# AC

## Advocacy:

Developing countries should accept the Precautionary Principle in contexts where environmental protection conflicts with resource extraction, consistent with the guidelines of clean production, zero discharge, and reverse onus.

**Aff gets RVIs** on I meets and counter-interps because

(a) 1AR time skew means I can’t cover theory and still have a fair shot at substance.

(b) no-risk theory gives her a free source of no-risk offense which allows her to moot the AC.

No generic defense. The plan rectifies any flaws with the PP.

**Collins 5**

Lynda M. Collins (environmental attorney). “Strange Bedfellows? The Precautionary Principle and Toxic Tort: A Tort Paradigm for the 21st Century.” 2005 Environmental Law Institute®, Washington, DC.

Prof. Thomas McGarity argues that the “essence” of the precautionary principle “can be captured in three familiar phrases: look before you leap; it is better to be safe than sorry; and when in doubt, err on the side of safety.”52 The precautionary principle rejects risk-based assumptions about the capacity of the environment (including human bodies) to assimilate contaminants on the one hand and the ability of science to predict this capacity on the other.53 From a normative perspective, the precautionary principle recognizes an ethical imperative to do the best we can to avoid creating risks of serious or irreversible harm. **The precautionary principle has been criticized for its amorphous nature**, and **Thornton** concedes that the principle is too vague to function as a regulatory standard.55 Accordingly, he **proposes three** additional **policy guidelines to provide specific guidance on** the **implementation** of the precautionary principle.56 **These are clean production, zero discharge, and reverse onus.** Clean Production mandates a shift from our current focus on pollution control to a proactive and precautionary rule favoring pollution prevention.57 It requires that industry make use of the most benign available methods and materials and seek to prevent the release of hazardous materials by preventing their production in the first place.58 The policy of zero discharge prohibits absolutely the release of persistent and/or bioaccumulative chemicals.59 Reverse onus would invert current regulatory policy by requiring proponents of synthetic chemicals to demonstrate their safety before being permitted to produce and release them.60 As Thornton points out, **this is the policy currently in force with** respect to **pharmaceuticals.**61 Further **precedent** for such an approach **can be found in environmental assessment statutes**, which require that proponents of physical projects elucidate the environmental impact of proposed projects before approval can be granted.62 In contrast to risk paradigm approaches, “[r]ather than asking how much toxic pollution we can live with, the precautionary principle [focuses on the] kind of . . . world we want to live in, and provides a decision-making framework for getting there.”63 **Thornton’s** ecological **paradigm provides a viable policy framework to guide** the **implementation** of the precautionary principle in statutory environmental law. The remainder of this Article examines the extent to which tort law can or should, in turn, embrace this ecological paradigm in its own treatment of environmental wrongs.

**The PP shifts the legal burden in environmental cases towards the plaintiffs. That’s key to justice in tort law.**

**Collins 5**

Lynda M. Collins (environmental attorney). “Strange Bedfellows? The Precautionary Principle and Toxic Tort: A Tort Paradigm for the 21st Century.” 2005 Environmental Law Institute®, Washington, DC.

**Scientists and legal scholars have made a compelling case for a precautionary approach to** statutory **environmental regulation. Common-law tort must** also **embrace a precautionary paradigm** if it is to realize its potential **to play a meaningful role in arresting environmental degradation and doing justice to the victims of environmental torts.** Given the important role of tort law in our hybrid legal system of environmental protection, it is incumbent on scholars and jurists alike to ensure that tort improves its ability to respond to environmental wrongdoing. Indeed, tort’s response to environmental pollution will play a significant role in determining the extent to which ordinary people are exposed to untested toxic chemicals in the coming decades.153 In order to contribute effectively to environmental protection, **tort needs to embrace the** ecological paradigm and abandon the **scientifically questionable assumptions of** the **risk** and injury **paradigms.** As a starting point, the single **cause-of-action** rule **should be relaxed in all toxic tort cases** in order to allow for early deterrence and full compensation of injured plaintiffs. **For cases in which a plaintiff has not yet fallen ill, toxic battery** (coupled with a relaxation of the single cause-of-action rule) **offers a**n appropriately **precautionary tort response.** For cases in which health damage has already occurred, **the ecological paradigm demands a reversal of the burden of proof** on generic causation where a defendant has failed to adequately research its product. **All three approaches vindicate the individual’s interest in bodily integrity, and, by proxy, our collective** interest in **ecological integrity.** Until trees have standing,154 these approaches may be the most effective way for the common law to do its part in safeguarding the environment.

## Contention 1 is Food Crises

Prioritizing resource extraction makes collapse of civilization and food crises inevitable. Sustainable development solves.

**Luntz 3-19** writes[[1]](#footnote-1)

**Our** industrial **civilization faces the same threats of collapse that** earlier versions such as **the Mayans experienced**, a study to be published in Ecological Economics has warned. The idea is far from new, but the authors have put new rigor to the study of how so many previous societies collapsed, and why ours could follow. Lead author Mr Safa Motesharrei is no wild-eyed conspiracy theorist. Motesharrei is a graduate student in mathematics at the National Socio-Environmental Synthesis Center, a National Science Foundation-supported institution, and the research was done with funding from NASA's Goddard Space Flight Center. "The fall of the Roman Empire, and the equally (if not more) advanced Han, Mauryan, and Gupta Empires, as well as so many advanced Mesopotamian Empires, are all testimony to the fact that advanced, sophisticated, complex, and creative civilizations can be both fragile and impermanent," the forthcoming paper states Two key social features are identified that contributed to the collapse of every civilization studied: “The stretching of resources due to the strain placed on the ecological carrying capacity," and "The economic stratification of society into Elites [rich] and Masses (or "Commoners") [poor]". If these look familiar, so do the factors that make up the resource side of the equation, with climatic change, and scarcity of water and energy key among them, although for others climate variation was a matter of bad luck, rather than their own actions. The model **Motesharrei** used, Human And Nature Dynamics (HANDY), **explores the relationship between population and resources**, drawing heavily on predator-prey models used by ecologists. Four key factors were included in the model: Elites, Commoners, nature and wealth. Equations of how these interact were created with varying inputs. The outcomes were not pretty. The timing and shape of collapses varied, but the **societies that most closely resembled our own doomed themselves, through overuse of resources** exacerbated by economic stratification. In one scenario many commoners do make it into the elite population at year 750, but the “scarcity of workers” caused a collapse by year 1000. In another so many of the Earth's resources are consumed that society, and the ecology of the planet, are doomed by the year 500. “It is important to note that in both of these scenarios, the Elites — due to their wealth — do not suffer the detrimental effects of the environmental collapse until much later than the Commoners,” the paper notes. If those year numbers seem comfortingly far off, be aware that the year zero in these models is well behind us. Nevertheless, contrary to much of the reporting, the model does not provide a useful timeline for when we can expect to see the world we live in turn into something that resembles a post-apocalyptic nightmare, although **studies of** the **convergence of climate and resource challenges suggest we may witness drastic food crises within a little over a decade**. In every economic bubble people looking back to past crashes are told “this time it is different”. Certainly some things have changed for modern civilization compared to the others Motesharrei has looked at. Technological developments that provide access to greater resources is the most frequently mentioned difference. Motesharrei responds, “**Tech**nological change **can raise** the **efficiency of resource use, but** it **also** tends to raise both **per capita** resource **consumption and the scale of** resource **extraction**, so that, absent policy effects, the increases in consumption often compensate for the increased efficiency of resource use.” One advantage we do have, however, is much greater knowledge of what has gone wrong in the past, and therefore the capacity to build models like HANDY. In a presentation of an earlier draft of this work in 2012 Motesharrei noted, “Simple models provide a great intuition and can teach us invaluable points. It is crucial to have a measure that can give us an early warning of collapse. Carrying Capacity tells us when overshoot happens, and this can be defined by noticing the decline in wealth.” Some coverage of the announcement has described disaster as inevitable, but that is not the paper's conclusion at all. “**Collapse can be avoided** and population can reach equilibrium **if** the **per capita** rate of **depletion** of nature **is reduced to a sustainable level**, and if resources are distributed in a reasonably equitable fashion,” it argues. Although the study has reportedly passed peer review it is yet to be published. It received global attention after a pre-release version was provided to The Guardian.

