Will Humanity Choose Its Future?

Guive Assadi | Center for the Governance of AI

Abstract

An evolutionary future is a future where, due to competition between actors, the world develops in a direction that almost no one would have chosen. This paper explores the possibility of an evolutionary future. Some of the most important changes in history, such as the rise of agriculture, were not chosen by anyone. They happened because of competitive pressures. I introduce a three stage model of the conditions that could prevent an evolutionary future. A world government, strong multilateral coordination, and strong defensive advantage each, in principle, could stop competitive pressures. It is difficult to see how an evolutionary future could be prevented in the absence of any of these three conditions; this suggests that one would need to be very confident that one or more of them will exist to be very confident that humanity will choose its future.

Table of Contents

Introduction[1]

Discussion of the future often focuses on the values that will guide society's development. Consider, for example, both the tradition of utopian science fiction (for instance, Iain Banks's *Culture* novels) and of dystopian science fiction (for instance, Aldous Huxley's *Brave New World*). Insofar as alternative possibilities are considered at all, many writers seem to contrast the possibility of humanity choosing its future with the possibility of human extinction (or, especially recently, the possibility of takeover by out of control AI systems).[2] But, though human extinction has yet to happen, some of the most important events in human history seem to have been heavily constrained by competitive pressures. It is therefore worth considering whether it is possible that humanity will fail to choose its future not because it goes extinct, but because the future is determined by competitive pressures. I call this the possibility of an evolutionary future.

In this paper, I briefly discuss the concept of an evolutionary future in the abstract, and then proceed to explore the circumstances under which an evolutionary future might occur or be avoided. An evolutionary future might be prevented in any of the following situations:

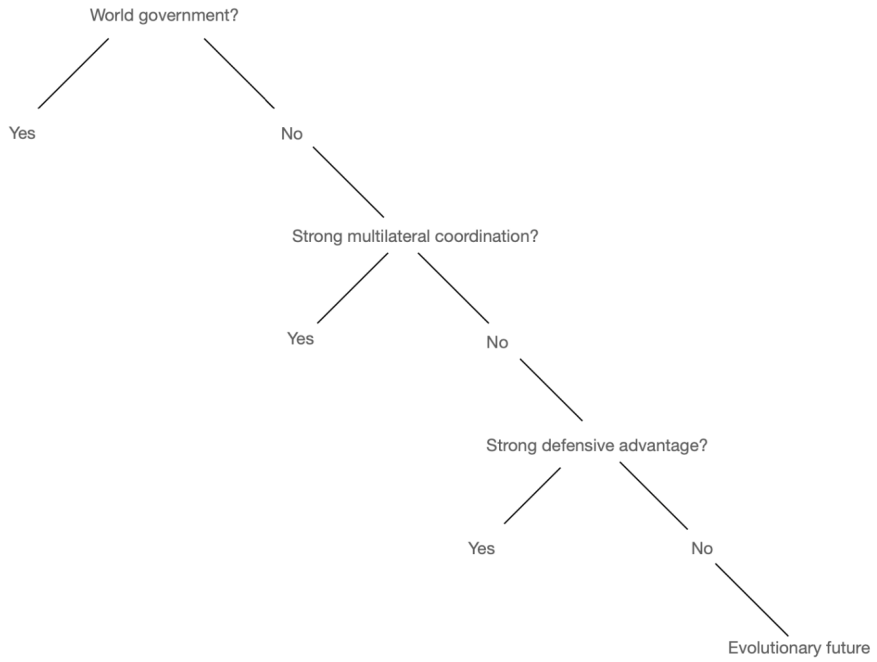
(1) A world government might ban militarily or economically advantageous practices, thereby enabling choice (by leaders) to determine the future rather than competitive pressures.
(2) Strong multilateral coordination might allow a group of actors (e.g. states) to work together to prevent competitively advantageous but undesired actions from being taken.
(3) Strong defensive advantage might allow local authorities to ignore competitive pressures and make choices without regard for them.[3]

[1] Thanks to Ben Garfinkel for supervising this research, and Michael Aird, Emma Bluemke, Allan Dafoe, Lukas Finnveden, Tom Davidson, Eric Drexler, Finn Hambly, Robin Hanson, Lennart Heim, Julian Hazell, Mckay Jensen, Anne Le Roux, Peter Mclaughlin, Richard Ngo, and seminar participants at Rethink Priorities and the Global Priorities Institute at Oxford University for comments and discussion.

[2] For example, though he does briefly discuss the related subject of "undesired dystopias", Ord (2020) mainly focuses on extinction risks. Some of the relatively few works that do extensively consider evolutionary futures include Alexander (2014), Bostrom (2004, 2014), Critch & Krueger (2020), Dafoe (2020), and, in particular, Hanson (1998, 2016, 2021a).

[3] Lukas Finnveden points out that these challenges will arise repeatedly, every time a civilization from one planet encounters a civilization from another. I ignore this issue in the following in order to control scope, not because it is unimportant.

It may be a bit clearer to consider these three preconditions of an evolutionary future in the form of a flowchart:



After discussing each of the stages of this chart, I will consider the implications of that analysis for the probability that humanity will choose its future. If one takes the view that an evolutionary future would be inevitable absent any of these conditions, then the probability that humanity will fail to choose its future cannot be lower than the probability that none of these conditions will obtain.[4] Therefore, if one has subjective probability estimates for each of the conditions (conditional on the previous conditions), one can multiply them through to get a lower bound on the probability of an evolutionary future. My goal is not to advocate for a particular probability estimate–rather it is to show that it seems that one needs a specific set of highly confident beliefs about the future's trajectory to be able to dismiss the possibility of an evolutionary future.

