### NC—AT: Deon

#### The most moral course of action is preventing extinction at all cost

Seth D. Baum and Anthony M. Barrett 18. Global Catastrophic Risk Institute. 2018. “Global Catastrophes: The Most Extreme Risks.” Risk in Extreme Environments: Preparing, Avoiding, Mitigating, and Managing, edited by Vicki Bier, Routledge, pp. 174–184.

2. What Is GCR And Why Is It Important? Taken literally, a global catastrophe can be any event that is in some way catastrophic across the globe. This suggests a rather low threshold for what counts as a global catastrophe. An event causing just one death on each continent (say, from a jet-setting assassin) could rate as a global catastrophe, because surely these deaths would be catastrophic for the deceased and their loved ones. However, in common usage, a global catastrophe would be catastrophic for a significant portion of the globe. Minimum thresholds have variously been set around ten thousand to ten million deaths or $10 billion to $10 trillion in damages (Bostrom and Ćirković 2008), or death of one quarter of the human population (Atkinson 1999; Hempsell 2004). Others have emphasized catastrophes that cause long-term declines in the trajectory of human civilization (Beckstead 2013), that human civilization does not recover from (Maher and Baum 2013), that drastically reduce humanity’s potential for future achievements (Bostrom 2002, using the term “existential risk”), or that result in human extinction (Matheny 2007; Posner 2004). A common theme across all these treatments of GCR is that some catastrophes are vastly more important than others. Carl Sagan was perhaps the first to recognize this, in his commentary on nuclear winter (Sagan 1983). Without nuclear winter, a global nuclear war might kill several hundred million people. This is obviously a major catastrophe, but humanity would presumably carry on. However, with nuclear winter, per Sagan, humanity could go extinct. The loss would be not just an additional four billion or so deaths, but the loss of all future generations. To paraphrase Sagan, the loss would be billions and billions of lives, or even more. Sagan estimated 500 trillion lives, assuming humanity would continue for ten million more years, which he cited as typical for a successful species. Sagan’s 500 trillion number may even be an underestimate. The analysis here takes an adventurous turn, hinging on the evolution of the human species and the long-term fate of the universe. On these long time scales, the descendants of contemporary humans may no longer be recognizably “human”. The issue then is whether the descendants are still worth caring about, whatever they are. If they are, then it begs the question of how many of them there will be. Barring major global catastrophe, Earth will remain habitable for about one billion more years 2 until the Sun gets too warm and large. The rest of the Solar System, Milky Way galaxy, universe, and (if it exists) the multiverse will remain habitable for a lot longer than that (Adams and Laughlin 1997), should our descendants gain the capacity to migrate there. An open question in astronomy is whether it is possible for the descendants of humanity to continue living for an infinite length of time or instead merely an astronomically large but finite length of time (see e.g. Ćirković 2002; Kaku 2005). Either way, the stakes with global catastrophes could be much larger than the loss of 500 trillion lives. Debates about the infinite vs. the merely astronomical are of theoretical interest (Ng 1991; Bossert et al. 2007), but they have limited practical significance. This can be seen when evaluating GCRs from a standard risk-equals-probability-times-magnitude framework. Using Sagan’s 500 trillion lives estimate, it follows that reducing the probability of global catastrophe by a mere one-in-500-trillion chance is of the same significance as saving one human life. Phrased differently, society should try 500 trillion times harder to prevent a global catastrophe than it should to save a person’s life. Or, preventing one million deaths is equivalent to a one-in500-million reduction in the probability of global catastrophe. This suggests society should make extremely large investment in GCR reduction, at the expense of virtually all other objectives. Judge and legal scholar Richard Posner made a similar point in monetary terms (Posner 2004). Posner used $50,000 as the value of a statistical human life (VSL) and 12 billion humans as the total loss of life (double the 2004 world population); he describes both figures as significant underestimates. Multiplying them gives $600 trillion as an underestimate of the value of preventing global catastrophe. For comparison, the United States government typically uses a VSL of around one to ten million dollars (Robinson 2007). Multiplying a $10 million VSL with 500 trillion lives gives $5x1021 as the value of preventing global catastrophe. But even using “just" $600 trillion, society should be willing to spend at least that much to prevent a global catastrophe, which converts to being willing to spend at least $1 million for a one-in-500-million reduction in the probability of global catastrophe. Thus while reasonable disagreement exists on how large of a VSL to use and how much to count future generations, even low-end positions suggest vast resource allocations should be redirected to reducing GCR. This conclusion is only strengthened when considering the astronomical size of the stakes, but the same point holds either way. The bottom line is that, as long as something along the lines of the standard riskequals-probability-times-magnitude framework is being used, then even tiny GCR reductions merit significant effort. This point holds especially strongly for risks of catastrophes that would cause permanent harm to global human civilization. The discussion thus far has assumed that all human lives are valued equally. This assumption is not universally held. People often value some people more than others, favoring themselves, their family and friends, their compatriots, their generation, or others whom they identify with. Great debates rage on across moral philosophy, economics, and other fields about how much people should value others who are distant in space, time, or social relation, as well as the unborn members of future generations. This debate is crucial for all valuations of risk, including GCR. Indeed, if each of us only cares about our immediate selves, then global catastrophes may not be especially important, and we probably have better things to do with our time than worry about them. While everyone has the right to their own views and feelings, we find that the strongest arguments are for the widely held position that all human lives should be valued equally. This position is succinctly stated in the United States Declaration of Independence, updated in the 1848 Declaration of Sentiments: “We hold these truths to be self-evident: that all men and 3 women are created equal”. Philosophers speak of an agent-neutral, objective “view from nowhere” (Nagel 1986) or a “veil of ignorance” (Rawls 1971) in which each person considers what is best for society irrespective of which member of society they happen to be. Such a perspective suggests valuing everyone equally, regardless of who they are or where or when they live. This in turn suggests a very high value for reducing GCR, or a high degree of priority for GCR reduction efforts.

#### Neuroscience proves util- pleasure and pain *are* intrinsic value and disvalue – everything else regresses.

Blum et al. 18 [Kenneth Blum, 1Department of Psychiatry, Boonshoft School of Medicine, Dayton VA Medical Center, Wright State University, Dayton, OH, USA 2Department of Psychiatry, McKnight Brain Institute, University of Florida College of Medicine, Gainesville, FL, USA 3Department of Psychiatry and Behavioral Sciences, Keck Medicine University of Southern California, Los Angeles, CA, USA 4Division of Applied Clinical Research & Education, Dominion Diagnostics, LLC, North Kingstown, RI, USA 5Department of Precision Medicine, Geneus Health LLC, San Antonio, TX, USA 6Department of Addiction Research & Therapy, Nupathways Inc., Innsbrook, MO, USA 7Department of Clinical Neurology, Path Foundation, New York, NY, USA 8Division of Neuroscience-Based Addiction Therapy, The Shores Treatment & Recovery Center, Port Saint Lucie, FL, USA 9Institute of Psychology, Eötvös Loránd University, Budapest, Hungary 10Division of Addiction Research, Dominion Diagnostics, LLC. North Kingston, RI, USA 11Victory Nutrition International, Lederach, PA., USA 12National Human Genome Center at Howard University, Washington, DC., USA, Marjorie Gondré-Lewis, 12National Human Genome Center at Howard University, Washington, DC., USA 13Departments of Anatomy and Psychiatry, Howard University College of Medicine, Washington, DC US, Bruce Steinberg, 4Division of Applied Clinical Research & Education, Dominion Diagnostics, LLC, North Kingstown, RI, USA, Igor Elman, 15Department Psychiatry, Cooper University School of Medicine, Camden, NJ, USA, David Baron, 3Department of Psychiatry and Behavioral Sciences, Keck Medicine University of Southern California, Los Angeles, CA, USA, Edward J Modestino, 14Department of Psychology, Curry College, Milton, MA, USA, Rajendra D Badgaiyan, 15Department Psychiatry, Cooper University School of Medicine, Camden, NJ, USA, Mark S Gold 16Department of Psychiatry, Washington University, St. Louis, MO, USA, “Our evolved unique pleasure circuit makes humans different from apes: Reconsideration of data derived from animal studies”, U.S. Department of Veterans Affairs, 28 February 2018, accessed: 19 August 2020, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6446569/>] R.S.

**Pleasure** is not only one of the three primary reward functions but it also **defines reward.** As homeostasis explains the functions of only a limited number of rewards, the principal reason why particular stimuli, objects, events, situations, and activities are rewarding may be due to pleasure. This applies first of all to sex and to the primary homeostatic rewards of food and liquid and extends to money, taste, beauty, social encounters and nonmaterial, internally set, and intrinsic rewards. Pleasure, as the primary effect of rewards, drives the prime reward functions of learning, approach behavior, and decision making and provides the **basis for hedonic theories** of reward function. We are attracted by most rewards and exert intense efforts to obtain them, just because they are enjoyable [10].

Pleasure is a passive reaction that derives from the experience or prediction of reward and may lead to a long-lasting state of happiness. The word happiness is difficult to define. In fact, just obtaining physical pleasure may not be enough. One key to happiness involves a network of good friends. However, it is not obvious how the higher forms of satisfaction and pleasure are related to an ice cream cone, or to your team winning a sporting event. Recent multidisciplinary research, using both humans and detailed invasive brain analysis of animals has discovered some critical ways that the brain processes pleasure [14].

Pleasure as a hallmark of reward is sufficient for defining a reward, but it may not be necessary. A reward may generate positive learning and approach behavior simply because it contains substances that are essential for body function. When we are hungry, we may eat bad and unpleasant meals. A monkey who receives hundreds of small drops of water every morning in the laboratory is unlikely to feel a rush of pleasure every time it gets the 0.1 ml. Nevertheless, with these precautions in mind, we may define any stimulus, object, event, activity, or situation that has the potential to produce pleasure as a reward. In the context of reward deficiency or for disorders of addiction, homeostasis pursues pharmacological treatments: drugs to treat drug addiction, obesity, and other compulsive behaviors. The theory of allostasis suggests broader approaches - such as re-expanding the range of possible pleasures and providing opportunities to expend effort in their pursuit. [15]. It is noteworthy, the first animal studies eliciting approach behavior by electrical brain stimulation interpreted their findings as a discovery of the brain’s pleasure centers [16] which were later partly associated with midbrain dopamine neurons [17–19] despite the notorious difficulties of identifying emotions in animals.

Evolutionary theories of pleasure: The love connection BO:D

Charles Darwin and other biological scientists that have examined the biological evolution and its basic principles found various mechanisms that steer behavior and biological development. Besides their theory on natural selection, it was particularly the sexual selection process that gained significance in the latter context over the last century, especially when it comes to the question of what makes us “what we are,” i.e., human. However, the capacity to sexually select and evolve is not at all a human accomplishment alone or a sign of our uniqueness; yet, we humans, as it seems, are ingenious in fooling ourselves and others–when we are in love or desperately search for it.

It is well established that modern biological theory conjectures that **organisms are** the **result of evolutionary competition.** In fact, Richard Dawkins stresses gene survival and propagation as the basic mechanism of life [20]. Only genes that lead to the fittest phenotype will make it. It is noteworthy that the phenotype is selected based on behavior that maximizes gene propagation. To do so, the phenotype must survive and generate offspring, and be better at it than its competitors. Thus, the ultimate, distal function of rewards is to increase evolutionary fitness by ensuring the survival of the organism and reproduction. It is agreed that learning, approach, economic decisions, and positive emotions are the proximal functions through which phenotypes obtain other necessary nutrients for survival, mating, and care for offspring.

Behavioral reward functions have evolved to help individuals to survive and propagate their genes. Apparently, people need to live well and long enough to reproduce. Most would agree that homo-sapiens do so by ingesting the substances that make their bodies function properly. For this reason, foods and drinks are rewards. Additional rewards, including those used for economic exchanges, ensure sufficient palatable food and drink supply. Mating and gene propagation is supported by powerful sexual attraction. Additional properties, like body form, augment the chance to mate and nourish and defend offspring and are therefore also rewards. Care for offspring until they can reproduce themselves helps gene propagation and is rewarding; otherwise, many believe mating is useless. According to David E Comings, as any small edge will ultimately result in evolutionary advantage [21], additional reward mechanisms like novelty seeking and exploration widen the spectrum of available rewards and thus enhance the chance for survival, reproduction, and ultimate gene propagation. These functions may help us to obtain the benefits of distant rewards that are determined by our own interests and not immediately available in the environment. Thus the distal reward function in gene propagation and evolutionary fitness defines the proximal reward functions that we see in everyday behavior. That is why foods, drinks, mates, and offspring are rewarding.

There have been theories linking pleasure as a required component of health benefits salutogenesis, (salugenesis). In essence, under these terms, pleasure is described as a state or feeling of happiness and satisfaction resulting from an experience that one enjoys. Regarding pleasure, it is a double-edged sword, on the one hand, it promotes positive feelings (like mindfulness) and even better cognition, possibly through the release of dopamine [22]. But on the other hand, pleasure simultaneously encourages addiction and other negative behaviors, i.e., motivational toxicity. It is a complex neurobiological phenomenon, relying on reward circuitry or limbic activity. It is important to realize that through the “Brain Reward Cascade” (BRC) endorphin and endogenous morphinergic mechanisms may play a role [23]. While natural rewards are essential for survival and appetitive motivation leading to beneficial biological behaviors like eating, sex, and reproduction, crucial social interactions seem to further facilitate the positive effects exerted by pleasurable experiences. Indeed, experimentation with addictive drugs is capable of directly acting on reward pathways and causing deterioration of these systems promoting hypodopaminergia [24]. Most would agree that pleasurable activities can stimulate personal growth and may help to induce healthy behavioral changes, including stress management [25]. The work of Esch and Stefano [26] concerning the link between compassion and love implicate the brain reward system, and pleasure induction suggests that social contact in general, i.e., love, attachment, and compassion, can be highly effective in stress reduction, survival, and overall health.

