Risk Analysis, Vol. 30, No. 10, 2010 DOI: 10.1111/j.1539-6924.2010.01460.x

Anthropic Shadow: Observation Selection Effects and Human Extinction Risks

Milan M. Ćirković,[1]* Anders Sandberg,[2] and Nick Bostrom[2]

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

KEY WORDS: Anthropic principle; astrobiology; existential risks; global catastrophes; impact hazard; natural hazards; risk management; selection effects; vacuum phase transition

1. INTRODUCTION: EXISTENTIAL RISKS AND OBSERVATION SELECTION EFFECTS

Humanity faces a series of major global threats, both in the near- and in the long-term future. These are of theoretical interest to anyone who is concerned about the future of our species, but they are also of direct relevance to many practical and policy decisions we make today. General awareness of the possibility of global catastrophic events has risen recently, thanks to discoveries in geochemistry, human evolution, astrophysics, and molecular biology.[1–6] In this study, we concentrate on the subset of catastrophes called *existential risks* (ERs): risks where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential.[7] Examples of potential ERs include global nuclear war, collision of Earth with a 10-km sized (or larger) asteroidal or cometary body, intentional or accidental misuse of bio- or nanotechnologies, or runaway global warming.

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

Our goal in this article is to study a specific observation selection effect that influences estimation of some ER probabilities, threatening to induce an anthropic bias into the risk analysis.[3] Anthropic bias

[1] Astronomical Observatory of Belgrade, Volgina, Belgrade, Serbia.
[2] Future of Humanity Institute, Faculty of Philosophy & James Martin 21st Century School, Oxford University, London, UK.
*Address correspondence to M. M. Ćirković, Astronomical Observatory of Belgrade, Volgina 7, 11160 Belgrade-74, Serbia; tel: +381-11-3089079; fax: +381-11-2419553; mcirkovic@aob.rs.

[3] For a summary of the vast literature on anthropic principles and anthropic reasoning in general, see Barrow and Tipler; Balashov; and Bostrom.[56–58]

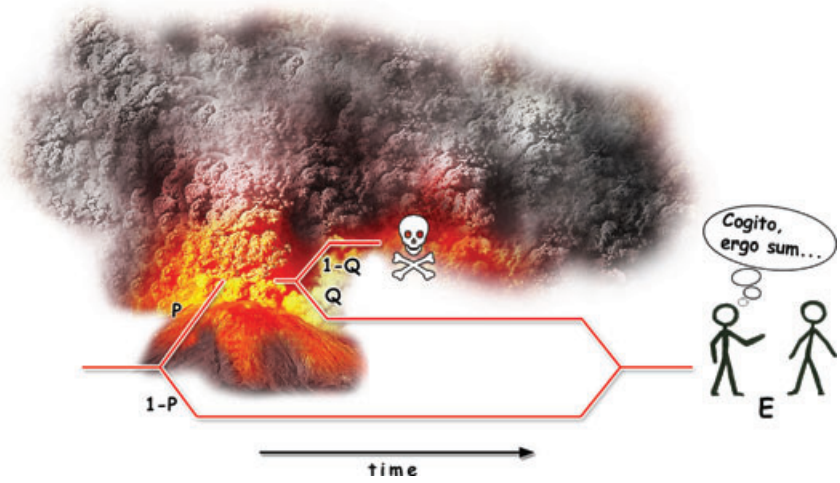

Fig. 1. A schematic representation of our single-event toy model. P is the *a priori* probability of a global catastrophe; Q is the probability of human survival given the catastrophe; E is the fact of our present-day existence.

can be understood as a form of sampling bias, in which the sample of observed events is not representative of the universe of all events, but only representative of the set of events compatible with the existence of suitably positioned observers. We show that some ER probabilities derived from past records are unreliable due to the presence of observation selection effects. Anthropic bias, we maintain, can lead to underestimation of the probability of a range of catastrophic events.

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

2. A TOY MODEL OF ANTHROPIC BIAS[4]

The basis of our approach is Bayes's formula for conditional probability:

$$P(B_i \mid E) = \frac{P(B_i)\,P(E \mid B_i)}{\sum_{j=1}^{n} P(B_j)\,P(E \mid B_j)}, \qquad (1)$$

where $P(B_i)$ is prior probability of hypothesis B_i being true, and $P(B_i|E)$ is the conditional probability of hypothesis B_i being true, given evidence E. The evidence we will consider is our existence as intelligent observers in the present epoch. Our existence entails a host of biological, chemical, and physical preconditions. In particular, our existence implies that the evolutionary chain of terrestrial evolution leading to our emergence was not broken by a terminally catastrophic event. We shall discuss some of the ambiguities related to this condition below. The hypotheses $B_1, B_2, \ldots, B_n$ that are of interest to us here are those dealing with the occurrence or nonoccurrence of a particular type of global catastrophic event in a given interval of time. For example, one hypothesis may be "There were at least five impacts of asteroids or comets of 10–20 km size during the last 10^8 years"; or "There was no supernova explosion closer than 10 parsecs from the Sun between 2×10^7 and 5×10^6 years before present (henceforth B.P.)."

Consider the simplest case of a single very destructive global catastrophe, such as a Toba-like supervolcanic eruption.[9] The evidence that we wish to conditionalize upon in a Bayesian manner is the fact of our existence at the present epoch. We can schematically represent the situation as in Fig. 1: the prior probability of catastrophe is P and the probability of human survival following the catastrophic event is Q. We shall suppose that the two probabilities are: (1) constant, (2) adequately normalized, and (3) apply to a particular interval of past time. Event B_2 is the occurrence of the catastrophe, event B_1 is the nonoccurrence of the catastrophe, and by E we denote the evidence of our present existence.

The direct application of Bayes's formula in the form:

$$P(B_2 \mid E) = \frac{P(B_2)\,P(E \mid B_2)}{P(B_1)\,P(E \mid B_1) + P(B_2)\,P(E \mid B_2)}, \qquad (2)$$

[4] Some earlier findings related to this section were presented in Ćirković.[59]

yields the posterior probability as:

$$P(B_2 \mid E) = \frac{PQ}{(1-P)\cdot 1 + PQ} = \frac{PQ}{1-P+PQ}\,. \tag{3}$$

We can define an *overconfidence parameter* as:

$$\eta \equiv \frac{P\,(\text{a priori})}{P\,(\text{a posteriori})}, \tag{4}$$

which in this special case becomes:

$$\eta = \frac{P}{P\,(B_2 \mid E)} = \frac{1-P+PQ}{Q}\,. \tag{5}$$

As η moves beyond 1, our inferences from the past become increasingly unreliable, and we underestimate the probabilities of future catastrophes. For instance, suppose $Q = 0.1$ and $P = 0.5$, corresponding to a fair-coin-toss chance that a Toba-scale event occurs once per 1 million (10^6) years (Myr) of human evolution, and that the probability of human survival following such an event is 0.1. The resulting value of the overconfidence parameter is $\eta = 5.5$, indicating that the actual probability of such an event is 5.5 times our initial estimate. Values of overconfidence as a function of severity (as measured by the extinction probability $1 - Q$) are shown in Fig. 2.