In the final section, I briefly discuss the normative implications of the possibility that humanity may not choose its future. Many recent discussions of the future involve the concept of existential risk: "where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential."[5] I consider whether an evolutionary future would amount to "value erosion," humanity's potential being permanently and drastically curtailed by competitive pressures. This depends on how good or bad an unchosen future would be, a question which I leave for future work. I merely argue that it is plausible that value erosion may be an existential risk–but also that reducing that risk may be intractable or pose serious risks of its own.

---

[4] That is, a necessary but not sufficient condition for avoiding an evolutionary future is that one or more of the items in the chart must exist. This is an example of J. L. Mackie's INUS condition: "[each step is an] *insufficient* but *necessary* part of a condition [i.e. three "no" answers in the flow chart] which is itself *unnecessary* but *sufficient* for the result [i.e. an evolutionary future]," Mackie (1965), p. 245.

[5] Bostrom (2003).

Conceptual clarification and examples

What is an evolutionary future?

An evolutionary future is a future where, due to competition between actors, the world develops in a direction that almost no one would have chosen.[6] It need not be a future that is unworthy of being chosen; it is not defined in normative terms. What is important is that the world is persistently and significantly different from how it would be if the dominant actors prior to the competitive process had been able to deliberately choose the course of the world's development.[7]

It might be objected that this definition is far too broad. Many, perhaps even all, extinction risks involve a collective action problem. A collective action problem prevents nuclear disarmament; therefore, extinction risks from nuclear weapons are in part a result of a collective action problem. Similar arguments could be made about the importance of collective action problems for risks from pandemics and climate change. Even a failure to mitigate naturally arising risks (from, e.g., supervolcanoes or asteroids) might be attributed to the fact that security against such risks is a global public good.[8] However, continued uncontrolled competition arguably remains reasonably probable conditional on no near-term extinction. That residual probability is the probability of an evolutionary future, as the term is used in this paper.

The mere fact that a future scenario involves competition does not necessarily make it an evolutionary future, in the sense used here. Competition of various kinds might be employed deliberately in non-evolutionary futures. In overall non-evolutionary futures businesses may compete to attract customers and politicians may compete for votes. The key difference is whether competition in some domain was instituted deliberately by some actor, for some purpose, or whether it emerged from uncoordinated actions serving no coherent purpose.

An evolutionary future might involve changes in what actors prefer, changes in the type of actors that are dominant, or large externalities. It is possible to get more of a sense of what is meant by going through a few examples, both historical and hypothetical.

A historical example: the rise of agriculture

Competition between political communities seems to have driven humanity's transition from hunting and gathering to farming.[9] Agriculture, once invented, allowed people to extract more food from a given amount of land than hunting and gathering. Farming societies therefore could support larger populations than

---

[6] One way that the future might not reflect what any one actor would have chosen is if the future's trajectory is a compromise between various visions that leaves no one completely satisfied. Such cases are not what I mean by an evolutionary future; any trajectory that would be in the "bargaining range" of the dominant actors prior to that trajectory's beginning is meant to be considered a chosen future.

[7] Choosing the future requires a great deal of knowledge about the long-run consequences of various courses of action. Therefore, for humanity to choose its future, it will need to not only become far better coordinated than it is now, but will need to gain a great deal of knowledge. The focus of this paper is coordination (top-down or multipolar) and the possibility of avoiding the need for coordination (through a strong defensive advantage). Whether the relevant level of knowledge is possible or likely is very much open to doubt. I leave this matter open, but for the beginning of a skeptical case, see Friedman (2019).

[8] Olson (1965); Posner (2004); Ord (2020).

[9] This is at least a plausible interpretation of results such as Sokal et al. (1991); for a popular exposition of this point of view on the rise of agriculture, see Diamond (1999).

hunter-gatherer bands. Thus farmers tended to win when growing populations led them to fight with hunter gatherers over land. Gradually, hunter-gatherers were forced to migrate, die, or adopt agriculture themselves. Thousands of years of territorial competition resulted in nearly all of the world's population living as farmers.

Stone age technology simply did not allow for humanity to consciously choose between agriculture and foraging; global coordination was not possible then. Had it somehow been possible to choose, the destruction of the hunter-gatherer way of life was probably not what most people would have chosen. While farming societies were more effective at capturing and holding territory, early farmers were more malnourished and more diseased.[10] Jared Diamond famously described the transition to agriculture as "the worst mistake in the history of the human race."[11] It was not really a *mistake*, though, since no one chose this path for humanity.

Hypothetical evolutionary future scenarios

Evolutionary future scenarios are not hard to imagine (though they are admittedly highly speculative). As I see it, there are two broad categories of mechanisms that might bring about an evolutionary future. First, competitive pressures might create a stable equilibrium that will last indefinitely. Agricultural civilizations which supplanted hunter-gatherers were themselves transformed in many ways by the industrial revolution. If long-term technological stagnation eventually sets in, there may cease to be future technological revolutions to transform what is advantageous. Therefore a set of practices may remain advantageous (and dominant) indefinitely.

Second, competition may directly change what actors value. Then, even if there is a later chance to stop competition and choose the future (from that point on), the future that gets chosen will already have been changed by competitive pressures. I will discuss two specific, hypothetical evolutionary future scenarios.