Food crisis causes conflict and instability which kills compromise over AI.

**Tomasik 13** writes[[2]](#footnote-2)

As a general rule, less crop cultivation now probably implies more food stability in the future. One clear example is in the area of topsoil loss as discussed above. John Crawford explains: water will reach a crisis point. This issue is already causing conflicts in India, China, Pakistan and the Middle East and before climate change and food security really hit, the next wars are likely to be fought over unsustainable irrigation. Even moderately degraded soil will hold less than half of the water than healthy soil in the same location. If you're irrigating a crop, you need water to stay in the soil close to the plant roots. [...] Soil erosion is most serious in China, Africa, India and parts of South America. **If the food supply goes down,** then obviously, **the price goes up. The crisis** points **will hit the poorest countries hardest, in particular those which rely on imports**: Egypt, for example, is almost entirely dependent on imports of wheat. The capacity of the planet to produce food is already causing conflict. A lot of people argue that food price hikes caused the Arab spring, and may even have contributed to the recent violence following the release of an anti-Islam film. In general, consumption of more food crops implies higher prices on the world market. From "Food Insecurity and Violent Conflict: Causes, Consequences, and Addressing the Challenges" by Henk-Jan Brinkman and Cullen S. Hendrix (p. 4): **is food insecurity** itself **a cause of conflict? Based on** a review of **recent research, the answer is a highly qualified yes**. **Food insecurity, especially when caused by higher food prices, heightens** the risk of **democratic breakdown, civil conflict**, protest, rioting, **and communal conflict**. The evidence linking food insecurity to interstate conflict is less strong, though **there is** some **historical evidence linking declining ag**ricultural **yields to** periods of **regional conflict in Europe and Asia**. That said, the effects of these rebellions on democracy can be both negative and positive (p. 7): Food insecurity, proxied by low availability of calories for consumption per capita, makes democratic breakdown more likely, especially in higher-income countries, where people expect there to be larger social surpluses that could be invested to reduce food insecurity (Reenock, Bernhard and Sobek, 2007). Though statistical evidence is lacking, rising food prices have been implicated in the wave of demonstrations and transitions from authoritarian rule to fledgling democracy in some countries across North Africa and the Middle East in 2011. There are some historical precedents for this: a bad harvest in 1788 led to high food prices in France, which caused rioting and contributed to the French revolution in 1789; and the wave of political upheaval that swept Europe in 1848 was at least in part a response to food scarcity, coming after three below-average harvests across the continent (Berger and Spoerer 2001). Most of these conflicts occur in poor countries and so are less likely to influence AGI arms races among major world powers. Still, it seems plausible that the **destabilizing consequences of environmental degradation are net harmful for compromise prospects among** the **big players in AGI development** in the long term.

International conflict risks an AI arms race involving military robotics.