Note that

$$\lim_{Q\to 0} \eta = \infty\,. \tag{6}$$

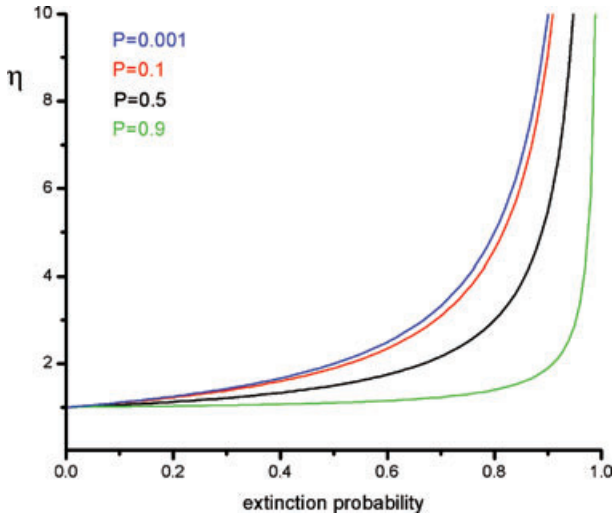

Fig. 2. Overconfidence parameter as a function of the extinction probability 1−Q in our single-event toy model. Different values of the real event probability P are color-coded (colors visible in online version). We notice that the overconfidence bias is strongest for low-probability events.

Overconfidence becomes very large for very destructive events. As a consequence, we should have no confidence in historically based probability estimates for events that would certainly extinguish humanity ($Q = 0$). While this conclusion may seem obvious, it is not widely appreciated. For instance, as we discuss below, a well-known argument of Hut and Rees dealing with the hypothetical risk to the stability of quantum vacuum due to the high-energy physics experiments is partially misleading because it fails to take into account the anthropic bias.[10]

The same reasoning applies to those extremely rare, but still definitely possible, physical disasters like various strange astronomical occurrences leading to the Earth becoming an unbound planet due to close passage of a normal star (see e.g., Laughlin and Adams for estimates how probable it is in the remaining lifetime of the solar system[11]), or even more exotic objects, like a neutron star or a black hole. The conclusion that the irreversible destruction of Earth in an encounter of the solar system with a passing star or a black hole is extremely improbable cannot be obtained *solely* from the inference from the past history of our planetary system. In this case, however, admission of additional information, based on our understanding of the solar neighborhood in the Milky Way and the mass function of stellar objects, for instance, could render the conclusion that we are safe from this particular risk rather bias-free and persuasive. On the other hand, the amount of additional admissible information is highly uneven when we are dealing with a wide spectrum of possible global hazards.

3. GENERALIZING THE MODEL

How to generalize this to a series of possible catastrophic events? We shall briefly sketch one possible approach here. We face a situation like the one shown in Fig. 3.

Let α be the inherent probability of a disaster, β the probability that it is lethal (in a sufficiently generalized sense, which we shall discuss in some detail in Section 5 below), and N the number of possible disasters that could occur. Let O be the existence of an observer (i.e., no lethal disasters) and k be the number of disasters observed. As far as both N and α are small,[5] the probability for an observer to see

[5] This assumption is convenient as a working hypothesis—but when we consider interpretation of our results (Section 5) for the real hazards, we shall find some broad physical

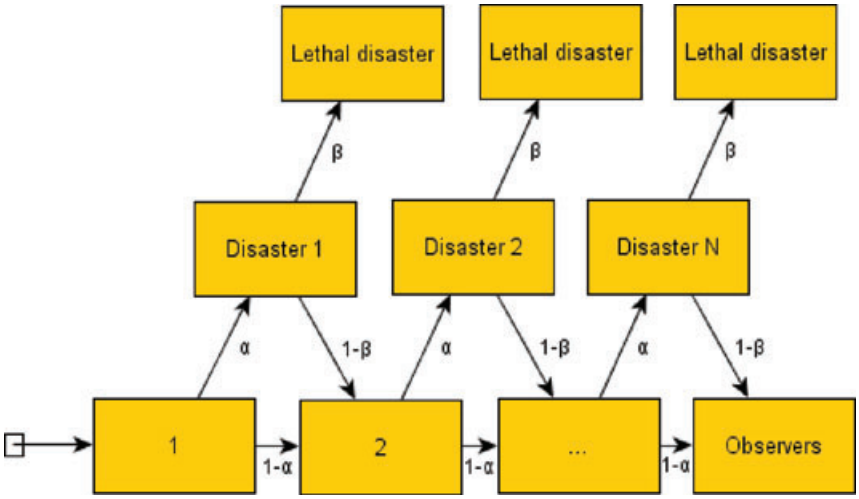

Fig. 3. A series of possibly lethal disasters in the observers' past—a generalization of the situation presented in Fig. 1.

k disasters in his past is:

$$P(k, O \mid \alpha, \beta) = \binom{N}{k} \alpha^k (1-\alpha)^{N-k} (1-\beta)^k . \tag{7}$$

Under the assumption of uniform prior distribution of the parameters, $P(\alpha, \beta) = 1$, it is possible to calculate the probability $P(O, k)$:

$$P(O,k) = \int_0^1 \int_0^1 P(\alpha,\beta) \binom{N}{k} \alpha^k (1-\alpha)^{N-k} (1-\beta)^k \times d\alpha\, d\beta = \frac{1}{(1+k)(1+N)}, \tag{8}$$

entailing the general formula:

$$P(\alpha, \beta \mid O, k) = \frac{1}{(1+k)(1+N)} \binom{N}{k} \alpha^k (1-\alpha)^{N-k} (1-\beta)^k . \tag{9}$$

Consequently, the probability of existence of an observer for a given pair of values α, β is given as:

$$P(O \mid \alpha, \beta) = \sum_{k=0}^{N} P(O, k \mid \alpha, \beta) = \sum_{k=0}^{N} \binom{N}{k} \alpha^k (1-\alpha)^{N-k} (1-\beta)^k . \tag{10}$$

justifications for it. One could argue that a small value of N is almost a prerequisite for speaking about global, extinction-level events. There are also important issues as to what degree it is justifiable to talk about temporal "slots" for catastrophic occurrences to be resolved in the course of the future work.