Understanding the role of neurotransmission and pleasurable states both positive and negative have been adequately studied over many decades [26–37], but comparative anatomical and neurobiological function between animals and homo sapiens appear to be required and seem to be in an infancy stage.

Finding happiness is different between apes and humans

As stated earlier in this expert opinion one key to happiness involves a network of good friends [38]. However, it is not entirely clear exactly how the higher forms of satisfaction and pleasure are related to a sugar rush, winning a sports event or even sky diving, all of which augment dopamine release at the reward brain site. Recent multidisciplinary research, using both humans and detailed invasive brain analysis of animals has discovered some critical ways that the brain processes pleasure.

Remarkably, there are pathways for ordinary liking and pleasure, which are limited in scope as described above in this commentary. However, there are **many brain regions**, often termed hot and cold spots, that significantly **modulate** (increase or decrease) our **pleasure or** even produce **the opposite** of pleasure— that is disgust and fear [39]. One specific region of the nucleus accumbens is organized like a computer keyboard, with particular stimulus triggers in rows— producing an increase and decrease of pleasure and disgust. Moreover, the cortex has unique roles in the cognitive evaluation of our feelings of pleasure [40]. Importantly, the interplay of these multiple triggers and the higher brain centers in the prefrontal cortex are very intricate and are just being uncovered.

Desire and reward centers

It is surprising that many different sources of pleasure activate the same circuits between the mesocorticolimbic regions (Figure 1). Reward and desire are two aspects pleasure induction and have a very widespread, large circuit. Some part of this circuit distinguishes between desire and dread. The so-called pleasure circuitry called “REWARD” involves a well-known dopamine pathway in the mesolimbic system that can influence both pleasure and motivation.

In simplest terms, the well-established mesolimbic system is a dopamine circuit for reward. It starts in the ventral tegmental area (VTA) of the midbrain and travels to the nucleus accumbens (Figure 2). It is the cornerstone target to all addictions. The VTA is encompassed with neurons using glutamate, GABA, and dopamine. The nucleus accumbens (NAc) is located within the ventral striatum and is divided into two sub-regions—the motor and limbic regions associated with its core and shell, respectively. The NAc has spiny neurons that receive dopamine from the VTA and glutamate (a dopamine driver) from the hippocampus, amygdala and medial prefrontal cortex. Subsequently, the NAc projects GABA signals to an area termed the ventral pallidum (VP). The region is a relay station in the limbic loop of the basal ganglia, critical for motivation, behavior, emotions and the “Feel Good” response. This defined system of the brain is involved in all addictions –substance, and non –substance related. In 1995, our laboratory coined the term “Reward Deficiency Syndrome” (RDS) to describe genetic and epigenetic induced hypodopaminergia in the “Brain Reward Cascade” that contribute to addiction and compulsive behaviors [3,6,41].

Furthermore, ordinary “liking” of something, or pure pleasure, is represented by small regions mainly in the limbic system (old reptilian part of the brain). These may be part of larger neural circuits. In Latin, hedus is the term for “sweet”; and in Greek, hodone is the term for “pleasure.” Thus, the word Hedonic is now referring to various subcomponents of pleasure: some associated with purely sensory and others with more complex emotions involving morals, aesthetics, and social interactions. The capacity to have pleasure is part of being healthy and may even extend life, especially if linked to optimism as a dopaminergic response [42].

Psychiatric illness often includes symptoms of an abnormal inability to experience pleasure, referred to as anhedonia. A negative feeling state is called dysphoria, which can consist of many emotions such as pain, depression, anxiety, fear, and disgust. Previously many scientists used animal research to uncover the complex mechanisms of pleasure, liking, motivation and even emotions like panic and fear, as discussed above [43]. However, as a significant amount of related research about the specific brain regions of pleasure/reward circuitry has been derived from invasive studies of animals, these cannot be directly compared with subjective states experienced by humans.

In an attempt to resolve the controversy regarding the causal contributions of mesolimbic dopamine systems to reward, we have previously evaluated the three-main competing explanatory categories: “liking,” “learning,” and “wanting” [3]. That is, dopamine may mediate (a) liking: the hedonic impact of reward, (b) learning: learned predictions about rewarding effects, or (c) wanting: the pursuit of rewards by attributing incentive salience to reward-related stimuli [44]. We have evaluated these hypotheses, especially as they relate to the RDS, and we find that the incentive salience or “wanting” hypothesis of dopaminergic functioning is supported by a majority of the scientific evidence. Various neuroimaging studies have shown that anticipated behaviors such as sex and gaming, delicious foods and drugs of abuse all affect brain regions associated with reward networks, and may not be unidirectional. Drugs of abuse enhance dopamine signaling which sensitizes mesolimbic brain mechanisms that apparently evolved explicitly to attribute incentive salience to various rewards [45].

Addictive substances are voluntarily self-administered, and they enhance (directly or indirectly) dopaminergic synaptic function in the NAc. This activation of the brain reward networks (producing the ecstatic “high” that users seek). Although these circuits were initially thought to encode a set point of hedonic tone, it is now being considered to be far more complicated in function, also encoding attention, reward expectancy, disconfirmation of reward expectancy, and incentive motivation [46]. The argument about addiction as a disease may be confused with a predisposition to substance and nonsubstance rewards relative to the extreme effect of drugs of abuse on brain neurochemistry. The former sets up an individual to be at high risk through both genetic polymorphisms in reward genes as well as harmful epigenetic insult. Some Psychologists, even with all the data, still infer that addiction is not a disease [47]. Elevated stress levels, together with polymorphisms (genetic variations) of various dopaminergic genes and the genes related to other neurotransmitters (and their genetic variants), and may have an additive effect on vulnerability to various addictions [48]. In this regard, Vanyukov, et al. [48] suggested based on review that whereas the gateway hypothesis does not specify mechanistic connections between “stages,” and does not extend to the risks for addictions the concept of common liability to addictions may be more parsimonious. The latter theory is grounded in genetic theory and supported by data identifying common sources of variation in the risk for specific addictions (e.g., RDS). This commonality has identifiable neurobiological substrate and plausible evolutionary explanations.

Over many years the controversy of dopamine involvement in especially “pleasure” has led to confusion concerning separating motivation from actual pleasure (wanting versus liking) [49]. We take the position that animal studies cannot provide real clinical information as described by self-reports in humans. As mentioned earlier and in the abstract, on November 23rd, 2017, evidence for our concerns was discovered [50]

In essence, although nonhuman primate brains are similar to our own, the disparity between other primates and those of human cognitive abilities tells us that surface similarity is not the whole story. Sousa et al. [50] small case found various differentially expressed genes, to associate with pleasure related systems. Furthermore, the dopaminergic interneurons located in the human neocortex were absent from the neocortex of nonhuman African apes. Such differences in neuronal transcriptional programs may underlie a variety of neurodevelopmental disorders.

In simpler terms, the system controls the production of dopamine, a chemical messenger that plays a significant role in pleasure and rewards. The senior author, Dr. Nenad Sestan from Yale, stated: “Humans have evolved a dopamine system that is different than the one in chimpanzees.” This may explain why the behavior of humans is so unique from that of non-human primates, even though our brains are so surprisingly similar, Sestan said: “It might also shed light on why people are vulnerable to mental disorders such as autism (possibly even addiction).” Remarkably, this research finding emerged from an extensive, multicenter collaboration to compare the brains across several species. These researchers examined 247 specimens of neural tissue from six humans, five chimpanzees, and five macaque monkeys. Moreover, these investigators analyzed which genes were turned on or off in 16 regions of the brain. While the differences among species were subtle, **there was** a **remarkable contrast in** the **neocortices**, specifically in an area of the brain that is much more developed in humans than in chimpanzees. In fact, these researchers found that a gene called tyrosine hydroxylase (TH) for the enzyme, responsible for the production of dopamine, was expressed in the neocortex of humans, but not chimpanzees. As discussed earlier, dopamine is best known for its essential role within the brain’s reward system; the very system that responds to everything from sex, to gambling, to food, and to addictive drugs. However, dopamine also assists in regulating emotional responses, memory, and movement. Notably, abnormal dopamine levels have been linked to disorders including Parkinson’s, schizophrenia and spectrum disorders such as autism and addiction or RDS.

Nora Volkow, the director of NIDA, pointed out that one alluring possibility is that the neurotransmitter dopamine plays a substantial role in humans’ ability to pursue various rewards that are perhaps months or even years away in the future. This same idea has been suggested by Dr. Robert Sapolsky, a professor of biology and neurology at Stanford University. Dr. Sapolsky cited evidence that dopamine levels rise dramatically in humans when we anticipate potential rewards that are uncertain and even far off in our futures, such as retirement or even the possible alterlife. This may explain what often motivates people to work for things that have no apparent short-term benefit [51]. In similar work, Volkow and Bale [52] proposed a model in which dopamine can favor NOW processes through phasic signaling in reward circuits or LATER processes through tonic signaling in control circuits. Specifically, they suggest that through its modulation of the orbitofrontal cortex, which processes salience attribution, dopamine also enables shilting from NOW to LATER, while its modulation of the insula, which processes interoceptive information, influences the probability of selecting NOW versus LATER actions based on an individual’s physiological state. This hypothesis further supports the concept that disruptions along these circuits contribute to diverse pathologies, including obesity and addiction or RDS.

#### Revisionary intuitionism is true and proves util

Yudkowsky 08 [Eliezer Yudkowsky (research fellow of the Machine Intelligence Research Institute; he also writes Harry Potter fan fiction). “The ‘Intuitions’ Behind ‘Utilitarianism.’” 28 January 2008. LessWrong. http://lesswrong.com/lw/n9/the\_intuitions\_behind\_utilitarianism/]