**Tomasik 4-3** writes[[3]](#footnote-3)

**Government AI development could go wrong** in several ways. Probably most on LW feel the prevailing scenario is that **governments would botch the process by not realizing** the **risks at hand. It's also possible that governments would use** the **AI for** malevolent, **totalitarian purposes**. It seems that **both** of these bad scenarios **would be exacerbated by international conflict. Greater hostility means countries are more inclined to use AI as a weapon**. Indeed, **whoever builds the first AI can take over the world, which makes building AI the ultimate arms race**. A USA-China race is one reasonable possibility. Arms races encourage risk-taking -- being willing to skimp on safety measures to improve your odds of winning ("Racing to the Precipice"). In addition, the weaponization of AI could lead to worse expected outcomes in general. CEV seems to have less hope of success in a Cold War scenario. ("What? You want to include the evil Chinese in your CEV??") With a pure CEV, presumably it would eventually count Chinese values even if it started with just Americans, because people would become more enlightened during the process. However, when we imagine more crude democratic decision outcomes, this becomes less likely. 5. Ways to avoid an arms race Averting an AI arms race seems to be an important topic for research. It could be partly informed by the Cold War and other nuclear arms races, 'President Reagan and General Secretary Gorbachev signing the INF Treaty in the East Room of the White House.' By White House Photographic Office [Public domain], via Wikimedia Commons: https://commons.wikimedia.org/wiki/File:Reagan\_and\_Gorbachev\_signing.jpg as well as by other efforts at nonproliferation of chemical and biological weapons. Forthcoming robotic and nanotech weapons might be even better analogues of AI arms races than nuclear weapons because these newer technologies can be built more secretly and used in a more targeted fashion. Apart from more robust arms control, other factors might help: Improved international institutions like the UN, allowing for better enforcement against defection by one state. In the long run, a scenario of global governance would likely be ideal for strengthening international cooperation, just like nation states reduce intra-state violence. Better construction and enforcement of nonproliferation treaties. Improved game theory and international-relations scholarship on the causes of arms races and how to avert them. (For instance, arms races have sometimes been modeled as iterated prisoner's dilemmas with imperfect information.) How to improve verification, which has historically been a weak point for nuclear arms control. (The concern is that if you haven't verified well enough, the other side might be arming while you're not.) Moral tolerance and multicultural perspective, aiming to reduce people's sense of nationalism. (In the limit where neither Americans nor Chinese cared which government won the race, there would be no point in having the race.) Improved trade, democracy, and other forces that historically have reduced the likelihood of war. 6. Are these efforts cost-effective? World peace is hardly a goal unique to effective altruists (EAs), so we shouldn't necessarily expect low-hanging fruit. On the other hand, projects like nuclear nonproliferation seem relatively underfunded even compared with anti-poverty charities. I suspect more direct MIRI-type research has higher expected value, but among EAs who don't want to fund MIRI specifically, encouraging donations toward international cooperation could be valuable, since it's certainly a more mainstream cause. I wonder if GiveWell would consider studying global cooperation specifically beyond its indirect relationship with catastrophic risks. 7. Should we publicize AI arms races? When I mentioned this topic to a friend, he pointed out that we might not want the idea of AI arms races too widely known, because then governments might take the concern more seriously and therefore start the race earlier -- giving us less time to prepare and less time to work on FAI in the meanwhile. From David Chalmers, "The Singularity: A Philosophical Analysis" (footnote 14): When I discussed these issues with cadets and staff at the West Point Military Academy, the question arose as to whether the US military or other branches of the government might attempt to prevent the creation of AI or AI+, due to the risks of an intelligence explosion. The consensus was that they would not, as such prevention would only increase the chances that AI or AI+ would first be created by a foreign power. One might even expect an AI arms race at some point, once the potential consequences of an intelligence explosion are registered. According to this reasoning, although AI+ would have risks from the standpoint of the US government, the risks of Chinese AI+ (say) would be far greater. We should take this information-hazard concern seriously and remember the unilateralist's curse. If it proves to be fatal for explicitly discussing AI arms races, we might instead encourage international cooperation without explaining why. Fortunately, it wouldn't be hard to encourage international cooperation on grounds other than AI arms races if we wanted to do so. Also note that a government-level arms race could easily be preferable to a Wild West race among a dozen private AI developers where coordination and compromise would be not just difficult but potentially impossible. Of course, if we did decide it was best for governments to take AI arms races seriously, this would also encourage private developers to step on the gas pedal. That said, once governments do recognize the problem, they may be able to impose moratoria on private development. How concerned should we be about accidentally accelerating arms races by talking about them? My gut feeling is it's not too risky, because It's hard to contain the basic idea. Super-powerful AI is already well known not just by governments but even in popular movies. Developing verification measures, technology restrictions, and so on require governments knowing what technology they're dealing with. If governments can think about these issues ahead of time (decades before strong AI becomes feasible), they're more likely to go for cooperation and less likely to panic and build up their own defenses, because they see that there's time for negotiations to potentially work before losing that much ground. Right now most AI research appears to be done in public, so there's not a huge cost for a given country in delaying at this point. Most risk analysts don't express concerns like these too much when talking about military arms races. Of course, there's selection bias; maybe most of the military does think it's dangerous to talk about these issues in public, and we only hear form the minority that defects from this view. But I've never heard criticism against people who talk too much about arms races in public, except this one comment from my friend. Talking about arms-race scenarios specifically makes it much more clear why we need global governance and improved cooperation. It's more persuasive than just saying, "Wouldn't it be great if the world could sing Kumbaya?" That said, I remain open to being persuaded otherwise, and it seems important to think more carefully about how careful to be here. The good news is that the information hazards are unlikely to be disastrous, because all of this material is already publicly available somewhere. In other words, the upsides and downsides of making a bad judgment seem roughly on the same order of magnitude. 8. How do our prospects look? In Technological change and nuclear arms control (1986), Ted Greenwood suggests that arms control has historically had little counterfactual impact: In no case has an agreement inhibited technological change that the United States both actually wanted to pursue at the time of agreement and was capable of pursuing during the intended duration of the agreement. Only in one area of technological innovation (i.e., SALT II constraints on the number of multiple independently-targetable reentry vehicles, or MIRVs, on existing missiles) is it possible that such agreements actually inhibited Soviet programs, although in another (test of new light ICBMs [intercontinental ballistic missiles]) their program is claimed by the United States to violate the SALT II Treaty that the Soviets have stated they will not undercut. In "Why Military Technology Is Difficult to Restrain" (1987), Greenwood adds that the INF Treaty was arguably more significant, but it still didn't stop technological development, just a particular application of known technology. In other domains we also see competition prevail over cooperation, such as in most markets, where usually there are at least several companies vying for customers. Of course, this is partly by social design, because we have anti-trust laws. Competition in business makes companies worse off while making consumers better off. Likewise, competition to build a quick, hacky AI makes human nations worse off while perhaps making the unsafe AIs better off. If we care some about the unsafe AIs for their own sakes as intelligent preference-satisfying agents, then this is less of a loss than it at first appears, but it still seems like there's room to expand the pie, and reduce suffering, if everyone takes things more slowly. Maybe the best hope comes from the possibility of global unification. There is just one US government, with a monopoly on military development. If instead we had just one world government with a similar monopoly, arms races would not be necessary. Nationalism has been a potent force for gluing countries together and if channeled into internationalism, perhaps it could help to bind together a unified globe. Of course, we shouldn't place all our hopes on a world government and need to prepare for arms-control mechanisms that can also work with the present-day nation-state paradigm. 9. Robot arms races **Robots require AI that contains clear goal systems and** an **ability to act effectively** in the world. **Thus,** they seem like a reasonable candidate for where artificial general intelligence will first emerge. Facebook's image-classification algorithms and Google's search algorithms don't need general intelligence, with many human-like cognitive faculties, as much as a smart robot does. **Military robotics seems** like **one of the most likely reasons that** a robot **arms race might develop**. Indeed, to some degree **there's already an arms race to build drones and autonomous weapons** systems. Mark Gubrud: Killer robots are not the only element of the global technological arms race, but they are currently the most salient, rapidly-advancing and fateful. If we continue to allow global security policies to be driven by advancing technology, then **the arms race** will continue, and it **may** even **reheat to Cold War levels, with multiple players** this time. Robotic armed forces controlled by AI systems too complex for anyone to understand will be set in confrontation with each other, and sooner or later, our luck will run out.

AI arms race causes extinction, outweighs nuclear war, and turns other moral theories.

