Version: 1.0.0
Last updated: 2021-07-07

# Nick Bostrom: The Future of Humanity (extracts)

https://www.nickbostrom.com/papers/future.html

**Abstract**

The future of humanity is often viewed as a topic for idle speculation. Yet our beliefs and assumptions on this subject matter shape decisions in both our personal lives and public policy – decisions that have very real and sometimes unfortunate consequences. It is therefore practically important to try to develop a realistic mode of futuristic thought about big picture questions for humanity. This paper sketches an overview of some recent attempts in this direction, and it offers a brief discussion of four families of scenarios for humanity's future: extinction, recurrent collapse, plateau, and posthumanity.

**The future of humanity as an inescapable topic**

In one sense, the future of humanity comprises everything that will ever happen to any human being, including what you will have for breakfast next Thursday and all the scientific discoveries that will be made next year. In that sense, it is hardly reasonable to think of the future of humanity as a topic: it is too big and too diverse to be addressed as a whole in a single essay, monograph, or even 100-volume book series. It is made into a topic by way of abstraction. We abstract from details and short-term fluctuations and developments that affect only some limited aspect of our lives. A discussion about the future of humanity is about how the important fundamental features of the human condition may change or remain constant in the long run.

What features of the human condition are fundamental and important? On this there can be reasonable disagreement. Nonetheless, some features qualify by almost any standard. For example, whether and when Earth-originating life will go extinct, whether it will colonize the galaxy, whether human biology will be fundamentally transformed to make us posthuman, whether machine intelligence will surpass biological intelligence, whether population size will explode, and whether quality of life will radically improve or deteriorate: these are all important fundamental questions about the future of humanity. Less fundamental questions – for instance, about methodologies or specific technology projections – are also relevant insofar as they inform our views about more fundamental parameters.

Traditionally, the future of humanity has been a topic for theology. All the major religions have teachings about the ultimate destiny of humanity or the end of the world. 1 Eschatological themes have also been explored by big-name philosophers such as Hegel, Kant, and Marx. In more recent times the literary genre of science fiction has continued the tradition. Very often, the future has served as a projection screen for our hopes and fears; or as a stage setting for dramatic entertainment, morality tales, or satire of tendencies in contemporary society; or as a banner for ideological mobilization. It is relatively rare for humanity's future to be taken seriously as a subject matter on which it is important to try to have factually correct beliefs. There is nothing wrong with exploiting the symbolic and literary affordances of an unknown future, just as there is nothing wrong with fantasizing

about imaginary countries populated by dragons and wizards. Yet it is important to attempt (as best we can) to distinguish futuristic scenarios put forward for their symbolic significance or entertainment value from speculations that are meant to be evaluated on the basis of literal plausibility. Only the latter form of "realistic" futuristic thought will be considered in this paper.

We need realistic pictures of what the future might bring in order to make sound decisions. Increasingly, we need realistic pictures not only of our personal or local near-term futures, but also of remoter global futures. Because of our expanded technological powers, some human activities now have significant global impacts. The scale of human social organization has also grown, creating new opportunities for coordination and action, and there are many institutions and individuals who either do consider, or claim to consider, or ought to consider, possible long-term global impacts of their actions. Climate change, national and international security, economic development, nuclear waste disposal, biodiversity, natural resource conservation, population policy, and scientific and technological research funding are examples of policy areas that involve long time-horizons. Arguments in these areas often rely on implicit assumptions about the future of humanity. By making these assumptions explicit, and subjecting them to critical analysis, it might be possible to address some of the big challenges for humanity in a more well-considered and thoughtful manner.

The fact that we "need" realistic pictures of the future does not entail that we can have them. Predictions about future technical and social developments are notoriously unreliable – to an extent that have lead some to propose that we do away with prediction altogether in our planning and preparation for the future. Yet while the methodological problems of such forecasting are certainly very significant, the extreme view that we can or should do away with prediction altogether is misguided. That view is expressed, to take one example, in a recent paper on the societal implications of nanotechnology by Michael Crow and Daniel Sarewitz, in which they argue that the issue of predictability is "irrelevant":

> preparation for the future obviously does not require accurate prediction; rather, it requires a foundation of knowledge upon which to base action, a capacity to learn from experience, close attention to what is going on in the present, and healthy and resilient institutions that can effectively respond or adapt to change in a timely manner. 2

Note that each of the elements Crow and Sarewitz mention as required for the preparation for the future relies in some way on accurate prediction. A capacity to learn from experience is not useful for preparing for the future unless we can correctly assume (predict) that the lessons we derive from the past will be applicable to future situations. Close attention to what is going on in the present is likewise futile unless we can assume that what is going on in the present will reveal stable trends or otherwise shed light on what is likely to happen next. It also requires non-trivial prediction to figure out what kind of institution will prove healthy, resilient, and effective in responding or adapting to future changes.