I haven’t said much about metaethics – the nature of morality – because that has a forward dependency on a discussion of the Mind Projection Fallacy that I haven’t gotten to yet. I used to be very confused about metaethics. After my confusion finally cleared up, I did a postmortem on my previous thoughts. I found that my object-level moral reasoning had been valuable and my **meta-level moral reasoning had been** worse than **useless**. And this appears to be a general syndrome – **people do much better when discussing whether torture is** good or **bad than**when they discuss **the meaning of “good” and “bad”. Thus, I deem it prudent to keep moral discussions on the object level** wherever I possibly can. Occasionally people object to any discussion of morality on the grounds that morality doesn’t exist, and in lieu of jumping over the forward dependency to explain that “exist” is not the right term to use here, I generally say, “But what do you do anyway?” and take the discussion back down to the object level. Paul Gowder, though, has pointed out that both the idea of choosing a googolplex dust specks in a googolplex eyes over 50 years of torture for one person, and the idea of “utilitarianism”, depend on “intuition”. He says I’ve argued that the two are not compatible, but charges me with failing to argue for the utilitarian intuitions that I appeal to. Now “intuition” is not how I would describe the computations that underlie human morality and distinguish us, as moralists, from an ideal philosopher of perfect emptiness and/or a rock. But I am okay with using the word “intuition” as a term of art, bearing in mind that “intuition” in this sense is not to be contrasted to reason, but is, rather, the cognitive building block out of which both long verbal arguments and fast perceptual arguments are constructed. **I see** the project of **morality as a project of renormalizing intuition.** We have intuitions about things that seem desirable or undesirable, intuitions about actions that are right or wrong, intuitions about how to resolve conflicting intuitions, intuitions about how to systematize specific intuitions into general principles. **Delete all** the **intuitions, and** you aren’t left with an ideal philosopher of perfect emptiness, **you’re left with a rock. Keep all your** specific **intuitions and** refuse to build upon the reflective ones, and you aren’t left with an ideal philosopher of perfect spontaneity and genuineness, **you’re left with a** grunting **caveperson** running in circles, due to cyclical preferences and similar inconsistencies. “Intuition”, as a term of art, is not a curse word when it comes to morality – there is nothing else to argue from. **Even modus ponens is an “intuition”** in this sense – **it**‘s **just** that modus ponens **still seems like a good idea after being** formalized, **reflected on**, extrapolated out to see if it has sensible consequences, etcetera. So that is “intuition”. However, Gowder did not say what he meant by “utilitarianism”. Does utilitarianism say… That right actions are strictly determined by good consequences? That praiseworthy actions depend on justifiable expectations of good consequences? That probabilities of consequences should normatively be discounted by their probability, so that a 50% probability of something bad should weigh exactly half as much in our tradeoffs? That virtuous actions always correspond to maximizing expected utility under some utility function? That two harmful events are worse than one? That two independent occurrences of a harm (not to the same person, not interacting with each other) are exactly twice as bad as one? That for any two harms A and B, with A much worse than B, there exists some tiny probability such that gambling on this probability of A is preferable to a certainty of B? If you say that I advocate something, or that my argument depends on something, and that it is wrong, do please specify what this thingy is… anyway, I accept 3, 5, 6, and 7, but not 4; I am not sure about the phrasing of 1; and 2 is true, I guess, but phrased in a rather solipsistic and selfish fashion: you should not worry about being praiseworthy. Now, what are the “intuitions” upon which my “utilitarianism” depends? This is a deepish sort of topic, but I’ll take a quick stab at it. First of all, it’s not just that someone presented me with a list of statements like those above, and I decided which ones sounded “intuitive”. Among other things, **if you try to violate** “**util**itarianism”, **you run into paradoxes, contradictions**, circular preferences, **and other** things that aren’t **symptoms of** moral wrongness so much as **moral incoherence**. After you think about moral problems for a while, and also find new truths about the world, and even discover disturbing facts about how you yourself work, you often end up with different moral opinions than when you started out. This does not quite define moral progress, but it is how we experience moral progress. As part of my experienced moral progress, I’ve drawn a conceptual separation between questions of type Where should we go? and questions of type How should we get there? (Could that be what Gowder means by saying I’m “utilitarian”?) The question of where a road goes – where it leads – you can answer by traveling the road and finding out. If you have a false belief about where the road leads, this falsity can be destroyed by the truth in a very direct and straightforward manner. When it comes to wanting to go to a particular place, this want is not entirely immune from the destructive powers of truth. You could go there and find that you regret it afterward (which does not define moral error, but is how we experience moral error). But, even so, wanting to be in a particular place seems worth distinguishing from wanting to take a particular road to a particular place. Our intuitions about where to go are arguable enough, but our intuitions about how to get there are frankly messed up. **After** the two hundred and eighty-seventh **research** study **showing that people will chop their own feet off if you frame the problem the wrong way, you start to distrust first impressions. When you’ve read** enough **research on scope insensitivity** – people will pay only 28% more to protect all 57 wilderness areas in Ontario than one area, **people will pay the same amount to save 50,000 lives as 5,000** lives… that sort of thing… Well, the worst case of scope insensitivity I’ve ever heard of was described here by Slovic: Other recent research shows similar results. Two Israeli psychologists asked people to contribute to a costly life-saving treatment. They could offer that contribution to a group of eight sick children, or to an individual child selected from the group. The target amount needed to save the child (or children) was the same in both cases. Contributions to individual group members far outweighed the contributions to the entire group. There’s other research along similar lines, but I’m just presenting one example, ’cause, y’know, eight examples would probably have less impact. If you know the general experimental paradigm, then the reason for the above behavior is pretty obvious – focusing your attention on a single child creates more emotional arousal than trying to distribute attention around eight children simultaneously. So people are willing to pay more to help one child than to help eight. Now, **you could** look at this intuition, and **think it was** revealing **some** kind of incredibly **deep moral truth** which shows that one child’s good fortune is somehow devalued by the other children’s good fortune. But what about the billions of other children in the world? Why isn’t it a bad idea to help this one child, when that causes the value of all the other children to go down? How can it be significantly better to have 1,329,342,410 happy children than 1,329,342,409, but then somewhat worse to have seven more at 1,329,342,417? **Or you could** look at that and **say: “The intuition is wrong: the brain can’t** successfully **multiply** by eight and get a larger quantity than it started with. **But it ought to**, normatively speaking.” And once you realize that the brain can’t multiply by eight, then the other cases of scope neglect stop seeming to reveal some fundamental truth about 50,000 lives being worth just the same effort as 5,000 lives, or whatever. You don’t get the impression you’re looking at the revelation of a deep moral truth about nonagglomerative utilities. It’s just that the brain doesn’t goddamn multiply. Quantities get thrown out the window. If you have $100 to spend, and you spend $20 each on each of 5 efforts to save 5,000 lives, you will do worse than if you spend $100 on a single effort to save 50,000 lives. Likewise if such choices are made by 10 different people, rather than the same person. As soon as you start believing that it is better to save 50,000 lives than 25,000 lives, that simple preference of final destinations has implications for the choice of paths, when you consider five different events that save 5,000 lives. (It is a general principle that Bayesians see no difference between the long-run answer and the short-run answer; you never get two different answers from computing the same question two different ways. But the long run is a helpful intuition pump, so I am talking about it anyway.) The aggregative valuation strategy of “shut up and multiply” arises from the simple preference to have more of something – to save as many lives as possible – when you have to describe general principles for choosing more than once, acting more than once, planning at more than one time. Aggregation also arises from claiming that the local choice to save one life doesn’t depend on how many lives already exist, far away on the other side of the planet, or far away on the other side of the universe. Three lives are one and one and one. No matter how many billions are doing better, or doing worse. 3 = 1 + 1 + 1, no matter what other quantities you add to both sides of the equation. And if you add another life you get 4 = 1 + 1 + 1 + 1. That’s aggregation. **When you’ve read** enough heuristics and **biases research, and**enough **coherence** and uniqueness **proofs for** Bayesian probabilities and **expected utility**, and you’ve seen the “Dutch book” and “money pump” effects that penalize trying to handle uncertain outcomes any other way, then **you don’t see** the **preference reversals** in the Allais Paradox **as** revealing some incredibly **deep moral truth** about the intrinsic value of certainty. **It** just **goes to show that the brain doesn’t** goddamn **multiply.** The primitive, perceptual intuitions that make a choice “feel good” don’t handle probabilistic pathways through time very skillfully, especially when the probabilities have been expressed symbolically rather than experienced as a frequency. So you reflect, devise more trustworthy logics, and think it through in words. When you see people insisting that no amount of money whatsoever is worth a single human life, and then driving an extra mile to save $10; or when you see people insisting that no amount of money is worth a decrement of health, and then choosing the cheapest health insurance available; then you don’t think that their protestations reveal some deep truth about incommensurable utilities. Part of it, clearly, is that **primitive intuitions don’t**successfully **diminish the emotional impact of** symbols standing for **small quantities** – anything you talk about seems like “an amount worth considering”. And part of it has to do with preferring unconditional social rules to conditional social rules. Conditional rules seem weaker, seem more subject to manipulation. If there’s any loophole that lets the government legally commit torture, then the government will drive a truck through that loophole. So it seems like there should be an unconditional social injunction against preferring money to life, and no “but” following it. Not even “but a thousand dollars isn’t worth a 0.0000000001% probability of saving a life”. Though the latter choice, of course, is revealed every time we sneeze without calling a doctor. The rhetoric of sacredness gets bonus points for seeming to express an unlimited commitment, an unconditional refusal that signals trustworthiness and refusal to compromise. So you conclude that moral rhetoric espouses qualitative distinctions, because espousing a quantitative tradeoff would sound like you were plotting to defect. On such occasions, people vigorously want to throw quantities out the window, and they get upset if you try to bring quantities back in, because quantities sound like conditions that would weaken the rule. But you don’t conclude that there are actually two tiers of utility with lexical ordering. You don’t conclude that there is actually an infinitely sharp moral gradient, some atom that moves a Planck distance (in our continuous physical universe) and sends a utility from 0 to infinity. You don’t conclude that utilities must be expressed using hyper-real numbers. Because the lower tier would simply vanish in any equation. It would never be worth the tiniest effort to recalculate for it. All decisions would be determined by the upper tier, and all thought spent thinking about the upper tier only, if the upper tier genuinely had lexical priority. As Peter Norvig once pointed out, if Asimov’s robots had strict priority for the First Law of Robotics (“A robot shall not harm a human being, nor through inaction allow a human being to come to harm”) then no robot’s behavior would ever show any sign of the other two Laws; there would always be some tiny First Law factor that would be sufficient to determine the decision. Whatever value is worth thinking about at all, must be worth trading off against all other values worth thinking about, because thought itself is a limited resource that must be traded off. When you reveal a value, you reveal a utility. I don’t say that morality should always be simple. I’ve already said that the meaning of music is more than happiness alone, more than just a pleasure center lighting up. I would rather see music composed by people than by nonsentient machine learning algorithms, so that someone should have the joy of composition; I care about the journey, as well as the destination. And I am ready to hear if you tell me that the value of music is deeper, and involves more complications, than I realize – that the valuation of this one event is more complex than I know. But that’s for one event. When it comes to multiplying by quantities and probabilities, complication is to be avoided – at least if you care more about the destination than the journey. **When you’ve reflected** on enough intuitions, **and corrected enough absurdities, you** start to **see a common denominator**, a meta-principle at work, **which one might phrase as “Shut up and multiply.”** Where music is concerned, I care about the journey. When lives are at stake, I shut up and multiply. It is more important that lives be saved, than that we conform to any particular ritual in saving them. And the optimal path to that destination is governed by laws that are simple, because they are math. **And that’s why I’m a utilitarian** – at least when I am doing something that is overwhelmingly more important than my own feelings about it – which is most of the time, because there are not many utilitarians, and many things left undone.

#### C] Mathematically outweighs.

MacAskill 14 [William, Oxford Philosopher and youngest tenured philosopher in the world, Normative Uncertainty, 2014]

The human race might go extinct from a number of causes: asteroids, supervolcanoes, runaway climate change, pandemics, nuclear war, and the development and use of dangerous new technologies such as synthetic biology, all pose risks (even if very small) to the continued survival of the human race.184 And different moral views give opposing answers to question of whether this would be a good or a bad thing. It might seem obvious that human extinction would be a very bad thing, both because of the loss of potential future lives, and because of the loss of the scientific and artistic progress that we would make in the future. But the issue is at least unclear. The continuation of the human race would be a mixed bag: inevitably, it would involve both upsides and downsides. And if one regards it as much more important to avoid bad things happening than to promote good things happening then one could plausibly regard human extinction as a good thing.For example, one might regard the prevention of bads as being in general more important that the promotion of goods, as defended historically by G. E. Moore,185 and more recently by Thomas Hurka.186 One could weight the prevention of suffering as being much more important that the promotion of happiness. Or one could weight the prevention of objective bads, such as war and genocide, as being much more important than the promotion of objective goods, such as scientific and artistic progress. If the human race continues its future will inevitably involve suffering as well as happiness, and objective bads as well as objective goods. So, if one weights the bads sufficiently heavily against the goods, or if one is sufficiently pessimistic about humanity’s ability to achieve good outcomes, then one will regard human extinction as a good thing.187 However, even if we believe in a moral view according to which human extinction would be a good thing, we still have strong reason to prevent near-term human extinction. To see this, we must note three points. First, we should note that the extinction of the human race is an extremely high stakes moral issue. Humanity could be around for a very long time: if humans survive as long as the median mammal species, we will last another two million years. On this estimate, the number of humans in existence in the The future, given that we don’t go extinct any time soon, would be 2×10^14. So if it is good to bring new people into existence, then it’s very good to prevent human extinction. Second, human extinction is by its nature an irreversible scenario. If we continue to exist, then we always have the option of letting ourselves go extinct in the future (or, perhaps more realistically, of considerably reducing population size). But if we go extinct, then we can’t magically bring ourselves back into existence at a later date. Third, we should expect ourselves to progress, morally, over the next few centuries, as we have progressed in the past. So we should expect that in a few centuries’ time we will have better evidence about how to evaluate human extinction than we currently have. Given these three factors, it would be better to prevent the near-term extinction of the human race, even if we thought that the extinction of the human race would actually be a very good thing. To make this concrete, I’ll give the following simple but illustrative model. Suppose that we have 0.8 credence that it is a bad thing to produce new people, and 0.2 certain that it’s a good thing to produce new people; and the degree to which it is good to produce new people, if it is good, is the same as the degree to which it is bad to produce new people, if it is bad. That is, I’m supposing, for simplicity, that we know that one new life has one unit of value; we just don’t know whether that unit is positive or negative. And let’s use our estimate of 2×10^14 people who would exist in the future, if we avoid near-term human extinction. Given our stipulated credences, the expected benefit of letting the human race go extinct now would be (.8-.2)×(2×10^14) = 1.2×(10^14). Suppose that, if we let the human race continue and did research for 300 years, we would know for certain whether or not additional people are of positive or negative value. If so, then with the credences above we should think it 80% likely that we will find out that it is a bad thing to produce new people, and 20% likely that we will find out that it’s a good thing to produce new people. So there’s an 80% chance of a loss of 3×(10^10) (because of the delay of letting the human race go extinct), the expected value of which is 2.4×(10^10). But there’s also a 20% chance of a gain of 2×(10^14), the expected value of which is 4×(10^13). That is, in expected value terms, the cost of waiting for a few hundred years is vanishingly small compared with the benefit of keeping one’s options open while one gains new information.

#### Revisionary intuitionism is true and proves util

Yudkowsky 08 [Eliezer Yudkowsky (research fellow of the Machine Intelligence Research Institute; he also writes Harry Potter fan fiction). “The ‘Intuitions’ Behind ‘Utilitarianism.’” 28 January 2008. LessWrong. http://lesswrong.com/lw/n9/the\_intuitions\_behind\_utilitarianism/]