**Shulman and Armstrong 11** write[[4]](#footnote-4)

II. **An AI arms race may be “winner-take-all”** The threat of an AI arms race does not appear to be primarily about the direct application of AI to warfare. While automated combat systems such as drone aircraft have taken on greatly increased roles in recent years (Singer, 2009; Arkin, 2009), they do not greatly disrupt the balance of power between leading militaries: slightly lagging states can use older weapons, including nuclear weapons, to deter or defend against an edge in drone warfare. Instead, the military impact of an intelligence explosion would seem to lie primarily in the extreme acceleration in the development of new capabilities. **A state might launch an AI Manhattan Project to gain** a few months or **years of sole access** to advanced AI systems**, and then initiate an intelligence explosion** to greatly increase the rate of progress. Even if rivals remain only a few months behind chronologically, they may therefore be left many technological generations behind until their own intelligence explosions. It is much more probable that such a large gap would allow the leading power to safely disarm its nuclear-armed rivals than that any specific technological generation will provide a decisive advantage over the one immediately preceding it. If states do take AI potential seriously, how likely is it that a government's “in-house” systems will reach the the point of an intelligence explosion months or years before competitors? Historically, there were substantial delays between the the first five nuclear powers tested bombs in 1945, 1949. 1952, 1960, and 1964. The Soviet Union's 1949 test benefited from extensive espionage and infiltration of the Manhattan Project, and Britain's 1952 test reflected formal joint participation in the Manhattan Project. If the speedup in progress delivered by an intelligence explosion were large, such gaps would allow the leading power to solidify a monopoly on the technology and military power, at much lower cost in resources and loss of life than would have been required for the United States to maintain its nuclear monopoly of 1945-1949. **To the extent that states distrust their rivals with** such **complete power**, or wish to exploit it themselves, **there would be strong incentives to vigorously push forward AI research**, and to ensure government control over systems capable of producing an intelligence explosion. In this paper we will discuss factors affecting the feasibility of such a localized intelligence explosion, particularly the balance between internal rates of growth and the diffusion of or exchange of technology, and consider historical analogs including the effects of the Industrial Revolution on military power and nuclear weapons. III. Accidental risks and negative externalities A second critical difference between the nuclear and AI cases is in the expected danger of development, as opposed to deployment and use. Manhattan Project scientists did consider the possibility that a nuclear test would unleash a self-sustaining chain reaction in the atmosphere and destroy all human life, conducting informal calculations at the time suggesting that this was extremely improbable. A more formal process conducted after the tests confirmed the earlier analysis (Konopinski, Marvin, & Teller, 1946), although it would not have provided any protection had matters been otherwise. The historical record thus tells us relatively little about the willingness of military and civilian leaders to forsake or delay a decisive military advantage to avert larger risks of global catastrophe. In contrast, **numerous scholars have argued that advanced AI poses a nontrivial risk of** catastrophic outcomes, including **human extinction.** (Bostrom, 2002; Chalmers, 2010; Friedman, 2008; Hall, 2007; Kurzweil, 2005; Moravec, 1999; Posner, 2004; Rees, 2004; Yudkowsky, 2008). Setting aside anthropomorphic presumptions of rebelliousness, a more rigorous argument (Omohundro, 2007) relies on the instrumental value of such behavior for entities with a wide variety of goals that are easier to achieve with more resources and with adequate defense against attack. Many decision algorithms could thus appear benevolent when in weak positions during safety testing, only to cause great harm when in more powerful positions, e.g. after extensive self-improvement. Given abundant time and centralized careful efforts to ensure safety, it seems very probable that these risks could be avoided: development paths that seemed to pose a high risk of catastrophe could be relinquished in favor of safer ones. However, the context of an arms race might not permit such caution. A risk of **accidental AI disaster would threaten all of humanity**, while the benefits of being first to develop AI would be concentrated, creating a collective action problem insofar as tradeoffs between speed and safety existed. A first-pass analysis suggests a number of such tradeoffs. Providing more computing power would allow AIs to either operate at superhumanly fast timescales or to proliferate very numerous copies. Doing so would greatly accelerate progress, but also render it infeasible for humans to engage in detailed supervision of AI activities. To make decisions on such timescales AI systems would require decision algorithms with very general applicability, making it harder to predict and constrain their behavior. Even obviously **risky systems might be embraced for competitive advantage**, and the powers with the most optimistic estimates or cavalier attitudes regarding risk would be more likely to take the lead. IV. Barriers to AI arms control Could an AI arms race be regulated using international agreements similar to those governing nuclear technology? In some ways, there are much stronger reasons for agreement: the stability of **nuclear deterrence, and** the **protection afforded by existing nuclear powers to their allies, mean that** the **increased threat of a new nuclear power is not overwhelming**. No nuclear weapons have been detonated in anger since 1945. **In contrast,** simply **developing AI capable of producing an intelligence explosion puts all states at risk** from the effects of accidental catastrophe, or the military dominance engendered by a localized intelligence explosion. However, AI is a dual-use technology, with incremental advances in the field offering enormous economic and humanitarian gains that far outweigh near-term drawbacks. Restricting these benefits to reduce the risks of a distant, novel, and unpredictable advance would be very politically challenging. Superhumanly intelligent AI promises even greater rewards: advances in technology that could vastly improve human health, wealth, and welfare while addressing other risks such as climate change. Efforts to outright ban or relinquish AI technology would seem to require strong evidence of very high near-term risks. However, agreements might prove highly beneficial if they could avert an arms race and allow for more controlled AI development with more rigorous safety measures, and sharing of the benefits among all powers. Such an agreement would face increased problems of verification and enforcement. Where nuclear weapons require rare radioactive materials, large specialized equipment, and other easily identifiable inputs, AI research can proceed with only skilled researchers and computing hardware. Verification of an agreement would require incredibly intrusive monitoring of scientific personnel and computers throughout the territory of participating states. Further, while violations of nuclear arms control agreements can be punished after the fact, a covert intelligence explosion could allow a treaty violator to withstand later sanctions. These additional challenges might be addressed in light of the increased benefits of agreement, but might also become tractable thanks to early AI systems. If those systems do not themselves cause catastrophe but do provide a decisive advantage to some powers, they might be used to enforce safety regulations thereafter, providing a chance to “go slow” on subsequent steps. V. Game-theoretic model of an AI arms race In the full paper, we present a simple game-theoretic model of a risky AI arms race. In this model, the risk of accidental catastrophe depends on the number of competitors, the magnitude of random noise in development times, the exchange rate between risk and development speed, and the strength of preferences for developing safe AI first. VI. Ethical implications and responses The above analysis highlights two important possible consequences of advanced AI: a disruptive change in international power relations and a risk of inadvertent disaster. From an ethical point of view, the accidental risk deserves special attention since it threatens human extinction, not only killing current people but also denying future generations existence. (Matheny, 2007; Bostrom, 2003). **While AI systems would outlive humanity, AI systems might lack key features contributing to moral value, such as** individual **identities, play, love, and happiness** (Bostrom, 2005; Yudkowsky, 2008). Extinction risk is a distinctive feature of AI risks: **even a catastrophic nuclear war or** engineered **pandemic that killed billions would still likely allow survivors** to eventually rebuild human civilization**, while AIs killing billions would likely not** leave survivors. (Sandberg & Bostrom, 2008). However, a national monopoly on an AI intelligence explosion could also have permanent consequences if it was used to stably establish its position. Permanent totalitarianism is one possibility (Caplan, 2008). We conclude by discussing some possible avenues for reducing these long-term risks.

## Contention 2 is Solvency

The PP is the best middle ground between industry and regulation. **Sachs 11** writes[[5]](#footnote-5)

Critics are overlooking that **the Principle can provide** a workable **accommodation between the needs of industry and** the need to ensure harm prevention and **adherence to ecological limits.** Specifically, **putting government in a** risk **gatekeeping role serves** several **important purposes, including:**  **Ensuring that the applicant** is competent to engage in the activity and **has** the **required expertise and resources;**  **Regulating** the **location of** potentially **risky activities** and ensuring that they occur in places where risks to the public are minimized;  **Ensuring that activities** presenting serious threats to public health or the environment can be prohibited (or **have safety precautions** placed on them) **before harm occurs;**  **Ensuring, through** establishing **a uniform review process** for every applicant**, that the cumulative amount of a risky activity does not exceed limits that would be damaging to the environment** or human health; **and**  **Minimizing risks while further research is conducted and making** that **research the responsibility of firms that will benefit the most** from the activity. I am not trying to defend every permitting and licensing scheme, of course. Government permitting programs can be burdensome and prone to political favoritism and rent-seeking behavior. They are often complex. If inadequately funded and staffed, a governmental review may be no more than a fig leaf of risk management (witness the Deepwater Horizon Oil Spill and the lax oversight of the Minerals Management Service). But the long-standing practice in U.S. law of establishing government agencies as ex ante gatekeepers for risk does suggest that the Strong Precautionary Principle cannot be so easily dismissed. It is not as alien to U.S. law and values as the critics would have us believe, and it hardly seems “paralyzing” in the many contexts in which it has been applied.