The reality is that predictability is a matter of degree, and different aspects of the future are predictable with varying degrees of reliability and precision. 3 It may often be a good idea to develop plans that are flexible and to pursue policies that are robust under a wide range of

contingencies. In some cases, it also makes sense to adopt a reactive approach that relies on adapting quickly to changing circumstances rather than pursuing any detailed long-term plan or explicit agenda. Yet these coping strategies are only one part of the solution. Another part is to work to improve the accuracy of our beliefs about the future (including the accuracy of conditional predictions of the form "if x is done, y will result"). There might be traps that we are walking towards that we could only avoid falling into by means of foresight. There are also opportunities that we could reach much sooner if we could see them farther in advance. And in a strict sense, prediction is always necessary for meaningful decision-making. 4

Predictability does not necessarily fall off with temporal distance. It may be highly unpredictable where a traveler will be one hour after the start of her journey, yet predictable that after five hours she will be at her destination. The very long-term future of humanity may be relatively easy to predict, being a matter amenable to study by the natural sciences, particularly cosmology (physical eschatology). And for there to be a degree of predictability, it is not necessary that it be possible to identify one specific scenario as what will definitely happen. If there is at least some scenario that can be ruled out, that is also a degree of predictability. Even short of this, if there is some basis for assigning different probabilities (in the sense of credences, degrees of belief) to different propositions about logically possible future events, or some basis for criticizing some such probability distributions as less rationally defensible or reasonable than others, then again there is a degree of predictability. And this is surely the case with regard to many aspects of the future of humanity. While our knowledge is insufficient to narrow down the space of possibilities to one broadly outlined future for humanity, we do know of many relevant arguments and considerations which in combination impose significant constraints on what a plausible view of the future could look like. The future of humanity need not be a topic on which all assumptions are entirely arbitrary and anything goes. There is a vast gulf between knowing exactly what will happen and having absolutely no clue about what will happen. Our actual epistemic location is some offshore place in that gulf. 5

**Technology, growth, and directionality**

Most differences between our lives and the lives of our hunter-gatherer forebears are ultimately tied to technology, especially if we understand "technology" in its broadest sense, to include not only gadgets and machines but also techniques, processes, and institutions. In this wide sense we could say that technology is the sum total of instrumentally useful culturally-transmissible information. Language is a technology in this sense, along with tractors, machine guns, sorting algorithms, double-entry bookkeeping, and Robert's Rules of Order. 6

Technological innovation is the main driver of long-term economic growth. Over long time scales, the compound effects of even modest average annual growth are profound. Technological change is in large part responsible for many of the secular trends in such basic parameters of the human condition as the size of the world population, life expectancy, education levels, material standards of living, and the nature of work, communication, health care, war, and the effects of human activities on the natural environment. Other aspects of society and our individual lives are also influenced by technology in many direct and indirect ways, including governance, entertainment, human relationships, and our views on morality,

mind, matter, and our own human nature. One does not have to embrace any strong form of technological determinism to recognize that technological capability – through its complex interactions with individuals, institutions, cultures, and environment – is a key determinant of the ground rules within which the games of human civilization get played out. 7

This view of the important role of technology is consistent with large variations and fluctuations in deployment of technology in different times and parts of the world. The view is also consistent with technological development itself being dependent on socio-cultural, economic, or personalistic enabling factors. The view is also consistent with denying any strong version of inevitability of the particular growth pattern observed in human history. One might hold, for example, that in a "re-run" of human history, the timing and location of the Industrial Revolution might have been very different, or that there might not have been any such revolution at all but rather, say, a slow and steady trickle of invention. One might even hold that there are important bifurcation points in technological development at which history could take either path with quite different results in what kinds of technological systems developed. Nevertheless, under the assumption that technological development continues on a broad front, one might expect that in the long run, most of the important basic capabilities that could be obtained through some possible technology, will in fact be obtained through technology. A bolder version of this idea could be formulated as follows:

> **Technological Completion Conjecture.** If scientific and technological development efforts do not effectively cease, then all important basic capabilities that could be obtained through some possible technology will be obtained.

The conjecture is not tautological. It would be false if there is some possible basic capability that could be obtained through some technology which, while possible in the sense of being consistent with physical laws and material constraints, is so difficult to develop that it would remain beyond reach even after an indefinitely prolonged development effort. Another way in which the conjecture could be false is if some important capability can only be achieved through some possible technology which, while it could have been developed, will not in fact ever be developed even though scientific and technological development efforts continue.

The conjecture expresses the idea that which important basic capabilities are eventually attained does not depend on the paths taken by scientific and technological research in the short term. The principle allows that we might attain some capabilities sooner if, for example, we direct research funding one way rather than another; but it maintains that provided our general techno-scientific enterprise continues, even the non-prioritized capabilities will eventually be obtained, either through some indirect technological route, or when general advancements in instrumentation and understanding have made the originally neglected direct technological route so easy that even a tiny effort will succeed in developing the technology in question. 8

One might find the thrust of this underlying idea plausible without being persuaded that the Technological Completion Conjecture is strictly true, and in that case, one may explore what exceptions there might be. Alternatively, one might accept the conjecture but believe that its antecedent is false, i.e. that scientific and technological development efforts will at some point effectively cease (before the enterprise is complete). But if one accepts both the conjecture and its antecedent, what are the implications? What will be the results if, in the

long run, all of the important basic capabilities that could be obtained through some possible technology are in fact obtained? The answer may depend on the order in which technologies are developed, the social, legal, and cultural frameworks within which they are deployed, the choices of individuals and institutions, and other factors, including chance events. The obtainment of a basic capability does not imply that the capability will be used in a particular way or even that it will be used at all.