I haven’t said much about metaethics – the nature of morality – because that has a forward dependency on a discussion of the Mind Projection Fallacy that I haven’t gotten to yet. I used to be very confused about metaethics. After my confusion finally cleared up, I did a postmortem on my previous thoughts. I found that my object-level moral reasoning had been valuable and my **meta-level moral reasoning had been** worse than **useless**. And this appears to be a general syndrome – **people do much better when discussing whether torture is** good or **bad than**when they discuss **the meaning of “good” and “bad”. Thus, I deem it prudent to keep moral discussions on the object level** wherever I possibly can. Occasionally people object to any discussion of morality on the grounds that morality doesn’t exist, and in lieu of jumping over the forward dependency to explain that “exist” is not the right term to use here, I generally say, “But what do you do anyway?” and take the discussion back down to the object level. Paul Gowder, though, has pointed out that both the idea of choosing a googolplex dust specks in a googolplex eyes over 50 years of torture for one person, and the idea of “utilitarianism”, depend on “intuition”. He says I’ve argued that the two are not compatible, but charges me with failing to argue for the utilitarian intuitions that I appeal to. Now “intuition” is not how I would describe the computations that underlie human morality and distinguish us, as moralists, from an ideal philosopher of perfect emptiness and/or a rock. But I am okay with using the word “intuition” as a term of art, bearing in mind that “intuition” in this sense is not to be contrasted to reason, but is, rather, the cognitive building block out of which both long verbal arguments and fast perceptual arguments are constructed. **I see** the project of **morality as a project of renormalizing intuition.** We have intuitions about things that seem desirable or undesirable, intuitions about actions that are right or wrong, intuitions about how to resolve conflicting intuitions, intuitions about how to systematize specific intuitions into general principles. **Delete all** the **intuitions, and** you aren’t left with an ideal philosopher of perfect emptiness, **you’re left with a rock. Keep all your** specific **intuitions and** refuse to build upon the reflective ones, and you aren’t left with an ideal philosopher of perfect spontaneity and genuineness, **you’re left with a** grunting **caveperson** running in circles, due to cyclical preferences and similar inconsistencies. “Intuition”, as a term of art, is not a curse word when it comes to morality – there is nothing else to argue from. **Even modus ponens is an “intuition”** in this sense – **it**‘s **just** that modus ponens **still seems like a good idea after being** formalized, **reflected on**, extrapolated out to see if it has sensible consequences, etcetera. So that is “intuition”. However, Gowder did not say what he meant by “utilitarianism”. Does utilitarianism say… That right actions are strictly determined by good consequences? That praiseworthy actions depend on justifiable expectations of good consequences? That probabilities of consequences should normatively be discounted by their probability, so that a 50% probability of something bad should weigh exactly half as much in our tradeoffs? That virtuous actions always correspond to maximizing expected utility under some utility function? That two harmful events are worse than one? That two independent occurrences of a harm (not to the same person, not interacting with each other) are exactly twice as bad as one? That for any two harms A and B, with A much worse than B, there exists some tiny probability such that gambling on this probability of A is preferable to a certainty of B? If you say that I advocate something, or that my argument depends on something, and that it is wrong, do please specify what this thingy is… anyway, I accept 3, 5, 6, and 7, but not 4; I am not sure about the phrasing of 1; and 2 is true, I guess, but phrased in a rather solipsistic and selfish fashion: you should not worry about being praiseworthy. Now, what are the “intuitions” upon which my “utilitarianism” depends? This is a deepish sort of topic, but I’ll take a quick stab at it. First of all, it’s not just that someone presented me with a list of statements like those above, and I decided which ones sounded “intuitive”. Among other things, **if you try to violate** “**util**itarianism”, **you run into paradoxes, contradictions**, circular preferences, **and other** things that aren’t **symptoms of** moral wrongness so much as **moral incoherence**. After you think about moral problems for a while, and also find new truths about the world, and even discover disturbing facts about how you yourself work, you often end up with different moral opinions than when you started out. This does not quite define moral progress, but it is how we experience moral progress. As part of my experienced moral progress, I’ve drawn a conceptual separation between questions of type Where should we go? and questions of type How should we get there? (Could that be what Gowder means by saying I’m “utilitarian”?) The question of where a road goes – where it leads – you can answer by traveling the road and finding out. If you have a false belief about where the road leads, this falsity can be destroyed by the truth in a very direct and straightforward manner. When it comes to wanting to go to a particular place, this want is not entirely immune from the destructive powers of truth. You could go there and find that you regret it afterward (which does not define moral error, but is how we experience moral error). But, even so, wanting to be in a particular place seems worth distinguishing from wanting to take a particular road to a particular place. Our intuitions about where to go are arguable enough, but our intuitions about how to get there are frankly messed up. **After** the two hundred and eighty-seventh **research** study **showing that people will chop their own feet off if you frame the problem the wrong way, you start to distrust first impressions. When you’ve read** enough **research on scope insensitivity** – people will pay only 28% more to protect all 57 wilderness areas in Ontario than one area, **people will pay the same amount to save 50,000 lives as 5,000** lives… that sort of thing… Well, the worst case of scope insensitivity I’ve ever heard of was described here by Slovic: Other recent research shows similar results. Two Israeli psychologists asked people to contribute to a costly life-saving treatment. They could offer that contribution to a group of eight sick children, or to an individual child selected from the group. The target amount needed to save the child (or children) was the same in both cases. Contributions to individual group members far outweighed the contributions to the entire group. There’s other research along similar lines, but I’m just presenting one example, ’cause, y’know, eight examples would probably have less impact. If you know the general experimental paradigm, then the reason for the above behavior is pretty obvious – focusing your attention on a single child creates more emotional arousal than trying to distribute attention around eight children simultaneously. So people are willing to pay more to help one child than to help eight. Now, **you could** look at this intuition, and **think it was** revealing **some** kind of incredibly **deep moral truth** which shows that one child’s good fortune is somehow devalued by the other children’s good fortune. But what about the billions of other children in the world? Why isn’t it a bad idea to help this one child, when that causes the value of all the other children to go down? How can it be significantly better to have 1,329,342,410 happy children than 1,329,342,409, but then somewhat worse to have seven more at 1,329,342,417? **Or you could** look at that and **say: “The intuition is wrong: the brain can’t** successfully **multiply** by eight and get a larger quantity than it started with. **But it ought to**, normatively speaking.” And once you realize that the brain can’t multiply by eight, then the other cases of scope neglect stop seeming to reveal some fundamental truth about 50,000 lives being worth just the same effort as 5,000 lives, or whatever. You don’t get the impression you’re looking at the revelation of a deep moral truth about nonagglomerative utilities. It’s just that the brain doesn’t goddamn multiply. Quantities get thrown out the window. If you have $100 to spend, and you spend $20 each on each of 5 efforts to save 5,000 lives, you will do worse than if you spend $100 on a single effort to save 50,000 lives. Likewise if such choices are made by 10 different people, rather than the same person. As soon as you start believing that it is better to save 50,000 lives than 25,000 lives, that simple preference of final destinations has implications for the choice of paths, when you consider five different events that save 5,000 lives. (It is a general principle that Bayesians see no difference between the long-run answer and the short-run answer; you never get two different answers from computing the same question two different ways. But the long run is a helpful intuition pump, so I am talking about it anyway.) The aggregative valuation strategy of “shut up and multiply” arises from the simple preference to have more of something – to save as many lives as possible – when you have to describe general principles for choosing more than once, acting more than once, planning at more than one time. Aggregation also arises from claiming that the local choice to save one life doesn’t depend on how many lives already exist, far away on the other side of the planet, or far away on the other side of the universe. Three lives are one and one and one. No matter how many billions are doing better, or doing worse. 3 = 1 + 1 + 1, no matter what other quantities you add to both sides of the equation. And if you add another life you get 4 = 1 + 1 + 1 + 1. That’s aggregation. **When you’ve read** enough heuristics and **biases research, and**enough **coherence** and uniqueness **proofs for** Bayesian probabilities and **expected utility**, and you’ve seen the “Dutch book” and “money pump” effects that penalize trying to handle uncertain outcomes any other way, then **you don’t see** the **preference reversals** in the Allais Paradox **as** revealing some incredibly **deep moral truth** about the intrinsic value of certainty. **It** just **goes to show that the brain doesn’t** goddamn **multiply.** The primitive, perceptual intuitions that make a choice “feel good” don’t handle probabilistic pathways through time very skillfully, especially when the probabilities have been expressed symbolically rather than experienced as a frequency. So you reflect, devise more trustworthy logics, and think it through in words. When you see people insisting that no amount of money whatsoever is worth a single human life, and then driving an extra mile to save $10; or when you see people insisting that no amount of money is worth a decrement of health, and then choosing the cheapest health insurance available; then you don’t think that their protestations reveal some deep truth about incommensurable utilities. Part of it, clearly, is that **primitive intuitions don’t**successfully **diminish the emotional impact of** symbols standing for **small quantities** – anything you talk about seems like “an amount worth considering”. And part of it has to do with preferring unconditional social rules to conditional social rules. Conditional rules seem weaker, seem more subject to manipulation. If there’s any loophole that lets the government legally commit torture, then the government will drive a truck through that loophole. So it seems like there should be an unconditional social injunction against preferring money to life, and no “but” following it. Not even “but a thousand dollars isn’t worth a 0.0000000001% probability of saving a life”. Though the latter choice, of course, is revealed every time we sneeze without calling a doctor. The rhetoric of sacredness gets bonus points for seeming to express an unlimited commitment, an unconditional refusal that signals trustworthiness and refusal to compromise. So you conclude that moral rhetoric espouses qualitative distinctions, because espousing a quantitative tradeoff would sound like you were plotting to defect. On such occasions, people vigorously want to throw quantities out the window, and they get upset if you try to bring quantities back in, because quantities sound like conditions that would weaken the rule. But you don’t conclude that there are actually two tiers of utility with lexical ordering. You don’t conclude that there is actually an infinitely sharp moral gradient, some atom that moves a Planck distance (in our continuous physical universe) and sends a utility from 0 to infinity. You don’t conclude that utilities must be expressed using hyper-real numbers. Because the lower tier would simply vanish in any equation. It would never be worth the tiniest effort to recalculate for it. All decisions would be determined by the upper tier, and all thought spent thinking about the upper tier only, if the upper tier genuinely had lexical priority. As Peter Norvig once pointed out, if Asimov’s robots had strict priority for the First Law of Robotics (“A robot shall not harm a human being, nor through inaction allow a human being to come to harm”) then no robot’s behavior would ever show any sign of the other two Laws; there would always be some tiny First Law factor that would be sufficient to determine the decision. Whatever value is worth thinking about at all, must be worth trading off against all other values worth thinking about, because thought itself is a limited resource that must be traded off. When you reveal a value, you reveal a utility. I don’t say that morality should always be simple. I’ve already said that the meaning of music is more than happiness alone, more than just a pleasure center lighting up. I would rather see music composed by people than by nonsentient machine learning algorithms, so that someone should have the joy of composition; I care about the journey, as well as the destination. And I am ready to hear if you tell me that the value of music is deeper, and involves more complications, than I realize – that the valuation of this one event is more complex than I know. But that’s for one event. When it comes to multiplying by quantities and probabilities, complication is to be avoided – at least if you care more about the destination than the journey. **When you’ve reflected** on enough intuitions, **and corrected enough absurdities, you** start to **see a common denominator**, a meta-principle at work, **which one might phrase as “Shut up and multiply.”** Where music is concerned, I care about the journey. When lives are at stake, I shut up and multiply. It is more important that lives be saved, than that we conform to any particular ritual in saving them. And the optimal path to that destination is governed by laws that are simple, because they are math. **And that’s why I’m a utilitarian** – at least when I am doing something that is overwhelmingly more important than my own feelings about it – which is most of the time, because there are not many utilitarians, and many things left undone.

#### 6. actor spec-

#### No intent-foresight distinction for states.

Enoch 07 Enoch, D [The Faculty of Law, The Hebrew Unviersity, Mount Scopus Campus, Jersusalem]. (2007). INTENDING, FORESEEING, AND THE STATE. Legal Theory, 13(02). doi:10.1017/s1352325207070048 https://www.cambridge.org/core/journals/legal-theory/article/intending-foreseeing-and-the-state/76B18896B94D5490ED0512D8E8DC54B2

The general difficulty of the intending-foreseeing distinction here stemmed, you will recall, from the feeling that attempting to pick and choose among the foreseen consequences of one’s actions those one is more and those one is less responsible for looks more like the preparation of a defense than like a genuine attempt to determine what is to be done. Hiding behind the intending-foreseeing distinction seems like an attempt to evade responsibility, and so thinking about the distinction in terms of responsibility serves 39. Anderson & Pildes, supra note 38. I will use this text as my example of an expressive theory here. 40. See id. at 1554, 1564. 41. For a general critique, see Mathew D. Adler, Expressive Theories of Law: A Skeptical Overview, 148 U. PA. L. REV. 1363 (1999–2000). 42. As Adler repeatedly notes, the understanding of expression Anderson & Pildes work with is amazingly broad, so that “To express an attitude through action is to act on the reasons the attitude gives us”; Anderson & Pildes, supra note 38, at 1510. If this is so, it seems that expression drops out of the picture and everything done with it can be done directly in terms of reasons. 43. This may be true of what Anderson and Pildes have in mind when they say that “expressive norms regulate actions by regulating the acceptable justifications for doing them”; id. at 1511. http://journals.cambridge.org Downloaded: 03 Aug 2014 IP address: 134.153.184.170 Intending, Foreseeing, and the State 91 to reduce even further the plausibility of attributing to it intrinsic moral significance. This consideration—however weighty in general—seems to me very weighty when applied to state action and to the decisions of state officials. For perhaps it may be argued that individuals are not required to undertake a global perspective, one that equally takes into account all foreseen consequences of their actions. Perhaps, in other words, individuals are entitled to (roughly) settle for having a good will, and beyond that let chips fall where they may. But this is precisely what stateswomen and statesmen—and certainly states—are not entitled to settle for.44 In making policy decisions, it is precisely the global (or at least statewide, or nationwide, or something of this sort) perspective that must be undertaken. Perhaps, for instance, an individual doctor is entitled to give her patient a scarce drug without thinking about tomorrow’s patients (I say “perhaps” because I am genuinely not sure about this), but surely when a state committee tries to formulate rules for the allocation of scarce medical drugs and treatments, it cannot hide behind the intending-foreseeing distinction, arguing that if it allows45 the doctor to give the drug to today’s patient, the death of tomorrow’s patient is merely foreseen and not intended. When making a policy-decision, this is clearly unacceptable. Or think about it this way (I follow Daryl Levinson here):46 perhaps restrictions on the responsibility of individuals are justified because individuals are autonomous, because much of the value in their lives comes from personal pursuits and relationships that are possible only if their responsibility for what goes on in the (more impersonal) world is restricted. But none of this is true of states and governments. They have no special relationships and pursuits, no personal interests, no autonomous lives to lead in anything like the sense in which these ideas are plausible when applied to individuals persons. So there is no reason to restrict the responsibility of states in anything like the way the responsibility of individuals is arguably restricted.47 States and state officials have much more comprehensive responsibilities than individuals do. Hiding behind the intending-foreseeing distinction thus more clearly constitutes an evasion of responsibility in the case of the former. So the evading-responsibility worry has much more force against the intending-foreseeing distinction when applied to state action than elsewhere.