In a world ensemble, this would translate into the density of observers. We can think about this situation as involving a set of Earth-like planets with well-defined ages, having biospheres, but subject to different quantitative and qualitative environmental hazards.[12] For example, in case of $N = 4$, Equation (10) gives the probability of survival as shown in Fig. 4. For $k = 0$, we have no information about the danger of disasters, so the probability distribution is constant along the β axis. For higher values of k, the probability mass for high β values decreases, since disasters are becoming common enough that they cannot be extremely severe. For a special case of this example, $N = 4$, $k = 2$, the distribution of probabilities of observing particular values of (α, β) is shown in Fig. 5. Cases with a higher N look similar.

An obvious next step in this direction is to implement a simulation model, creating a large number of worlds for each α, β and running N trials where they could suffer disasters. Detailed results of simulations done on this and related classes of toy models will be reported in a forthcoming study. It is already clear, however, that the distributions of the parameters among the survivors will be strongly biased. Considering that we have already amassed important knowledge on empirical and semi-empirical probability distributions for particular classes of risks and the enormous practical importance of searching for any kind of bias in risk analysis,[13] there is clearly a lot of room for integration of the existing knowledge in the analysis of the anthropic bias, once we clarify which specific processes this bias applies to.

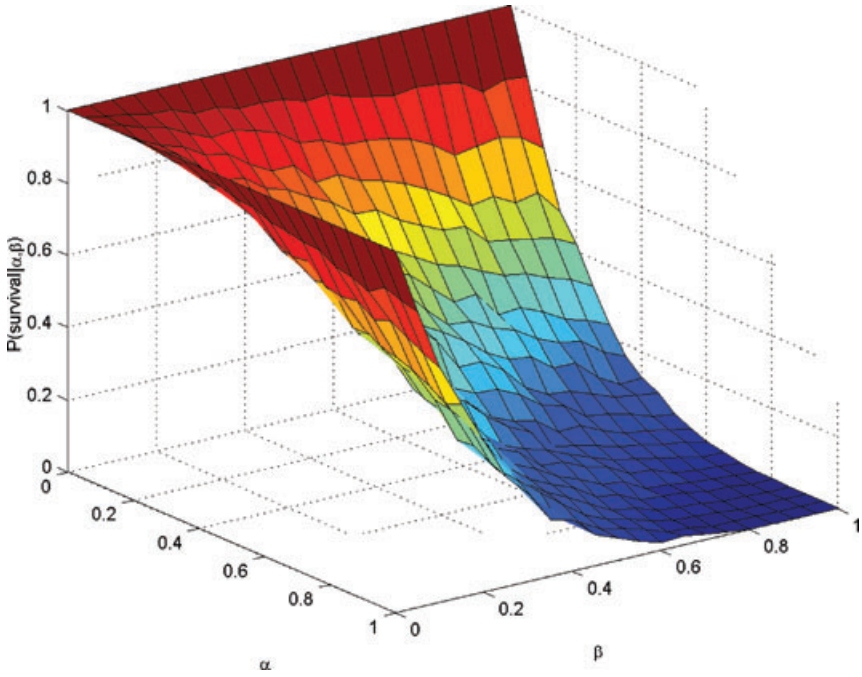

Fig. 4. The probability of observers $P(O \mid \alpha, \beta)$ for the $N = 4$ toy model as a function of the *a priori* probability of a global catastrophe α and the extinction probability β. For $\alpha = \beta = 0$, the world is safe and the density is maximal; as higher values of either parameter are approached, the number of observers decline.

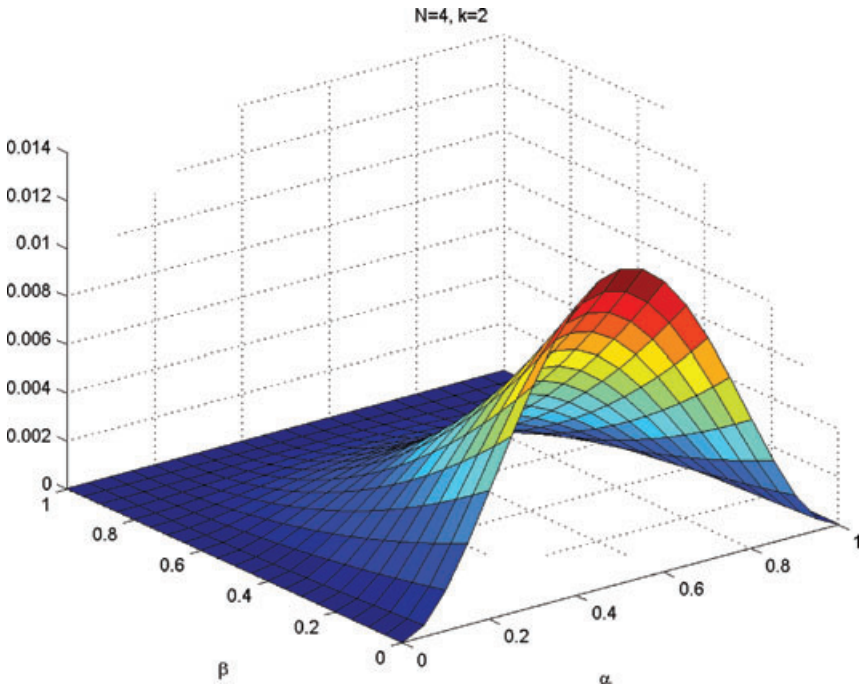

Fig. 5. Probability $P(\alpha, \beta \mid O, k)$ for the $N = 4,\ k = 2$ toy model.

4. ANTHROPIC BIAS: UNDERESTIMATING NATURAL RISKS?

Traditionally, in the analysis of natural hazards, scientists construct an empirical distribution function from evidence of past events, such as geological evidence of past extraterrestrial impacts, or supernova/γ-burst explosions, or supervolcanic eruptions. In the Bayesian approach, we can dub this distribution function the *a posteriori* distribution function.

