnology Initiative (NNI) handed out several billion dollars in grants to fund “nanotechnology”. However, their definition of nanotechnology was quite expansive, and included materials science and plain chemistry, among other areas. As a result, the term “nanotechnology” was overhyped. It is necessary to distinguish between the overextended definition of nanotechnology promulgated by the NNI and the original meaning of the term as Eric Drexler used it, which refers to the robotic manipulation of individual atoms and the construction of atomically precise products, including self-replicating molecular assemblers. To clarify this kind of nanotechnology, sometimes the phrase “Drexlerian nanotechnology” or “molecular nanotechnology” is used.

**4. Erroneous idea that nanorobots will be weaker than bacteria,   
because bacteria had billions years to adapt**

Automobiles are not slower than cheetahs, even though cheetahs have had thousands of years to adapt. Swords are not weaker than claws. Just because something had thousands of years to adapt does not mean it will be stronger. Machines eventually surpass biology, given enough time.

5. Erroneous representation that nanorobots cannot breed in the environment

If bacteria can breed in nature, there is no reason why nanobots cannot do so. They can even borrow the same chemical reactions, or replicate as organic-inorganic hybrids. S.W. Wilson coined the term “animats” to refer to the possibility of such artificial animals, a contraction of animal-materials.

6. Erroneous representation that nanorobots in an environment   
will be easy to destroy by bombing

If nanorobots do become a threat, self-replicating in the environment *a la* grey goo, there will be too many of them to bomb. Even a “saturation bombing” does not literally saturate the ground with explosions. If bombs are used, they will disperse nanorobots on the fringes of the blast area, where they will continue to replicate. Rather than targeting nanorobots themselves with bombs, it might be wiser to create “firebreaks” by saturation bombing all the organic material in a given area, removing the nanorobots' energy source. However, if nanorobots can be dispersed on the wind, as they are likely to be, they will just float over such firebreaks. A more effectively means of countering grey goo is through friendly nanobots, “blue goo”.

7. Erroneous representation that nanorobots will  
be too small to contain control computers

Though few detailed designs for nanorobots yet exist, what designs do exist, such as for the primitive nanofactory designed by Chris Phoenix, set aside volume for computers9. Technically, the term “nanorobot” is a misnomer, since these machines are more likely to be several microns across—the size of cells—rather than tiny like viruses. This will give them sufficient volume to contain sophisticated computers for navigation and swarm behavior. The DNA in human cells contains about 500 MB of information, and is constantly being processed by ribosomes, so we have an existence proof for small devices with substantial computing capabilities.

8. Our propensity to expect grandiose results only from grandiose causes

In technology, seemingly simple tricks can give rise to extreme consequences. Radioactivity was originally just discovered as a certain kind of rock that left marks on photopaper, and eventually the same principles were used to create bombs that destroy cities and reactors that power them. Drexler illustrates this error from the following examples: Boring fact: some electric switches can turn each other on and off. These switches can be made very small and consume little electricity. Grandiose consequence: if enough of these switches are put together, they can create computers, the foundation of the information revolution. Boring fact: mold and bacteria compete for food, therefore some molds have evolved to produce toxins that kill bacteria. Grandiose consequence: penicillin saved millions of lives by fighting infection.

**9. Other objections to molecular nanotechnology**

Experts in the field of nanotechnology have posed many reasons why they think molecular nanotechnology will not work. These include quantum uncertainty, thermal vibrations, van der Waals forces, the “sticky fingers” criticism, spontaneous rearrangement, and so on. We cite various sources of both the objections and their rebuttals. First, the objections: the Wikipedia article on the Drexler-Smalley debate on molecular nanotechnology (which summarizes a long debate), an email exchange between Philip Moriarty and Chris Phoenix10, and “six challenges for molecular nanotechnology” by Richard Jones11. Then, the responses: a rebuttal to Smalley co-authored by Eric Drexler and many others12, a brief open letter by Drexler13, and a response by the Nanofactory Collaboration of scientists14. There have been many other discussions in many other venues, but these are the most prominent, authoritative, and widely cited. Reviewing them all will give the reader a good understanding of the most common objections and counterarguments, and provide a platform for further discussion.

