INSTITUTE
FOR LAW & AI

Catastrophic risk, uncertainty, and agency analysis

LawAI Working Paper Series, No. 1-2022 | Alasdair Phillips-Robins | September 2022 law-ai.org

Catastrophic Risk, Uncertainty, and Agency Analysis

Alasdair Phillips-Robins

September 1, 2022

Summary

I propose three changes to the governance of federal policymaking: (1) an amendment to Circular A-4 that provides agencies with guidance on evaluating policies implicating catastrophic and existential risks, (2) principles for the assessment and mitigation of those risks, and (3) proposed language to be included in an executive order requiring agencies to report on such risks.

Each proposal is intended to serve a different goal, and each adopts policy choices that are explained in detail in the accompanying essay. First, the goals. The amendment to Circular A-4 aims to help agencies produce robust assessments of proposed agency action that implicates catastrophic and existential risks. The amendment balances two dangers: on the one hand, allowing agencies to taking risky actions without carefully considering their effects on potential catastrophic and existential risks; and on the other, requiring implausibly definite justifications for catastrophic-risk-mitigation policies when those policies will inevitably involve high levels of uncertainty. A related and important distinction is between agency action designed to reduce the threat of catastrophic public harms or human extinction and agency action that is taken for some other purpose but that may nevertheless affect catastrophic and existential risks.

Circular A-4 is important, but limited. It covers only a subset of agency action—official agency rulemakings. The proposed principles sweep more broadly. They encourage all departments and agencies to actively focus on catastrophic and existential risks and to apply the best practices of Circular A-4, including the use of quantitative estimates combined with qualitative analysis, beyond the scope of agency rulemakings.

The principles are broad, but non-binding. My final proposal, therefore, is for an executive order (EO) requiring agencies to produce reports on relevant catastrophic and existential risks, including the state of relevant expert knowledge and proposals for executive or legislative action. The proposed language is designed

to be included in an EO implementing President Biden's January 2021 memorandum on modernizing regulatory review.

Each of the proposals reflects policy choices that I explain in the accompanying essay. They divide into two areas: how to make sure federal agencies do not neglect catastrophic and existential risks and how to properly evaluate policies related to those risks.

The proposed principles and EO concern to the former problem, agency agenda-setting. Agencies are influenced explicitly by legislative mandates and more subtly by presidential policy preferences and external political pressures, such as public opinion, interest groups, and media coverage. The White House does not, in general, direct day-to-day agency action, but it can influence agency priorities and shape high-profile decisions. To that end, the principles and the EO are designed to get catastrophic and existential risks on the agency agenda—and to produce, through public reports, congressional attention and external political pressure.

The amendment to Circular A-4, for their part, reflect two major choices. First, it instructs agencies to apply quantified benefit-cost analysis (BCA) to policies implicating catastrophic and existential risks, despite a competing school of thought that calls for the use of the precautionary principle in such circumstances. The essay considers the arguments in favor of the latter approach in some detail, in particular the problems of uncertainty associated with extreme risks, but ultimately rejects the precautionary principle as theoretically unjustified, practically indeterminate, and, given the entrenched nature of BCA, politically implausible. Second, the amendment acknowledges the limits of quantitative analysis in this area and calls for agencies to offer, alongside quantitative BCA, robust qualitative justifications for policies affecting catastrophic and existential risks. As the essay explains, the problems of fat-tailed distributions and expected value calculations that rely on low probabilities of extremely large costs or benefits should lead policymakers to be wary of relying solely on the numbers.

Draft Amendment to Circular A-4

(Insert before "***Discount Rates***")

9. Treatment of Catastrophic and Extinction Risks

When proposed agency action implicates the threat of catastrophic, or "worst case" outcomes, special analytical considerations arise. Catastrophic risks are those that threaten loss of life or civilizational destruction on a far greater scale than other risks; prominent examples include pandemics, sudden and extreme climate change, nuclear war, asteroid and comet impacts, the accidental release or offensive use of novel biotechnology, and uncontrolled or hostile AI systems. A subset of the most extreme catastrophic risks, known as existential risks, threaten human extinction. Proposed agency action may aim to prepare for, or reduce the risk of, such threats, or such threats may be implicated by agency action taken for other reasons.

Wherever possible, you should use standard approaches to estimate benefits and costs related to catastrophic and existential risks. However, both cost-effectiveness analysis and benefit-cost analysis may be significantly more difficult thanks to special issues presented by such risks. Catastrophic and existential risks typically involve high levels of uncertainty. The costs of catastrophic outcomes such as pandemics may be difficult to estimate. Existential risks present the additional problem of irreversible losses. As a result, your analysis may need to rely more than normal on a qualitative explanation of the link between the proposed agency action and mitigating a catastrophic risk if estimates of key parameters and relationships are unreliable.

When a proposed action implicates a catastrophic or existential risk, the following considerations should be taken into account:

- If the proposed action is rationally related to the risk of a catastrophe, you should not ignore the possibility of catastrophic outcomes solely because their probabilities fall below a certain threshold. For example, the analysis of a proposed rule governing biosecurity procedures should consider not only the immediate costs and benefits of the rule, but also the effect that reducing the accidental release of pathogens would have on the probability of a major disease outbreak or a pandemic.
- Estimates of the effects of the proposed action and of underlying parameters and relationships may be highly uncertain. Your analysis should fully present the uncertainty involved. This may include identifying key assumptions and data sources, describing models on which your analysis relies, and presenting sensitivity analyses. You may wish to consider soliciting expert views on uncertain parameters and relationships as part of

a Delphi method analysis. (For more on this topic, see "*Treatment of Uncertainty*," below.)

- Be aware of the possibility of "fat tails" in the probability distributions of potentially catastrophic events. Fat tails occur when extreme outcomes are unusually frequent. Complex systems are especially likely to involve fat tails. Even when the underlying process is normally distributed, if there is high uncertainty about the shape and key statistics of the true distribution, a fat-tailed distribution may be appropriate for modelling purposes.
- When the case for a proposed action that exceeds the $100 million annual threshold is driven primarily by estimates of its effects on low-probability, high-impact outcomes, you should include a detailed qualitative analysis of the effects of the proposed rule alongside the quantitative analysis. Given the high levels of uncertainty inherent in estimating the effect of regulatory action on the probability of extreme events, such actions should be supported by a robust qualitative explanation of the mechanism by which the effects will occur, the track record of similar policies, and other relevant factors, such as expert assessments.
- In general, worst-case, or "precautionary" analyses are not appropriate, as they do not consider the full range of outcomes and may lead to inappropriately high levels of risk aversion. However, when a proposed action plausibly implicates an extinction risk, it may be proper to deviate from the general assumption of risk neutrality on the principle that society is risk averse when it comes to truly irreversible outcomes such as extinction.

Memorandum for the Heads of Executive Departments and Agencies

Principles for the Assessment and Mitigation of Catastrophic and Existential Risks

Certain risks present dangers that are far out of the ordinary. Often known as catastrophic risks, these threaten disasters severe enough to jeopardize the safety or welfare of a large proportion of the civilian population of the United States or to significantly harm, set back, or destroy human civilization on a global scale. Examples of such risks include pandemics, sudden and extreme climate change, nuclear war, asteroid and comet impacts, the accidental release or offensive use of novel biotechnology, and uncontrolled or hostile AI systems. Some catastrophic risks may also present existential risks if they are severe enough to threaten human extinction.