The PP is key to sustainability. **Grant and Quiggin 13** writes[[6]](#footnote-6)

6. Concluding comments Informally stated, **the P**recautionary **P**rinciple **has** strong **intuitive appeal**, particularly in the context of environmental regulation. **In dealing with** complex, **fragile and poorly understood natural systems, it seems to make sense ‘to err on the side of caution.’** However, this way of putting things points out the difficulties in formalizing the Precautionary Principle. ‘To err’ means to commit an error, and it is obviously difficult to include a prescription for error in a formal theory of decision under uncertainty. Yet **decisions are prone to errors arising from** an incomplete understanding of the problem at hand, and of the likelihood that some **contingencies** will **not** be **taken into account**. It seems desirable to take account of this reality in formulating a procedure for making decisions. In this paper, we have addressed the question in relation to the standard Bayesian model of decision theory, developed in the framework of an extensive-form game. We have argued that the Precautionary Principle is best understood, not as a modification of Bayesian decision theory, but rather as a heuristic constraint on the application of that theory; that is, as a response to the recognition that the outcomes of decisions may be affected by unforeseen contingencies. Heuristic constraints such as the Precautionary Principle must be satisfied before it is appropriate to apply the tools of Bayesian decision theory. **The P**recautionary **P**rinciple **is most commonly applied in relation to** interactive decisions, involving judgments as to whether or not to proceed with projects or innovations that may pose **unforeseen risks**. In this context, **the P**recautionary **P**rinciple may be regarded as a procedural rule that **places** the **burden of proof on proponents of activities subject to poorly-understood risks**. Under the Precautionary Principle, **proponents must convince policy makers** not only that the expected benefits exceed the expected costs but also that the project will not be subject to any significant unanticipated adverse outcomes.

## FW

Moral uncertainty is high now, but there’s room for improvement. **Parfit 84** writes[[7]](#footnote-7)

Some people believe that there cannot be progress in Ethics, since everything has been already said. Like Rawls and Nagel, I believe the opposite. How many people have made Non-Religious Ethics their life's work? Before the recent past, very few. In most civilizations, **most people have believed in** the existence of a **God**, or of several gods. A large minority were in fact Atheists, whatever they pretended. But, **before the recent past, very few Atheists made Ethics their life’s work.** Buddha may be among this few, as may Confucius, and a few Ancient Greeks and Romans. After more than a thousand years, there were a few more between the Sixteenth and Twentieth centuries. Hume was an atheist who made Ethics part of his life's work. Sidgwick was another. **After Sidgwick,** there were several **atheists** who were professional moral philosophers. But most of these **did not do Ethics. They did Meta-Ethics.** They did not ask which outcomes would be good or bad, or which acts would be right or wrong. They asked, and wrote about, only the meaning of moral language, and the question of objectivity. **Non-Religious Ethics has been systematically studied**, by many people, **only since the** 19**60s. Compared with the other sciences**, Non-Religious **Ethics is** the youngest and **the least advanced.**

Adopt a parliamentary model to account for moral uncertainty. This entails minimizing existential risks. **Bostrom 9** writes[[8]](#footnote-8)

It seems people are overconfident about their moral beliefs.  But **how should one** reason and **act if one** acknowledges that one **is uncertain about morality** – not just applied ethics but fundamental moral issues? if you don't know which moral theory is correct?

It doesn't seem **you can[’t] simply plug your uncertainty into expected utility** decision theory and crank the wheel; **because many** moral **theories** state that you **should not** always **maximize** expected **utility.**

Even if we limit consideration to consequentialist theories, it still is hard to see how to combine them in the standard decision theoretic framework.  For example, suppose you give X% probability to total utilitarianism and (100-X)% to average utilitarianism.  Now an action might add 5 utils to total happiness and decrease average happiness by 2 utils.  (This could happen, e.g. if you create a new happy person that is less happy than the people who already existed.)  Now what do you do, for different values of X?

The problem gets even more complicated if we consider not only consequentialist theories but also deontological theories, contractarian theories, virtue ethics, etc.  We might even throw various meta-ethical theories into the stew: error theory, relativism, etc.

I'm working on a paper on this together with my colleague Toby Ord.  We have some arguments against a few possible "solutions" that we think don't work.  On the positive side we have some tricks that work for a few special cases.  But beyond that, the best **we have managed** so far is **a** kind of **metaphor, which** we don't think is literally and exactly correct, and it is a bit under-determined, but it **seems to get things roughly right** and it might point in the right direction:

**The Parliamentary Model.**  Suppose that you have a set of mutually exclusive moral theories, and that you assign each of these some probability.  Now imagine that **each** of these **theorie**s **gets to send** some number of **delegates to The Parliament**.  The number of delegates each theory gets to send is **proportional to the probability of the theory.**  Then the delegates bargain with one another for support on various issues; and the Parliament reaches a decision by the delegates voting.  What you should do is act according to the decisions of this imaginary Parliament.  (Actually, we use an extra trick here: we imagine that the delegates act as if the Parliament's decision were a stochastic variable such that the probability of the Parliament taking action A is proportional to the fraction of votes for A.  This has the effect of eliminating the artificial 50% threshold that otherwise gives a majority bloc absolute power.  Yet – unbeknownst to the delegates – the Parliament always takes whatever action got the most votes: this way we avoid paying the cost of the randomization!)

The idea here is that moral theories get more influence the more probable they are; yet **even a** relatively **weak theory can still get its way on some issues** that the theory think are extremely important **by sacrificing** its influence **on other** i**s**sues that other theories deem more important.  For example, **suppose you assign 10% probability to** total **util**itarianism and 90% to moral egoism (just to illustrate the principle).  Then **the Parliament** would mostly take actions that maximize egoistic satisfaction; however it **would make some concessions to util**itarianism **on** issues that utilitarianism thinks is especially important.  In this example, the person might donate some portion of their income to **existential risks** research and otherwise live completely selfishly.

I think there might be wisdom in **this model**.  It **avoids the** dangerous and **unstable extremism** that would result **from letting one’s current favorite moral theory completely dictate action**, while still allowing the aggressive pursuit of some non-commonsensical high-leverage strategies so long as they don’t infringe too much on what other major moral theories deem centrally important.