These factors determining the uses and impacts of potential basic capabilities are often hard to predict. What might be somewhat more foreseeable is which important basic capabilities will eventually be attained. For under the assumption that the Technological Completion Conjecture and its antecedent are true, the capabilities that will eventually be include all the ones that could be obtained through some possible technology. While we may not be able to foresee all possible technologies, we can foresee many possible technologies, including some that that are currently infeasible; and we can show that these anticipated possible technologies would provide a large range of new important basic capabilities.

One way to foresee possible future technologies is through what Eric Drexler has termed "theoretical applied science". 9 Theoretical applied science studies the properties of possible physical systems, including ones that cannot yet be built, using methods such as computer simulation and derivation from established physical laws.10 Theoretical applied science will not in every instance deliver a definitive and uncontroversial yes-or-no answer to questions about the feasibility of some imaginable technology, but it is arguably the best method we have for answering such questions. Theoretical applied science – both in its more rigorous and its more speculative applications – is therefore an important methodological tool for thinking about the future of technology and, a fortiori, one key determinant of the future of humanity.

It may be tempting to refer to the expansion of technological capacities as "progress". But this term has evaluative connotations – of things getting better – and it is far from a conceptual truth that expansion of technological capabilities makes things go better. Even if empirically we find that such an association has held in the past (no doubt with many big exceptions), we should not uncritically assume that the association will always continue to hold. It is preferable, therefore, to use a more neutral term, such as "technological development", to denote the historical trend of accumulating technological capability.

Technological development has provided human history with a kind of directionality. Instrumentally useful information has tended to accumulate from generation to generation, so that each new generation has begun from a different and technologically more advanced starting point than its predecessor. One can point to exceptions to this trend, regions that have stagnated or even regressed for extended periods of time. Yet looking at human history from our contemporary vantage point, the macro-pattern is unmistakable.

It was not always so. Technological development for most of human history was so slow as to be indiscernible. When technological development was that slow, it could only have been detected by comparing how levels of technological capability differed over large spans of time. Yet the data needed for such comparisons – detailed historical accounts, archeological excavations with carbon dating, and so forth – were unavailable until fairly recently, as Robert Heilbroner explains:

> At the very apex of the first stratified societies, dynastic dreams were dreamt and visions of triumph or ruin entertained; but there is no mention in the papyri and cuniform tablets on which these hopes and fears were recorded that they envisaged, in the slightest degree, changes in the material conditions of the great masses, or for that matter, of the ruling class itself. 11

Heilbroner argued in *Visions of the Future* for the bold thesis that humanity's perceptions of the shape of things to come has gone through exactly three phases since the first appearance of Homo sapiens. In the first phase, which comprises all of human prehistory and most of history, the worldly future was envisaged – with very few exceptions – as changeless in its material, technological, and economic conditions. In the second phase, lasting roughly from the beginning of the eighteenth century until the second half of the twentieth, worldly expectations in the industrialized world changed to incorporate the belief that the hitherto untamable forces of nature could be controlled through the appliance of science and rationality, and the future became a great beckoning prospect. The third phase – mostly post-war but overlapping with the second phase – sees the future in a more ambivalent light: as dominated by impersonal forces, as disruptive, hazardous, and foreboding as well as promising.

Supposing that some perceptive observer in the past had noticed some instance of directionality – be it a technological, cultural, or social trend – the question would have remained whether the detected directionality was a global feature or a mere local pattern. In a cyclical view of history, for example, there can be long stretches of steady cumulative development of technology or other factors. Within a period, there is clear directionality; yet each flood of growth is followed by an ebb of decay, returning things to where they stood at the beginning of the cycle. Strong local directionality is thus compatible with the view that, globally, history moves in circles and never really gets anywhere. If the periodicity is assumed to go on forever, a form of eternal recurrence would follow.

Modern Westerners who are accustomed to viewing history as directional pattern of development may not appreciate how natural the cyclical view of history once seemed. 12 Any closed system with only a finite number of possible states must either settle down into one state and remain in that one state forever, or else cycle back through states in which it has already been. In other words, a closed finite state system must either become static or else start repeating itself. If we assume that the system has already been around for an eternity, then this eventual outcome must already have come about; i.e., the system is already either stuck or is cycling through states in which it has been before. The proviso that the system has only a finite number of states may not be as significant as it seems, for even a system that has an infinite number of possible states may only have finitely many perceptibly different possible states. 13 For many practical purposes, it may not matter much whether the current state of the world has already occurred an infinite number of times, or whether an infinite number of states have previously occurred each of which is merely imperceptibly different from the present state. 14 Either way, we could characterize the situation as one of eternal recurrence – the extreme case of a cyclical history.

In the actual world, the cyclical view is false because the world had a beginning a finite time ago. The human species has existed for a mere two hundred thousand years or so, and this

is far from enough time for it to have experienced all possible conditions and permutations of which the system of humans and their environment is capable.

