Sci Eng Ethics
DOI 10.1007/s11948-014-9559-3

ORIGINAL PAPER

Existential Risks: Exploring a Robust Risk Reduction Strategy

Karim Jebari

Received: 23 February 2014 / Accepted: 25 May 2014

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

Keywords Existential risk · Black swan · Engineering safety · Safety barriers · Uncertainty · Shelters · Global catastrophe

K. Jebari (✉)
Royal Institute of Technology, KTH, Teknikringen 78B, 114 28 Stockholm, Sweden
e-mail: jebarikarim@gmail.com; Jebari@kth.se

Published online: 03 June 2014

Existential Threats

There is a non-trivial risk that mankind will not survive this century.[1] This remarkable claim is presented by the astronomer royal Sir Martin Rees in his book *Our Final Century*, which argues that we live in an age of existential risk or an age of risks that we may suffer an existential catastrophe.[2] An existential catastrophe refers to a category of possible events that permanently and drastically reduce the ability of earth-originating intelligence to create or sustain value. Examples of such catastrophes could consist of the premature extinction of mankind or, even more dramatically, the eradication of all life on this planet. This category is contained within the broader category "global catastrophes", which refer to events that kill at least 10 million people on at least two continents. Mankind has a history of suffering from natural disasters such as pandemics, asteroids, and super volcanoes, and remains a hostage to these risks. In addition, in the last century of technological development, new anthropogenic risks have been introduced (Cirković et al. 2010). We now face new threats from human activity such as thermonuclear war, bioweapons, and geoengineering gone wrong. We should not assume that the risks mentioned above or in any other list compiled by the few scholars devoted to this issue are the only ones. It would be arrogant and naive to believe that we are now in a position to foresee the unfolding of events and technological change in such a way that can allow us to fully assess future risks to the existence of humanity. I will refer to unknown possible extinction events as "black swan extinction events." The term "black swan," coined by the economist Nassim Nicholas Taleb, denotes a high-impact, hard to predict, and rare event that is beyond the realm of normal expectations (Taleb 2010). Taleb argues that as the world becomes more interconnected and technology advances at an increasing rate, black swan events become more common and have a greater impact. He suggests that the risk of suffering a black swan extinction event may be higher than most of us think. The notion that some existential risks are black swans should make us aware of the daunting nature of risk reduction. We need to understand and reduce these risks. Given the cost effectiveness of a substantial reduction of a non-trivial existential risk, a mitigation strategy may be reasonable to pursue, even if the costs are significant. This point has been thoroughly discussed by others, most notably by Jason G. Matheny, Nick Bostrom, and Anders Sandberg (Matheny 2007).

How can we estimate the loss of value brought about by an existential disaster? The common sense assessment of its negative value is to equate the extermination of mankind with the disvalue of a single individual dying prematurely multiplied by the number of humans that exist at the moment of the catastrophe. According to Derek Parfit and others, this notion vastly underestimates disvalue of such an outcome, since it neglects the contributive value of future generations (Parfit 1984).

[1] In this paper, the terms "mankind" and "humanity" refer to the sapient animals that are members of our global civilization. These are not necessarily human in the strict biological sense.

[2] Experts at the Global Catastrophic Risk Conference suggested a 19 % chance of human extinction over the next century. Future of Humanity Institute. Global catastrophic risks survey, technical report [Internet]. 2008 Available from: http://www.fhi.ox.ac.uk/__data/assets/pdf_file/0020/3854/global-catastrophic-risks-report.pdf.

Population ethics, or the study of the is a novel field in moral philosophy, and most positions with regard to the value of potential people are associated with absurd conclusions and paradoxes (Arrhenius 2000). Yet, axiological actualism, or the view that only existing (actual) people matter, remains a minority view. Most people agree with Carl Sagan who pointed out the discontinuity between risks that threaten 99 and 100 % of humanity. Assuming our species survives for some 10 million years, about the average for a successful mammalian species, there could be trillions of people yet to come. Extinction is therefore much worse than a mere *near* extinction (Sagan 1983). Making some further assumptions about the long-term prospects of earth-originating intelligent life, the opportunity cost of extinction could be even more staggering, according to Nick Bostrom. He argues that, according to most prominent strands of consequentialism, the potential value lost to human extinction is equal to the aggregate value of all lives that could have existed in a non-extinction scenario (Bostrom 2003). Whereas not all normative theories are aggregative, as utilitarianism is, it would be plausible to assign these theories at least some probability of being correct, since we are deliberating under *moral uncertainty*. Were some aggregating theory of value to be "true" in a normatively relevant sense, then the case for preventing an extinction event seems overwhelming.