Digital minds may eventually become both the world's main inhabitants and its main decision-makers. Such entities might be able to easily copy themselves and to be modified to better suit their tasks. In an environment of uncontrolled competition these features would allow digital minds to evolve rapidly. Most obviously, they might evolve to become better at their jobs.[12] This might mean a greater ability to focus, greater commitment to work, more joy in their work, or more fear of underperforming. Just as cetaceans no longer have feet, future digital minds might lose vestigial structures that no longer serve adaptive purposes.[13] This evolutionary process could gradually change the goals or characteristics of these digital minds, which could eventually shift the world in directions that almost no one prior to that evolutionary process would have chosen.[14] In one particularly extreme hypothetical case (due to Nick Bostrom), digital minds might lose the ability to have conscious experiences if phenomenal consciousness is unhelpful in economic competition.[15]

Malthusian space colonization is another hypothetical evolutionary future scenario worth considering. Barring human extinction, civilization will probably eventually expand through space. The inhabitants of an uncoordinated civilization spreading through space might gradually adapt to maximally efficient space colonization. Agents that wish to use the resources of space for any purpose other than facilitating duplication of themselves might gradually lose out to agents that exclusively use their resources to reproduce themselves and claim more space.[16] The universe may eventually be dominated by agents whose primary goal is reproducing

---

[10] Lambert (2009); Larsen (1995).
[11] Diamond (1999).
[12] Hanson(2016); Bostrom (2004); Shulman & Bostrom (2021).
[13] Bostrom (2014), chap. 11.
[14] See Shulman (2010), and, arguably, "Part II" of Christiano (2019).
[15] Bostrom (2014).
[16] Hanson (1998).

themselves as much as possible. Just as in the case of digital minds, vestigial structures and vestigial goals might eventually disappear.[17]

Modern experience may lead us to underrate the long-run significance of competition

History affords examples of deliberate choice shaping the course of events, at least on the scale of centuries. Many seemingly competitively disadvantageous practices are very widespread.[18] For example, many countries restrict civilian nuclear energy more heavily than fossil fuels, despite nuclear power generally being safer. It seems that countries could become more economically competitive if they relaxed their restrictions on nuclear energy, which would allow them to increase energy consumption without having to pay the costs from pollution created by burning fossil fuels (lost work hours to medical problems, healthcare costs, etc). Though the nuclear power example may be controversial, most readers will probably be able to think of modern practices that are widespread but (in their view) competitively disadvantageous. Additionally, there is clear evidence of contingent decisions shaping subsequent history. The enduring nature of certain political constitutions and many of the features of world religions come to mind. Such things do not necessarily seem to be what is most competitively advantageous, yet they are lasting.[19]

However, there are a few arguments that suggest that, despite the prominence of unconstrained choice and blind imitation in the present, the modern weakening of competitive pressures may not last. First, per capita wealth is currently far above subsistence–and growing. However, if there is any limit to the amount of wealth that can be extracted from a fixed level of resources, per capita income growth must eventually cease. And population growth might then reduce per capita income to a subsistence level. The modern era may prove to be a brief non-malthusian interlude between the malthusian period before the industrial revolution and the malthusian period after fundamental limits to growth are reached.[20] If this argument is correct, future people who are struggling to survive might have less opportunity to use their surplus to directly pursue value the way that many modern people can.

Second, in recent history changes between levels of technology have been very rapid. Consider the saying that "generals always fight the last war." For this to make sense as a critique of generalship, it has to be the case that the next war and the last war are very different from each other. The rapidity of technological change prevents selective sweeps for institutions and practices that are advantageous at one level of technology but that are not advantageous at other levels. If technological change eventually slows down or stops, there may again be time for selective pressures to act on institutions at a given level of technology. So it would be wrong to conclude that an evolutionary future is unlikely from the (apparent) fact that the modern world is more characterized by choice and mimicry than selection and differential proliferation.[21] True though that may be, there is reason to believe that the future may be different.

---

[17] Hanson (2021d).

[18] Hanson (2020).

[19] Cf. John Adams' (1787) prophetic remarks: "The institutions now made in America will not wholly wear out for thousands of years. It is of the last importance, then, that they should begin right. If they set out wrong, they will never be able to return, unless it be by accident, to the right path", (p. 298).

[20] Hanson (2009).

[21] Dafoe (2015).

Why humanity nevertheless might choose its future

Despite all of the above, technological change might make it possible for humanity to choose its future. The technology to create a world government capable of solving global coordination problems simply did not exist for most of human history, which arguably limits the relevance of extrapolating from the fact that humanity did not choose much of its past to the prediction that humanity will not choose its future.[22] In the previous section, I discussed a few hypothetical technological changes that might enable an evolutionary future. Artificial general intelligence (AGI) and biotechnology are two technologies that may enable us to avoid an evolutionary future.

Definitions of AGI vary, but I define it as a level of AI technology that is capable of performing almost all human labor more cheaply or more effectively than humans themselves could.[23] If AGI is ever created, it might be tasked with performing the majority of useful work, making the most important decisions, and enforcing rules or laws. If AGI systems' goals do not change over time, it might be possible for a civilization in which AGIs perform these functions to prevent its own values from changing. This would count as humanity choosing its future if some group of humans at one point decided what goals to give the AGIs.