More fundamentally, the reason why the cyclical view is false is that the universe itself has existed for only a finite amount of time. 15 The universe started with the Big Bang an estimated 13.7 billion years ago, in a low-entropy state. The history of the universe has its own directionality: an ineluctable increase in entropy. During its process of entropy increase, the universe has progressed through a sequence of distinct stages. In the eventful first three seconds, a number of transitions occurred, including probably a period of inflation, reheating, and symmetry breaking. These were followed, later, by nucleosynthesis, expansion, cooling, and formation of galaxies, stars, and planets, including Earth (circa 4.5 billion years ago). The oldest undisputed fossils are about 3.5 billion years old, but there is some evidence that life already existed 3.7 billion years ago and possibly earlier. Evolution of more complex organisms was a slow process. It took some 1.8 billion years for eukaryotic life to evolve from prokaryotes, and another 1.4 billion years before the first multicellular organisms arose. From the beginning of the Cambrian period (some 542 million years ago), “important developments” began happening at a faster pace, but still enormously slowly by human standards. Homo habilis – our first “human-like ancestors” – evolved some 2 million years ago; Homo sapiens 100,000 years ago. The agricultural revolution began in the Fertile Crescent of the Middle East 10,000 years ago, and the rest is history. The size of the human population, which was about 5 million when we were living as hunter-gatherers 10,000 years ago, had grown to about 200 million by the year 1; it reached one billion in 1835 AD; and today over 6.6 billion human beings are breathing on this planet. 16 From the time of the industrial revolution, perceptive individuals living in developed countries have noticed significant technological change within their lifetimes.

All techno-hype aside, it is striking how recent many of the events are that define what we take to be the modern human condition. If compress the time scale such that the Earth formed one year ago, then Homo sapiens evolved less than 12 minutes ago, agriculture began a little over one minute ago, the Industrial Revolution took place less than 2 seconds ago, the electronic computer was invented 0.4 seconds ago, and the Internet less than 0.1 seconds ago – in the blink of an eye.

Almost all the volume of the universe is ultra-high vacuum, and almost all of the tiny material specks in this vacuum are so hot or so cold, so dense or so dilute, as to be utterly inhospitable to organic life. Spatially as well as temporally, our situation is an anomaly. 17

Given the technocentric perspective adopted here, and in light of our incomplete but substantial knowledge of human history and its place in the universe, how might we structure our expectations of things to come? The remainder of this paper will outline four families of scenarios for humanity’s future:

- Extinction
- Recurrent collapse
- Plateau
- Posthumanity

# Nick Bostrom: Humanity's Biggest Problems Aren't What You Think They Are (16m video)

https://youtu.be/Yd9cf_vLvil

*This Chrome extension makes it easy to watch videos at the speed you prefer.*

---

# Katarzyna de Lazari-Radek: On Sidgwick & The Point of View of The Universe

https://www.3-16am.co.uk/articles/from-the-point-of-view-of-the-universe

3:AM: So for Sidgwick, what is ethics, its methods and what is a philosopher up to when she's investigating them? And what does it mean to say there is a 'point of view of the universe'?

KLR: Sidgwick defines ethics as a study of what we ought to do as opposed to other studies such as psychology or biology that tell you what is the case. The methods of ethics are rational procedures which we, individual beings, use to determine what we ought to do. In everyday life we are often not very consistent: we use many different methods, and we mix them as well. But Sidgwick is a scholar and he wants to make them scientific. Therefore he will separate them carefully and underline differences between them. He will talk of egoism, intuitionism and utilitarianism.

As for the most important expression: "the point of view of the universe" that is to symbolize an impartial concern for everyone. Sidgwick calls for impartiality in ethics and thinks that when deciding what we ought to do, we should try to take an impartial perspective – not mine, not yours, not my children's but "the point of view of the universe". Rawls, Nagel and Parfit will all refer to that perspective in their works later on.

3:AM: What does rationality add up to in Sidgwick? Is this a return to Kantianism and a kick back against Hume and Hare and is it part of the reason why Sidgwick rejects common sense ethics?

KLR: I do think Sidgwick was influenced by Kant in this respect, but also by such English intuitionists as Thomas Reid or William Whewell as well as Coleridge. He did believe in reason and rationality. But he also saw a great crack in it. Claiming that both maximizing my own good and maximizing impartial good is rational, he could not reach a final answer to the most important question of his inquiry: what ought I to do? When in a tragic situation, should I save my own child or rather a few children of complete strangers? Sidgwick regretted to say that but he confesses at the end of The Methods that reason may not give us a final answer. That would be tragic indeed as it would open the door to subjectivism again. Peter

Singer and I tried to help Sidgwick to overcome that chaos. We claim that only impartial action is fully rational.

As for his rejection of common sense. He is not satisfied with rules given by common sense as he finds them unclear, vague, not self-evident.

3:AM: And does this mean that he is out of line with someone like Rawls who'd argue that we need to find a 'reflective equilibrium between theory and considered moral judgments? Where do you stand on this?

KLR: This is an interesting question but I treat it more as a problem of justification. First, unlike Rawls, both Sidgwick and we are interested in truth and finding true moral principles. Now the question is do we use coherentism or foundationalism to find out the truth. Reflective equilibrium seems a useful tool but as Hare, in his review of The Theory of Justice, recalled Plato saying: "If a man starts from something he knows not, and the end and middle of his argument are tangled together out of what he knows not, how can such a mere consensus ever turn into knowledge?" (Rep. 533 c). On the other hand, foundationalism can lead easily to dogmatism. We tried our best to stand somewhere in between those two.

