Crucial Phenomena

Daniel Dewey*

daniel.dewey@philosophy.ox.ac.uk
Oxford Martin Programme on the Impacts of Future Technology,
Future of Humanity Institute

Abstract

I give a case that, as a public good, societies and their governments should support and invest in scientific research on *crucial phenomena,* empirical features of the world that figure strongly in how humanity's choices influence the size of its future. In particular, I give reasons for thinking that (1) humanity's vulnerability or robustness to accidents arising from biological engineering, and (2) the future rates of improvement of artificial intelligence and its susceptibility to misuse, are phenomena that call strongly for our systematic attention.

1 Introduction

In his 1986 talk *You and Your Research*, Richard Hamming recounts a story about his time at Bell Labs:

> Over on the other side of the dining hall was a chemistry table. I had worked with one of the fellows, Dave McCall; furthermore he was courting our secretary at the time. I went over and said, "Do you mind if I join you?" They can't say no, so I started eating with them for a while. And I started asking, "What are the important problems of your field?" And after a week or so, "What important problems are you working on?" And after some more time I came in one day and said, "If what you are doing is not important, and if you don't think it is going to lead to something important, why are you at Bell Labs working on it?" I wasn't welcomed after that; I had to find somebody else to eat with! That was in the spring.[1]

The individual researcher will often have practical answers for Hamming's final question: funding may not be available for important problems in one's field, or one might not have the particular skills and resources that would be required to tackle those problems. However, it is not so easy to escape from Hamming's questions when they are asked of the entire scientific community, or of society as a whole: what are the important problems

*Supported by the Alexander Tamas Research Fellowship on Machine Superintelligence and the Future of AI.

[1] Hamming and Kaiser, "You and your research".

of our time, and what problems are we working on? If resources are misallocated or the required skills are not available, then we have nobody to blame for this but ourselves.

This essay puts forth the idea that one of the most important tasks facing us today is the scientific investigation of certain *crucial phenomena*, empirical features of the world that figure strongly in how humanity's choices influence the size of its future. Technical, publicly understandable knowledge of these phenomena is important in two senses: it has great value in terms of consequences, and it seems to be reasonably achievable from our current position (in Hamming's terminology, we "have an attack" on these phenomena). In particular, I will give reasons for thinking that (1) humanity's vulnerability or robustness to accidents arising from biological engineering, and (2) the future rates of improvement of artificial intelligence and its susceptibility to misuse, are phenomena that call strongly for our systematic attention.

I begin with a series of arguments culminating in an moral rule of thumb that we ought to maximize the chance that humanity's future is "Large" instead of "Small"; then, I show the relevance of crucial phenomena to this endeavor. Finally, I put forward the claim that since in many cases we do not know enough about crucial phenomena to make good decisions, we ought to be working towards scientific knowledge of crucial phenomena, and that we should focus on the most time-sensitive ones.

This essay is built of significant insights from several people, relying particularly heavily on the ideas of Nick Bostrom and Nick Beckstead. My incremental contribution is to compile these ideas into a form that makes the value and urgency of certain kinds of of scientific knowledge clear, and to argue that the acquisition of this knowledge is one of the best available policies for humanity today.

2 Aim for a large future

Significant credence must be given to the idea that many times more potential value lies in humanity's long-term future – let's say the year 2100 and onward – than lies in its short-term future, between now and 2100. One way of supporting this proposition is to combine the following three plausible premises, two moral and one empirical:

1. Value is likely to be aggregative: more of a good thing is more valuable, and returns do not diminish quickly. "Good things" could be happy experiences, virtuous people, beautiful works of art, etc.

2. Intrinsic value is likely to be time-insensitive: whether a thing exists in the year 571 BC, 2014 AD, or 20014 AD does not affect its intrinsic value.

3. Humanity's long-term future has the potential to contain vastly more good things than its near-term future.

The third premise is the empirical one, following from the idea that humanity's long-term future could plausibly be *many times larger* than its near-term future. It could be many times larger in duration, since the universe is expected to continue in a usable state for at least on the order of billions of years. It could also be many times larger in 'breadth', roughly the number of 'things' (people, artifacts, etc.) humanity influences at any given

time, as increasing technological ability grants us increasing access to the resources of Earth, our solar system, our galaxy, and so on outwards to larger stages.