There are several reasons why a civilization with AGI might be able to choose its future.[24] Currently, the best way of conveying information to the future is writing. Books (digital and paper) can be lost or destroyed, and in any case there is inevitable loss of information when something is written down. Through digital error correction and an ability to intelligently respond to physical threats to computers, AGI systems might be able to preserve highly complex valuational information indefinitely. One major reason that values change over time in human societies is that people die and their successors often disagree with them about valuational questions. AGI systems need not age or die. They have no natural lifespan and it may be possible to duplicate them, just as other software programs can easily be duplicated. Finally, if AGI systems do drift in their goals, it may be possible to reset them to an earlier state and undo that drift.

In addition to AGI, there are also a few, hypothetical, highly advanced biotechnologies that might someday allow humanity to choose its future. Indefinite life extension might solve succession problems. If it at some point becomes possible for individual human beings to avoid aging and deaths of old age, then it would be possible for leaders to continue indefinitely in positions of authority. It would therefore be possible to avoid the change in values that comes with cohort replacement.[25] And advanced lie detection, if it is possible, might promote future coordination.

Summing up

Based on the above, whether humanity will choose its future seems to be an open question. In the next three sections, I will discuss the path to an evolutionary future and various factors that make those preconditions more or less likely.

---

[22] Bostrom (2005).

[23] Grace et al. (2017) use an equivalent definition, but for the term "high level machine intelligence" rather than "artificial general intelligence."

[24] Finnveden et al. (2022).

[25] MacAskill (2022).

World government

A world government is the most obvious way that an evolutionary future could be avoided.[26] Perhaps the most commonly used definition of a government in the social sciences is Max Weber's. Weber wrote that: "the state is the form of human community that (successfully) lays claim to the monopoly of legitimate physical violence within a particular territory."[27] This definition has to be modified: a *world* state would not be limited to a *particular* territory.[28] But the key point, for present purposes, is that a world government would be able to use its monopoly on force to prevent the world from being driven along a competitive track by restricting behaviors that would otherwise be competitively advantageous but which have undesired long-run effects.[29] The selection pressures shaping a world under a single government at least have the possibility of being chosen.

When trying to predict the future, it is often useful to look at both background trends and specific mechanisms. In this section, I first briefly discuss two long-run trends relevant to predicting whether a world government will come about; I then turn to two mechanisms by which a world government might arise.

Trends

One relevant trend is that human history and the evolutionary history of life both seem to show development in the direction of the top level of organization growing higher over time. At the beginning of the history of life, the most complex organisms were single-celled. Single-celled organisms were followed as the most complex organisms by simple multicellular organisms that were internally undifferentiated (sponges), then multicellular organisms with structural and functional internal differentiation (nematodes, sea stars, trees), and then organisms with social behavior (schools of fish, ant colonies, wolf packs).[30] In human history, hunter-gatherer bands held together by biological relatedness were succeeded by small-scale agricultural societies (such as the neolithic cities of the near east), then much larger agricultural civilizations with an advanced division of labor (such as Ancient Rome or Ancient China), and, in the modern era, by a complex global civilization with an interdependent world economy dominated by one or a few superpowers.[31] Thus, it may be reasonable to think that by extrapolating further, we can predict a world government in the future–a world system with the highest possible top level of organization.

However, a simpler observation about human history and the history of life is that there has never been a single state that ruled the whole world or an organism that included all of life within itself.[32] All history–human and pre-human–is the history of many competing entities. And even if a world government were to arise, there is no guarantee that it would last long enough to prevent the vast bulk of the future from being shaped by unguided evolution. A world government that lasted for a million years before dissolving into

---

[26] The existence of a world government would not necessarily prevent an evolutionary future, but it would at least make it possible to avoid one; see note 5 above.

[27] Weber (1919), p. 33.

[28] For present purposes, the "world" may include places in outer space, if those places are controlled by humans (or some sort of successor to humans). I chose the term "world" over the term "global" because really what is meant is a government encompassing all of human civilization, not a government that rules over the globe in a more literal sense.

[29] Alternatively, a world state might compel behaviors that would otherwise be competitively disadvantageous but that have beneficial long-run consequences.

[30] Maynard Smith & Szathmáry (1995).

[31] Wright (1999).

[32] At least, there has not been such an organism since the very beginning of the history of life.

competing successor states would have ruled for only a tiny fraction of the time remaining before the earth will cease to support life.

Mechanisms

Voluntary formation of a world government

A world government might be brought about through a mostly voluntary process, in which elites in various countries cede power to global governance organizations (gradually or all at once).[33] This process might develop out of existing global governance institutions or it might arise from some currently unanticipated source. Recent history (since the end of the Cold War) has arguably been characterized by increasing convergence between countries, through a combination of official multilateral institutions, unofficial activist networks, and informal centralization of policy discussion. Consider the similarity of regulations on human cloning in diverse jurisdictions. Robin Hanson has argued that the merging of separate national elites into a single, coordinated global elite might, over time, turn into a kind of world government.[34]

As of now, global governance institutions generally lack coercive authority; they are not really able to use force against states that defy them. They must rely on various weaker carrots and sticks such as access to markets or aid. One important example is the International Atomic Energy Agency (IAEA). The IAEA helps countries access civilian nuclear power; it withholds aid from countries that it determines are attempting to build nuclear weapons in violation of the Nuclear Non-Proliferation Treaty.[35] If global governance organizations ever move beyond weak carrots and sticks and are able to directly enforce their decisions, they will have become a world government.