In order to mitigate the threat of global catastrophic and existential risks ("extreme risks"), and to prepare the Federal Government for such risks, this memorandum sets out the following principles to guide the development and analysis of agency action, which should be respected to the extent permitted by law:

Risk Focus: Agencies should actively seek opportunities to mitigate or prepare for catastrophic and existential risks, whether through regulatory action, internal government preparation, proposing legislation, incentivizing private action, or other means. Agencies should consider developing a register of extreme risks related to their area of regulatory focus. An extreme risk should not be ignored in policy development or analysis solely because its probability falls below a certain threshold.

Risk Assessment: Quantitative estimates of relevant risks and potential outcomes should be given wherever possible, even if the figures involved are highly uncertain. If reliable estimates are not available, agencies should consider conducting expert surveys to establish reasonable risk ranges. If quantitative estimates are truly impossible, qualitative assessments of the risks and the effects of agency action should be given alone. In all cases, given the high levels of uncertainty involved, quantitative analysis of extreme risks should be accompanied by qualitative evaluations.

Benefits and Costs: Agency action implicating extreme risks should be based on estimates of the benefits and costs of the action. Wherever possible, these estimates should be quantitative, but they should recognize the limited information that may be available. Agency analysis should not display a rigid adherence to quantitative benefit and cost estimates where those estimates are highly uncertain or likely to change.

Scientific Integrity: Federal regulation and agency action taken to mitigate extreme risks should be based on the best available scientific evidence. When

reliable information is unavailable, agencies should consider facilitating relevant scientific research and should incorporate new information when it becomes available. To the extent possible, scientific judgements should be separated from policy judgements.

Public Participation: To the extent possible under legal and public policy constraints, proposed agency action and analysis should be developed with opportunities for public comment, expert consultation, and stakeholder participation.

International Cooperation: Many extreme risks threaten not only the United States but the entire world. The Federal Government should encourage international cooperation on scientific research, risk mitigation, and disaster response. When appropriate, and to the extent permitted by U.S. law, departments and agencies should communicate the policy positions and understandings of the United States to other nations and should coordinate policy initiatives internationally to the greatest extent possible.

Proposed Language for Inclusion in an Executive Order

[Sec. X.] Catastrophic and Existential Risks

(a) Consistent with the President's Memorandum for the Heads of Executive Departments and Agencies, "Principles for the Assessment and Mitigation of Catastrophic and Existential Risks," each agency shall determine whether the agency's purview includes any catastrophic or existential risks and, if so, develop an annual Catastrophic and Existential Risk Plan to be submitted to the Administrator of OIRA. The Plan shall contain at a minimum:
    (i) An assessment of the catastrophic and existential risks that fall within the agency's subject-matter area;
    (ii) Expert-informed analyses of the risk of the catastrophic and existential threats identified, including expert estimates of their likelihood and impact as well as associated uncertainties;
    (iii) A review of relevant intelligence collection, early warning systems, and other programs necessary to detect and evaluate the risk of catastrophic and existential threats, including how such programs may be improved;
    (iv) Proposals for how the agency or the federal government as a whole may act to reduce the threat or mitigate the impact of the most concerning catastrophic and existential risks, including recommendations for legislative and regulatory action as appropriate.

(b) In producing the Plan, the agency shall consult with experts on relevant catastrophic and existential risks, including from academic, non-governmental, and private sector institutions.
(c) For the purposes of this order, "agency" shall have the definition set out in Executive Order 12,866 § 3(b).

Contents

INTRODUCTION

Imagine a policymaker who is confronted with a new technology—artificial intelligence, say—or a field of research—into novel pathogens, perhaps—that some experts warn could pose a catastrophic, even existential, risk. What regulations should she propose? Nothing? A total ban? Safety limits? Which ones? What if experts disagree vociferously on whether disaster is likely, or even possible?

Faced with such difficulties, a common policy response is simply to ignore the problem.[1] Some of the models of climate change the federal government uses to calculate its social cost of carbon do not consider extreme outcomes from global warming. There are no public regulatory cost-benefit assessments that attempt to calculate the value of pandemic mitigation, the regulation of misaligned or hostile general artificial intelligence, or the prevention of bioterrorist attacks. The Supreme Court has allowed agencies to round down small and uncertain risks of disaster to zero under some circumstances.[2]

There are many explanations for the tendency to neglect low probability, high impact risks. Psychologists point to biases in human thinking, such as our tendency to ignore small probabilities and events that we have not experienced.[3] Economists note that policies to mitigate climate change or prevent pandemics are public goods—that is, those paying for them will not capture all the benefits, so everyone has an incentive to free-ride on the efforts of others—and thus the market will not supply them.[4] Political scientists point out that voters will not reward politicians for spending money today to ward off disaster in half a century, so the rational leader will always pass the buck to her successor.[5] Whichever explanation

---

1 *See, e.g.*, Daniel Farber, *Uncertainty*, 99 GEO. L.J. 901, 950 (2011) (noting the federal government's approach of ignoring risks below a one in ten thousand probability in planning for nuclear waste storage); *see also* Arden Rowell, *Regulating Best-Case Scenarios*, 50 ENVT'L L. 1105, 1116-19 (2021) ("[F]or many years, catastrophe neglect was the default across multiple policy contexts.").

2 *See* Balt. Gas & Elec. Co. v. Natural Res. Def. Council, Inc., 462 U.S. 87, 102-03 (1983) (upholding the Nuclear Regulatory Commission's decision to treat the risk of an accidental release of nuclear waste as non-existent).

3 *See* Milica Vasiljevic & Mario Weick, *Reasoning about extreme events: A review of behavioural biases in relation to catastrophe risks*, Economic and Social Research Council 6, 8 (2013) (reviewing the literature); Eliezer Yudkowsky, *Cognitive biases potentially affecting judgement of affecting global risks*, in GLOBAL CATASTROPHIC RISKS 91-119 (Nick Bostrom, Milan M. Cirkovic, & Martin Rees, eds., 2011).

4 *See generally*, SCOTT BARRETT, WHY COOPERATE?: THE INCENTIVE TO SUPPLY GLOBAL PUBLIC GOODS (2007).

5 *See* Andrew Healy & Neil Malhotra, *Myopic Voters and Natural Disaster Policy*, 103(3) AM. POL. SCI. REV. 387 (2009).

is true—and they all likely have some merit—that neglect is a mistake. Its tragic consequences were revealed by the Covid-19 pandemic. It may bring worse yet.