It follows, then, that if we can meaningfully affect humanity's long-term future, then it is immensely important that we do so. It is plausibly much, much more important to influence humanity's future from 2100 onward, than it is to influence the mere 86 years we have remaining between now and 2100. I will not try to argue the point conclusively here, since there are many subtleties and others have done so much better; I refer the curious reader especially to Beckstead, "On the overwhelming importance of shaping the far future".

❧

How could we meaningfully affect humanity's long-term future? Nick Bostrom approaches the problem by pointing to the category of *existential risks*, risks that threaten "the premature extinction of Earth-originating life or the permanent and drastic reduction of its potential for desirable future development".[2] By definition, any risk of the loss of a significant part of humanity's long-term future is covered by Bostrom's concept of existential risk. Given this definition of existential risk, Bostrom argues that it may be useful to adopt a moral rule of thumb which he calls "maxipok": "Maximize the probability of an "OK outcome," where an OK outcome is any outcome that avoids existential catastrophe."

Technically, the second clause of Bostrom's definition of existential risk renders the first redundant; extinction is an example of an event that could permanently and drastically reduce humanity's potential for desirable future development. In another writing[3], Bostrom sheds some light on this choice of emphasis:

> The notion [of existential risk] is more useful to the extent that likely scenarios fall relatively sharply into two distinct categories – very good ones and very bad ones. To the extent that there is a wide range of scenarios that are roughly equally plausible and that vary continuously in the degree to which the trajectory is good, the existential risk concept will be a less useful tool for thinking about our choices. One would then have to resort to a more complicated calculation. However, extinction is quite dichotomous, and there is also a thought that many sufficiently good future civilizations would over time asymptote to the optimal track.

In other words, the concept of existential risk is most useful if futures can be roughly sorted into two categories, *Prosperous* and *Disastrous*; human extinction, being dichotomous, supports this sorting hypothesis.

❧

In the definition of existential risk, what is this "desirable future development" that could be drastically reduced? Though this depends on open questions in moral philosophy, it seems to me that we can approximate the desired quantity by the *size* of humanity's future, as defined in section 2 – humanity's duration in time, times its 'breadth' in terms

[2] Bostrom, "Existential risk prevention as global priority".

[3] This quote appeared as a comment on Beckstead, *A proposed adjustment to the astronomical waste argument.*

of the matter and energy (and other resources) controlled by humans[4]. In this case, one could simply consider the *size* of the future to get a fairly good guide for how desirable it is. Does this "size view" fit with the existential risk picture? Let us examine the four categories of Disastrous (existentially catastrophic) futures that Bostrom lists: human extinction, permanent stagnation, flawed realization, and subsequent ruination.

Extinction: Presumably since extinction eliminates anything that could be considered "future development" at all, Bostrom does not further explain extinction's impact on existential risk.

Permanent stagnation: The value lost through permanent stagnation is exemplified by these futures' inability to "produce astronomical numbers of extremely long and valuable lives" – to create a 'broad' future, in my term from above.

Flawed realization: Flawed realization comes in two varieties, "unconsummated" and "ephemeral". Unconsummated realization covers disasters in which something critical about value has been lost on the way to technological maturity; for example, humans may have replaced themselves with artificially intelligent machines, but accidentally failed to make these machines so that they could have phenomenal experience, resulting in a hugely broad, long future with "no morally relevant beings there to enjoy the wealth". In an ephemeral realization, humanity crashes down to extinction or permanent stagnation shortly after reaching technological maturity.

Subsequent ruination: Subsequent ruination refers to futures in which humans reach technological maturity with all critical aspects of value intact, but through failure of luck or wisdom do not realize much of the potential desirable future development. Bostrom emphasizes that this situation seems less likely, given the considerable resources at humanity's disposal, and that we are not in much of a position to help these future people in any case; they seem to have been given all of the advantages we could hope to give them.