3:AM: Are you sympathetic to the foundational self-evident axioms Sidgwick uses? Is this where the idea of 'rational intuition' comes in – and your use of Parfit's 'Future Tuesday Indifference'? Can you explain the argument here? And why wouldn't this be congenial to contemporary economists who might have expected to find a defence of their models of rationality in this approach?

KLR: Yes, we are sympathetic to Sidgwick's appeal to self-evident axioms, especially his axiom of rational benevolence, which is linked to taking "the point of view of the universe." We argue that this is a rational axiom, because, in contrast to many other moral intuitions, our acceptance of it cannot be debunked by an evolutionary explanation.

Parfit's uses the idea of "Future Tuesday Indifference" in a slightly different context, to argue against the subjectivist view that what is rational is always dependent on a person's ultimate desires, or ends. A person who is indifferent to what happens to him on any future Tuesday (and therefore, when offered a choice between being pinched today and hours of torture next Tuesday, chooses the torture) may be acting in accordance with his bizarre set of desires, but he is still irrational. Contemporary economists assume that a view of rationality that is subjectivist, or as they would call it, instrumentalist, so they won't find this argument congenial. It will force them to reexamine their fundamental assumptions about rationality.

...

3:AM: This is a form of hedonism isn't it? How does Sidgwick understand hedonism – and are you sympathetic?

KLR: Well, you can be a hedonist no matter whether you are a rule or an act utilitarian. A hedonist defines the good which you should maximize in terms of happiness or pleasure. For

Sidgwick the two were the same thing and he defines pleasure as desirable consciousness, that is a state of mind which you desire at the time of feeling it.

...

3:AM: The 'repugnant conclusion' argument of Parfit regarding optimal population growth seems on the face of it a pretty decisive one for rejecting utilitarianism doesn't it? How do you handle this issue so we can remain utilitarians?

KLR: I don't see this as a ground for rejecting utilitarianism at all. Parfit's "repugnant conclusion" is an objection to one way of answering the simple question that Sidgwick was the first to raise: if by increasing the population, the average level of welfare decreases, but because everyone still has lives that are, on balance, happy, the total amount of happiness in the world increases, is that a good thing? What Parfit has shown is that all of the answers that seem plausible – not just those offered by Sidgwick or other utilitarians – lead to either inconsistency or counter-intuitive judgments. Therefore it isn't as if non-utilitarians do any better in answering the question than utilitarians.

http://www.stafforini.com/blog/summary-of-the-point-of-view-of-the-universe-by-katarzyna-de-lazari-radek-and-peter-singer/

Sidgwick distinguished three different stages of intuitionism: perceptional intuitionism, common sense morality, and philosophical intuitionism. His examination of the morality of common sense is especially noteworthy and is here discussed using the examples of benevolence and truth-telling. Sidgwick concluded that only philosophical intuitionism constitutes a sufficiently precise method of ethics. This chapter considers all three forms of intuitionism and their contemporary or recent exponents. Particularism, as espoused by Dancy, is today the leading form of perceptional intuitionism, while Ross, Gert, and Bok are taken as defenders of the morality of common sense. The chapter defends Sidgwick's view that neither perceptional intuitionism nor the morality of common sense is philosophically adequate.

# Samuel Scheffler: Conservatism, Temporal Bias, and Future Generations (extracts)

http://podcasts.ox.ac.uk/2015-uehiro-lectures-conservatism-temporal-bias-and-future-generations

Over the past couple of days, I've maintained that we have reasons of at least four different kinds for caring about the fate of future generations. Reasons of love, reasons of interest, reasons of value and reasons of reciprocity. All of these reasons depend in one way or another on our existing values and attachments and on our associated disposition to preserve and sustain the things that we value.

---

The concern for the future of humanity flows naturally from a conservative concern for the things that we value. Now, it's our very attachment to the status quo that propels our concerns into the future. Without such attachments, it's not clear how much reason we would have to care about humanities survival.

---

Things look very different from the perspective of the beneficence based literature on future generations. The primary focus of that literature is on questions of population ethics and the standard method that's used to investigate those questions is to describe alternative worlds or alternative states of a particular world whose populations differ from one another in their size, composition and or levels of wellbeing.

The hope is that by collecting our judgments about such cases, we can arrive at a satisfactory population axiology, a principle or standard that would allow us to determine the relative value of total states of the world, even when their populations differ in one or more of the respects I mentioned. Such an axiology would in turn supply the basis for a principle of beneficence, which would spell out either by itself or in conjunction with some other principles, our responsibilities for promoting the best population outcome.

There's no consensus among those who hoped to find a satisfactory population axiology about which one is the best candidate. But even if there were such a consensus, what claim is the preferred axiology supposed to have on our motives? Why is it thought either that we do or that we should care, which population outcomes are judged superior by the lights of this or that axial logical principle?

---

My suspicion is not that the proposed rationale for population axiology overestimates the extent of our concern for our successors, but rather that it underestimates and misrepresents our concern.