### 1nc

#### Extinction is inevitable from future technology — nanotech, our simulation gets shut down, AI, biotech, particle accelerators, and black swans

Bruce **Sterling**, 6-1-20**18**, "When Nick Bostrom says “Bang”," WIRED, https://www.wired.com/beyond-the-beyond/2018/06/nick-bostrom-says-bang/

4.1 Deliberate misuse of nanotechnology

In a mature form, molecular nanotechnology will enable the construction of bacterium-scale self-replicating mechanical robots that can feed on dirt or other organic matter [22-25]. Such replicators could eat up the biosphere or destroy it by other means such as by poisoning it, burning it, or blocking out sunlight. A person of malicious intent in possession of this technology might cause the extinction of intelligent life on Earth by releasing such nanobots into the environment.[9]

The technology to produce a destructive nanobot seems considerably easier to develop than the technology to create an effective defense against such an attack (a global nanotech immune system, an “active shield” [23]). It is therefore likely that there will be a period of vulnerability during which this technology must be prevented from coming into the wrong hands. Yet the technology could prove hard to regulate, since it doesn’t require rare radioactive isotopes or large, easily identifiable manufacturing plants, as does production of nuclear weapons [23].

Even if effective defenses against a limited nanotech attack are developed before dangerous replicators are designed and acquired by suicidal regimes or terrorists, there will still be the danger of an arms race between states possessing nanotechnology. It has been argued [26] that molecular manufacturing would lead to both arms race instability and crisis instability, to a higher degree than was the case with nuclear weapons. Arms race instability means that there would be dominant incentives for each competitor to escalate its armaments, leading to a runaway arms race. Crisis instability means that there would be dominant incentives for striking first. Two roughly balanced rivals acquiring nanotechnology would, on this view, begin a massive buildup of armaments and weapons development programs that would continue until a crisis occurs and war breaks out, potentially causing global terminal destruction. That the arms race could have been predicted is no guarantee that an international security system will be created ahead of time to prevent this disaster from happening. The nuclear arms race between the US and the USSR was predicted but occurred nevertheless.

4.2 Nuclear holocaust[winter]

The US and Russia still have huge stockpiles of nuclear weapons. But would an all-out nuclear war really exterminate humankind? Note that: (i) For there to be an existential risk it suffices that we can’t be sure that it wouldn’t. (ii) The climatic effects of a large nuclear war are not well known (there is the possibility of a nuclear winter). (iii) Future arms races between other nations cannot be ruled out and these could lead to even greater arsenals than those present at the height of the Cold War. The world’s supply of plutonium has been increasing steadily to about two thousand tons, some ten times as much as remains tied up in warheads ([9], p. 26). (iv) Even if some humans survive the short-term effects of a nuclear war, it could lead to the collapse of civilization. A human race living under stone-age conditions may or may not be more resilient to extinction than other animal species.

4.3 We’re living in a simulation and it gets shut down

A case can be made that the hypothesis that we are living in a computer simulation should be given a significant probability [27]. The basic idea behind this so-called “Simulation argument” is that vast amounts of computing power may become available in the future (see e.g. [28,29]), and that it could be used, among other things, to run large numbers of fine-grained simulations of past human civilizations. Under some not-too-implausible assumptions, the result can be that almost all minds like ours are simulated minds, and that we should therefore assign a significant probability to being such computer-emulated minds rather than the (subjectively indistinguishable) minds of originally evolved creatures. And if we are, we suffer the risk that the simulation may be shut down at any time. A decision to terminate our simulation may be prompted by our actions or by exogenous factors.

While to some it may seem frivolous to list such a radical or “philosophical” hypothesis next the concrete threat of nuclear holocaust, we must seek to base these evaluations on reasons rather than untutored intuition. Until a refutation appears of the argument presented in [27], it would intellectually dishonest to neglect to mention simulation-shutdown as a potential extinction mode.

4.4 Badly programmed superintelligence

When we create the first superintelligent entity [28-34], we might make a mistake and give it goals that lead it to annihilate humankind, assuming its enormous intellectual advantage gives it the power to do so. For example, we could mistakenly elevate a subgoal to the status of a supergoal. We tell it to solve a mathematical problem, and it complies by turning all the matter in the solar system into a giant calculating device, in the process killing the person who asked the question. (For further analysis of this, see [35].)

4.5 Genetically engineered biological agent

With the fabulous advances in genetic technology currently taking place, it may become possible for a tyrant, terrorist, or ~~lunatic~~ to create a doomsday virus, an organism that combines long latency with high virulence and mortality [36].

Dangerous viruses can even be spawned unintentionally, as Australian researchers recently demonstrated when they created a modified mousepox virus with 100% mortality while trying to design a contraceptive virus for mice for use in pest control [37]. While this particular virus doesn’t affect humans, it is suspected that an analogous alteration would increase the mortality of the human smallpox virus. What underscores the future hazard here is that the research was quickly published in the open scientific literature [38]. It is hard to see how information generated in open biotech research programs could be contained no matter how grave the potential danger that it poses; and the same holds for research in nanotechnology.

Genetic medicine will also lead to better cures and vaccines, but there is no guarantee that defense will always keep pace with offense. (Even the accidentally created mousepox virus had a 50% mortality rate on vaccinated mice.) Eventually, worry about biological weapons may be put to rest through the development of nanomedicine, but while nanotechnology has enormous long-term potential for medicine [39] it carries its own hazards.

4.6 Accidental misuse of nanotechnology (“gray goo”)

The possibility of accidents can never be completely ruled out. However, there are many ways of making sure, through responsible engineering practices, that species-destroying accidents do not occur. One could avoid using self-replication; one could make nanobots dependent on some rare feedstock chemical that doesn’t exist in the wild; one could confine them to sealed environments; one could design them in such a way that any mutation was overwhelmingly likely to cause a nanobot to completely cease to function [40]. Accidental misuse is therefore a smaller concern than malicious misuse [23,25,41].

However, the distinction between the accidental and the deliberate can become blurred. While “in principle” it seems possible to make terminal nanotechnological accidents extremely improbable, the actual circumstances may not permit this ideal level of security to be realized. Compare nanotechnology with nuclear technology. From an engineering perspective, it is of course perfectly possible to use nuclear technology only for peaceful purposes such as nuclear reactors, which have a zero chance of destroying the whole planet. Yet in practice it may be very hard to avoid nuclear technology also being used to build nuclear weapons, leading to an arms race. With large nuclear arsenals on hair-trigger alert, there is inevitably a significant risk of accidental war. The same can happen with nanotechnology: it may be pressed into serving military objectives in a way that carries unavoidable risks of serious accidents.

In some situations it can even be strategically advantageous to deliberately make one’s technology or control systems risky, for example in order to make a “threat that leaves something to chance” [42].

4.7 Something unforeseen

We need a catch-all category. It would be foolish to be confident that we have already imagined and anticipated all significant risks. Future technological or scientific developments may very well reveal novel ways of destroying the world.

Some foreseen hazards (hence not members of the current category) which have been excluded from the list of bangs on grounds that they seem too unlikely to cause a global terminal disaster are: solar flares, supernovae, black hole explosions or mergers, gamma-ray bursts, galactic center outbursts, supervolcanos, loss of biodiversity, buildup of air pollution, gradual loss of human fertility, and various religious doomsday scenarios. The hypothesis that we will one day become “illuminated” and commit collective suicide or stop reproducing, as supporters of VHEMT (The Voluntary Human Extinction Movement) hope [43], appears unlikely. If it really were better not to exist (as Silenus told king Midas in the Greek myth, and as Arthur Schopenhauer argued [44] although for reasons specific to his philosophical system he didn’t advocate suicide), then we should not count this scenario as an existential disaster. The assumption that it is not worse to be alive should be regarded as an implicit assumption in the definition of Bangs. Erroneous collective suicide is an existential risk albeit one whose probability seems extremely slight. (For more on the ethics of human extinction, see chapter 4 of [9].)

4.8 Physics disasters

The Manhattan Project bomb-builders’ concern about an A-bomb-derived atmospheric conflagration has contemporary analogues.

There have been speculations that future high-energy particle accelerator experiments may cause a breakdown of a metastable vacuum state that our part of the cosmos might be in, converting it into a “true” vacuum of lower energy density [45]. This would result in an expanding bubble of total destruction that would sweep through the galaxy and beyond at the speed of light, tearing all matter apart as it proceeds.

Another conceivability is that accelerator experiments might produce negatively charged stable “strangelets” (a hypothetical form of nuclear matter) or create a mini black hole that would sink to the center of the Earth and start accreting the rest of the planet [46].

These outcomes seem to be impossible given our best current physical theories. But the reason we do the experiments is precisely that we don’t really know what will happen. A more reassuring argument is that the energy densities attained in present day accelerators are far lower than those that occur naturally in collisions between cosmic rays [46,47]. It’s possible, however, that factors other than energy density are relevant for these hypothetical processes, and that those factors will be brought together in novel ways in future experiments.

The main reason for concern in the “physics disasters” category is the meta-level observation that discoveries of all sorts of weird physical phenomena are made all the time, so even if right now all the particular physics disasters we have conceived of were absurdly improbable or impossible, there could be other more realistic failure-modes waiting to be uncovered. The ones listed here are merely illustrations of the general case.

#### Growth causes a global toxification crisis - risks extinction

Ehrlichand Ehrlich 13 [Paul R. Ehrlich, Professor of Biology and President of the Center for Conservation Biology at Stanford University, and Adjunct Professor at the University of Technology, Sydney, Anne H. Ehrlich, Senior Research Scientist in Biology at Stanford and focuses her research on policy issues related to the environment, “Can a collapse of global civilization be avoided?”, Proc Biol Sci. Mar 7, 2013; 280(1754), \\wyo-bb]

Another possible threat to the continuation of civilization is global toxification. Adverse symptoms of exposure to synthetic chemicals are making some scientists increasingly nervous about effects on the human population [77–79]. Should a global threat materialize, however, no planned mitigating responses (analogous to the ecologically and politically risky ‘geoengineering’ projects often proposed to ameliorate climate disruption [80]) are waiting in the wings ready for deployment. Much the same can be said about aspects of the epidemiological environment and the prospect of epidemics being enhanced by rapid population growth in immune-weakened societies, increased contact with animal reservoirs, high-speed transport and the misuse of antibiotics [81]. Nobel laureate Joshua Lederberg had great concern for the epidemic problem, famously stating, ‘The survival of the human species is not a preordained evolutionary program’ [82, p. 40]. Some precautionary steps that should be considered include forbidding the use of antibiotics as growth stimulators for livestock, building emergency stocks of key vaccines and drugs (such as Tamiflu), improving disease surveillance, expanding mothballed emergency medical facilities, preparing institutions for imposing quarantines and, of course, moving as rapidly as possible to humanely reduce the human population size. It has become increasingly clear that security has many dimensions beyond military security [83,84] and that breaches of environmental security could risk the end of global civilization.

#### War is inevitable---BUT, the longer we wait, the worse it gets.

Seth **Baum &** Anthony **Barrett 18**. Global Catastrophic Risk Institute. 2018. “A Model for the Impacts of Nuclear War.” SSRN Electronic Journal. Crossref, doi:10.2139/ssrn.3155983.

On the other end of the spectrum, the norm could be weaker. The Hiroshima and Nagasaki bombings provided a vivid and enduring image of the horrors of nuclear war—hence the norm can reasonably be described as a legacy of the bombings. Without this image, there would be less to motivate the norm. A weaker norm could in turn have led to a nuclear war occurring later, especially during a near-miss event like the Cuban missile crisis. A later nuclear war would likely be much more severe, assuming some significant buildup of nuclear arsenals and especially if “overkill” targeting was used. A new nuclear war could bring a similarly wide range of shifts in nuclear weapons norms. It could strengthen the norm, hastening nuclear disarmament. Already, there is a political initiative drawing attention to the humanitarian consequences of nuclear weapons use in order to promote a new treaty to ban nuclear weapons as a step towards complete nuclear disarmament (Borrie 2014). It is easy to imagine this initiative using any new nuclear attacks to advance their goals. Alternatively, it could weaken the norm, potentially leading to more and/or larger nuclear wars. This is a common concern, as seen for example in debates over low-yield bunker buster nuclear weapons (Nelson 2003). Given that the impacts of a large nuclear war could be extremely severe, a shift in nuclear weapons norms could easily be the single most consequential effect of a smaller nuclear war.

#### Nuke war won’t cause extinction---BUT, it’ll spur political will for meaningful disarmament.

Daniel **Deudney 18**. Associate Professor of Political Science at Johns Hopkins University. 03/15/2018. “The Great Debate.” The Oxford Handbook of International Security. www.oxfordhandbooks.com, doi:10.1093/oxfordhb/9780198777854.013.22. //reem

Although nuclear war is the oldest of these technogenic threats to civilization and human survival, and although important steps to restraint, particularly at the end of the Cold War, have been achieved, the nuclear world is increasingly changing in major ways, and in almost entirely dangerous directions. The third “bombs away” phase of the great debate on the nuclear-political question is more consequentially divided than in the first two phases. Even more ominously, most of the momentum lies with the forces that are pulling states toward nuclear-use, and with the radical actors bent on inflicting catastrophic damage on the leading states in the international system, particularly the United States. In contrast, the arms control project, although intellectually vibrant, is largely in retreat on the world political stage. The arms control settlement of the Cold War is unraveling, and the world public is more divided and distracted than ever. With the recent election of President Donald Trump, the United States, which has played such a dominant role in nuclear politics since its scientists invented these fiendish engines, now has an impulsive and uninformed leader, boding ill for nuclear restraint and effective crisis management. Given current trends, it is prudent to assume that sooner or later, and probably sooner, nuclear weapons will again be the used in war. But this bad news may contain a “silver lining” of good news. Unlike a general nuclear war that might have occurred during the Cold War, such a nuclear event now would probably not mark the end of civilization (or of humanity), due to the great reductions in nuclear forces achieved at the end of the Cold War. Furthermore, politics on “the day after” could have immense potential for positive change. The survivors would not be likely to envy the dead, but would surely have a greatly renewed resolution for “never again.” Such an event, completely unpredictable in its particulars, would unambiguously put the nuclear-political question back at the top of the world political agenda. It would unmistakeably remind leading states of their vulnerability It might also trigger more robust efforts to achieve the global regulation of nuclear capability. Like the bombings of Hiroshima and Nagasaki that did so much to catalyze the elevated concern for nuclear security in the early Cold War, and like the experience “at the brink” in the Cuban Missile Crisis of 1962, the now bubbling nuclear caldron holds the possibility of inaugurating a major period of institutional innovation and adjustment toward a fully “bombs away” future.