In all but one of these cases (unconsummated realization), the badness of existential risk, the "future development" that is lost, can be accounted for in terms of the size of humanity's future, its breadth and duration, without considering other qualities. Given this observation, one could propose a new way of characterizing Bostrom's "very good" and "very bad" categories of futures. The hypothesis would be that plausible futures of humanity naturally split into Prosperous "Large" futures, where humanity's future has a long duration and is broad in terms of resources controlled, and Disastrous "Small" futures, in which humanity is either short-lived or "narrow", having relatively little control over resources.

This in turn suggests a new rule of thumb, in parallel to Bostrom's maxipok rule:

> **Aim for Large:** Maximize the probability of humanity's future being Large instead of Small.

[4]This is not to say that size is intrinsically valuable. The assumption is that future humans will figure out how to do something good, given enough size.

Relative to maxipok, aim-for-large's advantage is that it cashes out normative language – "potential for desirable future development" – in terms of the relatively concrete duration and breadth of (i.e. resources controlled by) humanity's future[5].

This is certainly not to say that all Large futures are good, or are better than all Small futures, just that a future's size is an unusually useful piece of information about how good a future is. Choosing to focus only on whether humanity's future is Large or Small loses some nuance; the example of an unconsummated realization, in which humanity's future is long and broad but lacks value, makes this clear. However, for practical purposes, there is much to be gained by cashing out normative language in concrete terms.

3 Crucial phenomena

The aim-for-large rule leaves us with a new question: how *can* humanity act to maximize the probability of having a Large future? We cannot wish a chosen future into existence; instead, our choices interact with features of the world, and the fundamental and emergent laws that govern those features determine how our choices affect humanity's future's duration and breadth. Our ability to choose effectively depends on our knowledge of these empirical phenomena.

Crucial phenomena are empirical phenomena that play a key role in determining how humanity's actions influence whether its future is Large or Small. By *empirical phenomena*, I mean relationships that hold between sets of real-world conditions. "The moon waxes and wanes in such-and-such a pattern" is an empirical phenomenon. Physics-based phenomena such as the phase transitions of water, emergent phenomena such as the relationships between predator and prey populations, and mathematical "phenomena" that become realized in the world, such as the difficulty of factoring a large composite number found encoded on a hard-drive, are also empirical phenomena. This broad usage is meant to capture all kinds of patterns that are found in features and behaviours of the world.

Some phenomena are much more relevant than others in determining whether humanity's future is Large or Small. For example, while different laws of plate tectonics could result in dramatic differences in future arrangements of planetary oceans and landmasses, it seems unlikely that these differences would result, *ceteris paribus*, in significant differences in humanity's duration or breadth. On the other hand, humanity's future size could be dramatically impacted by the cosmological rate of expansion, which determines how much matter is ultimately reachable by humans. Different cosmologies have radically differently-sized futures of humanity.[6]

It may be instructive to imagine that humanity's future is determined by a game whose players are Humanity and Nature. Each player has a number of parameters that they are allowed to set; Humanity's parameters correspond to its choices, and Nature's parameters correspond to empirical phenomena. There are some of Nature's choices that will affect the outcome of the game greatly, and some that will do so in such a way that Humanity

[5]It would be misleading to say that aim-for-large does not contain any non-concrete or normative language; the definition of "humanity" plays a key role in defining the depth and breadth of humanity's future, and the definition of "humanity" is surely normatively loaded and subject to debate.

[6]Ćirković, "Cosmological forecast and its practical significance".

would benefit greatly from being given a peek at Nature's move; given the knowledge of how Nature sets its phenomena, Humanity could act to maximize the value of their play.