It underestimates it because it neglects the variety of reasons we have for concerning ourselves with the fate of future generations.

And it misrepresents it because what those reasons support is not a generic concern for the welfare interests of future people, but rather a more specific desire rooted in the values we affirm in our daily lives, that the chain of generations should be extended into the indefinite future, and that our successes should be able to live under conditions conducive to their flourishing.

---

According to the alternative perspective that I've been defending in these lectures, by contrast, we have a variety of reasons, all rooted in our existing attachments to humanity, and to valued forms of human activity and endeavor to care about the capacity of future generations to survive and to flourish.

From this perspective, it's a mistake to think that our reasons for caring about the fate of future generations are hostage to our ability to construct a satisfactory population axiology, a complete theory of the relative goodness of total states of the world. The contrast between these two views is both normative and motivational.

Normatively there is on the one side, a moral imperative to implement a general principle of beneficence. On the other side, there was a set of compelling reasons whether moral or non-moral to secure the ability of our successors to survive under conditions conducive to their flourishing.

Motivationally there is on the one side a generalized concern for the welfare interests of all people, including all future people. On the other side, there's a conservative disposition to sustain the humanity we love and the existing values we now cherish.

It will already be clear that I find the second perspective, more persuasive, both normatively and motivationally. As a normative matter I find the reasons that it highlights for concerning ourselves with the fate of future generations, more compelling than those suggested by the beneficence-based approach.

And at the motivational level, the fact that it grounds our concern for future generations in a conservative disposition to sustain our existing attachments, puts that concern on a more secure footing and integrates it into a unified stance we may take toward the diachronic dimension of our values.

---

The conservative disposition I've been discussing is not a form of political conservatism. It's a disposition to preserve or sustain the things that we value and both the things that we value in the steps necessary to preserve them.

One way to illuminate this kind of conservatism is to consider how it relates to the very similar form of conservatism defended by the late Jerry Cohen. In his wonderful essay, defending what he calls small C conservatism, Cohen advocates a bias in favor of existing value by which he means that we should regret the destruction of particular valuable things as such, even when it would lead to their replacement, by things of greater value.

He thinks that "everyone who is sane" has this bias to some degree and that it is quote "rational and right, that they should". For Cohen, the crucial distinction is between value in the abstract and the particular things that have value or alternatively between the value that things bear and the bearers of such value.

The conservatism that he defends holds that particular things that have value take priority over value itself in at least two related senses. First particular valuable things do not matter or count simply because of the amount of value that they bear or that resides in them. Second, we have at least some defeasible reason to preserve particular, valuable things as such, even if by sacrificing them, we could produce more value overall.

...

"to seek to maximize value is to see nothing wrong in the destruction of valuable things as long as there's no reduction in the total amount of value as a result. Unlike the conservative, the utilitarian is indifferent between adding to what we have now got at no cost, something that has 5 units of value and adding something worth 10 units of value at the expense of destroying something worth 5."

If the utilitarian is willing to sacrifice a particular valuable thing, whenever it can be replaced by another particular valuable thing with even slightly more value than the original item is being valued solely in proportion to the value that it bears. And to say that is just to say that the utilitarian, unlike the conservative does not value the bearers of value independently of the value that they bring.

…

Cohen's insight is best appreciated if we focus not on Cohen's category of particular value as a type of value or indeed on any other category of value, but rather on what it is for a person to value a given thing.

...

I regard the distinction between something's having value and one's valuing it as significant. Valuing something in my view involves a complex syndrome of attitudes and dispositions, including a belief that the thing is valuable, a susceptibility to experience a variety of context dependent emotions concerning the thing, and a disposition to treat considerations pertaining to the thing as providing one with reasons for action in relevant contexts. Here, I'm using thing in a broad sense that encompasses any object of our valuing attitudes.

It's possible to regard something as valuable (or in Cohen's terms as possessing particular value) without actually valuing it oneself in this sense. Indeed, most of us regard many things as valuable that we ourselves do not value.

Valuing something involves more than just believing that it's valuable. It involves a kind of attachment to or investment in or engagement with that thing. This sort of attachment or investment or engagement is constituted both by emotional vulnerability and by a disposition to see oneself as having reasons for action, with respect to the valued item that one does not have with respect to other comparably valuable items of the same kind.

If I value my relationship with you, for example, then I will typically be vulnerable to feelings of distress. If you are harmed. And I will see myself as having reasons for acting in your behalf in relevant deliberative contexts that I do not have for acting in behalf of other equally valuable people.

So, for example, if my, if I value my friendship with you, then I'm justified in thinking: I have reasons to act in your behalf that other people do not have. And that I do not have with regard to people who are not my friends. And if I value an antique rug that has been in my family for generations, then I'm justified in thinking that I have reasons to care for it or

preserve it that other people do not have. And that I do not have with regard to other antique rugs. This does not mean that I have no reason to do anything at all on behalf of people who are not my friends or indeed that I never have reasons to help preserve other antique rugs or other people's family heirlooms. It means only that by virtue of valuing particular valuable things, we have reasons for action that go beyond the reasons that we and others may have solely in virtue of the intrinsic value of those.