In forecasting future events, we are interested in the "real" distribution of chances of events (or their causes), which is "given by Nature" and is not

Table I. Examples of Natural Hazards Potentially Comprising Existential Risks and Their Two Types of Distribution Functions; Only the *a priori* Distribution Veritably Describes Nature and Can Serve as a Source of Predictions About Future Events

Type of Event	*A Priori* Distribution	Empirical (*A Posteriori*) Distribution
Impacts	Distribution of near-Earth objects and Earth-crossing comets	Distribution of impact craters, shock glasses, etc.
Supervolcanism	Distribution of geophysical "hot spots" producing supereruptions	Distribution of calderas, volcanic ash, ice cores, etc.
Supernovae and/or γ-ray bursts (see Appendix)	Distribution of progenitors and their motions in the solar neighborhood	Geochemical trace anomalies, distribution of stellar remnants

necessarily revealed in their *a posteriori* distribution observed in or inferred from the historical record. The underlying objective characteristic of a system is its *a priori* distribution function. In predicting future events, the *a priori* distribution is crucial, since it is not skewed by selection effects. The relationship between *a priori* and *a posteriori* distribution functions for some natural catastrophic hazards is shown in a simplified manner in Table I. Only the *a priori* distribution veritably describes nature and can serve as a source of predictions about future events. A sketch of inference from the past to the future including these two distributions is shown in Fig. 6.

Catastrophic events exceeding some threshold severity eliminate all observers and all ecological conditions necessary for subsequent emergence of observers, and are hence unobservable. Some types of catastrophes may also make the existence of observers on a planet impossible in some subsequent interval, the size of which might be correlated with the magnitude of the catastrophe.[6] Because of this anthropic bias, the events reflected in our historical record are not sampled from the full events space but rather from the part of the events space that lies beneath the "anthropic compatibility boundary" (illustrated in Fig. 7). The part of the parameter space above the boundary lies in what can be called *anthropic shadow*: the observation selection effect implicit in conditioning on our present existence prevents us from sharply discerning magnitudes of extreme risks close (in both temporal and evolutionary terms) to us. This shadow is the source of bias, which must be corrected when we seek to infer the objective chance distribution from the observed empirical distribution of events.

Anthropic shadow is cumulative with the "classical" selection effects applicable to any sort of event (e.g., removal of traces of old events by erosion or other instances of natural entropy increase). Even after these classical selection effects have been corrected in constructing an empirical (*a posteriori*) distribution, anthropic bias must also be corrected in order to derive the correct *a priori* distribution function.

Of course, the scheme in Fig. 7 is a simplification. The anthropic compatibility boundary need not be a straight line. But the general diagonal direction in the severity-time diagram is preserved. We see a possible illustration of this effect in the empirical data on terrestrial impact cratering in Fig. 8. For the data on impact structures, we use the 2010 Earth Impact Database.[14] Although the ages of many craters are poorly known, the trend similar to the one in Fig. 7 is visible. For example, it is obvious that we cannot ever discover traces of a 100 km impactor having hit Earth during the last million years (or, indeed, at any time in the Phanerozoic eon; see Appendix). Does this mean that such events have only a vanishing probability? No, it means instead that such events lie in the censored region from which the empirical record cannot sample. Any straightforward extension of the empirical distribution function into this region will be artificially suppressed in comparison to the objective chance distribution of possible impactor size. In other words, giant impactors may exist and be a

[6] As an illustration, some authors have, perhaps half-jokingly, suggested that some species of dinosaurs could have evolved intelligence prior to their extinction in 65 Myr B.P.[60,61] Without considering the merit of this speculation, we can state that in such an imagined situation, the Chixhulub impact (if it was indeed the physical causative agent of the end-Cretaceous mass extinction) did not only eliminate all observers present at that epoch, but was also likely to make the planet unsuited for evolution of observers at, say, 63 Myr B.P. It—obviously—did not prevent the evolution of observers at circa. 1 Myr B.P. The issue of *recovery* from mass extinctions has been recognized as one of the least understood in paleobiology and evolutionary biology; preliminary results indicate that the recovery timescales are long, measured in tens of Myr.[62,63]

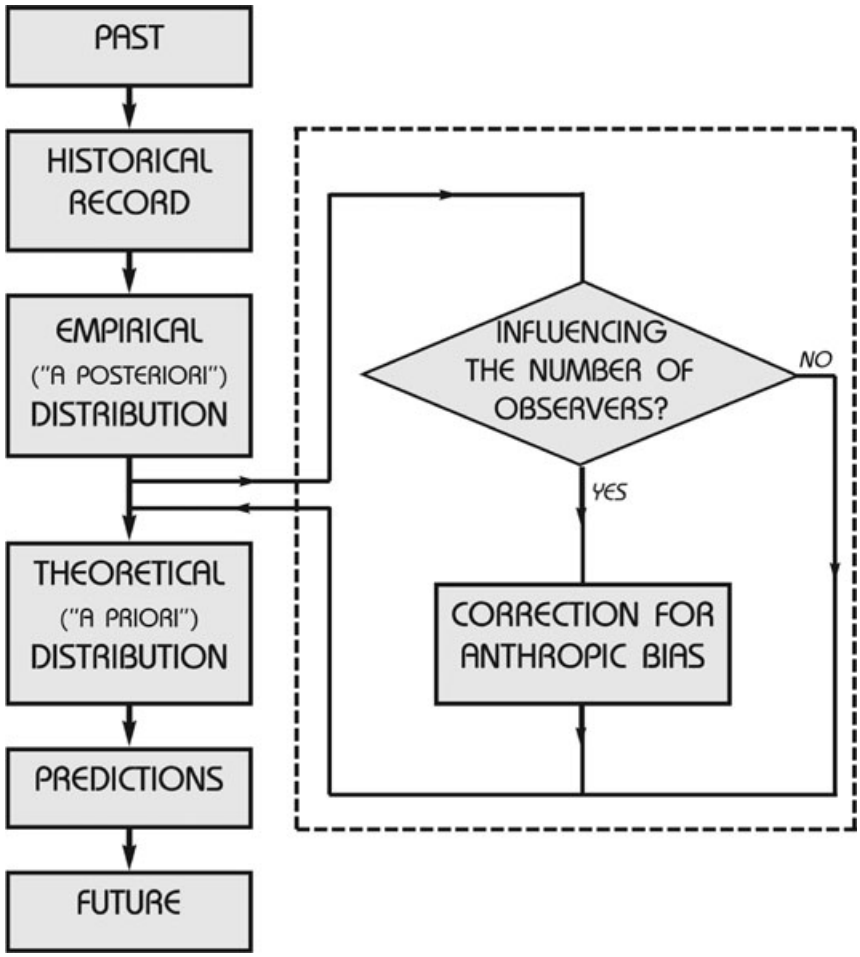

Fig. 6. A sketch of the common procedure for deriving predictions about the future from the past records. This applies to benign events as well as to existential risks (ERs), but only in the latter case do we need to apply the correction symbolically shown in dashed-line box. Steps framed by the dashed line are usually *not* performed in the standard risk analysis; they are, however, necessary in order to obtain unbiased estimates of the magnitude of natural ERs.