On January 20, 2021, President Biden issued a memorandum titled "Modernizing Regulatory Review."[6] The memorandum directed the head of the Office of Management and Budget (OMB) to produce recommendations for updating the regulatory review process and "concrete suggestions" for how it could promote several values, including "the interests of future generations."[7]

This essay suggests that one way to fulfill the Biden memorandum's promise to protect future generations is for policymakers to address catastrophic-risk neglect within the federal government. To that end, this essay proposes three things: First, amending OMB's Circular A-4, the "bible" of the regulatory state,[8] to include an explicit discussion of catastrophic and existential risks. Second, a set of guiding principles for the assessment and mitigation of such risks.[9] And finally, a proposed executive order (EO) requiring agencies to affirmatively consider and report on relevant catastrophic and existential risks. Broadly speaking, the principles and proposed EO aim to get agencies thinking about, and responding to, catastrophic risks, and the amendment to Circular A-4 aims to guide the evaluation of relevant actions once they are proposed. The principles are about agency agenda-setting and Circular A-4 is about the agency analytical process.

This essay proceeds in three Parts. Part I sets the stage by outlining the problem of catastrophic and existential risk. Part II explains the agency agenda-setting process and how the proposed principles and EO aim to influence it, as well as how both are designed to force agencies to give explicit reasons for their action or inaction. Part III contains a theoretical justification for the choices I have made in the amendment to Circular A-4. In particular, it justifies the reliance on quantified benefit-cost analysis (BCA) rather than the precautionary principle while acknowledging the need for qualitative analysis to accompany BCA in areas of extreme risk and high uncertainty.

---

6 WHITE HOUSE, MEMORANDUM: MODERNIZING REGULATORY REVIEW (2021).

7 *Id*.

8 CASS SUNSTEIN, AVERTING CATASTROPHE 24 (2021).

9 This proposal is in part modelled on a similar memorandum on the governance of emerging technologies issued by the Office of Science and Technology Policy and the Office of Information and Regulatory Affairs in 2011. *See* John. P. Holdren, Cass R. Sunstein, & Islam A. Siddiqui, Memorandum to Heads of Executive Departments & Agencies on Principles for Regulation and Oversight of Emerging Technologies (Mar. 11, 2011).

I. CATASTROPHES

Before diving into the details of my proposals, a brief survey of catastrophic risks is in order.[10] Although legal scholars have written extensively about the threat of catastrophe, the term has a flexible meaning, covering everything from events that threaten truly global destruction, such as pandemics and nuclear war, to others, such as terrorism, nuclear power plant accidents, and extreme weather events, that cause more limited devastation. One scholar has suggested that a disaster that kills ten thousand people would not qualify as a catastrophic risk, while one that killed ten million people would.[11] The Covid-19 pandemic—current estimated death toll, 22 million[12]—would thus count, and pandemics in general are a prominent catastrophic risk. Here, I use the term catastrophic risk narrowly, to refer to threats that have the potential to kill at least tens of millions of people.

Experts divide catastrophic risks into several categories, such as natural risks, anthropogenic risks, and future risks.[13] Natural risks include pandemics, supervolcanoes, massive asteroid or meteor strikes, and more remote possibilities such as a supernova close to our solar system. Anthropogenic risks include nuclear war, climate change, and other kinds of human-driven environmental damage. Future risks are those that have not seriously threatened humanity's survival yet but could do so in the future, thanks to societal or technological changes. These include biological warfare and bioterrorism, artificial intelligence, and technologically enabled authoritarianism. Although some individual risks may seem outlandish (and even for the more mainstream ones, such as climate change, extreme outcomes are quite unlikely), taken together, they pose a significant threat to humanity's future.

I note here another important distinction: between catastrophic and existential risks. There is a major difference between a catastrophe that kills 100 million people and one that causes human extinction—even between the death of several billion people and the death of everyone. Some researchers focus almost entirely on so-called existential risks, arguing that the destruction of humanity's entire future

---

[10] For overviews of the field of catastrophic risk, see generally, TOBY ORD, THE PRECIPICE (2020); GLOBAL CATASTROPHIC RISKS (Nick Bostrom, Milan M. Cirkovic, & Martin Rees, eds., 2011).

[11] Nick Bostrom & Milan M. Cirkovic, *Introduction* in Bostrom, Cirkovic, & Rees, *supra* note 10, at 24.

[12] *The pandemic's true death toll*, ECONOMIST, https://www.economist.com/graphic-detail/coronavirus-excess-deaths-estimates.

[13] *See, e.g.*, Ord, *supra* note 10, (proposing a categorization along these lines, with minor differences from the one I adopt here).

would be far worse even than a disaster that killed most people but left the possibility of recovery.[14] They have a least a plausible argument, but I will here consider catastrophic and existential risks together. For my purposes, the same analysis applies to both, as there is considerable overlap between the two. After all, several existential risks, such as nuclear war, biological weapons, artificial intelligence, and extreme climate change, are also catastrophic risks.[15] And policies that mitigate the risk of catastrophe will generally also guard against the risk of extinction. The two areas raise largely the same questions of public policy; often, only the numbers are different.

The literature on catastrophic risks is large and much of it concerns the details of individual risks, which I will not attempt to summarize here.[16] When addressing catastrophic risk as a whole, the scholarship makes three basic contentions: (1) humanity has acquired, in the past half century, the ability to destroy itself, or at least severely degrade its future potential;[17] (2) human extinction or collapse

---

[14] For a statement of this view, see Nick Bostrom, *Existential Risk Prevention as a Global Priority*, 4(1) Global Pol'cy 15 (2013).

[15] Ord, *supra* note 10.

[16] *See generally,* Ord, *supra* note 10; Stuart Russell, Human Compatible: Artificial Intelligence and the Problem of Control (2019) (an examination of catastrophic risks from AI); Hilary Greaves and William MacAskill, *The Case for Strong Longtermism*, Global Priorities Institute Working Paper Series (2019); Shahar Avin, *Classifying Global Catastrophic Risks*, 102 Futures 20 (2018); Marc Lipsitch, *Why Do Exceptionally Dangerous Gain-of-Function Experiments in Influenza?* 1836 Methods Molecular Biology 589 (2018) (warning of the risk of "a biosafety incident on a scale never before seen": "global spread of a virulent virus" accidentally released by a lab); Nick Bostrom, *Existential Risk Prevention as Global Priority*, 4 Global Pol'cy 15 (2013); Jason G. Matheny, *Reducing the Risk of Human Extinction*, 27 Risk Analysis 1335 (2007).

[17] The most commonly cited risks to human survival are nuclear war, climate change, biological threats, and artificial intelligence. *See* Ord, *supra* note 10, at 124-37 (surveying the evidence on the risks of natural pandemics, accidental lab releases, biological warfare, and bioterrorism); Brian Christian, The Alignment Problem (2020) (the risks of artificial intelligence); Russell, *supra* note 16 (same); Gernot Wagner & Martin Weitzman, Climate Shock: The Economic Consequences of a Hotter Planet (2015); Alan Robock, Luke Oman, & Georgiy L. Stenchikov, *Nuclear Winter Revisited with a Modern Climate Model and Current Nuclear Arsenals: Still Catastrophic Consequences* 112 J. Geophysical Res.: Atmospheres (2007) (providing the most recent rigorous modelling of the sequences of a large scale nuclear war). For a general discussion of existential threats, see Martin Rees, On the Future: Prospects for Humanity (2018).

would be an almost unimaginable catastrophe;[18] and (3) we can take concrete actions to quantifiably reduce the risk of disaster.[19] Each of these claims is plausible, although not uncontested, and together they add up to the conclusion that humanity should devote significantly more of its resources to mitigating the most extreme risks it faces.