These points about the relation between valuing and reasons for action are relevant to Cohen's defensive of conservatism, because in general, we cannot value things that do not exist and have never existed in the way we value existing things. Valuing involves attachment, attachment requires acquaintance, and non-existence makes the relevant form of acquaintance impossible.

So for example, one cannot value the friendships one has not yet formed in the way that one values one's existing friendships. One cannot value the projects one will someday develop in the way one values the projects one already has. One can not value the children one has not yet conceived in the way that one values ones existing children, nor can one now value the great works of art that artists will produce in the future in the way one values those great works that already exist. One can of course attach value to one prospects and plans before they have borne fruit and to one's hopes and dreams before they have been fulfilled. But in these cases, the prospects and plans and hopes and dreams already exist, one cannot in the same way, attach value to the plans one has not yet made or the dreams one does not yet have.

If this point is correct, then it's possible to identify a conservative attitude more or less along the lines suggested by Cohen's discussion, that goes beyond a temporarily neutral assignment of priority to the bearers of value over the value that they bear.

This form of conservatism includes, in addition, a bias in favor of certain particular bearers of value that already exist. The bias derives from the fact that we can form value based attachments to existing things in a way that we cannot form such attachments to things that do not yet exist. And things that we value are sources of distinctive reasons for action and distinctive patterns of emotional engagement.

The content of these reasons and the contours of these patterns of engagement will vary depending on the type of thing that's in question. But in most cases, the reasons will include reasons to care for and preserve the things that we value. And the emotions will include vulnerability to feelings of distress if those things are harmed or damaged or destroyed. In so far as this sort of bias is built into our valuing attitudes, every valuer must, to that extent possess the conservative disposition, this vindicates Cohen's assertion that everyone who is sane has something of this disposition.

At the same time, it's important not to exaggerate or misinterpret the normative significance of this disposition. Although we have special reasons for action pertaining to items that we already value, these reasons will not always be the strongest reasons we have in any given case. They may be outweighed by sufficiently strong reasons of other kinds. Furthermore, there will be many cases in which we can create new items of value without neglecting the reasons we have to care for the items we already value.

So conservatism is not incompatible with creativity. This is important because even if all sane people have something of the conservative disposition, all sane people also have something of the creative disposition.

This disposition is not limited to artists or to others who are colloquially described as creative people. It reveals itself in the impulse to make, to build, to invent, to change, to improve, to reform, to renew, to innovate and of course, to procreate. It reveals itself even in the impulse to act because each act is a novel intervention in the world, each act contributes something new to the course of human history.

In that sense, the conservative disposition to sustain and preserve the things that we value is itself a creative disposition. To be sure there are times when all it requires of us is that we refrain from performing actions that would harm or destroy those things. But there are also times when we must take affirmative steps, often requiring great imagination and tenacity, if we're to succeed in sustaining and preserving the things that we value—after all conservators are not people whose job is simply to do nothing. The conservative disposition and the creative disposition are not incompatible then not only because there are cases in which one of them applies and the other doesn't, but also because there's a sense in which the conservative disposition properly understood is itself a creative disposition.

This brings us back to the role of the conservative disposition in supporting our concern for the survival and flourishing of future generations. I've emphasized the extent to which our reasons to care about the fate of our successors are rooted in our value-based attachments to humanity and to the many different forms of human activity and endeavor that we cherish. Far from being a backward looking impulse that competes with, or inhibits a concern for the future of humanity, our conservative disposition to sustain and preserve the things that we value itself underwrites that concern. Nor does the fact that our concern for future generations depends on this conservative disposition, mean that it's incompatible with our creative impulses. To see this one has only to reflect on the creativity and imagination that will be required to overcome the challenges to human survival and secure the prospects of a decent future for our successors.

Moreover, since human beings are in essentially creative species, whose history has a history of change, experimentation and innovation, and who are always developing new modes of living and new dimensions of value, a concern to ensure the future of humanity is itself a concern to sustain the open-ended and unpredictable course of human creative activity.

As applied to the future of humanity, in other words, the conservative disposition is a disposition to ensure that human creativity and innovation will continue to flourish. Despite the ways in which the conservative disposition supports rather than competes with a concern for the future, skeptics may deny that the disposition is rational in so far as it gives existing valuable things priority over valuable things that do not yet exist, it may be said to amount to an irrational form of status quo bias. Although I'm sure there is such a thing as irrational status quo bias, I have a difficult time seeing that there's anything irrational about the conservative disposition as I've described it. That's because I have a difficult time seeing what the alternative to it might be.

The conservative disposition reflects the fact that our value-based attachments can only be directed at what is or has been actual. We could not have a temporarily neutral disposition to form attachments to things that do not yet exist in the same way that we do to existing things. What would it mean to be just as attached to our future friends or to the children we will one day have, or to the great paintings that will someday be produced or the great novels that will someday be written as we are to our actual friends or children, or to the great paintings in novels that have already been produced boost. When it comes to attachment temporal neutrality is not an option.

---

An alternate normative suggestion might be that attachment is always irrational. We should strive to realize an ideal of detachment and to free ourselves as far as possible from all of our attachments. Whatever may be said for or against such an ideal of detachment, however, it does not support the idea that our bias toward existing attachments in particular is irrational.