#### Industrial civilization wouldn’t recover.

Lewis **Dartnell 15**. UK Space Agency research fellow at the University of Leicester, working in astrobiology and the search for microbial life on Mars. His latest book is The Knowledge: How to Rebuild Our World from Scratch. 04-13-15. "Could we reboot a modern civilisation without fossil fuels? – Lewis Dartnell." Aeon. https://aeon.co/essays/could-we-reboot-a-modern-civilisation-without-fossil-fuels

Imagine that the world as we know it ends tomorrow. There’s a global catastrophe: a pandemic virus, an asteroid strike, or perhaps a nuclear holocaust. The vast majority of the human race perishes. Our civilisation collapses. The post-apocalyptic survivors find themselves in a devastated world of decaying, deserted cities and roving gangs of bandits looting and taking by force. Bad as things sound, that’s not the end for humanity. We bounce back. Sooner or later, peace and order emerge again, just as they have time and again through history. Stable communities take shape. They begin the agonising process of rebuilding their technological base from scratch. But here’s the question: how far could such a society rebuild? Is there any chance, for instance, that a post-apocalyptic society could reboot a technological civilisation? Let’s make the basis of this thought experiment a little more specific. Today, we have already consumed the most easily drainable crude oil and, particularly in Britain, much of the shallowest, most readily mined deposits of coal. Fossil fuels are central to the organisation of modern industrial society, just as they were central to its development. Those, by the way, are distinct roles: even if we could somehow do without fossil fuels now (which we can’t, quite), it’s a different question whether we could have got to where we are without ever having had them. So, would a society starting over on a planet stripped of its fossil fuel deposits have the chance to progress through its own Industrial Revolution? Or to phrase it another way, what might have happened if, for whatever reason, the Earth had never acquired its extensive underground deposits of coal and oil in the first place? Would our progress necessarily have halted in the 18th century, in a pre-industrial state? It’s easy to underestimate our current dependence on fossil fuels. In everyday life, their most visible use is the petrol or diesel pumped into the vehicles that fill our roads, and the coal and natural gas which fire the power stations that electrify our modern lives. But we also rely on a range of different industrial materials, and in most cases, high temperatures are required to transform the stuff we dig out of the ground or harvest from the landscape into something useful. You can’t smelt metal, make glass, roast the ingredients of concrete, or synthesise artificial fertiliser without a lot of heat. It is fossil fuels – coal, gas and oil – that provide most of this thermal energy. In fact, the problem is even worse than that. Many of the chemicals required in bulk to run the modern world, from pesticides to plastics, derive from the diverse organic compounds in crude oil. Given the dwindling reserves of crude oil left in the world, it could be argued that the most wasteful use for this limited resource is to simply burn it. We should be carefully preserving what’s left for the vital repertoire of valuable organic compounds it offers. But my topic here is not what we should do now. Presumably everybody knows that we must transition to a low-carbon economy one way or another. No, I want to answer a question whose interest is (let’s hope) more theoretical. Is the emergence of a technologically advanced civilisation necessarily contingent on the easy availability of ancient energy? Is it possible to build an industrialised civilisation without fossil fuels? And the answer to that question is: maybe – but it would be extremely difficult. Let’s see how. We’ll start with a natural thought. Many of our alternative energy technologies are already highly developed. Solar panels, for example, represent a good option today, and are appearing more and more on the roofs of houses and businesses. It’s tempting to think that a rebooted society could simply pick up where we leave off. Why couldn’t our civilisation 2.0 just start with renewables? Well, it could, in a very limited way. If you find yourself among the survivors in a post-apocalyptic world, you could scavenge enough working solar panels to keep your lifestyle electrified for a good long while. Without moving parts, photovoltaic cells require little maintenance and are remarkably resilient. They do deteriorate over time, though, from moisture penetrating the casing and from sunlight itself degrading the high-purity silicon layers. The electricity generated by a solar panel declines by about 1 per cent every year so, after a few generations, all our hand-me-down solar panels will have degraded to the point of uselessness. Then what? New ones would be fiendishly difficult to create from scratch. Solar panels are made from thin slices of extremely pure silicon, and although the raw material is common sand, it must be processed and refined using complex and precise techniques – the same technological capabilities, more or less, that we need for modern semiconductor electronics components. These techniques took a long time to develop, and would presumably take a long time to recover. So photovoltaic solar power would not be within the capability of a society early in the industrialisation process. Perhaps, though, we were on the right track by starting with electrical power. Most of our renewable-energy technologies produce electricity. In our own historical development, it so happens that the core phenomena of electricity were discovered in the first half of the 1800s, well after the early development of steam engines. Heavy industry was already committed to combustion-based machinery, and electricity has largely assumed a subsidiary role in the organisation of our economies ever since. But could that sequence have run the other way? Is there some developmental requirement that thermal energy must come first? On the face of it, it’s not beyond the bounds of possibility that a progressing society could construct electrical generators and couple them to simple windmills and waterwheels, later progressing to wind turbines and hydroelectric dams. In a world without fossil fuels, one might envisage an electrified civilisation that largely bypasses combustion engines, building its transport infrastructure around electric trains and trams for long-distance and urban transport. I say ‘largely’. We couldn’t get round it all together. When it comes to generating the white heat demanded by modern industry, there are few good options but to burn stuff While the electric motor could perhaps replace the coal-burning steam engine for mechanical applications, society, as we’ve already seen, also relies upon thermal energy to drive the essential chemical and physical transformations it needs. How could an industrialising society produce crucial building materials such as iron and steel, brick, mortar, cement and glass without resorting to deposits of coal? You can of course create heat from electricity. We already use electric ovens and kilns. Modern arc furnaces are used for producing cast iron or recycling steel. The problem isn’t so much that electricity can’t be used to heat things, but that for meaningful industrial activity you’ve got to generate prodigious amounts of it, which is challenging using only renewable energy sources such as wind and water. An alternative is to generate high temperatures using solar power directly. Rather than relying on photovoltaic panels, concentrated solar thermal farms use giant mirrors to focus the sun’s rays onto a small spot. The heat concentrated in this way can be exploited to drive certain chemical or industrial processes, or else to raise steam and drive a generator. Even so, it is difficult (for example) to produce the very high temperatures inside an iron-smelting blast furnace using such a system. What’s more, it goes without saying that the effectiveness of concentrated solar power depends strongly on the local climate. No, when it comes to generating the white heat demanded by modern industry, there are few good options but to burn stuff. But that doesn’t mean the stuff we burn necessarily has to be fossil fuels. Let’s take a quick detour into the pre-history of modern industry. Long before the adoption of coal, charcoal was widely used for smelting metals. In many respects it is superior: charcoal burns hotter than coal and contains far fewer impurities. In fact, coal’s impurities were a major delaying factor on the Industrial Revolution. Released during combustion, they can taint the product being heated. During smelting, sulphur contaminants can soak into the molten iron, making the metal brittle and unsafe to use. It took a long time to work out how to treat coal to make it useful for many industrial applications. And, in the meantime, charcoal worked perfectly well. And then, well, we stopped using it. In retrospect, that’s a pity. When it comes from a sustainable source, charcoal burning is essentially carbon-neutral, because it doesn’t release any new carbon into the atmosphere – not that this would have been a consideration for the early industrialists. But charcoal-based industry didn’t die out altogether. In fact, it survived to flourish in Brazil. Because it has substantial iron deposits but few coalmines, Brazil is the largest charcoal producer in the world and the ninth biggest steel producer. We aren’t talking about a cottage industry here, and this makes Brazil a very encouraging example for our thought experiment. The trees used in Brazil’s charcoal industry are mainly fast-growing eucalyptus, cultivated specifically for the purpose. The traditional method for creating charcoal is to pile chopped staves of air-dried timber into a great dome-shaped mound and then cover it with turf or soil to restrict airflow as the wood smoulders. The Brazilian enterprise has scaled up this traditional craft to an industrial operation. Dried timber is stacked into squat, cylindrical kilns, built of brick or masonry and arranged in long lines so that they can be easily filled and unloaded in sequence. The largest sites can sport hundreds of such kilns. Once filled, their entrances are sealed and a fire is lit from the top. The skill in charcoal production is to allow just enough air into the interior of the kiln. There must be enough combustion heat to drive out moisture and volatiles and to pyrolyse the wood, but not so much that you are left with nothing but a pile of ashes. The kiln attendant monitors the state of the burn by carefully watching the smoke seeping out of the top, opening air holes or sealing with clay as necessary to regulate the process. Brazil shows how the raw materials of modern civilisation can be supplied without reliance on fossil fuels Good things come to those who wait, and this wood pyrolysis process can take up to a week of carefully controlled smouldering. The same basic method has been used for millennia. However, the ends to which the fuel is put are distinctly modern. Brazilian charcoal is trucked out of the forests to the country’s blast furnaces where it is used to transform ore into pig iron. This pig iron is the basic ingredient of modern mass-produced steel. The Brazilian product is exported to countries such as China and the US where it becomes cars and trucks, sinks, bathtubs, and kitchen appliances. Around two-thirds of Brazilian charcoal comes from sustainable plantations, and so this modern-day practice has been dubbed ‘green steel’. Sadly, the final third is supplied by the non-sustainable felling of primary forest. Even so, the Brazilian case does provide an example of how the raw materials of modern civilisation can be supplied without reliance on fossil fuels. Another, related option might be wood gasification. The use of wood to provide heat is as old as mankind, and yet simply burning timber only uses about a third of its energy. The rest is lost when gases and vapours released by the burning process blow away in the wind. Under the right conditions, even smoke is combustible. We don’t want to waste it. Better than simple burning, then, is to drive the thermal breakdown of the wood and collect the gases. You can see the basic principle at work for yourself just by lighting a match. The luminous flame isn’t actually touching the matchwood: it dances above, with a clear gap in between. The flame actually feeds on the hot gases given off as the wood breaks down in the heat, and the gases combust only once they mix with oxygen from the air. Matches are fascinating when you look at them closely. Wartime gasifier cars could achieve about 1.5 miles per kilogram. Today’s designs improve upon this To release these gases in a controlled way, bake some timber in a closed container. Oxygen is restricted so that the wood doesn’t simply catch fire. Its complex molecules decompose through a process known as pyrolysis, and then the hot carbonised lumps of charcoal at the bottom of the container react with the breakdown products to produce flammable gases such as hydrogen and carbon monoxide. The resultant ‘producer gas’ is a versatile fuel: it can be stored or piped for use in heating or street lights, and is also suitable for use in complex machinery such as the internal combustion engine. More than a million gasifier-powered cars across the world kept civilian transport running during the oil shortages of the Second World War. In occupied Denmark, 95 per cent of all tractors, trucks and fishing boats were powered by wood-gas generators. The energy content of about 3 kg of wood (depending on its dryness and density) is equivalent to a litre of petrol, and the fuel consumption of a gasifier-powered car is given in miles per kilogram of wood rather than miles per gallon. Wartime gasifier cars could achieve about 1.5 miles per kilogram. Today’s designs improve upon this. But you can do a lot more with wood gases than just keep your vehicle on the road. It turns out to be suitable for any of the manufacturing processes needing heat that we looked at before, such as kilns for lime, cement or bricks. Wood gas generator units could easily power agricultural or industrial equipment, or pumps. Sweden and Denmark are world leaders in their use of sustainable forests and agricultural waste for turning the steam turbines in power stations. And once the steam has been used in their ‘Combined Heat and Power’ (CHP) electricity plants, it is piped to the surrounding towns and industries to heat them, allowing such CHP stations to approach 90 per cent energy efficiency. Such plants suggest a marvellous vision of industry wholly weaned from its dependency on fossil fuel. Is that our solution, then? Could our rebooting society run on wood, supplemented with electricity from renewable sources? Maybe so, if the population was fairly small. But here’s the catch. These options all presuppose that our survivors are able to construct efficient steam turbines, CHP stations and internal combustion engines. We know how to do all that, of course – but in the event of a civilisational collapse, who is to say that the knowledge won’t be lost? And if it is, what are the chances that our descendants could reconstruct it? In our own history, the first successful application of steam engines was in pumping out coal mines. This was a setting in which fuel was already abundant, so it didn’t matter that the first, primitive designs were terribly inefficient. The increased output of coal from the mines was used to first smelt and then forge more iron. Iron components were used to construct further steam engines, which were in turn used to pump mines or drive the blast furnaces at iron foundries. And of course, steam engines were themselves employed at machine shops to construct yet more steam engines. It was only once steam engines were being built and operated that subsequent engineers were able to devise ways to increase their efficiency and shrink fuel demands. They found ways to reduce their size and weight, adapting them for applications in transport or factory machinery. In other words, there was a positive feedback loop at the very core of the industrial revolution: the production of coal, iron and steam engines were all mutually supportive. In a world without readily mined coal, would there ever be the opportunity to test profligate prototypes of steam engines, even if they could mature and become more efficient over time? How feasible is it that a society could attain a sufficient understanding of thermodynamics, metallurgy and mechanics to make the precisely interacting components of an internal combustion engine, without first cutting its teeth on much simpler external combustion engines – the separate boiler and cylinder-piston of steam engines? It took a lot of energy to develop our technologies to their present heights, and presumably it would take a lot of energy to do it again. Fossil fuels are out. That means our future society will need an awful lot of timber. An industrial revolution without coal would be, at a minimum, very difficult In a temperate climate such as the UK’s, an acre of broadleaf trees produces about four to five tonnes of biomass fuel every year. If you cultivated fast-growing kinds such as willow or miscanthus grass, you could quadruple that. The trick to maximising timber production is to employ coppicing – cultivating trees such as ash or willow that resprout from their own stump, becoming ready for harvest again in five to 15 years. This way you can ensure a sustained supply of timber and not face an energy crisis once you’ve deforested your surroundings. But here’s the thing: coppicing was already a well-developed technique in pre-industrial Britain. It couldn’t meet all of the energy requirements of the burgeoning society. The central problem is that woodland, even when it is well-managed, competes with other land uses, principally agriculture. The double-whammy of development is that, as a society’s population grows, it requires more farmland to provide enough food and also greater timber production for energy. The two needs compete for largely the same land areas. We know how this played out in our own past. From the mid-16th century, Britain responded to these factors by increasing the exploitation of its coal fields – essentially harvesting the energy of ancient forests beneath the ground without compromising its agricultural output. The same energy provided by one hectare of coppice for a year is provided by about five to 10 tonnes of coal, and it can be dug out of the ground an awful lot quicker than waiting for the woodland to regrow. It is this limitation in the supply of thermal energy that would pose the biggest problem to a society trying to industrialise without easy access to fossil fuels. This is true in our post-apocalyptic scenario, and it would be equally true in any counterfactual world that never developed fossil fuels for whatever reason. For a society to stand any chance of industrialising under such conditions, it would have to focus its efforts in certain, very favourable natural environments: not the coal-island of 18th-century Britain, but perhaps areas of Scandinavia or Canada that combine fast-flowing streams for hydroelectric power and large areas of forest that can be harvested sustainably for thermal energy. Even so, an industrial revolution without coal would be, at a minimum, very difficult. Today, use of fossil fuels is actually growing, which is worrying for a number of reasons too familiar to rehearse here. Steps towards a low-carbon economy are vital. But we should also recognise how pivotal those accumulated reservoirs of thermal energy were in getting us to where we are. Maybe we could have made it the hard way. A slow-burn progression through the stages of mechanisation, supported by a combination of renewable electricity and sustainably grown biomass, might be possible after all. Then again, it might not. We’d better hope we can secure the future of our own civilisation, because we might have scuppered the chances of any society to follow in our wake.