---

[18] *See, e.g.*, Ord, *supra* note 10, at 20-21, 273-76 (arguing that the risk of human extinction presents a pressing moral problem); Piers Millett & Andrew Snyder-Beattie, *Existential Risk and Cost-Effective Biosecurity*, 15 HEALTH SECURITY 373 (2017) (providing quantitative estimates of how bad human extinction might be and the probability of different biosecurity catastrophes); Nick Bostrom, *Astronomical Waste: The Opportunity Cost of Delayed Technological Development*, 15 *Utilitas* 308 (2003) (attempting to quantify the number of potential lives that would be lost to human extinction).

[19] *See, e.g.*, Ord, *supra* note 10, at 277-81 (listing policy options to reduce existential risk); Russell, *supra* note 16, at 171-83 (outlining a proposal for reducing the risks from advanced artificial intelligence); Millett & Snyder-Beattie, *supra* note 18 (laying out a possible program to reduce existential biosecurity risks); Matheny, *supra* note 16, (suggesting possible interventions to reduce a range of severe risks).

II. Agency Agenda-Setting

If catastrophic and existential risks are neglected by the federal government, how can agencies be encouraged to focus on them? In general, three actors have the most sway over agency attention: Congress, the public, and the White House. I focus on the executive branch here, but many of the same ideas could be applied to potential legislative action. Around a third of agency regulations are the result of explicit congressional requirements.[20] Most of the remainder are workaday updates to prior regulations. Perhaps surprisingly, the White House appears to have little express influence over agency agenda-setting. The most explicit formal requirements come in the Reagan-era Executive Order 12,866, which requires agencies to hold annual priority-setting meetings, creates a regulatory working group to plan agency action, and forces agencies to submit information about their plans to the Office of Information and Regulatory Affairs (OIRA).[21] The EO's requirements, however, appear to have little practical effect, more "rote" rules than substantive influence on agency policy setting.[22]

The White House does display influence in some areas. Apart from the President's obvious role in selecting the heads of departments and agencies, the White House has a much greater influence over the creation and content of the most politically significant agency rules than over day-to-day agency action, much of which is probably of little interest to the President. [23] The White House can also shape how agencies go about their business by mandating rules of agency management.[24] Both President Obama and President Trump issued executive orders designed to shape the agency rulemaking process, including by specifying the kinds of evidence agencies could consider, setting reporting requirements, and conducting retrospective reviews of existing rules.[25] During the Obama administration, the White House's Office of Science and Technology Policy and

---

[20] Cary Coglianese & Daniel E. Walters, *Agenda-Setting in the Regulatory State: Theory and Evidence,* 68 Admin. L. Rev. 865, 871-73 (2016); William F. West & Connor Raso, *Who Shapes the Rulemaking Agenda? Implications for Bureaucratic Responsiveness and Bureaucratic Control*, 23 J. Pub. Admin. Res. & Theory 495, 504 (2013).

[21] West & Raso, *supra* note 20, at 510.

[22] Coglianese & Walters, *supra* note 20, at 880.

[23] *Id.* at 881.

[24] West & Raso, *supra* note 20, at 510.

[25] *See* Exec. Order No. 13,771, 82 Fed. Reg. 9,339 (Jan. 30, 2017); Exec. Order No. 13,563, 76 Fed. Reg. 3,821 (Jan. 18, 2011).

OIRA also issued a set of non-binding principles designed to shape regulatory policy for "emerging technologies."[26]

My proposed Principles for the Assessment and Mitigation of Catastrophic and Existential Risks, along with the proposed section of an EO, aim to play a similar role. The principles encourage agencies to affirmatively consider how they can mitigate catastrophic and existential risks related to their area of regulatory focus. They suggest that agencies develop a register of relevant risks and seek opportunities to address them. They also instruct agencies not to ignore a risk purely because it falls below an arbitrary probability threshold, something that is not unusual in some agencies.[27]

The proposed EO language is designed to be included in an order implementing President Biden's 2021 memorandum on modernizing regulatory review. The order would require each agency, if relevant, to submit to OIRA an annual Catastrophic and Existential Risk Plan that includes an assessment of risks relevant to the agency, expert estimates of the probability and magnitude of the threats along with associated uncertainties, and proposals for risk mitigation by the agency or the wider federal government. This last might include planned or proposed agency regulations or recommendations to Congress for legislative action. Along with the principles, the proposed EO is intended both to inform policymakers in OMB and the White House and to make sure that agencies pay attention to extreme risks, both in affirmatively developing proposals to mitigate them and in considering other agency actions that might implicate extreme risks.

---

[26] *See* John. P. Holdren, Cass R. Sunstein, & Islam A. Siddiqui, Memorandum to Heads of Executive Departments & Agencies on Principles for Regulation and Oversight of Emerging Technologies (Mar. 11, 2011).

[27] *See supra* note 1 and accompanying text.

III. Guiding Agency Policy Evaluation

Once the White House or OMB has succeeded in getting agencies to pay attention to extreme risks, a second problem arises: How should agencies deal with the tricky analytical questions they raise? Under EO 12,866 and Circular A-4, agencies are required to evaluate proposed regulations using either cost-effectiveness analysis (CEA) or benefit-cost analysis (BCA). The two areas encompass a wide range of approaches, but in short, CEA involves setting a pre-defined target, such as a level of emissions reduction, and determining the most cost-effective method of reaching it. BCA, on the other hand, takes a proposed agency action and evaluates its costs and benefits de novo; a regulation passes BCA only if the benefits exceed the costs.

In both cases, catastrophic and existential risks raise difficult issues. As this Part explains, such risks typically involve high levels of uncertainty. This uncertainty is often given as a justification for adopting a version of the precautionary principle to replace or supplement CEA or BCA.[28] My proposed amendment to Circular A-4 does not adopt the precautionary principle, for two reasons: first, whatever the precautionary principle's benefits in other situations, it is theoretically unjustified and practically indeterminate when dealing with extreme risks, and second, BCA has become so deeply entrenched in regulatory analysis that attempting to replace it is likely a fool's errand.

Extreme risks, however, also reveal the limits of quantitative analysis. The proposed amendment highlights the need to consider "fat tails" in BCA and the importance of including a rigorous qualitative analysis when the BCA assessment relies on uncertain estimates of very low probabilities of very large costs or benefits.[29] This Part explains those choices and provides more detail on the guidelines for evaluating policy for extreme risks.

Along with the principles and the EO, the amendment's focus on dual quantitative and qualitative analysis draws on a common theme in administrative law: the importance of reason-giving. Requiring decision makers to explain themselves, as Ashley Deeks has written, can improve decisions, promote efficiency, constrain policy choices, increase public legitimacy, and foster

---

[28] *See, e.g.*, Sunstein, *supra* note 8.