The peculiar nature of human psychology makes imagining unseen events very difficult. Often, the "worst-case scenario" is imagined as the worst historical scenario, even though we surely understand that before this scenario happened, it should have been planned for. Unfortunately, most approaches to risk reduction have focused on specific risks. This is not surprising. It is much easier to think about how to reduce the risk of nuclear war than to reduce the risk of an unspecified black swan extinction event. While preventing nuclear proliferation, developing global institutions to address global warming and to respond rapidly to pandemics are certainly worth pursuing, the seriousness of these risks compels us to consider any strategy that is plausibly effective in preventing an extinction event. However, an explicit approach to explore strategies to reduce the risk of black swan extinction events has to my knowledge not been proposed.

In section two, a number of heuristics, or rules of thumb, to reduce the risk of such events are proposed. Although not theoretically developed in the framework of existential risks or black swan events, *engineering safety* has great promise to this field. This approach to risk reduction has a great similarity with two approaches in the analysis of risk under uncertainty: *robust decision making,* developed by Lempert, Popper and Bankes (Bankes et al. 2003) and the info-gap method developed by Ben-Haim (2006). Engineers know that complex systems with many complicated and poorly understood components that have to be operated by many people can go wrong in unpredictable ways. In other words, engineers *expect* a system to fail, and try to reduce this risk and the damage involved in failure by applying the heuristics of safety engineering. This is a characteristic that is shared with the above mentioned approaches, where strategies that minimize harm across a wide range of possibilities are preferred to those that maximize utility under expected conditions (Hall et al. 2012). In section three, this article will also present an example of an implementation of one of these strategies, sometimes referred to as

safety barriers. The article will explore isolated, self-sufficient, and continuously manned underground refuges. These, it is argued, could be designed to prevent total extinction in a large number of possible scenarios. As such, these refuges exemplify how the risk of a black swan extinction event can be reduced by applying some heuristics inspired from engineering safety. While this is merely a conceptual sketch of such a strategy, some practical issues of the uncertainties and the feasibility of such refuges are discussed and addressed.

Engineering Safety

The principles of engineering safety have recently been systematized and developed by a number of philosophers and engineers (Hansson 2009; Möller and Hansson 2008; Linkov et al. 2011). Although it is possible to assess the robustness of engineering systems in many contexts, as argued by Baker et al. (2008) this requires lots of observational data, and is not always possible when dealing with new technology. Hansson has argued that the kind of uncertainty that is prevalent when assessing the safety of a complex engineering system, such as a nuclear power plant, can hardly be addressed by traditional probabilistic risk analysis. Failure in such systems is typically due to unforeseen circumstances, imperfect theoretical understanding of the systems involved, higher stress loads than foreseen and human error (either in design or in use) (Moses 1997). These sources of failure are also plausible candidates for potential black swan extinction events. Poor theoretical understanding of the climate system may for example lead to a disastrous result of an effort to reduce greenhouse gases by geoengineering. Or, a nuclear war could have unexpected consequences if the volcanic activity was at that moment more active than usual. Human errors, such as failure to predict how consumers will use a certain product, are also potential sources of catastrophic failure. In general, the heuristics are characterized by sacrificing some performance for less vulnerability to failed assumptions. Among the many heuristics that have been used for this purpose, two are particularly well suited in this context: safety factors and safety barriers.

Safety Factors

Safety factors are the ratio between a measure of the maximal load not leading to a specific failure and a corresponding measure of the allowed load. For example, calculations show that a certain bridge can support 100 tons at any given moment. To apply safety factors requires that we allow the possibility of an unforeseen circumstance that reduces the capacity of the bridge. We could therefore restrict the maximum stress that could be put on the bridge to 50 or 10 tons.

This methodology is of particular interest because it does not rely on knowing the source of the failure. So how can they be adapted to reduce the risk of a black swan extinction event? As noted by Tonn (2007), not many conceivable (and plausible) catastrophes would kill all humans. This is true in part because there are so many of us, and because we inhabit most of the land mass of this planet. In this sense, the large number and geographical distribution of human settlements works as a safety

factor. In the context of *global* catastrophes, population growth is seen as a great problem. From an *existential* risk point of view, more people means that humanity is more likely to survive a catastrophic event.