#### Existential threats outweigh

GPP 17 (Global Priorities Project, Future of Humanity Institute at the University of Oxford, Ministry for Foreign Affairs of Finland, “Existential Risk: Diplomacy and Governance,” Global Priorities Project, 2017, <https://www.fhi.ox.ac.uk/wp-content/uploads/Existential-Risks-2017-01-23.pdf>

1.2. THE ETHICS OF EXISTENTIAL RISK In his book Reasons and Persons, Oxford philosopher Derek Parfit advanced an influential argument about the importance of avoiding extinction: I believe that if we destroy mankind, as we now can, this outcome will be much worse than most people think. Compare three outcomes: (1) Peace. (2) A nuclear war that kills 99% of the world’s existing population. (3) A nuclear war that kills 100%. (2) would be worse than (1), and (3) would be worse than (2). Which is the greater of these two differences? Most people believe that the greater difference is between (1) and (2). I believe that the difference between (2) and (3) is very much greater. ... The Earth will remain habitable for at least another billion years. Civilization began only a few thousand years ago. If we do not destroy mankind, these few thousand years may be only a tiny fraction of the whole of civilized human history. The difference between (2) and (3) may thus be the difference between this tiny fraction and all of the rest of this history. If we compare this possible history to a day, what has occurred so far is only a fraction of a second.65 In this argument, it seems that Parfit is assuming that the survivors of a nuclear war that kills 99% of the population would eventually be able to recover civilisation without long-term effect. As we have seen, this may not be a safe assumption – but for the purposes of this thought experiment, the point stands. What makes existential catastrophes especially bad is that they would “destroy the future,” as another Oxford philosopher, Nick Bostrom, puts it.66 This future could potentially be extremely long and full of flourishing, and would therefore have extremely large value. In standard risk analysis, when working out how to respond to risk, we work out the expected value of risk reduction, by weighing the probability that an action will prevent an adverse event against the severity of the event. Because the value of preventing existential catastrophe is so vast, even a tiny probability of prevention has huge expected value.67 Of course, there is persisting reasonable disagreement about ethics and there are a number of ways one might resist this conclusion.68 Therefore, it would be unjustified to be overconfident in Parfit and Bostrom’s argument. In some areas, government policy does give significant weight to future generations. For example, in assessing the risks of nuclear waste storage, governments have considered timeframes of thousands, hundreds of thousands, and even a million years.69 Justifications for this policy usually appeal to principles of intergenerational equity according to which future generations ought to get as much protection as current generations.70 Similarly, widely accepted norms of sustainable development require development that meets the needs of the current generation without compromising the ability of future generations to meet their own needs.71 However, when it comes to existential risk, it would seem that we fail to live up to principles of intergenerational equity. Existential catastrophe would not only give future generations less than the current generations; it would give them nothing. Indeed, reducing existential risk plausibly has a quite low cost for us in comparison with the huge expected value it has for future generations. In spite of this, relatively little is done to reduce existential risk. Unless we give up on norms of intergenerational equity, they give us a strong case for significantly increasing our efforts to reduce existential risks. 1.3. WHY EXISTENTIAL RISKS MAY BE SYSTEMATICALLY UNDERINVESTED IN, AND THE ROLE OF THE INTERNATIONAL COMMUNITY In spite of the importance of existential risk reduction, it probably receives less attention than is warranted. As a result, concerted international cooperation is required if we are to receive adequate protection from existential risks. 1.3.1. Why existential risks are likely to be underinvested in There are several reasons why existential risk reduction is likely to be underinvested in. Firstly, it is a global public good. Economic theory predicts that such goods tend to be underprovided. The benefits of existential risk reduction are widely and indivisibly dispersed around the globe from the countries responsible for taking action. Consequently, a country which reduces existential risk gains only a small portion of the benefits but bears the full brunt of the costs. Countries thus have strong incentives to free ride, receiving the benefits of risk reduction without contributing. As a result, too few do what is in the common interest. Secondly, as already suggested above, existential risk reduction is an intergenerational public good: most of the benefits are enjoyed by future generations who have no say in the political process. For these goods, the problem is temporal free riding: the current generation enjoys the benefits of inaction while future generations bear the costs. Thirdly, many existential risks, such as machine superintelligence, engineered pandemics, and solar geoengineering, pose an unprecedented and uncertain future threat. Consequently, it is hard to develop a satisfactory governance regime for them: there are few existing governance instruments which can be applied to these risks, and it is unclear what shape new instruments should take. In this way, our position with regard to these emerging risks is comparable to the one we faced when nuclear weapons first became available. Cognitive biases also lead people to underestimate existential risks. Since there have not been any catastrophes of this magnitude, these risks are not salient to politicians and the public.72 This is an example of the misapplication of the availability heuristic, a mental shortcut which assumes that something is important only if it can be readily recalled. Another cognitive bias affecting perceptions of existential risk is scope neglect. In a seminal 1992 study, three groups were asked how much they would be willing to pay to save 2,000, 20,000 or 200,000 birds from drowning in uncovered oil ponds. The groups answered $80, $78, and $88, respectively.73 In this case, the size of the benefits had little effect on the scale of the preferred response. People become numbed to the effect of saving lives when the numbers get too large. 74 Scope neglect is a particularly acute problem for existential risk because the numbers at stake are so large. Due to scope neglect, decision-makers are prone to treat existential risks in a similar way to problems which are less severe by many orders of magnitude. A wide range of other cognitive biases are likely to affect the evaluation of existential risks.75

#### Rigorous climate simulations prove that hydrophilic black carbon would cause to atmospheric precipitation – results in a rainout effect that quickly reverses nuclear cooling

Reisner et al. 18 (Jon Reisner – Climate and atmospheric scientist at the Los Alamos National Laboratory. Gennaro D’Angelo – Climate scientist at the Los Alamos National Laboratory, Research scientist at the SETI institute, Associate specialist at the University of California, Santa Cruz, NASA Postdoctoral Fellow at the NASA Ames Research Center, UKAFF Fellow at the University of Exeter. Eunmo Koo - Scientist at Applied Terrestrial, Energy, and Atmospheric Modeling (ATEAM) Team, in Computational Earth Science Group (EES-16) in Earth and Environmental Sciences Division and Co-Lead of Parallel Computing Summer Research Internship (PCSRI) program at the Los Alamos National Laboratory, former Staff research associate at UC Berkeley. Wesley Even - Computational scientist in the Computational Physics and Methods Group at Los Alamos National Laboratory. Matthew Hecht – Atmospheric scientist at the Los Alamos National Laboratory. Elizabeth Hunke - Lead developer for the Los Alamos Sea Ice Model (CICE) at the Los Alamos National Laboratory responsible for development and incorporation of new parameterizations, model testing and validation, computational performance, documentation, and consultation with external model users on all aspects of sea ice modeling, including interfacing with global climate and earth system models. Darin Comeau – Climate scientist at the Los Alamos National Laboratory. Randy Bos - Project leader at the Los Alamos National Laboratory, former Weapons Effects program manager at Tech-Source. James Cooley – Computational scientist at the Los Alamos National Laboratory specializing in weapons physics, emergency response, and computational physics. <MKIM> “Climate impact of a regional nuclear weapons exchange:An improved assessment based on detailed source calculations”. 3/16/18. DOA: 7/13/19. <https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/2017JD027331>)

