2021 ASEE ANNUAL CONFERENCE
Virtual Meeting | July 26–29, 2021 | Pacific Daylight Time
Paper ID #34439

Engineering Existential Risks

Prof. Carl Mitcham, Colorado School of Mines

Carl Mitcham is International Distinguished Professor of Philosophy of Technology at Renmin University of China, Beijing, and Emeritus Professor of Humanities, Arts, and Social Sciences at Colorado School of Mines, Golden, Colorado. His publications include Thinking through Technology: The Path between Engineering and Philosophy (1994), Encyclopedia of Science, Technology, and Ethics (4 vols., 2005), Oxford Handbook of Interdisciplinarity (2010, with Robert Frodeman and Julie Thompson Klein), Ethics and Science: An Introduction (2012, with Adam Briggle), and Steps toward a Philosophy of Engineering: Historico-Philosophical and Critical Essays (2020). Additionally he served as a member of the Committee on Scientific Freedom and Responsibility of the American Association for the Advancement of Science (1994-2000) and on expert study groups for the European Commission (2009 and 2012). Awards include the International World Technology Network (WTN) award for Ethics (2006) and a Doctorate Honoris Causa from the Universitat Internacional Valenciana, Spain (2010). He holds the BA and MA in Philosophy from the University of Colorado and the PhD in Philosophy from Fordham University.

Engineering Existential Risks

Introduction

During the middle of the 20th century a number of new concepts emerged in ethical discourse. Supplementing traditional ethical-political concerns for distinguishing piety from impiety, the honorable from the shameful, virtue from vice, justice from injustice, good and bad, right and wrong, there emerged into ethical discourse such notions as autonomy and participation, responsibility, privacy, and sustainability. From its origins in military affairs and finance, risk (a word with Greek and Latin roots indicating nautical danger) was another such new entry in the moral lexicon. In all cases, the challenges of living in an increasingly engineered and engineering world were significant influences. Autonomy as participation took on new salience in biomedical research, responsibility was called for by nuclear weapons, privacy developed in conjunction with digital information technologies, and sustainability arose in response to industrial exploitation and environmental contamination. Questions of risk and its coordinate concept, safety, took on distinctly ethical relevance initially in civil and mechanical engineering (construction and tool or machine use safety) but from there spread throughout the engineering disciplines. In the form of risk-cost-benefit analysis it now poses a special challenge for any consequentialist moral theory.

Quantitative and Qualitative Risk Inflation

Since the 1970s the literature on risk and its challenges has ballooned. Literature (and risk work) is commonly parsed into categories dealing with the practices and problematics of (1) risk identification, (2) risk assessment, (3) risk management, and (4) risk communication. In all cases, however, risk issues are mostly assumed to be bounded: that is, to apply only to particular projects, locations, processes, or people. Concerns about the Cold War risks of thermonuclear warfare broke the boundaries to consider more comprehensive or catastrophic, global risks: in the famous phrase of engineer physicist and military strategist Herman Kahn [1], it forced "thinking about the unthinkable." Although nuclear related discourse subsided during *détante* and the end of the Cold War, the 1980s witnessed the emergence of new worries about global environmental risks, as exemplified in the stratospheric ozone depletion problem. Environmental to ecological mutation is clearly a cross-border issue [2]. As often noted, this problem was both caused and solved by chemical engineers who invented Freon and then, when its ozone depleting properties were discovered, engineered a replacement refrigerant. Adoption of the technical engineering fix nevertheless required a degree of international political engineering as well in order to facilitate the 1987 signing of the Montreal Protocol.

As has also regularly been noted, one complicating feature of risk is unintended consequences. With any human activity, as interactive complexities increase, it becomes progressively difficult to anticipate or predict a full spectrum of outcomes. This is true not only in human affairs, where political counsel has traditionally stressed the virtues of prudence and moderation, it is equally applicable to engineering. But just as in public affairs ambition often trumps moderation, so too in engineering. Use and convenience problem solvers subservient to military and civilian capitalist enterprises serving a consumer hungry public can find it hard to act with probity. It is difficult to know how to enact the fundamental canon of professional engineering ethics to

protect public safety, health, and welfare when the public itself lacks clarity and commitment to these values, the perceptions of which are subject to what Edward Bernays [3], the inventor of public relations, termed the "engineering of consent." The precautionary principle adopted by the European Commission [4] is an attempt to institutionalize some level of probity, but precaution has faced broad corporate public relations funded resistance in the United States, with its ideological commitment to "market-based solutions."