One relatively natural categorization of crucial phenomena I have found splits the full set into four subsets, induced by two binary qualities: each phenomenon affects either primarily *duration* or *breadth*, and does so in a way that is either *limiting* or *transformative*. These distinctions can be best understood by enumerating the four categories:

Duration-limiting phenomena: Duration-limiting phenomena are crucial by virtue of their potential to limit the possible duration of humanity's future; they set bounds on how long or short our future could be, often in ways that our choices cannot affect. For example, vacuum collapse could act as a duration-limiting phenomenon, as could physical factors such as the decay time of protons, or cosmological factors such as the time until a big crunch or similar universe-ending event.

Breadth-limiting phenomena: Breadth-limiting phenomena are crucial by virtue of their potential to limit the possible breadth of humanity's future. For example, phenomena that determine the material cost of taking control of additional solar systems (the density of interstellar dust, failure rates of relevant technologies, etc.) affect the potential breadth of humanity's future, as do some cosmological factors such as the rate of expansion. Laws of physical computing efficiency could also act as limiting factors on breadth, if computation is particularly relevant to the kinds of things we'd want to create in our future.

In most cases, crucial *limiting* phenomena – whether duration-limiting or breadth-limiting – don't interact *directly* with humanity's choices so much as they place boundaries on the stage on which human choices will be played out. We cannot choose in ways that change these basic limitations, but we can react to limitations to make sure the future is Large instead of Small. For example, the optimal trade-off between speed and caution of development could depend on how long we think we have and how widely we should expect to spread. If one overestimates the amount of time left, then some investments may be left to gather interest for too long, resulting in a suboptimal payout of value.

Duration-transformative phenomena: Duration-transformative phenomena are crucial because they directly determine some mapping between choices and durations in a dramatic way. For example, the phenomena harnessed by technologies that carry extinction risks are duration-tranformative; the details of physics phenomena determine whether the action of activating a particle accelerator will yield useful scientific knowledge, or alternatively create a strangelet that converts the Earth into a lifeless lump of strange matter. Preventable natural extinction risks also fall into this category.

Breadth-transformative phenomena: These are phenomena that are crucial because they directly determine some mapping between choices and breadths in a dramatic way. For example, the potential of von Neumann probes to waste large chunks of the cosmic resource pool and the effects of anti-space-colonization memes both create mappings between some of our choices and humanity's future breadth. Whatever phenomena ultimately explain Fermi's Paradox may turn out to be breadth-transformative phenomena.

4 Steering the future

I have derived the term "crucial phenomena" from Bostrom's *crucial considerations*, "idea[s] or argument[s] that might plausibly reveal the need for not just some minor course adjustment in our practical endeavours but a major change of direction or priority".[7] Since crucial phenomena are so important to our future, knowledge of the existence of a crucial phenomenon, of the laws that govern it, or of the ways that it interacts with our choices, will sometimes be crucial considerations.

Crucial phenomena relate to the aim-for-large rule in a simple way: whenever we face a choice, we ought to use whatever knowledge we have of crucial phenomena in order to choose the option that maximizes the chance of a Large future. By the time a given choice is presented to us, we should do our best to have the required knowledge of whatever crucial phenomena will be relevant to that choice well in hand.

Given that some piece of knowledge arrives in time to inform a particular choice, what properties would it be best for knowledge of crucial phenomena to have? An obvious first step would be that it should be *reliable*. Another desirable property is that the relevant knowledge should be *permissible as grounds for decisions that affect the common good*. While it may be acceptable to act on hunches or private evidence when making decisions on one's own behalf, it would be best if knowledge that guides significant decisions about humanity's shared future could be based on publicly verifiable evidence. This criterion is especially important given that governments will likely play significant roles in decisions that would benefit from knowledge of crucial phenomena; it would thus be desirable that our knowledge of crucial phenomena be available to them in a form that they can use legitimately.

Fortunately, we have societal means to secure reliable knowledge that is publicly verifiable and usable in common-good decision-making: the scientific community. The formal, professional, and social structures that make up the modern practice of science have been extremely effective in advancing our knowledge of many phenomena and in allowing us to harness and control those phenomena to improve our lives. Science can achieve the high reliability we need, and can be publicly examined and sanctioned as evidence to be used in making decisions that affect humanity's future in significant ways.