Diversity is sometimes considered to be a factor that contributes to the ability of a system to cope with failure. For example, poor genetic diversity increases the extinction risk for species (Frankham 2005). Diversity is a typical safety factor that is often rationalized away if a probabilistic risk analysis is used. Diversity can sometimes be very costly, and the benefits conferred by diversity could be very obscure. It is for example inefficient and costly to keep farm animals that are less predisposed to grow fast and produce tender meat. Therefore, it is likely that most farmers will prefer to buy similar animals. However, the added cumulative benefit of having these optimized animals may be countered by the increased risk that a large proportion of domesticated animals could die from some unknown pathogen.

Could human diversity be equally important in preventing an existential black swan event, and if so, what kind of diversity? This question cannot be answered here, but the engineering safety heuristic suggests that human diversity is important. It also suggests that the kind of diversity that is important cannot be predicted. The lesson from engineering safety is that *we just don't know*. Therefore, prima facie, any possible way in which a system is diverse, strengthens that system. The advice from safety engineering is thus *not* to limit our focus on a particular threat, and maximize diversity with that threat in mind. For example, we might be tempted to think about a pandemic when thinking about diversity. This potential catastrophe would make diversity in immune system profiles plausible. However, thinking only about infectious disease would perhaps leave humanity dangerously exposed to other threats if other dimensions of diversity are reduced.

Safety Barriers

Multiple *safety barriers* are physical or institutional barriers that prevent failure or reduce the damage of such failure. Importantly, the barriers are designed so that each barrier should be sufficient to prevent the failure. It is also crucial that the barriers are independent of each other, i.e. the failure of one barrier should not lead to the failure of other barriers. Also, barriers should be as diverse as possible, to insure that if a source of failure "breaches" the first barrier it would not automatically breach the second and third barriers. For example, medieval fortresses had a moat that could prevent the enemy from closing in on the fortress. Would the enemy be able to cross the moat, there were often high walls that would make an assault difficult. In addition to this, soldiers were trained to pour boiling oil or tar on assailants, further discouraging an assault. Finally, if the walls were breached, some fortresses would include secret escape routes for the nobility.

In modern contexts, this methodology is sometimes as *safe fail* engineering (as opposed to systems that claim to be *failsafe*) (Möller and Hansson 2008). The idea behind safety barriers is to accept the fact that systems fail, even when appropriate safety factors are in use. By being aware of this fact, a system could be made to fail in a less harmful way. Modern examples of this method are nuclear reactors that passively (without the intervention of an operator) shut down in the event of a

serious emergency. An evacuation routine or lifeboats are also examples of this methodology. As a strategy to reduce the risk of human extinction, this method holds much promise. Maher and Baum (2013) suggest constructing decentralized food stockpiles to mitigate some of the decline in human population that is associated with a global catastrophe. This strategy is promising because it is relatively cheap and probably effective against a range of potential disasters.

Another strategy that has been proposed in the literature consists in building permanent extra planetary settlements (Highfield 20:37). While some studies have been made regarding the possibility to colonize nearby planets and moons (Launius 2010), other strategies, that would yield comparable benefits, perhaps with less science fiction appeal, have hardly been explored (Hanson 2008). A "science fiction bias" is perhaps prevalent among those who take an interest in this issue. However, unless we investigate alternatives to extraterrestrial colonies strategies further, we will never know if they are worth pursuing. The costs assessments of a self-sufficient moon base vary. According to the Center for Strategic and International Studies (CSIS), the four man lunar base envisioned by the Bush administration would cost roughly 35 billion USD with annual operating costs of 7, 35 billion, excluding the cost of developing a rocket that can land on the Moon (Sabathier et al. 2009). This is arguably a very optimistic assessment, considering that the cost of the Apollo program was 170 billion in 2005 USD, and that this project involved lifting less weight from Earth's gravity well than a permanent moon base would require (Butts and Linton 2009). To achieve a colony that could survive a global catastrophe and eventually repopulate earth, we would need *at least* 80 colonists since this number is believed to be the minimal number required for the human species to persist. Studies have shown that this is roughly the number of people that colonized the Americas (Hey 2005). Assuming that costs increase linearly with the number of colonists, this would result in a rough estimate of 700 billion USD, plus 150 billion per year in operating costs for a Moon colony with 80 colonists. It is not unlikely that it might end up being two or three times more expensive, given the optimistic character of the CSIS assessment. To make the colony self-sufficient would require creating an industrial infrastructure. This might cost an additional and uncertain amount of money. It is also uncertain whether it is biologically possible for humans to settle the Moon, due to the likely adverse health effects of living in a reduced gravity environment. Although we are uncertain about the harm to the body from living on the Moon, the health issues associated with microgravity have been extensively studied on astronauts. These include bone decalcification, muscle atrophy, a slowing of cardiovascular system functions, decreased production of red blood cells, balance disorders, and a weakening of the immune system (Kanas and Manzey 2008). These effects have been observed on astronauts living for relatively short periods in these conditions. Human infants growing in this environment have not been studied. But it is not unlikely that they would suffer from a wide variety of health problems, if gestation would at all be possible under these conditions.