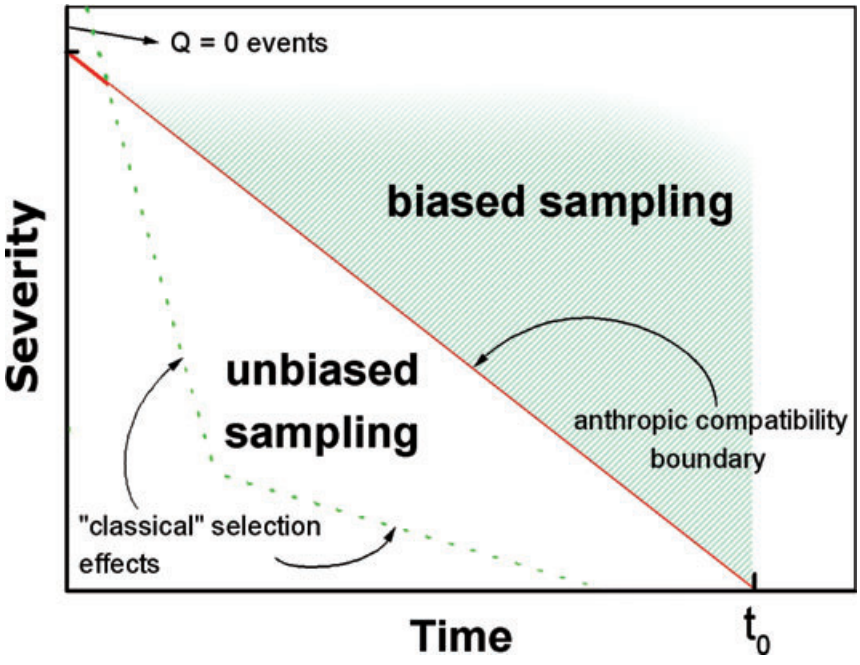

Fig. 7. A sketch of the anthropic bias: we do not fairly sample the entire time-severity plane, only a region compatible with our existence at this particular epoch (the rest is in the "anthropic shadow"—shaded region, see text). The current epoch is denoted by t_0 and we count time from the formation of our planet.

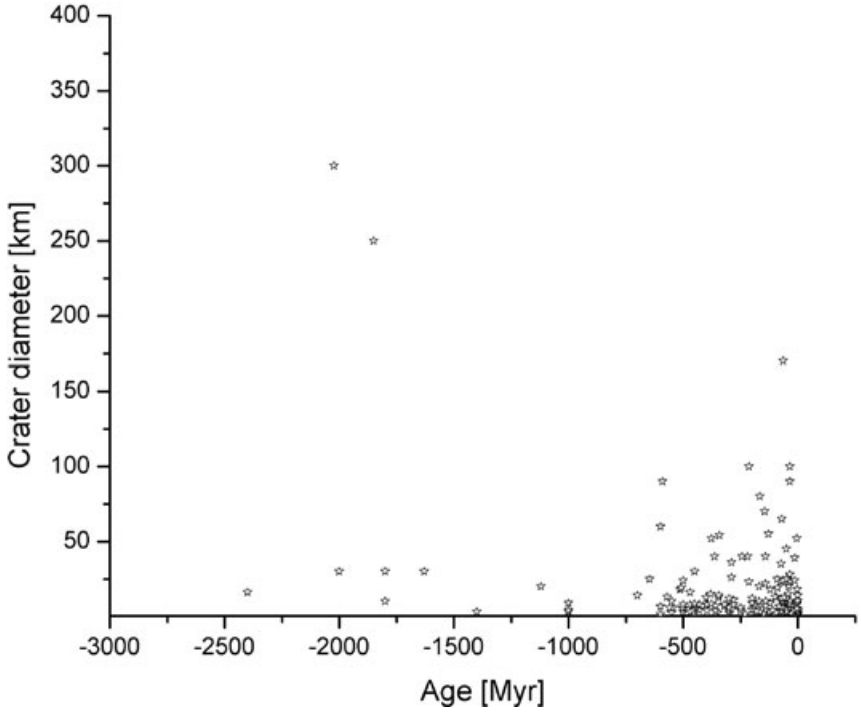

Fig. 8. The diagram showing the size of known impact craters as a function of their age according to the Earth Impact Database. The absence of points in the upper right area of the diagram is visible; the only clear outlier corresponds to the Chixhulub crater at 65 Myr B.P., a confirmed instance of global catastrophe.

significant threat for the *future*, but they leave no traces in the recent *past* of observers.[7] The past-future symmetry is broken by the anthropic shadow.

5. WHICH ERs ARE SUBJECT TO ANTHROPIC SHADOW?

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

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

Various hazards can be distinguished by the degree to which they satisfy these criteria. For instance, the asteroidal and cometary impact history of the solar system is, in theory, easier to obtain for the Moon, where the erosion is orders of magnitude weaker than on Earth.[8] In practice, this is still not feasible for obtaining the fair sampling of the impactors because: (1) precise dating of a large set of lunar craters is beyond our present capacities[9] and (2) most of the large lunar craters are known to originate in a highly special epoch of the so-called Late Heavy Bombardment,[15,16] ca. 4.0–3.8 billion years B.P., thus strongly skewing any attempt to plot the empirical distribution function of impacts for

[7] Of course, this is a highly simplified example. In practice, direct observational searches for NEOs are today more important for estimating the present and future impact hazards than the counting and dating of craters and other purely geophysical traces.[64–66] There is still a faint possibility, however, that a population of large dark impactors exists virtually unnoticed and detectable only through its past traces, and the data on terrestrial cratering rates still help to discriminate among rival hypotheses.[67,18,19]

[8] Though not completely nonexistent, as is often claimed in the popular literature; micrometeorites, as well as cosmic-ray bombardment necessarily cause some loss of information.