Existential Catastrophic Risk

Post-Cold War efforts to extend concern for global catastrophic risk can be traced back to Canadian philosopher John Leslie's unusual book *The End of the World: The Science and Ethics of Human Extinction* [5]. This was the first systematic parsing of global catastrophic risks from potential natural disasters (e.g., asteroid impacts) and from human activities (e.g., nuclear warfare) with an effort to catalogue those already well recognized while identifying many more largely unrecognized. Interestingly, Leslie added a set of risks from religion and philosophy. It would, he wrote, be a risk "to choose as Secretary for the Environment some politician convinced that, no matter what anyone did, the world would end soon with a Day of Judgement [or someone] who felt that God would keep the world safe for us for ever" [5, p. 10]. Schopenhauerian pessimism and ethical relativism pose similar risks.

Subsequent stimulus came from a book of popular scientific journalism by British cosmologist Martin Rees titled *Our Final Hour: A Scientist's Warning* with a second subtitle *How Terror, Error, and Environmental Disaster Threaten Humankind's Future in This Century — On Earth and Beyond* [6]. As Roger Pielke Jr. [7] noted in a review in *Science*, Rees makes multiple references to pop culture catastrophism in fiction and film but fails to deal seriously with issues of engineering governance, instead opting to promote space exploration and colonization as the ultimate answer to the risk of total human annihilation on the Earth, whether unintentionally engineered or not.

Philosopher Nick Bostrom, as founding director of an interdisciplinary Future of Humanity Institute (FHI) at Oxford University in 2005, has invested the most intellectual capital in advancing discourse on "existential risk." (FHI should not be confused with the Future of Life Institute or FLI founded by engineer entrepreneur Elon Musk at an MIT conference in 2014, but does ally itself with the Centre for the Study of Existential Risk and the Machine Intelligence Research Institute co-founded by Martin Rees at Cambridge University in 2012.) Bostrom argues that threats or risks have become not simply global but existential, with an "existential risk" defined as one in which "an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential" [8, section 1.2]. In a 2008 volume on *Global Catastrophic Risks* Bostrom and colleague Milan Ćirković [9] collect leading essays on a spectrum of existential risks, natural and anthropogenic. (An illustrated edition followed.)

An Existential Risk Precipice

The most recent summarizing and popularizing FHI effort at risk communication is Toby Ord's *The Precipice: Existential Risk and the Future of Humanity* [10], which argues that safeguarding

"humanity's future potential" is among the most important moral issues of our time. Although *The Precipice* has a style of analytic philosophy it is not the standard analytic philosophy fare. To begin, it is long: 468 pages. But the core argument is in the first 249 pages; the remainder is seven appendices, recommendations for further reading, 133 pages of technical notes, and bibliography.

The Precipice begins, in part one, with an argument for its importance: the very practical one that human beings are facing the precipitous prospect of numerous possible global, existential catastrophes. Ord reiterates and elaborates on Bostrom's definition of an existential risk by distinguishing two types of existential catastrophe: (1) absolute extinction or simply failed continuation of human progress, with failed continuation further distinguished into (2) unrecoverable collapse and (3) unrecoverable dystopia. In all three cases, however, Ord describes *existential catastrophe* as "the destruction of humanity's longterm potential" and *existential risk* as "a risk that threatens the destruction of humanity's longterm potential." Ord is thus concerned not simply with extinction of the human species but with a failure of what he terms "humanity" to achieve what he designates its "longterm potential" defined as "the set of all possible futures that remain open to us" [10, p. 37].

Part two identifies existential risks in three categories — currently existing natural risks, currently existing anthropogenic risks, and risks involved with future technological developments — referencing abundant empirical studies in support of each and exercising a lot of analytic muscle to develop quantitative risk assessments for different scenarios. Part three makes arguments for ways to manage the risks, one of which is to raise consciousness, including philosophical consciousness, about them. The book exemplifies not just analytic philosophy but — in its concern for what William James called the "cash value" of ideas — an emphatic version of pragmatism. Given the alliance with analysis, empirical data, and the attempt to design concrete problem solutions, it could even more accurately be described as turning philosophy into engineering or as engineering philosophy. I set aside here comment on or criticism of the replacement of philosophical questions with engineering problems (but see Cera [11]) as well as questions about Ord's philosophical anthropology and conception of human flourishing (which privilege potential over actuality), in order to focus instead on a singular failure to thematize engineering.

Chapters four (anthropogenic risks) and five (future risks) can serve as cases in point. The risks or threats examined include those engaged with nuclear engineering (weapons and power), chemical engineering (climate change and other environmental damages), geoengineering (overreach when response to climate mutation), bioengineered pandemics, and artificial intelligence. Threats involved can be either intended (from warfare, terrorism, and economic competition) or unintended (side effects, secondary effects, dual-use, etc.). Engineers often present their work as solutions to risks rather than as their causes. Yet engineering also creates risks, just of a different order from those it reduces. The emergence of existential risks from engineering prowess suggests a need for critical examination of the risk-cost-benefits, not just of particular engineering projects and processes, but of engineering more generally, including the risk-talk that has become endemic to our engineering world (see a classic essay by Langdon Winner [12, ch. 8]). Engineering is the foundation of our techno-human condition and deserves

assessment not only in terms of manifest benefits but also in relationship to its own fragilities, threats, and global mutations.