I don’t need to win that weighing values is possible. Extinction precludes all values, so it is wrong under any moral code. **Seeley 86**[[9]](#footnote-9)

In moral reasoning prediction of consequences is nearly always impossible. One balances the risks of an action against its benefits; one also considers what known damage the action would do. Thus a surgeon in deciding whether to perform an operation weighs the known effects (the loss of some nerve function, for example) and risks (death) against the benefits, and weighs also the risks and benefits of not performing surgery. Morally, however, **human extinction is unlike any other risk. No conceivable human good could be worth** the **extinction** of the race, **for** in order **to be a human good it must be experienced by human beings.** Thus extinction is one result we dare not-may not-risk. Though not conclusively established, **the risk of extinction is real enough to make nuclear war** utterly **impermissible under any** sane **moral code.**

Cost-benefit analysis is feasible. Ignore any util calc indicts. **Hardin 90** writes[[10]](#footnote-10)

**One** of the **cute**r **charge**s **against util**itarianism **is that** it is irrational in the following sense. **If I take the time to calculate** the consequences of various courses of action before me, **then** I will ipso facto have chosen the course of action to take, namely, to sit and calculate, because while I am calculating the other **courses of action will cease to be open to me. It should embarrass philosophers that they have ever taken this** objection **seriously. Parallel considerations in other realms are dismissed** with eminently good sense. Lord Devlin notes, “If the reasonable man ‘worked to rule’ by perusing to the point of comprehension every form he was handed, the commercial and administrative life of the country would creep **to** a standstill.” James March and Herbert Simon **escape** the quandary of **unending calculation** by noting that often we satisfice, **we do not maximize: we stop calculating** and considering **when we find a merely adequate choice** of action. **When**, in principle, **one cannot know what is** the **best** choice, **one can nevertheless be sure that** sitting and **calculating is not the best choice.** But, one may ask, How do you know that another ten minutes of calculation would not have produced a better choice? And one can only answer, You do not. At some point the quarrel begins to sound adolescent. It is ironic that **the point** of the quarrel **is almost never at issue in practice** (as Devlin implies, **we are** almost all **too reasonable** in practice **to bring the world to a standstill**) but only in the principled discussions of academics.

The standard is **maximizing happiness**.

First, revisionary intuitionism is true and leads to util.

**Yudkowsky 8** writes[[11]](#footnote-11)

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.

Second, util is epistemologically necessary. Everyone values happiness whether they want to or not. Even a skeptic wouldn’t shoot themselves in the foot.

Third, the ultimate human good is happiness. **Darwish 9**[[12]](#footnote-12)

Let’s start with knowledge. It is clear that **those who value knowledge for its own sake** (for instance, highly motivated professionals) **find pleasure in** both the **pursuit** and attainment **of knowledge**, however exhausted they become in either case. So, granted that **knowledge**, for them, is a value that **has intrinsic worth** in itself, and is thus sought for itself, **[because] it** is a value that **yields** their pleasure or **happiness. The same can be said about** the **other values.** Let’s take autonomy in the sense expressed by Hooker as “control over one’s own life” as another example. Here one needs to say no more than that **the mere fact that people seek autonomy explains** the satisfaction or the **pleasure autonomy brings.** Those who value autonomy, thus seek it for itself, cannot feel happy when their decisions are not in their hands, or when they do not have control over their own life. In short, they cannot be happy otherwise. Moore, who explicitly differs from the classical utilitarians in holding that pleasure is not the sole good, 20 says that “the most valuable things… are pleasures of human intercourse and the enjoyment of beautiful objects,” 21 which seems to mean that such things are valued for the pleasures and the enjoyment they bring. These examples clearly show that **though these values have intrinsic worth, they** bring or **constitute our pleasure.**

And fourth, sentimentalism begs the question. Must use util. **Debes 9**[[13]](#footnote-13)

On the other hand, the first question (#1 above) risks glossing over a further contentious debate in ethics, namely, what is the domain of “the moral”? That is, **what constitutes** or qualifies as a **“genuine”** moral judgment (e.g. of **moral goodness**) **versus** a judgment about the **“merely” valuable** (e.g. of non-moral goodness)? It is fairly common, for example, to claim that prudential judgments about one’s personal welfare don’t qualify as moral. But other value judgments have a more ambiguous status, e.g. judgments about beauty, wit, friendship, love, and the array of values classically connected with virtues like courage or temperance. Thus, if we are tasked with distinguishing which emotions genuinely have moral shape, e.g. because they can be described as having distinctly “moral” content, we will eventually also need to take a position with respect to the further question, **what constitutes the domain of “the moral”?** Indeed, **for those who wish to draw** an especially tight **connection between value and sentiment,** as historical or “neo” sentimentalists do, **this** further question **is** especially **pressing.** For, if left unanswered, she is open to the charge of “flattening” the terrain of value. The flattening objection is (roughly) the charge that one has illegitimately amalgamated the domain of “the moral” to all other value, i.e. to the “merely” valuable. These two domains, a detractor can argue, are intuitively distinct – perhaps even, as the Kantian deontologist would say, rationally distinct (see DEONTOLOGY). Thus, at issue here is not simply the potential failure to identify which emotions like guilt or resentment are distinctly, descriptively, “moral”. Instead, the flattening objection is a challenge to the foundational ambitions of certain metaethical views like **sentimentalism**, which seek to make sentiment the basis of morality (see also De Sousa 2006). At a minimum, the flattening objection is a response to the way such viewsimplicitly or **explicitly seek[s] to vindicate the** equal **ethical import of certain values that are not** (or not obviously) **distinctly moral**, e.g. values like “the enviable,” “the joyful,” “the amusing,” and so on (see also Schroeter 2006). Whatever its spur, however, the flattening objection looms over any theory of moral sentiments that fails to distinguish the domain of “the moral.”