[9] This, of course, applies even more forcefully to other bodies in the solar system with a discernible cratering record, for example, Mars.

"normal" times. In practice, in the current debates about the rates of cometary and asteroidal impacts, it is the terrestrial cratering rates that are used as an argument for or against the existence of a dark impactor population,[17–21] thus offering a good case on which the anthropic model bias can, at least potentially, be tested.[10] The amount of bias of the cratering record, in principle, can be decreased through extrapolation from the smaller sizes and comparing such extrapolation with the size-frequency distribution on other solar system bodies, which could be obtained without the need for technically unfeasible measurements of the age of craters. In practice, however, not only is it unclear where the extrapolation should start—since we know little about contingencies of biological evolution leading to the emergence of observers—but the size-frequency distribution expresses only *temporal averages* of the relevant relationships (between velocities, angles, sizes, and consistencies of impactors vs. crater size). The loss of information in averaging is important if the impactor population may significantly vary in time.

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

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

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

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

6. ANTHROPIC SHADOW AND RISKS FROM PHYSICS DISASTERS

An example *par excellence* of a $Q = 0$ event is a vacuum phase transition or a comparable quantum field collapse. Such an event would not only extinguish humanity but also completely and permanently destroy the terrestrial biosphere. Coleman and De Luccia first mentioned the possibility that such a disaster might be caused by the operation of high-energy particle colliders used in physics research.[40] This possibility has since been widely discussed,[10,41–46] and has motivated objections to the operation of high-energy particle colliders, including most recently the Large Hadron Collider.[46,47]

Three specific threats are relevant: (1) triggering vacuum phase transition through creation of an expanding bubble of "new" vacuum state, (2) accidental production of charged strangelets, which could transform all Earth's mass into strange matter, and (3) accidental production of a mini black hole falling into Earth's center and subsequently destroying our planet. Although smacking of science fiction, this idea has been seriously considered even by high-level

[10] In addition to the impact craters, there is a host of other traces one attempts to find in the field work, which contribute to the building of the empirical distribution function of impacts—notably, chemical anomalies or shocked glasses.[68]

administrators of modern particle-accelerator laboratories.[48] This is not only an eschatological issue for humanity: a vacuum phase transition would also destroy the habitability of the universe for any other observers in our future light cone. Even if the chance of such a disaster is remote, its catastrophic impact would be so enormous that it deserves close scrutiny.

Hut and Rees, in an important pioneering study of the problem of high-energy physics risks, suggested that concerns about particle colliders can be reasonably dismissed because high-energy particle collisions occurring in nature, such as those between cosmic-rays and the Earth's atmosphere or the solid mass of the Moon, are still orders of magnitude higher than those achievable in human laboratories in the near future.[10] With plausible general assumptions on the scaling of the relevant reaction cross-sections with energy, Hut and Rees concluded that the fact that the Earth (and the Moon) have survived cosmic-ray bombardment for about 4.5 Gyr implies that we are safe for the foreseeable future. For example, if the probability of a high-energy physics disaster in nature is 10^{-50} per year, then a doubling or even 10-fold increase of the risk through deliberate human activities is arguably trivial.

The Hut-Rees argument should provide us no comfort, however, as it fails to correct for anthropic bias. A vacuum phase transition is an event for which $Q = 0$. Probability estimates based on observations of the Earth's and Moon's existence are thus completely unreliable. Moreover, the unreliability of these estimates applies to both naturally occurring and human-induced vacuum phase transitions. (Hut and Rees also conclude, completely justifiably, that the number of potentially risky events in any conceivable human accelerator is much smaller than in the cosmic-ray interactions in nature.) Unfortunately, the same error is repeated in the recent Large Hadron Collider (LHC) safety study, where the duration of the solar system thus far is invoked as part of the arguments for accelerator safety.[46]

Tegmark and Bostrom manage to circumvent the observation selection effect by using data on the planetary age distribution and the relatively late formation date of Earth[12] to infer the *a priori* distribution of events that destroy or permanently sterilize a planet.[49] Based on their results, the rate of vacuum phase transitions within the volume of the Milky Way is less than 10^{-9} per year. This shows that awareness of anthropic shadow effects can enable more reliable estimation of catastrophic risk.

7. CONCLUSIONS

Smolin, among others, has claimed that the anthropic principle lacks predictive power and practical importance.[50] By contrast, our results suggest that correcting for the anthropic shadow bias can significantly affect probability estimates for catastrophic events, such as supervolcanic eruptions or asteroidal impacts. Moreover, recognizing this bias can help us to avoid pitfalls and errors in risk analysis, such as those in Hut-Rees's argument or the LHC Safety Assessment Group (SAG) study for the safety of particle colliders. The main lesson, therefore, lies in the direction of greater caution we need to exercise in facing the spectrum of global catastrophic risk and ERs. The dearth of research on biases in ERs is lamentable in light of both the natural hazards considered here and the more probable anthropogenic hazards we face with the advent of powerful technologies. It is hardly necessary to emphasize that improvements in the quantitative risk assessment are likely to lead to improved policies of risk mitigation and management.[6]

Further research on shape of the anthropic shadow and the magnitude of the resulting anthropic bias is needed, especially related to the changing of survival probability with time, superposition of various ER mechanisms, and the secular evolution of the *a priori* distribution function itself. Except for $Q = 0$ events like a vacuum phase transition, accurate corrections for anthropic bias will require more complex and realistic models. Catastrophic events of varying magnitude can influence the evolutionary chain leading to our emergence as observers at many points. Charting such influences is a difficult challenge, since the evolutionary impact of even a single large (but not sterilizing) catastrophe remains controversial even for relatively well-established cases, like the Chixhulub impact, in light of the ubiquitous biological contingency.[51–54] For distinct states of evolutionary development separated by stochastic catastrophes, some quite complex modeling formalism, perhaps using probabilistic cellular automata, might be needed to fully capture all the factors that can influence the magnitude of the bias.[55]

APPENDIX: GLOSSARY

ER—existential risks, a subset of global catastrophic risks where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential.[7]

GRB—Gamma-ray (or γ-ray) bursts, flashes of gamma-rays, lasting typically a few seconds, associated with the most energetic class of cosmic explosions ever detected. All detected GRBs have originated from outside the Milky Way galaxy, although a related class of phenomena, soft gamma repeater flares, are associated with magnetized neutron stars within our galaxy. It has been hypothesized that a gamma-ray burst in the Milky Way could cause a mass extinction on Earth.[30]

LHC—Large Hadron Collider, the world's largest and highest-energy particle accelerator, located in a tunnel 27 kilometers in circumference, and up to 175 meters beneath the Franco-Swiss border near Geneva, Switzerland. LHC was built by the European Organization for Nuclear Research and became operational in late 2009.