Alternatively, a world government might develop from a currently unanticipated source. Many people, now and in the past, have found the idea of world government attractive. Some medieval Christians, early Muslims, and twentieth century Marxist-Leninists hoped to unite the world under a single state based on their beliefs.[36] New religions or ideologies may arise in the future and try with more success to implement global governance. Imagine being alive in 600 AD and considering whether any one state would come to control the whole of the Middle East and North Africa. It might have been tempting to approach this as a question about whether existing states could expand over that area (the relevant ones would have been the Sassanian Persians and the Byzantine Empire). That approach would have missed the Rise of Islam. We should be wary of a similar neglect of unanticipated ideological forces.

World conquest

A world government could also be created involuntarily, i.e. through world conquest.[37] In this section, I discuss two processes that might enable world conquest: uneven growth acceleration, which could make one country

[33] It is much easier to imagine a world government voluntarily coming to include nearly all countries than 100% of countries. In practice, voluntary and violent processes may both be included in the formation of a world government.
[34] Hanson (2020, 2021b).
[35] Koppell (2010); Stafford & Trager (2022).
[36] Burnham (1943); Aligheri (1559); Ansary (2009); Kelsen (1948).
[37] In practice, a world government is probably more likely to come about through a combination of voluntary and involuntary means than by either means exclusively, but it is useful to separate them as ideal types.

more powerful than all others, and changes in the offense-defense balance, which might allow conquest without radical changes in relative wealth.

Uneven growth acceleration

The singularity thesis

The idea that growth will radically accelerate in the relatively near future has recently been explored both by futurist writers and by some mainstream growth economists. The futurist version of the argument runs as follows. Human beings are not really able to directly improve their intelligence; they can learn new information, concepts, and skills but they cannot deliberately redesign their brains. However, this constraint would not apply to AGI. An AGI would be able to redesign itself with, at first, human level AI engineering ability. Soon, it would achieve superhuman AI engineering ability. As it improved itself, the rate of improvement would increase (or so the argument goes).[38] The AGI would only stop improving itself when its investments in augmenting its own intelligence started to return diminished improvement.

Growth economists are generally much more reluctant than futurists to predict that a singularity is near or likely. Still, a singularity is one possible outcome of the progress of automation given some conventional models of economic growth. Automation since the industrial revolution has not made labor worthless because automation has historically created new applications of labor and increases the relative value of tasks whose productivity does not increase.[39] However, if full automation is ever achieved, growth may be proportional only to investment, and further growth might make more investment possible. This could potentially trigger a massive increase in growth rates–which might last until fundamental limits on growth are reached. If there were a singularity, that might allow one country to get such a large advantage over all of the others that it would be able to conquer the world.[40] On the other hand, multiple entities or nations might undergo singularities simultaneously or in quick succession, preventing any one of them from gaining an overwhelming advantage.

Moderate growth acceleration

Year on year economic growth rates have changed radically across human history.[41] They were much lower during the agricultural period than during the industrial period. It therefore seems imprudent to rule out the possibility that, in the future, there may be another transition to a faster growth mode (which nevertheless is slower than a growth singularity). A variety of technologies might be associated with this transition, just as the steam engine signaled the shift from agricultural era growth rates to industrial era growth rates.[42]

Moderate growth acceleration also might be enough to enable one country to conquer the world. The industrial revolution began in Britain, and took Britain from about 1% of world GDP to about 15% of world GDP and from political marginality to a leading diplomatic and military position. A similar acceleration in the future might move some nation from about 15% of world GDP (roughly the position of China or the United

---

[38] Yudkowsky (2013); Bostrom (2014).

[39] Aghion et al. (2019); Nordhaus (2021); Trammell & Korinek (2020).

[40] It is possible that the entity empowered to take over the world would not be a government but rather an AI system acting against the wishes of its human creators, see Bostrom (2014).

[41] See Pritchett (1997) for clear evidence that pre-modern growth rates must have been much slower than modern ones, despite the low quality of pre-modern economic data (this interpretation of the significance of Pritchett is due to Ben Jones's commentary on Tom Davidson's work on the possibility of future growth acceleration).

[42] Hanson (2000).

States today) to, say, 95%.[43] However, like a singularity, it is also possible that moderate growth acceleration could occur close to simultaneously in many countries and not fundamentally change the balance of power.

Stagnation

Economic stagnation, or something very close to it, has been the norm in human history. The past few centuries of consistent growth have been an aberration.[44] If some countries were to undergo prolonged economic stagnation (perhaps due to the "middle income trap"[45]) while their rivals continued to grow at rates typical of modern economies, that could enable world conquest by the still-growing coalition in much the same way as concentrated growth acceleration.

Non-economic strategic change

It is unclear what level of economic disparity is necessary for world conquest. For example, a coalition with only 5% of gross world product (GWP) might be able to exert significant leverage over a coalition with 95% if it has nuclear weapons. North Korea today has about 1/20th of South Korea's GDP, but it is far more than 1/20th as politically influential because it has nuclear weapons and the ability to unleash a massive artillery barrage on Seoul.[46] The more future technology favors defense over offense, the more rapid an uneven growth acceleration would have to be to give one actor the ability to overwhelm the others.[47]

Conversely, future technologies may be so offensive-advantaged that no large change in relative wealth is necessary to conquer the world. One simple example may be if it suddenly became easy to block nuclear missiles. If one nuclear power had this technology and the rest of the world did not, that power might be able to use nuclear weapons to conquer the world. If all countries had the ability to block nuclear weapons, then a country with an advantage in conventional weaponry might be able to conquer the world.