Instead what's alleged to be irrational is attachment itself rather than the temporal sensitivity of our disposition to form attachments. It's true that if all attachments are irrational, then it follows trivially that a temporally sensitive pattern of attachments is irrational. But if all attachments are rational, then it also follows trivially that a temporally neutral pattern of attachments would be irrational.

The ideal of detachment does not show that there's anything irrational about temporal sensitivity, per se.

---

In general, the interactions between our values and our attitudes toward time are complex. And we should be cautious about assuming that every manifestation of temporal bias in our valuing attitudes must be irrational. Indeed, to the extent that the very term bias suggests irrationality or lack of justification, it's, undiscriminating use to refer to all forms of temporal preferences is unfortunate.

Our values and desires are shaped by our self-understanding as temporarily extended creatures and by our experience of temporality. We would not have the values we have if we did not understand the temporal dimension of our lives and the ways that we do. And the direction of influence also runs the other way, the values that we form serve in turn to shape our attitudes toward time, we would not have the temporal attitudes that we have, if we did not have the values that we do. We need to try to understand these reciprocal influences and not to assume that every manifestation of temporal bias in our valuing attitudes is irrational. As with studies of rational judgment and decision-making in other areas, the trick is to navigate between the complacent assumption that our ordinary thinking must be in good order and the revisionist application of oversimplified models that lack any authority over our actual practices and tendencies of thought.

---

I've tried to illustrate these broad themes by showing how a conservative disposition to sustain existing bearers of value, which some might take to involve a form of irrational status quo bias, is built into our valuing attitudes and cooperates rather than competes with a concern for the future.

The conservative disposition strongly supports a concern for the survival of humanity and the flourishing of future generations. To state my view in a way that is only superficially paradoxical: our concern for the future of humanity and for the flourishing of future generations, depends on a conservative disposition that applies directly only to presently existing and past bearers of value.

---

And although I've argued that our bias toward the future is limited, I'm skeptical of the neutralist claim that rationality requires us to eliminate or overcome it entirely. I believe this is one of those cases in which when confronted with the complexity of our actual thought, we should be wary of the prescriptive application of a simplified model of rationality that would classify any recalcitrant attitudes as being normatively deficient.

**There is of course a well-known parallel between the view that temporal neutrality is the default rational stance, departures from which stand in need of special justification, and the view that impartial beneficence is the default moral stance, departures from which stand in need of special justification.

The first view treats temporal neutralism as presumptively authoritative, and is suspicious of any tendency people may have to be more concerned about what happens at some times than at others. The second view treats an equal concern for the welfare of all people as the presumptively authoritative moral position, and is suspicious of any tendency people may have to attach special value to their relationships with particular people or to be specially concerned about what happens to some people rather than to others. I reject both of these views. I'm comfortable with the thought that our temporal attitudes are complex and that we lack any single master attitude toward time that we are not uniformly biased toward the past or the future or uniformly neutral.

I'm also comfortable with the thought that we have strong value-laden attachments to particular people and projects and relationships, and that these attachments are sources of differential reasons for action and differential forms of emotional vulnerability. This bears on the contrast that I mentioned toward the beginning of this lecture, between two different ways of thinking about questions concerning future generations. As applied to those questions, a combination of temporal and moral neutralism leads more or less directly to the quest for a principle of beneficence that would solve the puzzles of population.

At first glance, such a principal might seem to be the perfect antidote to the kind of temporal parochialism that I discussed in lecture one. I've tried to make clear throughout these lectures, I'm convinced that this solution is illusory and that once one focuses on the rich variety of human values and attachments and on the complexity of our actual attitudes toward time, it begins to lose its charms.

It's tempting to think that once that happens, our reasons for concerning ourselves with the fate of future generations simply drain away. The beneficence based literature tacitly though, no doubt, unwittingly, encourages this thought. In these lectures however, I've tried to show the reverse is true. Once we free ourselves from the thought that the basis for any concern about the future of humanity must lie in a principle of beneficence of some as yet unspecified sort, we can see that we have reasons of a number of different kinds, all rooted in our actual attachments as flesh and blood human beings for wanting future generations to survive and to flourish. In so far as these reasons depend on our existing values and attachments and on our conservatism about value, they depart from moral and temporal neutralism.

Yet it is to these very departures rather than to any form of neutralist beneficence that we must look in order to identify our strongest and deepest reasons for caring about the fate of our successors. Or so I have been trying to show. At the very least, I hope to have persuaded you, that there is an alternative to thinking about problems of future generations in exclusively or primarily beneficence-based terms, or indeed in exclusively moral terms of any kind.

If we broaden our horizons, we may find that we have even more reasons than we realize to worry about the fate of future generations.

# Suggested reading (optional):

- Stewart Brand: The Clock of The Long Now (especially chapters 1, 2, 5, 6 & 7; each chapter is just a couple of pages)
- Nick Bostrom: The Fable of the Dragon Tyrant (link) (6K words)
- Nick Bostrom: Astronomical Waste (link) (2K words)
- Peter Hartree: Nick Bostrom—An Introductory Reader (link) (2.5K words)

A longer, evolving list can be found here.