Myr—million (10^6) years, the most important unit of geological and evolutionary "deep time."

NEO—Near-Earth object, a solar system object, typically asteroid or comet whose orbit brings it in the vicinity of Earth, thus potentially presenting terrestrial impact hazard. (Very small objects, with sizes <50 meters, belonging to this category are often called meteoroids, and even some objects of anthropogenic origin, such as Sun-orbiting spacecraft, are classified as such.)

pc—parsec (from "parallactic second"), the main unit of length used in astronomy and related sciences. 1 pc = 3.085668×10^{16} meters = 3.262 light years. Stars in the vicinity of the solar system are typically ~ 1 pc apart.

SN—supernova (plural SNe, supernovae), terminal explosion of either a massive star (larger than about nine solar masses), or a white dwarf star in close binary system.

Phanerozoic (eon)—the current eon in the geological timescale, characterized by the existence of abundant plant and animal fossil record. It is usually taken as starting with the beginning of the Cambrian epoch (roughly 545 Myr B.P.).

ACKNOWLEDGMENTS

Three anonymous reviewers for *Risk Analysis* are hereby acknowledged for their thoughtful comments and pertinent criticisms of a previous version of this article. Our foremost thanks go to Gaverick Jason Matheny for his comments on an earlier version, and to Rebecca Roache whose close reading led to significant improvements of the article. We are also grateful for helpful discussions with Jelena Andrejić, Seth Baum, Fred C. Adams, Tatjana Jakšić, Cosma R. Shalizi, Bojana Pavlović, Bill Napier, and Zoran Knežević. We thank Richard B. Cathcart, Aleksandar Zorkić, Maja Bulatović, Dušan Inić, Srdjan Samurović, Branislav K. Nikolić, Samir Salim, Nikola Milutinović, and the KoBSON consortium of libraries for their kind technical assistance. One of the authors (M.M.Ć.) has been partially supported by the Ministry of Science and Technological Development of the Republic of Serbia through Grant ON146012, and thanks the Future of Humanity Institute at Oxford University for its kind hospitality during his work on this project.

REFERENCES

1. Leslie J. The End of the World: The Ethics and Science of Human Extinction. London: Routledge, 1996.
2. Huggett R. Catastrophism. London: Verso, 1997.
3. McGuire B. A Guide to the End of the World: Everything You Never Wanted to Know. Oxford: Oxford University Press, 2002.
4. Rees MJ. Our Final Hour. New York: Basic Books, 2003.
5. Palmer T. Perilous Planet Earth: Catastrophes and Catastrophism Through the Ages. Cambridge: Cambridge University Press, 2003.
6. Bostrom N, Ćirković MM (eds). Global Catastrophic Risks. Oxford: Oxford University Press, 2008.
7. Bostrom N. Existential risks. Journal of Evolution and Technology, 9, 2002 (http://www.jetpress.org/volume9/risks.html).
8. Bostrom N. Unpublished data, 2010.
9. Rampino MR, Self S. Volcanic winter and accelerated glaciation following the Toba super-eruption. Nature, 1992; 359:50–52.
10. Hut P, Rees MJ. How stable is our vacuum? Nature, 1983; 302:508–509.
11. Laughlin G, Adams FC. The frozen earth: Binary scattering events and the fate of the solar system. Icarus, 2000; 145: 614–627.
12. Lineweaver CH. An estimate of the age distribution of terrestrial planets in the universe: Quantifying metallicity as a selection effect. Icarus, 2001; 151: 307–313.
13. Woo G. The Mathematics of Natural Catastrophes. London: Imperial College Press, 1999.
14. Earth Impact Database, 2010. Available at: http://www.unb.ca/passc/ImpactDatabase/.
15. Kring DA, Cohen BA. Cataclysmic bombardment throughout the inner solar system 3.9–4.0 Ga. Journal of Geophysical Research—Planets, 2002; 107: 4–10.
16. Gomes R, Levison HF, Tsiganis K, Morbidelli A. Origin of the cataclysmic late heavy bombardment period of the terrestrial planets. Nature, 2005; 435, 466–469.
17. Nurmi P, Valtonen MJ, Zheng JQ. Periodic variation of Oort Cloud flux and cometary impacts on the Earth and Jupiter. Monthly Notices of the Royal Astronomical Society, 2001; 327: 1367–1376.
18. Napier WM, Wickramasinghe JT, Wickramasinghe NC. Extreme albedo comets and the impact hazard. Monthly Notices of the Royal Astronomical Society, 2004; 355:191–195.
19. Napier WM. Evidence for cometary bombardment episodes. Monthly Notices of the Royal Astronomical Society, 2006; 366: 977–982.
20. Fernández JA, Morbidelli A. The population of faint Jupiter family comets near the Earth. Icarus, 2006; 185: 211–222.