[29] Extreme risks can also require long-term thinking, which raises the issue of appropriate discount rates. I do not discuss discount rates in my proposals, as Circular A-4 adopts a broadly sensible approach and revisions seem unlikely, but in short, the theoretically justified approach in the long-run future, when discount rates are uncertain, is, as Circular A-4 notes, to adopt the "minimum discount rate having any substantial positive probability." *See* Circular A-4 at 36 (citing Martin Weitzman, *"Just Keep Discounting, But…"*, in Discounting and Intergenerational Equity (Paul R. Portney & John P. Weyant, eds.) (1999)).

accountability.[30] In the extreme risk context, a requirement that agencies justify their decisions both through traditional BCA and through qualitative explanations should encourage agencies to think through their choices in more detail, consult more closely with outside experts, and receive greater congressional and public input into their decisions.

A. Uncertainty

This Section considers, and rejects, one of the strongest objections to BCA: the problem of "deep uncertainty." Deep uncertainty, I argue, is an incoherent concept in practical policymaking and thus poses no support for the use of the precautionary principle and no obstacle to the use of BCA.

i. Explaining Deep Uncertainty

Every policy decision involves risk. From setting flood wall requirements to imposing mask mandates, regulators must weigh the probabilities of competing benefits and harms. No outcomes are guaranteed. CEA and BCA recognize this and require modelling and quantifying the relevant risks.

But what if the uncertainty is so great that the probabilities are unquantifiable? Some decision theorists distinguish "deep uncertainty" (where numerical probabilities cannot be given) from risk (where they can).[31] Under deep uncertainty, BCA cannot be undertaken.[32]

A famous intuitive explanation of deep uncertainty comes from John Maynard Keynes, who suggested that there are some events for which "there is no scientific basis on which to form any calculable probability whatever."[33] "The sense in which I am using the term [uncertainty]," Keynes wrote in 1937, "is that in which the prospect of a European war is uncertain, or the price of copper and the rate of interest twenty years hence, or the obsolescence of a new invention, or the position of private wealth owners in the social system in 1970."[34] On such questions, Keynes argued, we do not merely have low confidence in our probability estimates; we are unable to come up with any numbers at all.

---

30 Ashley Deeks, *Secret Reason-Giving*, 129 YALE L.J. 612, 626-34 (2020).

31 Deep uncertainty is also known as "Knightian uncertainty," after the economist Frank Knight, who formalized the concept. *See* FRANK KNIGHT, RISK, UNCERTAINTY, AND PROFIT (1921).

32 *See* JOHN KAY & MERVYN KING, RADICAL UNCERTAINTY: DECISION-MAKING BEYOND THE NUMBERS 57-65 (2020) [hereinafter "Kay & King"].

33 JOHN MAYNARD KEYNES, THE GENERAL THEORY OF EMPLOYMENT, INTEREST AND MONEY 113-14 (1936) (cited in Sunstein, *Averting Catastrophe* at 3).

34 *Id.*

If deep uncertainty exists, catastrophic risks seem especially likely to be subject to it. Climate change is a common example. Long term climate projections involve several steps, each of them subject to doubt. First, scientists must predict future emissions of carbon dioxide and other greenhouse gasses. Then they must create models of the climate that can estimate how those emissions will affect global temperatures and weather patterns. Those outputs must then be fed into yet more models to translate changes in the climate into changes in economic growth, human health, and other social outcomes. Some experts suggest that the final numbers, in the form of economic damages from higher temperatures, or, for example, the U.S. government's social cost of carbon, are "little more than a guess."[35] The economists John Kay and Mervyn King believe that deep uncertainty applies to catastrophic risks across the board: "to describe catastrophic pandemics, or environmental disasters, or nuclear annihilation, or our subjection to robots, in terms of probabilities is to mislead ourselves and others."[36] We can, they say, "talk only in terms of stories."[37]

Where numbers fail, one option is to turn to the precautionary principle. The principle, which has gained significant popularity in international and environmental law, comes in many forms, but most of them boil down to a single idea: act aggressively to ward off extreme threats even when their probability cannot be known.[38] This apparently bright-line rule cuts through the messiness of BCA to give policymakers an immediate answer. Better to build in a margin of safety, the thinking goes, by moving to avert catastrophe than trust models whose reliability we cannot judge until it is too late.

ii. Rejecting Deep Uncertainty

Yet the concept of deep uncertainty has come under withering critique from economists. Milton Friedman, for example, argued that Knight's distinction between risk and uncertainty was invalid. Even if people decline to quantify risks, Friedman noted, "we may treat [them] as if they assigned numerical probabilities

---

[35] David Weisbach, *Introduction: Legal Decision Making under Deep Uncertainty*, 44 J. Leg. Stud. 319, 320 (2015); *see also* Robert Pindyck, *Climate Change Policy: What Do the Models Tell Us?*, NBER Working Paper 19244 (2013), https://www.nber.org/system/files/working_papers/w19244/w19244.pdf.

[36] Kay & King at 73; *see also* Sunstein, *Averting Catastrophe* at 77-79 (endorsing the concept of Knightian uncertainty but positing a more limited set of circumstances in which it applies).

[37] *Id*.

[38] *See, e.g.,* Cass Sunstein, *Beyond the Precautionary Principle*, 151 U. Penn. L. Rev. 1003 (2003).

to every conceivable event."[39] People may claim that it is impossible to put a probability on a terrorist attack on the New York subway tomorrow,[40] but they go to work nonetheless; if they thought the probability was anything other than negligible, they would surely refuse. People must decide whether to take a flight, go to a café during a pandemic, or cross the road before the light has changed. In each case, they are making a probability determination, even if they don't admit it.[41]

An extension of this line of thinking, known as a Dutch book argument, attempts to demonstrate that refusing to assign probabilities is irrational.[42] The idea is that one can extract probabilities on any given question even from those who deny they have them by observing which bets on the issue they will and won't accept. A rational person, so the argument goes, should always be willing to take one side of a bet if she thinks she can win. But if she agrees to a set of bets that together violate the axioms of probability theory, whoever takes the other side will make a Dutch book against her—that is, make money off her. And following the axioms requires keeping track of one's numerical predictions (to make sure they sum to one in the appropriate places, and so on).

A central premise of the Dutch book argument—that any rational agent will be willing to take at least one side of any bet—has struck many commentators as far-fetched.[43] A natural response to decision theorists proposing bets on outlandish events is to back slowly away. Yet recall Friedman's argument: people act as if they have numerical probabilities because *they have to*. If a friend suggests we go to a restaurant during a Covid-19 outbreak, refusing to decide whether to go is not an option. I must weigh the risks and make a choice. If you are standing at a crosswalk, you can choose to cross or not, but you can't choose not to decide at all. You can decline to take up a bet, but you can't decline to make decisions in everyday life.

The same point applies to policymakers. When choosing among policies that will affect climate change, for example, a policymaker must pick an option, even if that option is "do nothing." And the choice comes with an implicit bet: that the costs associated with the policy will outweigh its benefits. There is no backing away

---

39 Milton Friedman, Price Theory 282 (2007).

40 *See* Kay & King at 63 (arguing that assigning a probability to airliners striking the twin towers before the 9/11 attacks was impossible).