Thus, we come finally to a recommendation: as a public good, societies and their governments should support and invest in scientific research on crucial phenomena, prioritized according to the estimated size of their impact and the nearness of the relevant decisions we will need to make. I have taken such trouble to give my reasons for supporting this position because on the surface, it may sound familiar: unsurprisingly, researchers often declare that "funding for further research is needed"! To the extent that you have found my arguments convincing, however, this psychological explanation should not debunk the real need for this *particular* kind of "further research". If we are to make reliable, effective policy decisions in the future, then we must make a policy decision now to invest in our understanding of crucial phenomena.

To be as concrete as possible, I will describe two crucial phenomena that are relevant to decisions that are happening either soon or in the present time. These phenomena are (1) humanity's vulnerability or robustness to accidents arising from biological engineering,

[7] Bostrom, *Crucial considerations*.

which I will call "biological instability", and (2) the future rates of improvement of artificial intelligence and its susceptibility to misuse, which I will call "AI improvement and misuse properties".

Biological instability: Humanity is constituted of and embedded in biological systems, and biological engineering is advancing at a rapid rate. How unstable is humanity, or the ecosystem that we depend on, in the face of novel agents that could be produced by biological engineers? It seems from the historical record that it would be relatively difficult for natural mutation to stumble on an organism that could render humanity extinct[8]; is this because the space of biological organisms contains few of such threats, or is this merely a property of the part of the space that Nature can easily explore?

Factors that determine humanity's robustness against artificial biological system shocks – facts about the difference between the spaces of probable natural and artificial agents, facts about epidemiology of artificial agents, facts about the difficulty or ease of an ecosystem "takeover" by engineered organisms – are duration-transformative crucial phenomena, which could determine whether many actions we take are relatively harmless, or whether they could render humanity's future Small through extinction.

AI improvement and misuse properties: Though we cannot claim knowledge of specific future techniques in the field of artificial intelligence, there are reasons to look at the general "landscape" of artificial intelligence and conclude that large, sudden jumps in AI capabilities and rates of improvement are plausible. First, if several conjunctive factors all must reach a certain level before some capability is achieved, then some factors may reach many times the required level before the final factor reaches the critical level; at that point, the system will fall up an "overhang", suddenly achieving a much more effective, efficient, or economical version of the desired ability.[9] Second, there is a set of cognitive skills, exemplified in human children, that can be used to learn many other cognitive skills directly from human cultural artifacts such as books and websites; we should expect a large jump in capability as AI systems quickly acquire any cultural skills that have not yet been automated. Finally, and perhaps most significantly, it is clear that AI research and development is a cognitive skill like any other, and should be subject to automation; when it is, there are reasons to think that the rate of improvement of AI capabilities would accelerate sharply upward in an "intelligence explosion".[10] It is plausible that AI could reach levels of capability far beyond any human or group of humans at any number of tasks.

Furthermore, it has been argued that if we attempt to use such "superintelligent" AI, or if an AI system were to improve to a high level while in use, it would be easy to accidentally *misuse* such a system, and that such accidental misuse could lead to results as extreme as the extinction of humanity. This idea has been given a few

[8] Though see Ćirković, Sandberg, and Bostrom, "Anthropic shadow: Observation selection effects and human extinction risks" for commentary on the relevant concept of "anthropic shadow".

[9] Yudkowsky, "Artificial intelligence as a positive and negative factor in global risk"; Shulman and Sandberg, "Implications of a software-limited singularity".

[10] Good, "Speculations concerning the first ultraintelligent machine"; Yudkowsky, *Intelligence explosion microeconomics*.

supporting arguments, including the existence of "convergent instrumental goals" that would cause AIs with many different tasks to take actions that could harm humanity[11] and the difficulty of designing tasks for superintelligent AIs that would result in non-Disastrous outcomes.[12]

The phenomena that govern AI improvement rates and its susceptibility to misuse, which I have outlined above, could lead to human extinction, and so they are duration-transformative crucial phenomena. Depending on the true nature of these phenomena, certain kinds of AI research and development in the medium-term future could threaten human extinction, or could be purely beneficial ways of creating helpful new technology.