21. Baillie M. The case for significant numbers of extraterrestrial impacts through the late Holocene. Journal of Quaternary Science, 2007; 22:101–109.
22. Schindewolf O. Neokatastrophismus? Deutsch Geologische Gesellschaft Zeitschrift Jahrgang, 1962; 114: 430–445.
23. Ruderman MA. Possible consequences of nearby supernova explosions for atmospheric ozone and terrestrial life. Science, 1974; 184: 1079–1081.
24. Hunt GE. Possible climatic and biological impact of nearby supernovae. Nature, 1978; 271: 430–431.
25. Brakenridge GR. Terrestrial paleoenvironmental effects of a late quaternary-age supernova. Icarus, 1981; 46: 81–93.
26. Thorsett SE. Terrestrial implications of cosmological gamma-ray burst models. Astrophysical Journal, 1995; 444: L53–L55.
27. Annis J. An astrophysical explanation for the great silence. Journal of the British Interplanetary Society, 1999; 52:19–22 (preprint astro-ph/9901322).
28. Dar A, De Rújula A. The threat to life from Eta Carinae and gamma-ray bursts. Pp. 513–523 in Morselli A, Picozza P (eds). Astrophysics and Gamma Ray Physics in Space. Rome: Frascati Physics Series Vol. XXIV, 2002.
29. Scalo J, Wheeler JC. Astrophysical and astrobiological implications of gamma-ray burst properties. Astrophysical Journal, 2002; 566: 723–737.
30. Melott AL, Lieberman BS, Laird CM, Martin LD, Medvedev MV, Thomas BC et al. Did a gamma-ray burst initiate the late Ordovician mass extinction? International Journal of Astrobiology, 2004; 3: 55–61.
31. Vukotić B, Ćirković MM. Neocatastrophism and the Milky Way astrobiological landscape. Serbian Astronomical Journal, 2008; 176: 71–79.
32. Ćirković MM, Vukotić B. Astrobiological phase transition: Towards resolution of Fermi's paradox. Origin of Life and Evolution of the Biosphere, 2008; 38: 535–547.
33. Dreschhoff GAM, Laird CM. Evidence for a stratigraphic record of supernovae in polar ice. Advances in Space Research, 2006; 38: 1307–1311.
34. Stothers R. Giant solar flares in Antarctic ice. Nature, 1980; 287: 365.
35. White RV. Earth's biggest "whodunnit": Unravelling the clues in the case of the end-Permian mass extinction. Philosophical Transactions of the Royal Society A, 2002; 360: 2963–2985.
36. Benton MJ. When Life Nearly Died: The Greatest Mass Extinction of All Time. London: Thames and Hudson, 2003.
37. Ambrose SH. Late Pleistocene human population bottlenecks, volcanic winter, and differentiation of modern humans. Journal of Human Evolution, 1998; 34: 623–651.
38. Roscoe HK. The risk of large volcanic eruptions and the impact of this risk on future ozone depletion. Natural Hazards, 2001; 23: 231–246.
39. Rampino MR. Supereruptions as a threat to civilizations on Earth-like planets. Icarus, 2002; 156: 562–569.
40. Coleman S, De Luccia F. Gravitational effects on and of vacuum decay. Physical Review D, 1980; 21: 3305–3315.
41. Turner MS, Wilczek F. Is our vacuum metastable? Nature, 1982; 298: 633–634.
42. Sher M, Zaglauer HW. Cosmic-ray induced vacuum decay in the standard model. Physics Letters B, 1988; 206: 527–532.
43. Crone MM, Sher M. The environmental impact of vacuum decay. American Journal of Physics, 1991; 59: 25–32.
44. Dar A, De Rújula A, Heinz U. Will relativistic heavy-ion colliders destroy our planet? Physics Letters B, 1999; 470: 142–148.
45. Kent A. A critical look at risk assessments for global catastrophes. Risk Analysis, 2004; 24:157–168.
46. Ellis J., Giudice G, Mangano M, Tkachev I, Wiedemann U. Review of the Safety of LHC Collisions, 2008. Available at: http://lsag.web.cern.ch/lsag/LSAG-Report.pdf (LHC Safety Assessment Group).
47. Ord T, Hillerbrand R, Sandberg A. Probing the improbable: Methodological challenges for risks with low probabilities and high stakes. Journal of Risk Research, 2010; 13:191–205.
48. Jaffe L, Busza W, Wilczek F, Sandweiss J. Review of speculative "disaster scenarios" at RHIC. Reviews of Modern Physics, 2000; 72: 1125–1140.
49. Tegmark M, Bostrom N. Is a doomsday catastrophe likely? Nature, 2005; 438: 754.
50. Smolin L. Scientific alternatives to the anthropic principle. Pp. 323–366 in Carr B (ed). Universe or Multiverse. Cambridge: Cambridge University Press, 2005.
51. Gould SJ. The paradox of the first tier: An agenda for paleobiology. Paleobiology, 1985; 11: 2–12.
52. Gould SJ. Wonderful Life. New York: W. W. Norton, 1989.
53. Gould SJ. Full House: The Spread of Excellence from Plato to Darwin. New York: Three Rivers Press, 1996.
54. McShea DW. Possible largest-scale trends in organismal evolution: Eight "live hypotheses." Annual Review of Ecology and Systematics, 1998; 29: 293–318.
55. Kaneko K, Akutsu Y. Phase transitions in two-dimensional stochastic cellular automata. Journal of Physics A, 1986; 19: L69–L75.
56. Barrow JD, Tipler FJ. The Anthropic Cosmological Principle. New York: Oxford University Press, 1986.
57. Balashov YV. Resource letter: AP-1: The anthropic principle. American Journal of Physics, 1991; 59: 1069–1076.
58. Bostrom N. Anthropic Bias: Observation Selection Effects in Science and Philosophy. New York: Routledge, 2002.
59. Ćirković MM. Evolutionary catastrophes and the Goldilocks problem. International Journal of Astrobiology, 2007; 6: 325–329.
60. McKay CP. Time for intelligence on other planets. Pp. 405–419 in Doyle LR (ed). Circumstellar Habitable Zones, Proceedings of the First International Conference. Menlo Park, CA: Travis House Publications, 1996.
61. Russell DA. Speculations on the evolution of intelligence in multicellular organisms. Pp. 259–275 in Billingham J (ed). Life in the Universe. Cambridge: MIT Press, 1981.
62. Sahney S, Benton MJ. Recovery from the most profound mass extinction of all time. Procedings of the Royal Society B, 2008; 275: 759–765.
63. Bowring SA, Erwin DH, Isozaki Y. The tempo of mass extinction and recovery: The end-Permian example. Proceedings of the National Academy of Sciences USA, 1999; 96: 8827–8828.
64. Binzel RP, Rivkin AS, Stuart JS, Harris AW, Bus SJ, Burbine TH. Observed spectral properties of near-Earth objects: Results for population distribution, source regions, and space weathering processes. Icarus, 2004; 170: 259–294.
65. Stuart JS, Binzel RP. Bias-corrected population, size distribution, and impact hazard for the near-Earth objects. Icarus, 2004; 170: 295–311.
66. Szabó GyM, Csák B, Sárneczky K, Kiss LL. Photometric observations of 9 near-Earth objects. Astronomy and Astrophysics, 2001; 375: 285–292.
67. Emel'Yanenko VV, Bailey ME. Capture of Halley-type comets from the near-parabolic flux. Monthly Notices of the Royal Astronomical Society, 1998; 298: 212–222.
68. Schultz PH, Zárate M, Hames B, Koeberl C, Bunch T, Storzer D et al. The quaternary impact record from the Pampas, Argentina. Earth and Planetary Science Letters, 2004; 219: 221–238.