This point deserves reiteration. As Kristin Shrader-Frechette [13] has argued at length, there is a persistent tendency for technical appeals to experts and expertise to both underestimate risk. There is an equally persistent among technical experts to continue engineering technologies that they themselves assess as posing significant risks. One post-Ord illustration is Nicole Perlroth's *This Is How They Tell Me the World Ends: The Cyberweapons Arms Race* [14], a journalist's account of what experts who are developing it themselves perceive as the risk-laden vulnerabilities being introduced into human affairs by an increasingly engineered dependence on a cyber-infrastructure.

Neglect of Engineering

Ord fails to name engineering as the significant element it is (the term does not even occur in the index), instead preferring to reference the products of engineering, that is, technology. Still, his attitude toward it is modestly nuanced. In his words,

> Growing up, I had always been strongly pro-technology. If not for the plausibility of … catastrophic risks, I'd remain so. But instead, I am compelled towards a much more ambivalent view. I don't for a moment think we should cease technological progress — indeed if some well-meaning regime locked in a permanent freeze on technology, that would probably itself be an existential catastrophe, preventing humanity from ever fulfilling its potential.
>
> But we do need to treat technological progress with maturity. We should continue our technological developments to make sure we receive the fruits of technology. Yet we must do so very carefully, and if needed, use a significant fraction of the gains from technology to address the potential dangers, ensuring the balance stays positive. Looking ahead and charting the potential hazards on our horizon is a key step. [10, p. 123]

He even goes so far as to put a monetary value on the fraction of gains from technology to invest in thinking about potential dangers so as to exercise better engineering governance: at least as much as we spend each year on ice cream [10, p. 63].

Ord reiterates the issue in his penultimate chapter, "Safeguarding Humanity," with a version of the cultural lag thesis. As he puts it, "the current predicament stems from the rapid growth of humanity's power outstripping the slow and unsteady growth of our wisdom" so that

> a more patient and prudent humanity would … try to limit this divergence. Most importantly, it would try to increase its wisdom. But if there were limits to how quickly it could do so, it would also make sense to slow the rate of increase in its power — not necessarily putting its foot on the brake, but at least pressing more lightly on the accelerator. [10, p. 206]

One way to restate Ord's belief in terms of the activity of designing, constructing, and operating technologies, that is, of engineering, might go something like this: Engineering existential risks

can implicate two different but not unrelated activities. It can mean either the engineering creation of existential risks or engineering in response to existential risks. For the existential risks created by engineers (to whatever unintentional extent) the only reasonable solution is more engineering. Existentially risky engineering must be complemented with existential risk reduction engineering. Such would constitute a new form of what engineer philosopher Samuel Florman once termed "the existential pleasures of engineering" [15].

The pursuit of pleasure can be dangerous. Adapting Ord (by replacing his royal "we" with "engineers"), he argues that engineers should devote their energies "to promoting the responsible deployment and governance of new technologies [because] unprecedented power from technological progress requires unprecedented responsibility" [10, p. 207]. This has been called the Spiderman principle, which may also demand unprecedented restraint.

Engineering Limits

One implication of this claim about existential engineering in the positive sense as a necessary solution to existential engineering in a negative sense is that risk creating engineers have an ethical obligation to ensure risk reduction engineers will be around in the future to deal with unforeseen developments. Since engineering as a knowledge-based skill set is like everything subject to historical generation, corruption, and mutation, engineers need to be critical of unreasonably long-term, large-scale design specifications. So far as I know, only philosopher of engineering Michael Davis has attempted to address this professional responsibility, one that could contribute to reducing pressure on the engineering accelerator.

In a recent article Davis begins with the observation that planning for future outcomes is inherent to the engineering design activity. Indeed, in many instances, "designing" and "planning" are convertible terms. He then ventures to ask whether there might be any temporal limits on such planning. As he writes,

> The temporal limits of engineers' planning is an important topic for engineers. Engineers are increasingly concerned with "life-cycle planning" and "sustainable development." Life-cycle planning is a guide for treating an artifact (or process) that begins with its conception and ends when the artifact has ceased to have a distinct existence — which can, in practice, be quite far in the future. [16, p. 1610]

The engineering commitment to sustainable development defined as "meeting the needs of the present without compromising the ability of future generations to meet their own needs" (a principle in many engineering codes of ethics) requires long-range projections into the future. Moreover, planning is necessarily based not on knowledge (only the past and present are truly knowable) but probabilities. Taking the engineering of the Yucca Mountain Nuclear Waste Repository as a case in point, Davis then argues that the design criteria were so long-term that the project "seemed more like science fiction than like engineering, especially to engineers."