41 *See* Sunstein *Averting Catastrophe* at 75-76 (outlining this argument).

42 The Dutch book idea originated with the economist Frank Ramsey. *See* Frank P. Ramsey, *Truth and probability*, in The Foundations of Mathematics and other Logical Essays 156-98 (1931). For an overview of Dutch book arguments, see Susan Vineberg, *Dutch Book Arguments*, Stanford Encyclopedia of Philosophy (2016), available at https://plato.stanford.edu/entries/dutch-book/.

43 *See, e.g.,* Kay & King at 65.

from the wager. If our policymakers choose wrong, reality will make a book against them—that is, society will be worse off.

One final objection: Cognitive irrationalities, such as the conjunctive fallacy[44] or loss aversion,[45] may mean that even if subjective probabilities can be extracted "by brute force," they will, as Jacob Gersen and Adrian Vermuele have put it, lack "any credible epistemic warrant."[46] Jon Elster sums up the case for skepticism: "One could certainly elicit from a political scientist the subjective probability that he attaches to the prediction that Norway in the year 3000 will be a democracy rather than a dictatorship, but would anyone even contemplate acting on the basis of this numerical magnitude?"[47] When faced with such uncertainty, Gersen and Vermuele conclude that maxmin (a variety of the precautionary principle that chooses the policy with the least bad worst-case outcome) and maxmax (choosing the policy with the best best-case outcome) are "equally rational" and the choice between them is "rationally arbitrary."[48]

But, as we have seen, probability estimates of some kind are inevitable, so, the right response to flaws in human reasoning is not to give up on making probability estimates; it is to make better ones.[49] It is true that question framing, loss aversion, the conjunctive fallacy, and other psychological biases make it difficult for humans to reason probabilistically. But intuitions and heuristics like the precautionary principle are just as subject to those biases. And policymakers are not relying on unreflective guesses made by someone who has just been confronted with a problem designed to produce an irrational answer. They can approach policy questions systematically, with full knowledge of the defects in human reasoning, and improve their estimates as they learn more. Policymakers will do best if, as the amendment to Circular A-4 suggests, they apply, as far as possible, standard quantitative methods to extreme risks.

---

44 *See* Amos Tversky & Daniel Kahneman, *Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment* 90 Psychological Rev. 293 (1983).

45 *See* Jon Elster, Explaining Technical Change 193 (1983).

46 Jacob Gersen & Adrian Vermeule, *Thin Rationality Review*, 114 Mich. L. Rev. 1355, 1385 (2016).

47 Elster, *Explaining Technical Change* 199 (quoted in Gersen & Vermuele, *supra* note 46).

48 Gersen & Vermuele, *supra* note 46, at 1385-86.

49 For a powerful case that better probability estimates are possible, *see* Philip E. Tetlock & Dan Gardner, Superforecasting: The Art and Science of Prediction (2015); Philip E. Tetlock, Expert Political Judgement (2005).

B. Indeterminacy

This Section explains why the precautionary principle is both often unduly risk-averse and, even when risk-aversion is called for, indeterminate in practice.

A common charge against the precautionary principle is that it does not know when to stop. The version of the precautionary principle known as maxmin suggested by John Rawls, for example, which seeks to eliminate the worst worst-case outcomes, was widely rejected by economists on the grounds that it called for infinite risk aversion. The same objection applies to all categorical formulations of the principle. Any possibility of harm to health or the environment, no matter how small, requires taking precautionary action.

To illustrate the problem, imagine a regulator who is presented with a new drug that lowers the risk of heart disease. Clinical trials show the drug is safe, and the most likely outcome is that approving the drug will save thousands of lives over the next ten years. But experts estimate there is a 0.001 percent chance that the trials have missed a major long-term safety problem.[50] In the worst-case scenario, millions of people take the drug and experience significant negative health effects. Maxmin requires our regulator to reject the drug. Because absolute safety is impossible, that cannot be the right answer.

Because policy choices frequently have catastrophic risks on both sides, the precautionary principle becomes paralyzing. Research into novel pathogens might provide tools to stop the next pandemic—or it might cause one.[51] Missile defense technology might make nuclear war less deadly—or it might set off an escalation spiral.[52] Artificial intelligence might design cures to humanity's worst diseases—or it might destroy humanity instead.[53] As Cass Sunstein has pointed out, while it is tempting to interpret the precautionary principle as warning against adopting risky technologies or policies, it cannot do so.[54] The principle cannot tell us whether it is riskier to conduct dangerous research or to ban it, to invest in missile technology or not to, to build AI systems or to refrain. Faced with such dilemmas, the precautionary principle leaves us in a kind of policy Bermuda Triangle, where risk is on every side and the compass needle spins.

---

[50] The numbers are stylized, but the situation is analogous to that of a doctor considering prescribing a drug that is overwhelmingly likely to help the patient but has "death" listed as a rare side effect.

[51] *See* Marc Lipsitch, *Why Do Exceptionally Dangerous Gain-of-Function Experiments in Influenza?*, 1836 Methods in Molecular Biology 589 (2018).

[52] *See generally,* Robert Jervis, The Illogic of American Nuclear Strategy 31 (1984).

[53] *See* Russell, *supra* note 16.

[54] Cass Sunstein, *Beyond the Precautionary Principle* 151 U. Penn. L. Rev. 1003, 1020-29.

One response to is to limit the use of the precautionary principle to situations of deep uncertainty. If no probability, even a subjective one, exists, then the cautious policymaker cannot be accused of letting a miniscule probability of a catastrophe dominate her decision making. But as we have seen, deep uncertainty is an incoherent concept for policymakers, who are always working with implicit probabilities. Another common response is to apply the precautionary principle threshold only above some threshold of plausibility.[55] Yet this brings us straight back to probabilities: the only way to set the threshold right (and presumably it must be set at different points for different risks: a 0.01 percent risk of a nuclear power plant accident might be acceptable while a 0.001 percent risk of a novel pathogen leaking from a lab might not) is to resort to numbers.[56] At that point, we might as well use the numbers to conduct BCA.

Even if we adopted a threshold probability level for the precautionary principle, there is an even more practical problem: the principle is useless for day-to-day policymaking. Consider a policymaker faced with the threat of a pandemic who adopts maxmin and asks which policy option forecloses the worst-case outcome. There is none. Various policies might mitigate the threat, but none can eliminate the possibility of disaster. Since every policy choice leaves at least some probability of the worst-case outcome, maxmin provides no guidance.

Perhaps our policymaker should adopt a more flexible form of the principle, one perhaps one that merely requires her to take precautions against the possibility of a pandemic. But which precautions? And how much effort and funding should go into them? There are many proposals and limited resources.[57] Worse still, some options are incompatible with one another. The precautionary principle can say little more than "Do something!" To decide what should be done, a more rigorous decision process is needed.