**Citations**

1. Jurgen Altmann. *Military Nanotechnology: Potential Applications and Preventive Arms Control*. New York: Routledge. 2006.
2. Mark Gubrud. “Nanotechnology and International Security”. Draft paper for the Fifth Foresight Conference on Molecular Nanotechnology, 1997. <http://www.foresight.org/Conferences/MNT05/Papers/Gubrud/>
3. Steven D. Perrault, William M. Shih. “Virus-Inspired Membrane Encapsulation of DNA Nanostructures To Achieve In Vivo Stability”. ACS Nano, 2014; 140422010051005 DOI: 10.1021/nn5011914
4. Tim Liedl, Bjorn Hogberg, Jessica Tytell, Donald E. Ingber, William M. Shih. “**Self-assembly of 3D prestressed tensegrity structures from DNA”**. Nature Nanotechnology, 2010; DOI: [10.1038/nnano.2010.107](http://dx.doi.org/10.1038/nnano.2010.107#_blank)
5. Kyle Lund et al. “Molecular robots guided by prescriptive landscapes”. *Nature* 465, 206–210 (13 May 2010). <http://www.nature.com/nature/journal/v465/n7295/full/nature09012.html>
6. Ido Bachelet. “DNA robots work in a live cockroach”. *Nature* **508**,153 (10 April 2014). <http://www.nature.com/nature/journal/v508/n7495/full/508153e.html>
7. Eric Drexler. *Engines of Creation*. New York: Anchor Books, 1986. <http://e-drexler.com/p/06/00/EOC_Cover.html>
8. Chris Phoenix and Eric Drexler. “Safe exponential manufacturing”. Nanotechnology15 (2004) 869–872 PII: S0957-4484(04)78839-X <http://crnano.org/IOP%20-%20Safe%20Exp%20Mfg.pdf>
9. Chris Phoenix. “Design of a Primitive Nanofactory”. *Journal of Evolution and Technology* vol. 13, 2003.
10. Chris Phoenix and Philip Moriarty. “Is mechanosynthesis feasible? The debate continues”. *Soft Machines* blog. <http://www.softmachines.org/wordpress/?p=70>
11. Richard Jones. “Six challenges for molecular nanotechnology”. *Soft Machines* blog. <http://www.softmachines.org/wordpress/?p=175>
12. Drexler, K. Eric; Forrest, David; Freitas, Robert A.; Hall, J. Storrs; Jacobstein, Neil; McKendree, Tom; Merkle, Ralph; Peterson, Christine (2001). "On Physics, Fundamentals, and Nanorobots: A Rebuttal to Smalley’s Assertion that Self-Replicating Mechanical Nanorobots Are Simply Not Possible". Institute for Molecular Manufacturing. Retrieved 25 April 2010. <http://www.imm.org/publications/sciamdebate2/smalley/>
13. Eric Drexler. “[An Open Letter on Assemblers](http://www.foresight.org/nano/Letter.html)”. April 2003, Foresight Institute. <http://www.foresight.org/nano/Letter.html>
14. Robert A. Freitas Jr. and Ralph C. Merkle. “Remaining Technical Challenges for Achieving Positional Diamondoid Molecular Manufacturing and Diamondoid Nanofactories”. Nanofactory Collaboration. June 2006. <http://www.molecularassembler.com/Nanofactory/Challenges.htm>

## Chapter 7. Conclusions from the analysis of cognitive biases in the estimate of global risks and possible rules for rather effective estimate of global risks

Need for open discussion

The scale influence of errors on reasoning on global risks can be estimated by comparing opinions of different experts, scientists and politicians of possibility of definitive global catastrophe and its possible reasons. It is easy to be convinced, that the disorder of opinions is huge. Some consider total risk insignificant, others are confident in inevitability of human extinction. As the possible reasons the set of different technologies and scenarios is called, and different experts offer the sets of possible scenarios and sets of impossible scenarios.

It is obvious, that roots of such disorder of opinions - in a variety of movements of thought which, in absence of any visible reference point, appears it is subject to various biases. As we cannot find a reference point concerning global risks in experiment, it is represented desirable that open discussion about methodology of research of global risks on which basis the uniform and conventional picture of global risks could be generated, became such reference point.

Important maintenance of open discussion by all kinds of risks. It means consideration of any objection as true sufficient time, that it to estimate before deciding it to reject. Not to reject any objections to a descent and to support presence of opponents.

Precaution principle

It means preparation for the worst realistic scenario in all situations of uncertainty. Realistic it is necessary to consider any scenario which does not contradict known laws of physics and has precisely measurable probability above some threshold level. It corresponds to the principle of a conservative engineering estimate. However precaution should not have irrational character, that is should not exaggerate a situation. One of formulations of a principle of precaution sounds so: «the precaution principle is a moral and political principle which asserts, that if a certain action or the politician can cause a severe or irreversible damage to a society, that, in absence of the scientific consent that harm will not happened, weight of the proof lays on those who offers the given actions».

Doubt principle

The principle of doubt demands to suppose possibility of an inaccuracy of any idea. However the doubt should not lead to instability of the course of thought, blind trust to authorities, absence of the opinion and uncertainty in it if it is proved enough.

Introspection

The continuous analysis of own conclusions about possible errors from all list.

Independent repeated calculations

Here enters independent calculation by different people, and also comparison of direct and indirect estimates.

An indirect estimate of degree of an error

We can estimate degree of underestimate of global catastrophe, studying that, how much people underestimate similar risks - that is risks of unique catastrophes. For example, spaceships Shuttle have been calculated on one failure more than on 1000 flights, but the first failure has occurred on 25th flight. That is the initial estimate 1 to 25 would be more exact. Nuclear stations were under construction counting upon one failure in one million years, but Chernobyl failure has occurred approximately after less than 10 000 stations-years of operation (this number turns out from multiplication of number of stations by that moment for average term of their operation, and demands specification). So, in the first case real stability has appeared in 40 times worse, than the design estimate, and in the second - in 100 times is worse. From here we can draw a conclusion, that in case of unique difficult objects people underestimate their risks in tens times.