With regard to such a project, Davis asks whether engineers as such should limit the timeframe of their planning to a hundred years, a thousand years, ten thousand years? "Is there a time beyond which engineers as such cannot plan? Is there a time beyond which they should not

plan?" [16, p. 1611]. Davis's argues that since engineering projects necessarily require engineers to operate and manage them,

> engineers as such can at most plan only a little farther into the future than they can reasonably expect engineers of the right sort to be present (two or three generations). Beyond that time, engineers can still plan, perhaps even successfully, but not as engineers. [16, p. 1611]

They have stepped outside the bounds of their professional engineering competence. Most provocatively with regard to the particular case a nuclear waste repository, Davis maintains that "today's engineers should refuse to work on a project like the Yucca Mountain" (p. 1620).

There are many nuances to Davis argument. He is a more careful philosopher than Ord. Suffice to suggest here that his ethical provocation can have salutary implications for the cultivation of probity on what Ord calls the precipice of existential risk.

Conclusion

I have taken Toby Ord's *The Precipice* as an occasion, first, for calling attention to existential risk as a new category of risk, which is itself a new morally salient concept in ethics. Existential risk, or risk that threatens the global catastrophic collapse of civilization if not extinction of the human species, obviously raises the stakes in the risk work of identification, assessment, management, and communication. Second, I have criticized Ord (and other philosophers working in the new philosophy of existential risk, especially those associated with the interdisciplinary Future of Humanity Institute at Oxford University, where Ord is based) for failing to thematize engineering as a key element in both the engineering creation of existential risks and engineering efforts to mitigate or adapt to existential risks. Finally, third, I have suggested that, in accord with Ord's brief for slowing the rate of increase of engineering prowess in order to narrow the gap between expanding engineering power and ethical political wisdom, it would be useful to consider an argument advanced by Michael Davis regarding the temporal limits on engineer planning. Were engineers to adopt as a professional ethical responsibility the refusal as engineers to undertake projects that depend on planning specifications that fall outside the reasonable temporal bounds of their professional competence, this might help introduce a small measure of probity into engineering practice that might in turn militate against existential risk creating engineering.

References

[1] Kahn, Herman. (1962) *Thinking about the Unthinkable*. New York: Horizon Press.

[2] Latour, Bruno. (2018) *Down to Earth: Politics in the New Climatic Regime*. Trans. Catherine Porter. Medford, MA: Polity.

[3] Bernays, Edward L. (1947). "The Engineering of Consent," *Annals of the American Academy of Political and Social Science*, vol. 250, no. 1 (March), pp. 113–120.

[4] Commission of the European Communities. (2000) "Communication from the Commission on the Precautionary Principle." COM 2000 (1) Brussels.

[5] Leslie, John [A]. (1996) *The End of the World: The Science and Ethics of Human Extinction.* London: Routledge.

[6] Rees, Martin. (2003) *Our Final Hour: A Scientist's Warning: How Terror, Error, and Environmental Disaster Threaten Humankind's Future In This Century—On Earth and Beyond*. New York: Basic Books.

[7] Pielke, Roger A. Jr. (2003) "Which Future for Humanity?" *Science*, vol. 301, issue 5639 (12 September), pp. 1483-1484.

[8] Bostrom, Nick. (2002) "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards," *Journal of Evolution and Technology*, vol. 9, no. 1 (March).

[9] Bostrom, Nick, and Milan Ćirković, eds. (2008) *Global Catastrophic Risks*. Oxford, UK: Oxford University Press.

[10] Ord, Toby. (2020) *The Precipice: Existential Risk and the Future of Humanity*. London: Bloomsbury.

[11] Cera, Agostino. (2020) "Beyond the Empirical Turn: Elements for an Ontology of Engineering," *Információs Társadalom*, vol. 20, no. 4, pp. 74-89.

[12] Winner, Langdon. (1986) *The Whale and the Reactor: A Search for Limits in an Age of High Technology*. Chicago: University of Chicago Press.

[13] Kristin Shrader-Frechette. (1991) *Risk and Rationality: Philosophical Foundations for Populist Reforms*. Berkeley, CA: University of California Press.

[14] Perlroth, Nicole. (2021) *This Is How They Tell Me the World Ends: The Cyberweapons Arms Race*. London: Bloomsbury.

[15] Florman, Samuel C. (1976) *The Existential Pleasures of Engineering*. New York: St. Martin's Press.

[16] Davis, Michael. (2019) "Temporal Limits on What Engineers Can Plan," *Science and Engineering Ethics*, vol. 25, pp. 1609-1624.