Strong multilateral coordination

In the previous section, a world government was defined as an entity with a monopoly on the legitimate use of force throughout the entire world. One can think of the task of preventing an evolutionary future as a collective action problem which might be solved by a world government. Collective action problems arise when the benefit from an action exclusively accrues to one individual, but at least some of the costs are borne by a group.[48] Selfish individuals will engage in the action more than would be socially optimal because the cost to them is smaller than the total cost. Perhaps the most influential paper in the literature on collective action problems is Garret Hardin's "The Tragedy of the Commons." Hardin assumed that the only possible solution to the putative tragedy is state action, either through direct state management or privatization and the enforcement of property rights.[49]

---

[43] I first heard this point made by Carl Shulman.
[44] Clark (2007).
[45] Eichengreen et al. (2011).
[46] This idea is due to Ben Garfinkel.
[47] Lynn-Jones (1995).
[48] This is in the case of public bads. In the case of public goods, all the costs accrue to an individual but some of the benefits are shared by the group; Olson (1965).
[49] Hardin (1968).

But, as Elinor Ostrom showed, these two solutions do not exhaust the possibilities. Collective action problems are often solved through decentralized coordination. It thus may be possible for multilateral coordination to prevent an evolutionary future without a world government.[50] One example of strong multilateral coordination is the fishing community of Alanya, Turkey. In the 1970s, Alanya developed a decentralized way of sustainably managing common resources and settling disputes. To prevent overfishing, the fishers of Alanya agreed to limit the number of fishing licenses distributed. The fishing area was divided into numbered locations, spaced so that they did not interfere with each other. Then, each licensed fisher drew lots and was assigned a location. The fishers moved one spot over each day, rotating through the entire set of spots over the course of a year. This system relied on the fishers themselves for monitoring and enforcement–no outside power was required.[51] Similarly, successful coordination sometimes happens in the anarchic international system. Consider the successful global campaign to limit the use of CFCs because of their effects on the ozone layer.[52] Finally, and particularly relevant for present purposes, technology can facilitate decentralized coordination schemes. The internet, through such schemes as border gateway protocol routing, coordinates many actors even though it lacks any central authority capable of compelling obedience by force.[53]

Future technology may enable greater coordination with unshared preferences

Obstacles to coordination tend to be most serious when actors do not share preferences. Coordination failure between actors with dissimilar preferences can be caused both by irrationality and by structural impediments to bargaining for rational agents. Because these impediments to coordination are distinct, I will discuss the implications of each separately.

Changes to rationality

Many failures of coordination are caused by an inability to correctly interpret bargaining relevant information. World War I may not have happened if the leaders of Germany, Russia, and Austria had understood that they were starting a total war rather than a limited war such as the Franco-Prussian War or the Crimean War. In addition to issues of bounded rationality, bargaining itself can be costly and difficult. Outright irrationality also plays a role in bargaining failure. Wars, famously, can be caused by hotheadedness.

The ability to determine the likely consequences of one's actions, engage in highly complex negotiations, and control one's emotions are all, in principle, subject to technological improvement. For example, if leaders can delegate some aspects of negotiations to AI assistants, that might mitigate some of these

[50] The boundary between world government and strong multilateral coordination may be fuzzy. For instance, Weber (1919) held that "all other organizations or individuals can assert the right to use physical violence *only insofar as the state permits them to do so*" (p. 33, emphasis added). It is not clear to me whether the United States of America was one or many states according to Weber's definition between 1781-1789, when the Articles of Confederation were in effect. If a similar (but worldwide) confederation exists in the future, it is therefore not clear whether it should be seen as a world government or as a case of strong multilateral coordination.

[51] Farmers in Switzerland and Japan use non-state mechanisms for avoiding the overuse of common land that have lasted for centuries. Farmers in Spain and the Philippines used decentralized coordination to share scarce water; Ostrom (1990) (compare Ellickson (1994) and Scott (1999)).

[52] Gonzalez et al. (2015).

[53] Feigenbaum et al. (2007).

problems. Therefore, given that the ability to bargain is generally advantageous, and that technology opens up a wide variety of new possibilities, we should (at least weakly) expect that rationality in bargaining will improve in the future.

Changes to obstacles to coordination for rational agents

Not all coordination failures among actors with disparate preferences are failures of rationality. In particular, a mutually beneficial bargain might not be reached if the parties do not have the same information about the strength of their relative positions.[54] Without shared information, it may not be possible to find an alternative that all parties prefer to the breakdown of negotiations.[55]

A second major reason that rational agents may fail to reach mutually beneficial agreements is a commitment problem. A commitment problem arises when there is no way for the parties to ensure each other's compliance with an agreement. This problem is solved in business with contracts enforced by national governments. However, it significantly contributes to interstate war because there is currently no way to extract binding commitments from sovereign states.[56]

Asymmetric information

In the absence of asymmetric information, lingering uncertainty about the strength of the parties' bargaining positions need not prevent bargaining. To see why, consider the example of two states that might go to war. War is usually the result of a bargaining failure. Fighting a war brings death and destruction that might have been avoided if the sides could decide beforehand who would win and allowed that side to extract concessions without fighting. If both states shared all relevant information, they should (if they are rational) be able to agree on the probability that each side would win. They then might be able to compromise, emphasizing the predicted winning side's interests proportionally to the probability that it would win.[57]