Since the planet Mars has significantly higher surface gravity than the Moon, these problems *may* be avoided if a colony were to be built there. However, the gravity of Mars is still only one-third of what humans are used to, which may or may not incur harm to the potential colonists. Furthermore, the initial cost of a Mars

colony is probably much higher than a Moon colony, although estimates vary widely. A notable concern for such a mission is the radiation and microgravity that colonists would suffer on the 8–9 months voyage to Mars. Another concern is that the pilots or computers landing on Mars would have to do so without much assistance from Earth, a task made more difficult by the thin atmosphere of Mars, which makes aero breaking very inefficient. The track record of Mars missions is not encouraging, considering that it is more difficult to land larger payloads, that many landings will be required, and that a single casualty is likely to derail (or at least hamper) the whole effort.[3] To sum up, to build extraterrestrial colonies is a very expensive strategy to reduce existential risk and is not likely to be politically, economically or technologically feasible in the next few decades.

Although extra planetary colonies should not be ruled out as a long term risk reduction strategy, another, more "down to earth" alternative of the *safety barriers* strategy is feasible. I propose that isolated and secure shelters on Earth could confer similar benefits as permanent extra planetary settlements at a fraction of the cost and risk. This makes these shelters plausible medium-term projects that may ensure the long-term survival of terrestrial intelligence. Since the shelters combine isolation and physical barriers, they are a paradigmatic example of a *safety barrier*. Properly equipped these shelters would form an effective barrier that could protect us from a wide range of black swan extinction events.

A Concrete Risk Reduction Strategy

During the Cold War, many countries built fallout shelters for high-ranking officials and crucial military personnel. In Switzerland and many other countries, vast shelters with the purpose of saving as many lives as possible were built to shield people from the blast and the initial radioactive fallout of a nuclear war. While these shelters could be of major use during a limited nuclear war, or a major conventional war, they would not suffice to increase the chances of mankind surviving an existential catastrophe. This was never their explicit purpose.

Robin Hanson discusses shelters as a means to create an information market about global catastrophic risks (Hanson 2008). His idea is to build shelters and to sell the right to seek refuge in one. If the market for these refuge futures is big enough, Hanson argues, we could have a means of predicting whether or not global catastrophes are imminent or not. These prediction markets would be very useful, he suggests, in preventing such catastrophes. Yet, the shelters that Hanson proposes are not nearly as safe as they could be. Since Hanson's suggested shelters are not continuously manned, they would not protect against a microbial agent that lies dormant in a patient for a long time. Also, the population of such a shelter is not necessarily sufficiently large enough to repopulate earth if a catastrophe would eliminate all the other remaining people.

[3] About 43 flybys, orbiting and landing missions have been sent to Mars by NASA and other nations in the past 40 years, and only 12 have been fully successful (Kaufman 2011).

Location and Resources

The central aspect behind the concept of the proposed shelter is to recreate some of the benefits to risk reduction conveyed by an extra-planetary permanent and self-sufficient settlement. Although the benefit of distance cannot be achieved on this planet, other benefits are possible, for example, physical isolation, a controlled environment and self-sufficiency. Although these shelters are only a minor modification on existing refuges, the idea of an effective shelter with the explicit purpose to reduce existential risk has, to my knowledge, not been proposed in the academic literature. An effective shelter would differ in various ways from existing refuges in order to maximize their protective capacity. The geographical location of the shelter should be as remote from population centers as possible, in order to maximize the benefit of isolation and separation. A thermonuclear war would for example target population centers and military installations. An effective shelter would seek to be far from these places. Social collapse due to a catastrophic event may create hostile groups that may pose a danger to the inhabitants of the shelter. These groups are also likely to be found in areas that are urban today. Some catastrophes, such as asteroid impacts, nuclear war or some possible ETI attack scenarios involve huge amounts of radiation. The proposed shelter would therefore, just as typical fallout shelters, be underground to protect against such harm. Such locations are abundant and easy to find. The EU and the US have many large and unused shelters that could be refurbished too serve this purpose. Old mines and large natural cave formations could also be used.