## Next is Theory Preempts

1. Debating about the PP is key to topic education.

**Schettler and Raffensperger 4** write[[14]](#footnote-14)

**Proof is a value-laden concept that integrates** statistics, **empirical observation, inference**, research design **and research** agendas **in**to **a political** and social **context.** This section discusses the uses and misuses of some of the criteria commonly used to establish proof. Strict criteria may be useful for establishing “facts”, but by the time a fact or causal relationship has been established by rigorous standards of proof, considerable avoidable harm may already have occurred. The effects of lead exposure on children’s brain development or asbestos on lung cancer risk are examples. In each case, people were damaged over many decades, long after substantial evidence of serious health effects was established, while lead and asbestos advocates contested epidemiological “proof” of causation. **Guided by the p**recautionary **p**rinciple**, people are** as **concerned with** the weight of the available evidence as they are with establishing facts by **rigorous standards of proof**. The weight of the evidence can guide preventive action, whereas waiting for proof may allow damage to occur. By convention, a considerable amount of consistent evidence is necessary to establish factual “proof” of a cause-and-effect relationship. Traditionally, in a study of the relationship between two variables, a correlation is said to be statistically significant only if the results show the two to be linked, 5. Why is a precautionary approach needed? 71 independent of other factors, with greater than 95% likelihood that the positive results of the study did not occur by chance. But correlation does not establish causation. In epidemiology, a series of additional criteria, for example, those of Hill (1965), are usually added before causation can be claimed. Hill criteria include not only establishment of a statistically significant correlation between two variables but also require that the causal variable precede the effect, a dose–response relationship, elimination of sources of bias and confounding, coherence with other studies and understanding of a plausible biological mechanism. Tobacco smoking, for example, was known to be associated with lung cancer for more than 50 years before a plausible biological mechanism was finally described. At that point, denying that tobacco “causes” cancer became impossible. People’s adherence to conventions or choices among criteria expresses their willingness to make type I or type II errors. A type I error is the mistake of concluding that an association or phenomenon exists when, in truth, it does not. Conversely, a type II error is the mistake of failing to recognize an association or phenomenon when it does exist. Each kind of error has consequences. Type II errors may, for example, lead people to allow a harmful activity to go forward and are the inevitable result of a consistent bias towards avoiding type I errors. Type I errors will result in invalid concerns about a product or activity and may lead to unnecessary restrictions. Establishing type I and type II error rates is a choice that reflects certain biases and is largely done by convention, often without considering the consequences. For example, by convention, interpretations of scientific data generally favour type II over type I errors. People generally require strong evidence that something is scientifically “true” before being willing to say so. An historical basis for error bias **A general theme that has gained currency** in many countries **is that people** are autonomous individuals who **are free** to live as they wish and **do as they want, provided that they do not cause harm to others.** This concept has set up a tension between the individual and society at large in terms of establishing the limits of tolerance and defining harm. In On Liberty, first published in 1859, John Stuart Mill (1978 (1859)) explored the nature and limits of power that can be legitimately exercised by society over the individual. He concluded that the only purpose for which power can be rightfully exercised over any member of a civilized community, against his or her will, is to prevent harm to others. Mill was concerned that, in a democratic society, the majority would set the limits to tolerance – that the majority would interfere with the creative individual’s inclination to invent and develop and to explore new frontiers. He also worried that the majority would go so far as to define “harm”, using inappropriate assertions of “harm” as a blockade to progress. In short, he feared the “tyranny of the majority” and their inclination to favour the status quo. **This** tension **is at the heart of many of today’s policy debates.** Not only must harm be defined but **people** also **have to decide how to act** or how to legitimately exercise power **when** the probability of harm (**risk**) **is uncertain**. Though decisions must be based on what is known at the time, **if “proof”** of harm **is required before** limiting an activity or **choosing an alternative**, as Mill would have, **there is a risk of failing to prevent harm**. Seeing how Mill’s fears are reflected in today’s policies in many countries throughout the world is easy. In general, the burden of proof of harm falls on the general public or individuals who assert that another party has injured them. High standards of “proof” add to this burden, even when the weight of the evidence suggests that harm has occurred or is likely. In other words, a **bias towards type II errors** – established by convention in interpreting scientific data – **has** also **crept into** social, political and judicial **policy. Asking whether such a bias is appropriate for preventing harm** or for choosing among optional human activities **is fully legitimate**. Further, it may be legitimately ask how such a bias is likely to influence the ways that human activities alter complex ecological systems that define the world to be left to future generations **– a consideration at the core of sustainability.**

2. Err aff on theory because of time skew and neg side bias. Negs won 12% more rounds at VBT according to Tabroom, Fantasy Debate confirms 7% neg side bias, and there was an 8% neg side bias in TOC prelims. Even if we both get 13 minutes, the distribution of time is the source of the abuse. Even if I could get faster or more efficient, she could too.

# 1AR

## AT “Extinction Scenarios Must Have a Single Author” Theory

Experts are right about their subject areas. Tetlock assumes subjects that experts are unfamiliar with. **Caplan 5** writes[[15]](#footnote-15)

Is my confidence in experts completely misplaced? I think not. **Tetlock**'s sample **suffers from** severe **selection bias**. **He deliberately asked** relatively **difficult and controversial questions**. As his methodological appendix explains, **questions had to "Pass the 'don't bother me** too often **with dumb questions' test." Dumb according to who?** The implicit answer is "Dumb according to **the typical expert** in the field." **What Tetlock really shows** is **that experts are overconfident if you exclude the questions where they have reached** a **solid consensus.** This is still an important finding. Experts really do make overconfident predictions about controversial questions. We have to stop doing that! However, this does not show that experts are overconfident about their core findings. It's particularly important to make this distinction because **Tetlock's work is so good that a lot of crackpots will want to highjack it:** "Experts are scarcely better than chimps, so why not give intelligent design and protectionism equal time?" But what **Tetlock** really **shows** is **that experts can raise** their **credibility if they stop overreaching**.

Tetlock concedes that experts are still key on political predictions. **WSJ 6** writes[[16]](#footnote-16)

Such predictions are good fun. But in general, the prognostications of political pundits are about as accurate as a chimp throwing darts. At least that's the finding of "Expert Political Judgment," a new book by University of California, Berkeley, political psychologist Philip **Tetlock**. From 1987 to 2003, Prof. Tetlock coaxed 284 political experts of all stripes -- academics, journalists and think-tankers from across the political spectrum -- to make specific, verifiable forecasts. He looked at more than 27,000 predictions in all. Prof. Tetlock's innovation was to elicit numerical predictions. As he noted in an interview with me, political punditry tends toward the oracular: statements vague enough to encompass all eventualities. But by promising confidentiality and relying on the "curiosity value of the project," he **was able to get pundits to provide** probability **estimates for** such questions as **whether certain countries' legislatures would see shifts in their** **ruling** parties, whether inflation or unemployment would rise and whether nations would go to war. **Without numerical predictions, "it's much easier to fudge**," Prof. Tetlock told me. "When you move from words to numbers, it's a really critical transition." What he found is that people with expertise in explaining events that have happened aren't very successful at predicting what will happen. He demonstrated this by checking their predictions against reality, and then comparing the humans' performance with that of several automated prediction schemes. The simplest type was chimp-like. No chimps were harmed in the experiment; Prof. Tetlock essentially used random numbers. More complex sets of predictions were based on the frequency of similar events in the past. The virtual chimps did about as well as humans, while several of the more-complex schemes exceeded the best human forecasters. As a psychologist, Prof. Tetlock's primary interest lies in understanding why certain forecasters did better than others, and to do so he found it useful to divide his panel of experts into two groups, hedgehogs and foxes, a classification scheme devised by the late British philosopher Isaiah Berlin. Hedgehogs know one big idea and confidently stick with it, applying it to all scenarios. Foxes are more flexible and skeptical. Prof. Tetlock assigned his experts to either group by asking them to agree or disagree with statements about their cognitive style, such as, "It is annoying to listen to people who cannot make up their minds." (A hedgehog would, in general, agree with the statement, while a fox would disagree.) Foxes outperformed hedgehogs in nearly every measure, Prof. Tetlock said. That suggested that those with flexible thinking were better equipped to make forecasts. Foxes were better, by about three times, at adapting their predictions to world events they hadn't expected. This relates to another failing of hedgehogs: they tended to be more likely to engage in "hindsight bias." In other words, hedgehogs were more likely to err in their own favor when asked to recall past predictions -- and if they couldn't recall their own mistakes, how could they be expected to learn from them? The media prefers hedgehogs, Prof. Tetlock writes, since "simple, decisive statements are easier to package in sound bites." Perhaps as a result, he found that fame, in both the public and academic spheres, correlated negatively with the accuracy of predictions. Writing a book about experts' fallibility, and reflecting on the pitfalls in thought processes that cause it, are good ways to stay humble about one's own expertise. Prof. Tetlock played along with his pundits by making his own set of forecasts, and "it certainly made me somewhat more humble about my own ability to predict." He thought the first President Bush would defeat Bill Clinton in the 1992 election, expected China's economic growth rates to stumble and predicted Saddam Hussein would fall soon after the first Gulf War. Much of Prof. Tetlock's fox-like book addresses potential criticism of his methods and conclusions, and acknowledges the study's weaknesses. And in our interview, Prof. Tetlock was open to my questions about his methods. The **panel's anonymity** seems to **create two** potential **problems** for readers. First, **we have to trust that the participants were** indeed **qualified** and representative of political pundits. **Second, the participants seem to have had little at stake.** Prof. Tetlock, who has studied accountability and how it affects thinking, argues that making his pundits accountable would have exacerbated their tendencies. "Accountability pressure means you are more likely to follow your theory of what good judgment is ... the cognitive equivalent of putting on your Sunday best." For hedgehogs, that would mean an even more steadfast, single-issue approach, which appears to have been their downfall. The New Yorker's review9 of his book surveyed the grim state of expert political predictions and concluded by advising readers, "Think for yourself." Prof. Tetlock isn't sure he agrees with that advice. He pointed out an exercise he conducted in the course of his research, in which he gave Berkeley undergraduates brief reports from Facts on File10 about political hot spots, then asked them to make forecasts. **Their predictions** -- **based on far less background knowledge** than his pundits called upon -- **were the worst he encountered**, even less accurate than the worst hedgehogs. "**Unassisted human intuition is a bomb** here," Prof. **Tetlock told me**.