C. The Limits of Quantitative Analysis

That process should involve quantitative analysis. But although some form of BCA is the right starting point for assessing policies related to catastrophic and existential risk, it is not the whole ball game. As the proposed amendment to Circular A-4 suggests, when policymakers are considering catastrophic and existential risks, quantitative methods should be supplemented with detailed

---

[55] *See* Sunstein, *Averting Catastrophe.*

[56] *See generally*, H. Orri Stefánsson, *On the Limits of the Precautionary Principle*, 39 RISK ANALYSIS 1204 (2019) (demonstrating that the precautionary principle with a threshold violates plausible decision-theoretic principles).

[57] *See, e.g.,* Piers Millett & Andrew Snyder-Beattie, *Existential Risk and Cost-Effective Biosecurity*, 15 Health Security 373 (2017).

qualitative analysis. This Section explains why. One reason is that fat-tailed distributions are common in catastrophic risk and cause difficulties for expected value theory. Another is the problem of fanaticism, which arises when policy recommendations are dominated by uncertain estimates of very low probabilities of very large impacts. The solution is to begin with BCA but not to end there. A robust qualitative case for the policy is always necessary.

i. Fat Tails

Many things we encounter in daily life follow a normal distribution, in which most observations cluster around the center and extremes are rare. Take height. The average American man is 5 ft 9, and 95% of men are between 5 ft 3 and 6 ft 3. People above 8 feet are extraordinarily rare and those above 9 feet are nonexistent (the tallest human ever recorded was 8 ft 11).

In a fat-tailed distribution, by contrast, extreme events are much more common. Consider stock prices.[58] From 1871 to 2021, the largest monthly rise in the S&P 500 was 50.3 percent and the largest monthly fall was -26.4 percent.[59] If stock prices followed a normal distribution with the same mean and standard deviation as real stock prices, we'd expect the biggest monthly gain over 150 years to be around 16 percent and the biggest loss to be around -15 percent. Monthly price changes as large as those we observe would not happen if the stock market ran for the entire life of the universe. Stock prices are fat-tailed. Pandemics also follow a fat-tailed distribution, specifically a power law, in which those with the highest death tolls dominate all the others. Earthquakes, wars, commodity price changes, and individual wealth display the same pattern.[60] Other catastrophic risks, such as climate change, may also feature power law distributions.

Fat tails pose a problem for standard benefit-cost analysis. To see why, it is worth considering the case of climate change. The Intergovernmental Panel on Climate Change estimates a 90 percent chance that a doubling in atmospheric carbon dioxide levels will lead to between two and five degrees of warming, a factor known as climate sensitivity.[61] This implies a five percent chance of warming

---

[58] For more detail, see the explanation in William Nordhaus, *The Economics of Tail Events with an Application to Climate Change*, 5(2) REV. ENV. ECON. POL'CY 240, 243 (2011).

[59] Data from http://www.econ.yale.edu/~shiller/; analysis by the author.

[60] For an intuitive overview of fat-tailed distributions and their implications, see William Nordhaus, *The Economics of Tail Events with an Application to Climate Change*, 5(2) REV. ENV. ECON. POL'CY 240, 243 (2011).

[61] IPCC, *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* TS-130 (2021).

greater than five degrees. But as the economist Martin Weitzman has pointed out, we do not know the probability distribution of climate sensitivity.[62] If it is normally distributed, most of that five percent is clustered just above the five-degree mark, and there is little chance of warming above about six degrees. If climate sensitivity follows a fat-tailed distribution, the five percent is much more spread out, and there is a greater than one percent chance of warming above ten degrees.[63] There are physical mechanisms that could plausibly trigger such warming, including the release of methane trapped beneath arctic permafrost or the ocean floor, but we do not know how likely they are.

Weitzman has argued that the fat tails in climate impacts can break the standard tools of BCA. Weitzman's proposed "dismal theorem" suggests that under certain assumptions about uncertainty and societal preferences, the expected value of climate change risks is infinitely negative and normal economic tools cannot be used.[64] Economic analysis typically assumes that society is risk averse and that consumption has declining marginal utility (that is, an extra dollar is worth less when you are already rich). When facing a catastrophic risk, if the tails of the relevant probability distribution are fat enough, those assumptions can cause the expected cost of seemingly normal policy choices to become infinite, implying society should pay almost 100% of GDP to prevent the possibility of catastrophe. The result is counter-intuitive, but it comes quite straightforwardly from the utility functions typically used in economic analysis.[65]

---

[62] Martin Weitzman, *On Modeling and Interpreting the Economics of Catastrophic Climate Change*, 91 REV. ECON. STAT. 1 (2009).

[63] *Id.*

[64] *Id.*

[65] A simplified version of the analysis goes as follows. Recall two assumptions common in BCA. First, that society has some degree of risk aversion, meaning that we will accept somewhat lower growth in exchange for a lower risk of catastrophe (think of an investor who holds government bonds even though they earn a lower return than riskier debt). Economic analyses typically assume a constant relative risk aversion (CRRA)—a utility function in which as consumption falls (the agent gets poorer), risk aversion rises proportionally. The second assumption is that consumption has declining marginal utility, meaning that we value an additional dollar of consumption more when we are poor than when we are rich.

Put the two assumptions together with fat-tailed distributions and things start to get strange. CRRA utility functions take the form Uc=k1c1-, where c is consumption, the parameter $\alpha > 1$ is a measure of risk aversion (larger means more risk averse), and k1 is an arbitrary constant. If c follows a power law distribution, a common kind of fat-tailed distribution, then for small values of c (those than concern us here, as they indicate we are in the region of catastrophic loss), the probability distribution will be of the form fc=k2cfor values of $\beta>0$ (smaller means fatter tails) where k2is again an arbitrary constant.

Of course, expected utility cannot actually be negative infinity. The lower bound of utility is set by human extinction. While putting a value on humanity's continued survival is a tricky proposition, treating extinction as infinitely negative is implausible. On an individual level, humans are willing to trade off some risk of death against other values. Yet putting bounds on expected utility does violence to the cost-benefit calculation in other ways, as we are arbitrarily cutting off part of the relevant distribution, and the expected utility calculation will depend heavily on where exactly we set the bound.[66] Doing so is better than simply throwing up our hands at the problem of catastrophe, but the need to fudge the numbers suggests we should not rely exclusively on the standard expected utility calculations of welfare economics.

The bottom line is that when faced with threats of extinction, our standard tools of cost-benefit analysis are liable to produce strange results, and we should not take the numbers they produce too literally. As both the Circular A-4 amendment and the proposed principles suggest, policymakers should start with BCA, but they should properly incorporate uncertainties about key parameters, and they should recognize that the numbers those models produce are less definitive answers than suggestive indicators of the direction policy should take.[67]

ii. Fanaticism

Attempts to evaluate policy that rely at least in part on expected value calculations raise another worry: that very small probabilities of very bad (or very good) outcomes may dominate much higher probabilities of less extreme outcomes. This

---

The conditional utility function is thus given by the utility multiplied by the probability: $Ec=Ucfc=k1c1-\alpha k2ck=k1k2c1-\alpha+\beta$. The expected utility is given by the integral of $E(c)$ over the interval between zero and the maximum level of consumption. Discarding the constants k1and k2, the integral, and thus expected utility, converges to a finite number as c tends to zero only when $2-+\beta>0$. Otherwise, expected utility is unbounded. The upshot is that when risk aversion is very high (large α) or the tails are very fat (small β), expected utility is negative infinity. For a fuller explanation of the mathematics behind this result, see Weitzman 2009; Weitzman 2014; William Nordhaus, *An Analysis of the Dismal Theorem*, Cowles Foundation Discussion Paper (2009).