Natural resources that could assist a small group of survivors to recover such as fossil fuels, metals and chemicals, are another concern. We have already depleted many of the natural resources that are easy to extract. Re-industrialization without easily extractable coal or crude oil could be a daunting task. In particular, fossil fuels are much less abundant in locations where extraction requires little knowhow and capital. The shelter should therefore be equipped with tools to rapidly establish an industrial infrastructure when returning to the surface. Photovoltaic cells, hydroponic facilities, habitation modules, self-assembly manufacturing plants, and other infrastructure that is easy to deploy once the shelter population has emerged would shorten the period of risk associated with being a small and exposed group of humans. Thus, the source of power for the survivors is also relevant when choosing the location. Additive manufacturing could also be a key technology to ensure that utensils could be produced locally. Practical manuals to manufacture basic goods and to reestablish a stable society would be necessary as well as securing databases with as much of the current body of science, technology, and culture as possible. Only with the means to avoid reliance on abundant fossil fuels could a new human civilization reestablish itself.

Access to water from confined aquifers could be a crucial asset if surface water was contaminated. A proper location would also be in an area with low seismic activity. Also, the needs of the re-colonization efforts need to be considered. Therefore, a location with enough surface water, moderate temperatures and rich soil would be optimal. Such areas are for obvious reasons rarely distant from

population centers, so a careful evaluation of each location must be made, and take both the desirability of the location and its vulnerability into account.

Population

Given that the main purpose of the proposed shelter would not be to save as many lives as possible, or to guarantee the continuation of government, these shelters would differ in the number of people and the selection of their inhabitants. Previous fallout shelters aimed to grant protection to as many people as possible. This means sacrificing comfort, limiting the long-term prospects of the shelter inhabitants and increasing the risk of social conflict. The kind of shelter proposed here would be manned by a limited number of people of fertile age (18-40), unlike the most likely demographic in shelters for government officials. Dunbar's number, proposed by the anthropologist Robin Dunbar, is a suggested cognitive limit to the number of persons with whom an individual can maintain stable relationships. The commonly used value for this number is 150, with a range between 100 and 230 (Hernando et al. 2009). This should therefore be the average number of persons at each individual shelter. From the perspective of creating the potential for repopulation of earth once it is safe to return to the surface, it may seem like a good idea to bias the sex ratio to include more women than men. However, there is evidence that a biased operational sex ratio increases reproductive competition and may thus cause additional social friction (Clutton-Brock 2007).

To further increase genetic robustness, the proposed shelter could have a large number of fertilized eggs in cryostorage, to maximize the potential genetic variation. The cost of this supplementary effort would be minimal. Modern shelters do not provide sufficient protection against a bioengineered microbe that combines a high fatality rate, a long incubation period, a short latency period, and transmission via airborne droplets. Such a microbe could easily infect someone seeking refuge before entering the shelter. To mitigate such threats, an effective shelter would have to host a population continuously. Since this proposed shelter is quite affordable, rich and middle income economies would afford to build at least one. The choice of who gets drafted for "shelter service" could follow the US system of jury duty. A number of persons are randomly drafted and reimbursed with a small compensation for lost income. A rudimentary vetting procedure could be carried out to exclude persons who would have a strong adverse psychological reaction to be in the shelter, as well as individuals who display strong antisocial tendencies. People with medical conditions that require expert medical attention (for example dialysis) would also have the right to abstain. Since in practice it would be very difficult to find anyone willing to spend their whole life in a shelter, the population should be rotated every year or so. To avoid dangerous pathogens and to screen potential shelter inhabitants for emotional and psychosocial problems, the new inhabitants should be appropriately quarantined for at least a few weeks before entering the shelter. Both individuals and families could be accommodated to make the "shelter service" as attractive as possible for average citizens. The shelter should also include a small skeleton crew of engineers, doctors, psychologists and horticulturalists.