1. Stephen Luntz (staff writer). Citing Safa Motesharrei, graduate student in mathematics at the National Socio-Environmental Synthesis Center. “According to a NASA Funding Study, We’re Pretty Much Screwed.” I Fucking Love Science. March 19th, 2014. http://www.iflscience.com/environment/according-nasa-funded-study-were-pretty-much-screwed#cWych02cH9mk6Tm6.16 [↑](#footnote-ref-1)
2. Brian Tomasik (“I work on researching the best ways to reduce suffering in the future, examining crucial considerations in science, politics, sociology, and philosophy that bear on this topic. You can read more on my website: Essays On Reducing Suffering. I graduated from Swarthmore College in 2009, where I studied computer science, mathematics, statistics, and economics. I wrote a thesis on multitask feature selection and published additional papers on machine learning and international trade. From 2009-2013, I worked in the relevance division of Bing at Microsoft, improving ranking algorithms for web results through feature engineering, data mining, and architectural innovation. I built end-to-end three of Bing's production ranking models, which served over 5 billion searches on Bing and Yahoo every month.”). “Crop Cultivation and Wild Animals.” Essays on Reducing Suffering. 2008-2013, Last update: December 12th, 2013. http://www.utilitarian-essays.com/crop-cultivation-and-wild-animals.html [↑](#footnote-ref-2)
3. Brian Tomasik (graduated from Swarthmore, former programmer at Bing). “International Cooperation vs. AI Arms Race.” Last updated April 3rd, 2014. http://utilitarian-essays.com/ai-arms-race.html#section7 [↑](#footnote-ref-3)
4. Carl Shulman (Singularity Institute of Artificial Intelligence) and Stuart Armstrong (InhibOx, an organization dedicated to developing and delivering the best services and technologies in computer-aided drug discovery ). “Singularity Hypotheses: A Scientific and Philosophical Assessment.” April 13th, 2011. http://singularityhypothesis.blogspot.com/2011/04/arms-races-and-intelligence-explosions.html [↑](#footnote-ref-4)
5. Noah Sachs (Associate Professor, University of Richmond School of Law and Director, Robert R. Merhige, Jr. Center for Environmental Studies). “Rescuing the Strong Precautionary Principle from its Critics.” University of Illinois Law Review. August 1st, 2011. [↑](#footnote-ref-5)
6. Simon Grant and John Quiggin (University of Queensland, School of Economics). “Bounded awareness, heuristics, and the Precautionary Principle.” Journal of Economic Behavior and Organization. July 17th, 2013. [↑](#footnote-ref-6)
7. Derek Parfit, Reasons and Persons (Oxford: Clarendon, 1984). P. 453. [↑](#footnote-ref-7)
8. Bostrom, Nick (*Existential*ist of a different sort). “Moral uncertainty – toward a solution?” 1 January 2009. <http://www.overcomingbias.com/2009/01/moral-uncertainty-towards-a-solution.html> [↑](#footnote-ref-8)
9. Robert A., Central Committee for Conscientious Objectors, The Handbook of Non-Violence, p. 269-70 [↑](#footnote-ref-9)
10. Hardin, Russell (Helen Gould Shepard Professor in the Social Sciences @ NYU). May 1990. Morality within the Limits of Reason. University Of Chicago Press. pp. 4. ISBN 978-0226316208. [↑](#footnote-ref-10)
11. 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/ [↑](#footnote-ref-11)
12. “Rethinking Utilitarianism,” Bahaa Darwish (Qatar University). *Teaching Ethics* 10 (1):87-109 (2009). http://www.uvu.edu/ethics/seac/Darwish%20-%20Rethinking%20Utilitarianism.pdf [↑](#footnote-ref-12)
13. Remy Debes (PhD in Philosophy, UMemphis). “Moral Sentiments.” January 20th, 2009. http://www.memphis.edu/philosophy/pdfs/Debes\_IEE\_12.pdf [↑](#footnote-ref-13)
14. Ted Schettler (SEHN's Science Director, received his MD from Case-Western Reserve University and a masters degree in public health from the Harvard School of Public Health. He practiced medicine for many years in New England.Ted has worked extensively with community groups and non-governmental organizations throughout the US and internationally, addressing many aspects of human health and the environment. He has served on advisory committees of the US EPA and National Academy of Sciences.) and Carolyn Raffensperger (executive director of the Science and Environmental Health Network. In 1982 she left a career as an archaeologist in the desert Southwest to join the environmental movement. She first worked for the Sierra Club where she addressed an array of environmental issues, including forest management, river protection, pesticide pollutants, and disposal of radioactive waste. She began working for SEHN in December 1994. As an environmental lawyer she specializes in the fundamental changes in law and policy necessary for the protection and restoration of public health and the environment). “Why is a precautionary approach needed?” The precautionary principle: protecting public health, the environment and the future of our children. WHO. 2004. http://www.euro.who.int/\_\_data/assets/pdf\_file/0003/91173/E83079.pdf [↑](#footnote-ref-14)
15. Bryan Caplan (econ professor at George Mason). “Tackling Tetlock.” December 26th, 2005. <http://econlog.econlib.org/archives/2005/12/tackling_tetloc_1.html> [↑](#footnote-ref-15)
16. Wall Street Journal, 1/6/6, Evaluating Political Pundits, http://online.wsj.com/public/article\_print/SB113631113499836645.html?mod=djm\_HAWSJSB\_Welcome [↑](#footnote-ref-16)