[66] *See* Weitzmann (2009), *supra* note 62; Martin Weitzman, *Fat-Tailed Uncertainty in the Economics of Catastrophic Climate Change*, 5 REV. ENV. ECON. & POL'CY 275 (2011); Martin Weitzman, *Fat Tails and the Social Cost of Carbon*, 104 AM. ECON. REV. 544 (2014).

[67] *See* Martin L. Weitzmann, *Fat Tails and the Social Cost of Carbon*, 104(5) AM. ECON. REV. 544 (2014) (arguing that the dismal theorem was best understood as a warning not to rely too heavily on quantitative models and that cost-benefit analysis of climate change policy should take into account the worst possible outcomes from climate change, rather than to literally imply that the damage could be infinite).

is a more general problem than the concern around fat tails. Imagine a policymaker who can fund either (1) a program that is highly likely to save a few thousand lives over the next five years—an air pollution reduction effort, for example—or (2) a program that has a very small (and hard to pin down) chance of averting human extinction—research into preventing the development of particularly effective bioweapons by terrorists, say. If the value of averting extinction is high enough, then almost any non-zero probability of success for option (2) will be enough to outweigh option (1).[68] And extinction could be very bad indeed. Richard Posner has given a "minimum estimate" of its cost as $600 trillion, and many other estimates go far higher.[69]

That can lead to some surprising results. Posner gives the example of particle accelerators, which some scientists suggested had a tiny chance of destroying the earth by creating a "strangelet," a specific structure of quarks that could collapse the planet and its contents into a hyperdense sphere.[70] The chances of that happening were extremely slim ex ante, but if extinction is bad enough, then perhaps regulators should have banned all particle accelerator development. This leads to an apparent reductio ad absurdum: the mere suggestion that an action could lead to human extinction should be enough to stymie it, a claim sometimes known as Pascal's Mugging.[71]

It may seem obvious that we should reject such logic, but doing so can lead to serious problems.[72] Imagine two policy options: (1), which has probability one of saving one life and (2), which has probability 0.99999 of saving 1,000,000 lives and

---

[68] *See generally,* Hayden Wilkinson, *In Defense of Fanaticism*, 132 ETHICS 445 (2022) (explaining the problem and arguing that rejecting expected value thinking leads to even bigger problems).

[69] RICHARD POSNER, CATASTROPHE: RISK AND RESPONSE 168-69 (2004). Posner's estimate values only those people who are already living and does so at a dollar figure of $50,000 per person. One attempt to estimate how many future people would not exist thanks to extinction puts the number at 10^{13} on Earth and 10^{27} if humans expand into the rest of the Solar System. Toby Newberry, *How many lives does the future hold?*, Technical Report, GLOBAL PRIORITIES INSTITUTE (2021), https://globalprioritiesinstitute.org/wp-content/uploads/Toby-Newberry_How-many-lives-does-the-future-hold.pdf.

[70] Posner, supra note 69, at 3-4; *see also* MARTIN REES, ON THE FUTURE 111-13 (2018) (considering this possibility in more detail).

[71] The idea, drawing on Pascal's wager, is that Pascal is accosted in the street by a man who demands $10 and promises in return not to carry out some terrible threat. If the philosopher relies on expected value calculations, he is required to hand over the $10, since no matter how unlikely the threat, there is some threat bad enough that the expected value of paying the money is positive. *See* Nick Bostrom, *Pascal's Mugging*, 69 ANALYSIS 443 (2009).

[72] Adapted from the explanation in Wilkinson, *supra* note 68, at 11-13.

zero value otherwise. Clearly, (2) is preferable. Now imagine (2)' which has a 0.99999^2 chance of saving $1{,}000{,}000^{10}$ lives and zero value otherwise. Our new option (2)' appears better than (2), or at least there is some number of lives saved for which it would be better. We could continue gradually reducing the probability of success and increasing the size of the payoff, such that each step along the way is better than the previous one, until the probability of a good outcome is arbitrarily small. At that point, we have either to accept the fanatical conclusion we resisted earlier or to claim that somewhere in the series of steps the policy option became unacceptable.[73]

In practice, there are good reasons to think that the probabilities we assign to outlandish claims should be small enough to avoid fanaticism problems. In a Bayesian framework, one has a general prior on the effects of certain kinds of actions—regulations on scientific research, say, or environmental clean-up efforts. Naïve expected value calculations for specific programs, on the other hand, will likely have a high variance. A local environmental cleanup, for example, may have well-known but limited health benefits, while an investment in speculative medical research may appear vastly more valuable thanks a small chance of an enormous benefit. But that variance should lead us to put greater weight on our prior and less on the specific estimate. On some plausible models of Bayesian updating, in fact, sufficiently high variance in the estimates of policy effects leads the associated probabilities to fall so fast that the expected value of the action actually declines as the claimed payoff increases.[74] That result makes intuitive sense: a claim that a policy can save ten lives may be plausible; a claim that it can save billions suggests that something has gone wrong in the evaluation process. That is one reason why my proposed amendment to Circular A-4 and the principles ask regulators to take a step back from any BCA that relies heavily on low probabilities of enormous costs or benefits and provide a plausible qualitative case for the proposed action. Of course, longshot bets will sometimes be worth it, but a policymaker should always be able to back up the numbers with a more intuitive argument.

---

[73] Another option is to reject the principle of transitivity.

[74] For a fuller explanation of one such model, see Holden Karnofsky, *Why we can't take expected value estimates literally (even when they're unbiased)*, GIVEWELL (July 25, 2016), https://blog.givewell.org/2011/08/18/why-we-cant-take-expected-value-estimates-literally-even-when-theyre-unbiased/.

Conclusion

Those who want the executive branch to address catastrophic and existential risks face two big problems: preventing neglect and ensuring reliable policy analysis. Each problem arises in the agency rulemakings covered by Circular A-4, but each is also far broader. Much federal action implicating catastrophic and existential risks, from setting guidelines on funding for research using potential pandemic pathogens to maintaining safety procedures on nuclear weapons, does not involve rulemaking. My proposals thus center on the regulatory process but extend well beyond it. The executive order and the principles attempt to shift agency attention and guide agency practice across government. They are also designed to prompt congressional and public attention through the reporting requirement.

The amendment to Circular A-4, meanwhile, deals with the second problem, guiding agency analysis once the agency is considering actions to mitigate catastrophic or existential risks, or when it is evaluating the impact of other actions on those risks. The amendment makes two overarching claims—that quantified benefit-cost analysis will produce better results than the precautionary principle even in situations of extreme risk and uncertainty, and that quantitative analysis nevertheless needs to be supplemented with rigorous qualitative explanations when dealing with complex or fat-tailed phenomena or other low-probability, high-impact risks.