Creating a Psychologically Sound Environment

Importantly, the proposed shelter would have provisions for an extended period of time, at least a few decades, since some existential catastrophes could render the world uninhabitable for many years.[4] This mean that food must, to at least some degree, be produced in the shelter. Hydroponic systems could be combined with fish tanks to provide maximal variety in food supply. This also stresses the importance of an environment that is as psychologically sound as possible. Needless to say, if the persons in the shelter are the only remaining people, it is crucial that the social and physical infrastructure of this place is construed as to minimize conflict among the inhabitants. Data on the psychosocial challenges from experiments such as the MARS-500 mission as well as from studies of personnel on Arctic research stations and other similar facilities could be of great use in designing the shelter (Wang et al. 2014; Basner et al. 2014). We know from observing small isolated communities in artificial environments that being in such an environment is a major psychosocial hazard (Palinkas 2003). Famously, the participants to the Biosphere-2 project were subjected to a caloric restriction diet while the oxygen levels in the isolated dome were substantially reduced. These circumstances had likely an adverse effect on the morale of the crew in this experiment. Depression, factionalism and communication failures were reported (MacCallum et al. 2004).

Indoor gardens and farms reduce psychosocial stress and depression and should be a crucial component of the shelter (Pretty et al. 2005). A rigorous exercise schedule would yield similar benefits (Cornil et al. 1965). It is also very important that the inhabitants of the shelter are engaged in various aspects of maintenance, and practice in gardening, medicine and mechanics, since boredom and passivity is a psychosocial hazard (Blaszczynski et al. 1990). It also is vital that the shelter has ample room for its inhabitants, for two reasons: (1) crammed living conditions are likely to exacerbate the unavoidable social tension that would occur if the inhabitants of the shelter were isolated for an extended and uncertain amount of time. (2) If the inhabitants of the shelter are forced to remain underground for an extended period, the population is likely to increase, and would need room to expand.

These specifications differ quite substantially from existing bomb-shelters, where little or no attention was directed to creating a hospitable and sound environment. However, large scale food production indoors requires lots of electricity, to provide artificial light, regulate temperature and humidity and so on. For the refuge to be able to provide electricity for such an extended period, diesel generators are quite impractical as a primary power source. Nuclear power would be necessary. In particular, the small nuclear reactors that have recently been developed by a number of companies would be well-suited for this task. These run on highly enriched uranium and thus have high power density in a small volume. Furthermore, these

[4] For example the "Doomsday device" from Stanley Kubrick's film *Dr. Strangelove* is a nuclear warhead with an amount of cobalt at its core. This metal is transmuted in the explosion into the radioactive isotope cobalt-60, which would be vaporized. This device perfectly possible to manufacture for any nation with access to nuclear weapons capacity, and would contaminate an area with lethal levels of radiation for about 60 years, since cobalt has a long half-life and a lethal levels of radiation.

reactors have long core lives. For example the Toshiba 4S design does not need refueling in the entire 30-year lifecycle of the plant. The design of these reactors thus combines a compact design with a high level of safety (Ingersoll 2009).

Taking as much research from the psychology of isolated and confined environments into account is necessary to devise strategies, locations and institutions to deal with the likely psychological stress associated with living in the proposed shelter. However, as the shelter is proposed to be continuously manned, with crews rotating every year or so, the first experimental years would provide us with ample observations and psychological insights on what could facilitate a prolonged stay in the shelter. While data on similar experiments is of course valuable to the design of this facility, data from on-site experiments would give action-guiding and strongly valid information on how to prevent discord or depression. This would be a major benefit compared with existing shelters, and could prove invaluable for future potential extra planetary expeditions or settlements.