One reason that asymmetric information issues currently persist is because it is hard to release all bargaining-relevant information without giving opponents an advantage.[58] For example, it might be hard for the U.S. to prove it has nuclear second strike capacity without potentially compromising second strike capacity. This problem might be amenable to some technological solution. In the future, it could conceivably become possible to use zero-knowledge proofs or privacy-preserving machine learning to release all–and only–bargaining relevant information.[59]

Some forms of information, such as a leader's level of resolve in a standoff, cannot be released because they are inherently hard to transmit given human biology and absence of reliable lie-detection technology.[60] However, if reliable lie detectors are someday invented, it may become much easier for leaders to demonstrate

---

[54] In this section I draw heavily on Fearon (1995); as Taylor & Singleton (1992) and Blattman (2022) observe, the literature on the transaction costs of coordination can be seen as an extension of the argument of Coase (1960). Coase held that, in the absence of transaction costs, rational actors should always be able to find efficient bargains. When efficient bargains are manifestly not being found (e.g., if there is a war), the Coase Theorem focuses our attention on transaction costs.

[55] Fearon (1995).

[56] Ibid.

[57] States could also use weighted randomization for issues that are inherently indivisible (Fearon uses the example of a conflict over who will sit on the throne of Spain).

[58] Fearon (1995); Garfinkel (2022)

[59] Garinkel (2020, 2021, 2022); Trask et al. (2020)

[60] Current lie detection techniques are apparently not very reliable; Iacono & Ben-Shahar (2019).

their resolve. An even more speculative possibility is that artificial entities whose capabilities and intentions are more transparent than those of human beings may be making the decisions in the future, and it may be easier for them to bargain because they can share more information about their mental states.[61]

Other things being equal, one might expect future actors to pay significant costs to develop and implement technologies that reduce information asymmetries for the purpose of aiding collective action. People today are willing to behave in costly ways to make it easier for others to work with them. For example, they go out of their way to maintain good reputations or achieve institutional transparency. But it is, to say the least, unclear what the balance between transparency and obfuscation will turn out to be in the long run. Extreme confidence on this score seems quite inappropriate: technology might enable more effective deception, rendering bargaining-relevant claims untrustworthy.

Commitment problems

Commitment problems arise when it is difficult or impossible to credibly bind one's future self to follow an agreement. Consider, again, war as a paradigmatic example of bargaining failure. Even if the two sides could agree on the probability that each of them would win, commitment problems create an additional hurdle to bargaining. Suppose that two countries agree that the chance that the first country would win a war is 60%. It will still be hard to come to an agreement if they have no way of being sure that the favored country will not make fresh demands as soon as it is strengthened by control of 60% of the disputed resources.

There are some institutional and technological changes that might reduce the significance of commitment problems in the future. Jon Hovi and collaborators have argued that environmental treaties should require that signing countries place large sums of money in escrow as security for their compliance rather than relying on countries to voluntarily pay penalties for violating agreements after the fact.[62] This mechanism could be generalized to all manner of commitment problems–peace treaties, for instance, are no less amenable in principle to being secured with escrowed funds than environmental treaties.

Furthermore, certain technologies might strengthen the ability to make credible commitments. Decentralized cryptographic escrow services already exist; as of now, they are typically used for ransomware and other criminal activities that are not allowed by legitimate financial institutions. In the future, they might become robust enough for use as an aid to bargaining under international anarchy.[63] Another speculative possibility is that future agents bargaining with each other might collaboratively build "treaty bots": robots designed to autonomously enforce agreements and thereby to allow their makers to more credibly bind themselves.[64] This idea depends on it being possible for adversaries to work together on engineering projects and trust that the other party won't be able to subvert the ostensibly shared goals of the project. This is not necessarily a safe assumption, as an example from nuclear arms control efforts shows:

> The annual Underhanded C Contest challenges participants to write a program in the C computer programming language that appears [to expert judges who read the code] to be benign and straight-forward, but that contains malicious code. In the 2015 contest, Underhanded C partnered with the Nuclear Threat Initiative to develop a nuclear disarmament scenario, in which two state parties have agreed to an inspection regime in which measurements of objects representing nuclear weapons would be compared to a reference

[61] Vidal Bustamante et al. (2022); Hanson (2016).
[62] Hovi et al. (2012).
[63] Garfinkel (2021).
[64] Bostrom & Shulman (2020).

> measurement of a trusted object. An information barrier would be implemented to distill the results from a potentially sensitive measurement to a simple "yes" if the measured object was sufficiently similar to be deemed a nuclear weapon; otherwise "no". Contestants were asked to design a code that would make it appear that the detectors and were performing as expected during testing, but would then proceed to confirm as "yes" objects that were not sufficiently similar, so that the inspected party would could get credit for dismantling objects that were not actually nuclear weapons (thus allowing them to maintain larger nuclear weapons reserves than the other treaty party). The winner of the contest developed a program that would change the number of bytes being analyzed in a verification scenario, allowing an object with a very small amount of fissile material to trigger a false positive result.[65]

One somewhat common view among futurist writers is that, relatively soon, extremely powerful AGI systems will be created that will have a much easier time coordinating with each other than we do today. The argument here runs that AGIs should be able to modify themselves to credibly commit to taking particular actions such as rewarding those who help them and punishing those who harm them. Further, it is often thought that AGIs will be able to share their code to prove that they will follow through on any commitments they have made.[66]

Someone who is confident that AGIs will be able to coordinate near-perfectly might reasonably think that an evolutionary future is very unlikely. However, I do not think the evidence marshaled in favor of AGI coordination maximalist view thus far is fully convincing. If the relevant AI systems resemble vastly scaled up versions of contemporary machine learning systems, their "code" may not be much more transparent to them or modifiable at will than our brains are to us.[67] Aside from specific concerns about neural networks, it just seems premature to be extremely confident that future technology will strongly favor transparency over deception.