Of these two, it is plausible that biological instability is the more urgent crucial phenomenon. While AI still appears to be far from the two thresholds that I mention, biological engineering is creating novel, harmful agents today, and escapes from BSL-4 (highest security) labs are shockingly common.[13] Additionally, opportunities to monitor or regulate new biological technologies before they become too widespread for effective control may soon slip through our grasp. On the other hand, threats from AI improvement and misuse have a more deeply puzzling character; even if we understood them well, it is not clear what appealing policy courses we could take to mitigate them. Since AI improvement and misuse may require significantly more work to solve (should it prove to be a true problem), it should also be treated with some urgency.

❧

In this essay, I have sought to explain why societies and their governments should support and invest in scientific research on crucial phenomena. In particular, there are common-good issues where we lack sufficient understanding to take more proactive policy action; in these cases, such as the case of biological instability, engaging in scientific research may be the best policy choice available. Hamming, once again:

> Our society frowns on people who set out to do really good work. You're not supposed to; luck is supposed to descend on you and you do great things by chance. Well, that's a kind of dumb thing to say. I say, why shouldn't you set out to do something significant. You don't have to tell other people, but shouldn't you say to yourself, "Yes, I would like to do something significant."

This quote could be applied not just to individuals, but to generations and societies. In the case of humanity's long-term prospects, our collective humility and our duty to the future are in conflict with with one another. "So much the worse for our collective humility" seems, to me, the only acceptable response.

[11] Omohundro, "The basic AI drives"; Bostrom, "The superintelligent will: Motivation and instrumental rationality in advanced artificial agents".

[12] Yudkowsky, "Complex value systems in friendly AI".

[13] Klotz and Sylvester, "The unacceptable risks of a man-made pandemic".

References

Beckstead, Nick. *A proposed adjustment to the astronomical waste argument*. LessWrong post. 2013. URL: `http://lesswrong.com/lw/hjb/a_proposed_adjustment_to_the_astronomical_waste/`.

— "On the overwhelming importance of shaping the far future". PhD thesis. Rutgers University, 2013.

Bostrom, Nick. *Crucial considerations*. Website. URL: `http://nickbostrom.com/`.

— "Existential risk prevention as global priority". In: *Global Policy* 4.1 (2013), pp. 15–31.

— "The superintelligent will: Motivation and instrumental rationality in advanced artificial agents". In: *Minds and Machines* 22.2 (2012), pp. 71–85.

Ćirković, Milan M. "Cosmological forecast and its practical significance". In: *Journal of Evolution and Technology* 12 (2002).

Ćirković, Milan M, Anders Sandberg, and Nick Bostrom. "Anthropic shadow: Observation selection effects and human extinction risks". In: *Risk analysis* 30.10 (2010), pp. 1495–1506.

Good, Irving John. "Speculations concerning the first ultraintelligent machine". In: *Advances in computers* 6.31 (1965), p. 88.

Hamming, Richard W and JF Kaiser. "You and your research". In: *Transcription of the Bell Communications Research Colloquium Seminar* (1986).

Klotz, LC and EJ Sylvester. "The unacceptable risks of a man-made pandemic". In: *Bulletin of the Atomic Scientists* 7 (2012).

Omohundro, Stephen M. "The basic AI drives". In: *Frontiers in Artificial Intelligence and applications* 171 (2008), p. 483.

Shulman, Carl and Anders Sandberg. "Implications of a software-limited singularity". In: *Proceedings of the European Conference of Computing and Philosophy*. 2010.

Yudkowsky, Eliezer. "Artificial intelligence as a positive and negative factor in global risk". In: *Global catastrophic risks* 1 (2008), p. 303.

— "Complex value systems in friendly AI". In: *Artificial General Intelligence*. Springer, 2011, pp. 388–393.

— *Intelligence explosion microeconomics*. Tech. rep. Technical Report, 2013–1. Machine Intelligence Research Institute, Berkeley, CA. http://intelligence. org/files/IEM. pdf, 2013.