There are also some concerns that could affect the sealing of a shelter. During a catastrophe, many people may want to invade the shelter, either with hostile intent or to escape from whatever caused the catastrophe. This could be potentially fatal in some scenarios. For example if someone among these people is infected with the microbe that is the cause of the catastrophe, this person's intrusion is likely to undermine the purpose of the shelter. It is therefore necessary that the inhabitants of the shelter should be able to seal the shelter from inside, and that the location of the shelter remains a state secret. It may also be the case that inhabitants in the shelter would feel an urge to return to the surface, in particular if they have relatives that are on the surface during the catastrophe, even though this might put the rest of the shelter population at risk. It is possible that the information about the nature of the catastrophe is distorted or incomplete. It might therefore be a good idea to take measures to avoid rash action in opening the shelter from inside. A passive and temporary lock-down system in the event of some signal that would correlate strongly with a global catastrophe, perhaps a dramatic drop in the global stock index, could prevent the residents of the shelter from opening the doors prematurely. This would allow the residents of the shelter some time to make an appropriate risk assessment before opening the door. However, the potential stress that residents could suffer from being trapped in the shelter should also be considered. A system that requires a simple majority of shelter inhabitants to open the door would be another possibility to avoid rash and potentially fatal decisions. Too little is known about how people react in these situations to be confident on what would work. This makes the on-site observations of human behavior under a variety of constraints and regulatory frameworks extremely valuable.

Costs

While this measure would be quite expensive, it would probably be much cheaper than even the most optimistic assessments of colonizing the moon. There are already shelters that could be refitted for this purpose. A nuclear reactor with highly enriched uranium, similar to that which powers large submarines, would probably

be the most costly item. Thus, a comparison with an Ohio-class submarine, with a crew of 155, seems reasonable. This submarine costs 2 billion USD. Even if this shelter would be an order of magnitude more expensive, it would still cost only a fraction of what a Moon colony would cost on the most optimistic cost assessment. Furthermore, this facility would reduce the risk of black swan extinction events with existing and proven technology. It could also be implemented at a very short notice, compared with even the most optimistic plans to colonize the Moon.

Conclusion

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

Acknowledgments I would like to thank Seth Baum, Nick Beckstead, Jacob Haqq-Misra, Niklas Möller, Aron Vallinder and two anonymous reviewers for *Risk Analysis* for their comments on earlier versions of this manuscript.

Conflict of interest None.

References

Arrhenius, G. (2000). An impossibility theorem for welfarist axiologies. *Economics and Philosophy, 16*, 247–266.

Baker, J. W., Schubert, M., & Faber, M. H. (2008). On the assessment of robustness. *Structural Safety, 30*, 253–267. doi:10.1016/j.strusafe.2006.11.004.

Bankes, S. C., Lempert, R. J., & Popper, S. W. (2003). *Shaping the next one hundred years: New methods for quantitative, long-term policy analysis* (186th ed.). Santa Monica, CA: Rand Publishing.

Basner, M., Dinges, D. F., Mollicone, D. J., Savelev, I., Ecker, A. J., Di Antonio, A., et al. (2014). Psychological and behavioral changes during confinement in a 520-day simulated interplanetary mission to Mars. *PLoS One, 9*, e93298. doi:10.1371/journal.pone.0093298.

Ben-Haim, Y. (2006). *Info-gap decision theory: Decisions under severe uncertainty*. Waltham: Academic Press.

Blaszczynski, A., McConaghy, N., & Frankova, A. (1990). Boredom proneness in pathological gambling. *Psychological Reports, 67*, 35–42.

Bostrom, N. (2003). Astronomical waste: The opportunity cost of delayed technological development. *Utilitas, 15*, 308–314. doi:10.1017/S0953820800004076.

Butts, G., & Linton, K. (2009). The joint confidence level paradox: A history of denial. Presented at the NASA Cost Symposium, 28–30 April 2009.

Cirković, M. M., Sandberg, A., & Bostrom, N. (2010). Anthropic shadow: Observation selection effects and human extinction risks. *Risk Analysis, 30*, 1495–1506. doi:10.1111/j.1539-6924.2010.01460.x.

Clutton-Brock, T. (2007). Sexual selection in males and females. *Science, 318*, 1882–1885. doi:10.1126/science.1133311.

Cornil, A., Coster, A. D., Copinschi, G., & Franckson, J. R. M. (1965). Effect of muscular exercise on the plasma level of cortisol in man. *Acta Endocrinologica (Copenh), 48*, 163–168. doi:10.1530/acta.0.0480163.

Frankham, R. (2005). Genetics and extinction. *Biological Conservation, 126*, 131–140. doi:10.1016/j.biocon.2005.05.002.

Hall, J. W., Lempert, R. J., Keller, K., Hackbarth, A., Mijere, C., & McInerney, D. J. (2012). Robust climate policies under uncertainty: A comparison of robust decision making and info-gap methods. *Risk Analysis, 32*, 1657–1672. doi:10.1111/j.1539-6924.2012.01802.x.