\*BC = Black Carbon

The no-rubble simulation produces a significantly more intense fire, with more fire spread, and consequently a significantly stronger plume with larger amounts of BC reaching into the upper atmosphere than the simulation with rubble, illustrated in Figure 5. While the no-rubble simulation **represents the worst-case scenario** involving vigorous fire activity, **only a relatively small amount of carbon makes its way into the stratosphere** during the course of the simulation. But while small compared to the surface BC mass, stratospheric BC amounts from the current simulations are significantly higher than what would be expected from burning vegetation such as trees (Heilman et al., 2014), e.g., the higher energy density of the building fuels and the initial fluence from the weapon produce an intense response within HIGRAD with initial updrafts of order 100 m/s in the lower troposphere. Or, in comparison to a mass fire, wildfires will burn only a small amount of fuel in the corresponding time period (roughly 10 minutes) that a nuclear weapon fluence can effectively ignite a large area of fuel producing an impressive atmospheric response. Figure 6 shows vertical profiles of BC multiplied by 100 (number of cities involved in the exchange) from the two simulations. The total amount of BC produced is in line with previous estimates (about 3.69 Tg from no-rubble simulation); however, the majority of BC resides **below the stratosphere** (3.46 Tg below 12 km) and can be **readily impacted by scavenging from precipitation** either via pyro-cumulonimbus produced by the fire itself (not modeled) or other synoptic weather systems. While the impact on climate of these more realistic profiles will be explored in the next section, it should be mentioned that **these estimates are** still **at the high end**, considering the inherent simplifications in the combustion model that lead to **overestimating BC production**. 3.3 Climate Results Long-term climatic effects critically depend on the initial injection height of the soot, with larger quantities reaching the upper troposphere/lower stratosphere inducing a greater cooling impact because of longer residence times (Robock et al., 2007a). Absorption of solar radiation by the BC aerosol and its subsequent radiative cooling tends to heat the surrounding air, driving an initial upward diffusion of the soot plumes, an effect that depends on the initial aerosol concentrations. **Mixing and sedimentation** tend to **reduce this process**, and low altitude emissions are also significantly impacted by precipitation if aging of the BC aerosol occurs on sufficiently rapid timescales. But once at stratospheric altitudes, aerosol dilution via coagulation is hindered by low particulate concentrations (e.g., Robock et al., 2007a) and lofting to much higher altitudes is inhibited by gravitational settling in the low-density air (Stenke et al., 2013), resulting in more stable BC concentrations over long times. Of the initial BC mass released in the atmosphere, most of which is emitted below 9 km, **70% rains out within the first month** and 78%, or about 2.9 Tg, is removed within the first two months (Figure 7, solid line), with the remainder (about 0.8 Tg, dashed line) being transported above about 12 km (200 hPa) within the first week. This outcome differs from the findings of, e.g., Stenke et al. (2013, their high BC-load cases) and Mills et al. (2014), who found that most of the BC mass (between 60 and 70%) is lifted in the stratosphere within the first couple of weeks. This can also be seen in Figure 8 (red lines) and in Figure 9, which include results from our calculation with the initial BC distribution from Mills et al. (2014). In that case, only 30% of the initial BC mass rains out in the troposphere during the first two weeks after the exchange, with the remainder rising to the stratosphere. In the study of Mills et al. (2008) this percentage is somewhat smaller, about 20%, and smaller still in the experiments of Robock et al. (2007a) in which the soot is initially emitted in the upper troposphere or higher. In Figure 7, the e-folding timescale for the removal of tropospheric soot, here interpreted as the time required for an initial drop of a factor e, is about one week. This result compares favorably with the “LT” experiment of Robock et al. (2007a), considering 5 Tg of BC released in the lower troposphere, in which 50% of the aerosols are removed within two weeks. By contrast, the initial e-folding timescale for the removal of stratospheric soot in Figure 8 is about 4.2 years (blue solid line), compared to about 8.4 years for the calculation using Mills et al. (2014) initial BC emission (red solid line). The removal timescale from our forced ensemble simulations is close to those obtained by Mills et al. (2008) in their 1 Tg experiment, by Robock et al. (2007a) in their experiment “UT 1 Tg”, and © 2018 American Geophysical Union. All rights reserved. by Stenke et al. (2013) in their experiment “Exp1”, in all of which 1 Tg of soot was emitted in the atmosphere in the aftermath of the exchange. Notably, the e-folding timescale for the decline of the BC mass in Figure 8 (blue solid line) is also close to the value of about 4 years quoted by Pausata et al. (2016) for their long-term “intermediate” scenario. In that scenario, which is also based on 5 Tg of soot initially distributed as in Mills et al. (2014), the factor-of2 shorter residence time of the aerosols is caused by particle growth via coagulation of BC with organic carbon. Figure 9 shows the BC mass-mixing ratio, horizontally averaged over the globe, as a function of atmospheric pressure (height) and time. The BC distributions used in our simulations imply that the upward transport of particles is substantially less efficient compared to the case in which 5 Tg of BC is directly injected into the upper troposphere. The semiannual cycle of lofting and sinking of the aerosols is associated with atmospheric heating and cooling during the solstice in each hemisphere (Robock et al., 2007a). During the first year, the oscillation amplitude in our forced ensemble simulations is particularly large during the summer solstice, compared to that during the winter solstice (see bottom panel of Figure 9), because of the higher soot concentrations in the Northern Hemisphere, as can be seen in Figure 11 (see also left panel of Figure 12). Comparing the top and bottom panels of Figure 9, the BC reaches the highest altitudes during the first year in both cases, but the concentrations at 0.1 hPa in the top panel can be 200 times as large. Qualitatively, the difference can be understood in terms of the air temperature increase caused by BC radiation emission, which is several tens of kelvin degrees in the simulations of Robock et al. (2007a, see their Figure 4), Mills et al. (2008, see their Figure 5), Stenke et al. (2013, see high-load cases in their Figure 4), Mills et al. (2014, see their Figure 7), and Pausata et al. (2016, see one-day emission cases in their Figure 1), due to high BC concentrations, but it amounts to only about 10 K in our forced ensemble simulations, as illustrated in Figure 10. Results similar to those presented in Figure 10 were obtained from the experiment “Exp1” performed by Stenke et al. (2013, see their Figure 4). **In that scenario as well, somewhat less that 1 Tg of BC remained in the atmosphere after the initial rainout**. As mentioned before, the BC aerosol that remains in the atmosphere, lifted to stratospheric heights by the rising soot plumes, undergoes sedimentation over a timescale of several years (Figures 8 and 9). This mass represents the effective amount of BC that can force climatic changes over multi-year timescales. In the forced ensemble simulations, it is about 0.8 Tg after the initial rainout, whereas it is about 3.4 Tg in the simulation with an initial soot distribution as in Mills et al. (2014). Our more realistic source simulation involves the worstcase assumption of no-rubble (along with other assumptions) and hence serves as an upper bound for the impact on climate. As mentioned above and further discussed below, our scenario induces perturbations on the climate system similar to those found in previous studies in which the climatic response was driven by roughly 1 Tg of soot rising to stratospheric heights following the exchange. Figure 11 illustrates the vertically integrated mass-mixing ratio of BC over the globe, at various times after the exchange for the simulation using the initial BC distribution of Mills et al. (2014, upper panels) and as an average from the forced ensemble members (lower panels). All simulations predict enhanced concentrations at high latitudes during the first year after the exchange. In the cases shown in the top panels, however, these high concentrations persist for several years (see also Figure 1 of Mills et al., 2014), whereas the forced ensemble simulations indicate that the BC concentration starts to decline after the first year. In fact, in the simulation represented in the top panels, mass-mixing ratios larger than about 1 kg of BC © 2018 American Geophysical Union. All rights reserved. per Tg of air persist for well over 10 years after the exchange, whereas they only last for 3 years in our forced simulations (compare top and middle panels of Figure 9). After the first year, values drop below 3 kg BC/Tg air, whereas it takes about 8 years to reach these values in the simulation in the top panels (see also Robock et al., 2007a). Over crop-producing, midlatitude regions in the Northern Hemisphere, the BC loading is reduced from more than 0.8 kg BC/Tg air in the simulation in the top panels to 0.2-0.4 kg BC/Tg air in our forced simulations (see middle and right panels). The more rapid clearing of the atmosphere in the forced ensemble is also signaled by the soot optical depth in the visible radiation spectrum, which drops below values of 0.03 toward the second half of the first year at mid latitudes in the Northern Hemisphere, and everywhere on the globe after about 2.5 years (without never attaining this value in the Southern Hemisphere). In contrast, the soot optical depth in the calculation shown in the top panels of Figure 11 becomes smaller than 0.03 everywhere only after about 10 years. The two cases show a similar tendency, in that the BC optical depth is typically lower between latitudes 30º S-30º N than it is at other latitudes. This behavior is associated to the persistence of stratospheric soot toward high-latitudes and the Arctic/Antarctic regions, as illustrated by the zonally-averaged, column-integrated mass-mixing ratio of the BC in Figure 12 for both the forced ensemble simulations (left panel) and the simulation with an initial 5 Tg BC emission in the upper troposphere (right panel). The spread in the globally averaged (near) surface temperature of the atmosphere, from the control (left panel) and forced (right panel) ensembles, is displayed in Figure 13. For each month, the plots show the largest variations (i.e., maximum and minimum values), within each ensemble of values obtained for that month, relative to the mean value of that month. The plot also shows yearly-averaged data (thinner lines). The spread is comparable in the control and forced ensembles, with average values calculated over the 33-years run length of 0.4-0.5 K. This spread is also similar to the internal variability of the globally averaged surface temperature quoted for the NCAR Large Ensemble Community Project (Kay et al., 2015). These results imply that surface air temperature differences, between forced and control simulations, which lie within the spread may not be distinguished from effects due to internal variability of the two simulation ensembles. Figure 14 shows the difference in the globally averaged surface temperature of the atmosphere (top panel), net solar radiation flux at surface (middle panel), and precipitation rate (bottom panel), computed as the (forced minus control) difference in ensemble mean values. The sum of standard deviations from each ensemble is shaded. Differences are qualitatively significant over the first few years, when the anomalies lie near or outside the total standard deviation. Inside the shaded region, differences may not be distinguished from those arising from the internal variability of one or both ensembles. The surface solar flux (middle panel) is the quantity that appears most affected by the BC emission, with qualitatively significant differences persisting for about 5 years. The precipitation rate (bottom panel) is instead affected only at the very beginning of the simulations. The red lines in all panels show the results from the simulation applying the initial BC distribution of Mills et al. (2014), where the period of significant impact is much longer owing to the higher altitude of the initial soot distribution that results in longer residence times of the BC aerosol in the atmosphere. When yearly averages of the same quantities are performed over the IndiaPakistan region, the differences in ensemble mean values lie within the total standard deviations of the two ensembles. The results in Figure 14 can also be compared to the outcomes of other previous studies. In their experiment “UT 1 Tg”, Robock et al. (2007a) found that, when only 1 Tg of soot © 2018 American Geophysical Union. All rights reserved. remains in the atmosphere after the initial rainout, temperature and precipitation anomalies are about 20% of those obtained from their standard 5 Tg BC emission case. Therefore, the largest differences they observed, during the first few years after the exchange, were about - 0.3 K and -0.06 mm/day, respectively, comparable to the anomalies in the top and bottom panels of Figure 14. Their standard 5 Tg emission case resulted in a solar radiation flux anomaly at surface of -12 W/m2 after the second year (see their Figure 3), between 5 and 6 time as large as the corresponding anomalies from our ensembles shown in the middle panel. In their experiment “Exp1”, Stenke et al. (2013) reported global mean surface temperature anomalies not exceeding about 0.3 K in magnitude and precipitation anomalies hovering around -0.07 mm/day during the first few years, again consistent with the results of Figure 14. In a recent study, Pausata et al. (2016) considered the effects of an admixture of BC and organic carbon aerosols, both of which would be emitted in the atmosphere in the aftermath of a nuclear exchange. In particular, they concentrated on the effects of coagulation of these aerosol species and examined their climatic impacts. The initial BC distribution was as in Mills et al. (2014), although the soot burden was released in the atmosphere over time periods of various lengths. Most relevant to our and other previous work are their one-day emission scenarios. They found that, during the first year, the largest values of the atmospheric surface temperature anomalies ranged between about -0.5 and -1.3 K, those of the sea surface temperature anomalies ranged between -0.2 and -0.55 K, and those of the precipitation anomalies varied between -0.15 and -0.2 mm/day. All these ranges are compatible with our results shown in Figure 14 as red lines and with those of Mills et al. (2014, see their Figures 3 and 6). As already mentioned in Section 2.3, the net solar flux anomalies at surface are also consistent. This overall agreement suggests that the **inclusion of organic carbon aerosols, and** ensuing **coagulation** with BC, **should not dramatically alter the climatic effects** resulting from our forced ensemble simulations. Moreover, aerosol growth would likely **shorten the residence time of the BC particulate in the atmosphere** (Pausata et al., 2016), possibly **reducing the duration of these effects.**

#### Isolated island populations repopulate Earth after radiation and nuclear winter – bunkers and submarines expand the likelihood of survival

Turchin and Green 18 (Alexey Turchin – Scientist for the Foundation Science for Life Extension in Moscow, Russia, Founder of Digital Immortality Now, author of several books and articles on the topics of existential risks and life extension. Brian Patrick Green – Director of technology ethics at the Markkula Center for Applied Ethics, teaches AI ethics in the Graduate School of Engineering at Santa Clara University. <MKIM> “Islands as refuges for surviving global catastrophes”. September 2018. DOA: 7/20/19. https://www.emerald.com/insight/content/doi/10.1108/FS-04-2018-0031/full/html?fullSc=1&mbSc=1&fullSc=1)

Different types of possible catastrophes suggest different scenarios for how survival could happen on an island. What is important is that the island should have properties which protect against the specific dangers of particular global catastrophic risks. Specifically, different islands will provide protection against different risks, and their natural diversity will contribute to a higher total level of protection: **Quarantined island survives pandemic** . An island could impose effective quarantine if it is sufficiently remote and simultaneously able to protect itself, possibly using military ships and air defense. **Far northern aboriginal people survive an ice age**. Many far northern people have adapted to survive in extremely cold and dangerous environments, and under the right circumstances could potentially survive the return of an ice age. However, their cultures are endangered by globalization. If these people become dependent on the products of modern civilization, such as rifles and motor boats, and lose their native survival skills, then their likelihood of surviving the collapse of the outside world would decrease. Therefore, preservation of their survival skills may be important as a defense against the risks connected with **extreme cooling**. Remote polar island with high mountains survives brief global warming of median surface temperatures, up to 50˚C. There is a theory that the climates of planets similar to the Earth could have several semi-stable temperature levels (Popp et al., 2016). If so, because of climate change, the Earth could transition to a second semi-stable state with a median global temperature of around 330 K, about 60˚C, or about 45˚C above current global mean temperatures. But even in this climate, **some regions of Earth could still be survivable for humans**, such as the Himalayan plateau at elevations above 4,000 m, but below 6,000 (where oxygen deficiency becomes a problem), or on polar islands with mountains (however, global warming affects polar regions more than equatorial regions, and northern island will experience more effects of climate change, including thawing permafrost and possible landslides because of wetter weather). In the tropics, the combination of increased humidity and temperature may increase the wet bulb temperature above 36˚C, especially on islands, where sea moisture is readily available. In such conditions, proper human perspiration becomes impossible (Sherwood and Huber, 2010), and there will likely be increased mortality and morbidity because of tropical diseases. If temperatures later returned to normal – either naturally or through climate engineering – **the rest of the Earth could be repopulated**. ‘‘Swiss Family Robinsons’’ survive on a tropical island, unnoticed by a military robot ‘‘mutiny’’. Most AI researchers ignore medium-term AI risks, which are neither near-term risks, like unemployment, nor remote risks, like AI superintelligence. But a large drone army – if one were produced – could receive a wrong command or be infected by a computer virus, leading it to attack people indiscriminately. Remote islands without robots could provide protection in this case, allowing survival until such a drone army ran out of batteries, fuel, ammunition or other supplies: Primitive tribe survives civilizational collapse. The inhabitants of **North Sentinel Island**, near the Andaman Islands in the Indian Ocean, are hostile and uncontacted. **The Sentinelese survived the 2004 Indian Ocean tsunami apparently unaffected** (Voanews, 2009), and if the rest of humanity disappear, **they might well continue their existence without change.** Tropical Island survives extreme global nuclear winter and glaciation event. Were a **nuclear**, bolide impactor or volcanic “**winter**” scenario to unfold, these islands would remain surrounded by Warm Ocean, and local volcanism or other energy sources might provide heat, energy and food. Such island refuges may have helped life on Earth survive during the **“Snowball Earth”** event in Earth’s distant past (Hoffman et al., 1998). Remote island base for project “Yellow submarine”. Some catastrophic risks such as a gamma ray burst, a global nuclear war with high radiological contamination or multiple pandemics might be best survived **underwater in nuclear submarines** (Turchin and Green, 2017). However, after a catastrophe, the submarine with survivors would eventually need a place to dock, and an island with some prepared amenities would be a reasonable starting point for rebuilding civilization. Bunker on remote island. For risks which include multiple or complex catastrophes, such as a bolide impact, extreme volcanism, tsunamis, multiple pandemics and nuclear war with radiological contamination, **island refuges could be strengthened with bunkers**. Richard Branson survived hurricane Irma on his own island in 2017 by seeking refuge in his concrete wine cellar (Clifford, 2017). Bunkers on islands would have higher survivability compared to those close to population centers, as they will be neither a military target nor as accessible to looters or unintentionally dangerous (e.g. infected) refugees. These bunkers could potentially be connected to water sources by underwater pipes, and passages could provide cooling, access and even oxygen and food sources.