Hanson, R. (2008). Catastrophe, social collapse, and human extinction. In N. Bostrom & M. M. Ćirković (Eds.), *Global catastrophic risks* (p. 554). Oxford: Oxford University Press.

Hansson, S. O. (2009). From the casino to the jungle. *Synthese, 168*, 423–432. doi:10.1007/s11229-008-9444-1.

Hernando, A., Villuendas, D., Vesperinas, C., Abad, M., & Plastino, A. (2009). Unravelling the size distribution of social groups with information theory on complex networks. *The European Physical Journal B*, *76*(1), 87–97.

Hey, J. (2005). On the number of New World founders: A population genetic portrait of the peopling of the Americas. *PLoS Biology, 3*, e193. doi:10.1371/journal.pbio.0030193.

Highfield, R., (2001) 20:37. Colonies in space may be only hope, says Hawking. Telegraph.co.uk.

Ingersoll, D. T. (2009). Deliberately small reactors and the second nuclear era. *Progress in Nuclear Energy, 51*, 589–603. doi:10.1016/j.pnucene.2009.01.003.

Kanas, N., & Manzey, D. (2008). Basic issues of human adaptation to space flight. In *Space psychology and psychiatry*. The Space Technology Library (Vol. 22, pp. 15–48). Netherlands: Springer.

Kaufman, M. (2011). *Landing on Mars is hard, but another mission to the Red Planet is about to begin*. Washington: Washington Post, November 22, sec. National.

Launius, R. D. (2010). Can we colonize the solar system? Human biology and survival in the extreme space environment. *Endeavour, 34*, 122–129. doi:10.1016/j.endeavour.2010.07.001.

Linkov, I., Bates, M., Loney, D., Sparrevik, M., & Bridges, T. (2011). Risk management practices. In I. Linkov & T. S. Bridges (Eds.), *Climate, NATO science for peace and security series C: Environmental security* (pp. 133–155). Netherlands: Springer.

MacCallum, T., Poynter, J., & Bearden, D. (2004). Lessons learned from biosphere 2: When viewed as a ground simulation/analog for long duration human space exploration and settlement (SAE Technical Paper No. 2004-01-2473). SAE International, Warrendale, PA.

Maher, T. M., & Baum, S. D. (2013). Adaptation to and recovery from global catastrophe. *Sustainability, 5*, 1461–1479. doi:10.3390/su5041461.

Matheny, J. G. (2007). Reducing the risk of human extinction. *Risk Analysis, 27*, 1335–1344. doi:10.1111/j.1539-6924.2007.00960.x.

Möller, N., & Hansson, S. O. (2008). Principles of engineering safety: Risk and uncertainty reduction. *Reliability Engineering and System Safety, 93*, 798–805. doi:10.1016/j.ress.2007.03.031.

Moses, F. (1997). Problems and prospects of reliability-based optimization. *Engineering Structures, 19*, 293–301. doi:10.1016/S0141-0296(97)83356-1.

Palinkas, L. A. (2003). The psychology of isolated and confined environments. Understanding human behavior in Antarctica. *American Psychologist, 58*, 353–363.

Parfit, D. A. (1984). *Reasons and persons*. Oxford: Oxford University Press.

Pretty, J., Peacock, J., Sellens, M., & Griffin, M. (2005). The mental and physical health outcomes of green exercise. *International Journal of Environmental Health Research, 15*, 319–337. doi:10.1080/09603120500155963.

Sabathier, V.G., Weppler, J., & Bander, A. (2009). Costs of an International Lunar Base | [WWW Document]. Center for Strategic and International Studies. http://csis.org/publication/costs-international-lunar-base. Accessed April 14, 2014.

Sagan, C. (1983). Nuclear war and climatic catastrophe: Some policy implications. *Foreign Affairs*, 257–292.

Taleb, N. N. (2010). *The black swan: The impact of the highly improbable*. New York: Random House Trade Paperbacks.

Tonn, B. E. (2007). Futures sustainability. *Futures, 39*, 1097–1116. doi:10.1016/j.futures.2007.03.018.

Wang, Y., Jing, X., Lv, K., Wu, B., Bai, Y., Luo, Y., et al. (2014). During the long way to Mars: Effects of 520 days of confinement (Mars500) on the assessment of affective stimuli and stage alteration in mood and plasma hormone levels. *PLoS One, 9*, e87087. doi:10.1371/journal.pone.0087087.