Future preference convergence may make coordination easier

Other things being equal, if different people's preferences become more similar then collective action problems will be easier to solve.[68] If a group shares a common goal (i.e. if for each member private benefit = public benefit) there can be no collective action problem in the strict sense.[69]

---

[65] Gastelum (2020), p. 175. Lennart Heim points out that the winning strategy (Åkesson 2016a, 2016b) may not have worked in a formally verifiable programming language–but also that formal verifiability does not remove the possibility of errors of interpretation of co-written software programs enabling deceptive behavior.

[66] Yudkowsky (2022): "any system of sufficiently intelligent agents can probably behave as a single agent, even if you imagine you're playing them against each other." The idea that sufficiently advanced AIs will have an extremely easy time coordinating with each other is sometimes associated with functional decision theory (Yudkowsky & Soares 2017). However, whichever decision theory you accept, it's possible to see why transparent code and the ability to self-modify to lock-in commitments might enhance coordination.

[67] It is also possible that learned architectures would be most efficient but that dominant systems in the future will redesign themselves to be more transparent and self-modifiable in order to aid coordination.

[68] There could still be coordination problems in a looser sense, such as problems with sharing information or overseeing group work, regardless of whether preferences are shared or not. In principle, such problems might be serious enough to create an uncoordinated future.

[69] This is why Olson (1965) for the most part set aside "philanthropic and religious organizations" (p. 6). See the treatment of related themes in Cohen (2008): "Liberally minded economists take for granted that economic agents are self-seeking [...] and then they want people as political agents to act against the grain of their self-interest: pile up your earthly goods on the

It is therefore highly relevant to the future prospects of strong multilateral coordination whether preferences will become more shared or will remain various. In this subsection I will consider the strengths and weaknesses of two arguments that there will be less diversity of preferences in the future than there is today. First, if the future has much more per capita wealth than the present, conflict over scarce resources for consumption might become less common. Second, non-selfish moral, political, or religious preferences may eventually become much more widely shared than they are now because people converge to the (in some sense) right answers on those questions.

Conflict between selfish preferences may become less significant

Generally speaking, people are thought to have sharply diminishing marginal utility in wealth–often, sublogarithmic utility. Per capita wealth might radically expand in the future due to a sudden increase in growth rates or a long continuation of present growth rates. If utility is logarithmic or sublogarithmic in wealth and refusing to bargain carries substantial risk, then greater per capita wealth will make a wider array of potential bargains acceptable.[70] As Ben Garfinkel writes:

> Prosperity has two opposing effects on conflict that do not entirely balance out. First, prosperity increases the potential spoils of conflict: the wealthier your neighbor is, the more you could win by robbing or extorting him. Second, prosperity makes people more satisfied with what they have: the wealthier you are, the less you value each additional dollar you could take from your neighbor. The relative strengths of these effects depend on just how quickly people become satiated. The paper uses empirical evidence to argue that people become satiated quickly enough for conflict to lose appeal.[71]

This argument that coordination will improve in the future depends crucially on the function mapping wealth to utility for decision-makers. However, the logarithmic or sublogarithmic character of that function is a contingent fact about modern human psychology. As such, it may change in the future. If it does change, then there is no reason to expect that greater per capita wealth will improve coordination.

Non-selfish preferences may converge

In addition to the above argument that conflict between selfish preferences may become less significant in the future, there are a few arguments that suggest non-selfish preferences may also converge in the future. To understand the relationship between shared preferences, unshared preferences, and collective action, consider the following example. Imagine a socialist who wants to benefit the public by distributing socialist literature and thereby spreading socialist ideas. Imagine, also, a conservative who wants to distribute conservative literature for similar reasons. These two will not work together to distribute political literature because they have incompatible preferences about the kind of political education that should be supported. If there are economies of scale in the distribution of political literature (which is plausible), then less total political literature will be distributed than if they were to work together. A society with two socialists or two conservatives would have a

---

mundane plane of civil society but be a saint in the heaven of politics. One way out of the apparent contradiction is to generalize *Homo economicus*: hence the work of theorists like James Buchanan and David Gauthier. I am engaged in an exploration of the reverse generalization" (p. 2).

[70] Drexler (2018); Aschenbrenner (2020).

[71] Garfinkel (2023), p. 3.

greater degree of total political education, and the preferences of the two socialists or the two conservatives would be better fulfilled.

I will discuss three related but independent arguments for the view that non-selfish preferences will become more widely shared in the future than they are today. These are: (1) disagreements may be more empirical than evaluative, (2) realist moral convergence, and (3) subjectivist moral convergence.

Disagreements may be more empirical than evaluative

In the case of the socialist and conservative discussed above, it may be that much of their disagreement about public policy is ultimately caused by different ideas about how best to solve social and economic problems.[72] If this is the case, then they might come to agree if more reliable information about the social sciences becomes available. And it may be reasonable to expect that, with more time for study and more powerful research tools, more reliable information will indeed become available. If apparent evaluative disagreements are in fact largely empirical disagreements, and if more reliable empirical information becomes available in the future, impartial preferences may converge and reduce the significance of coordination problems.

Realist moral convergence

It may also be that the most important differences in apparent values are not caused by differing empirical beliefs. However, even if this is true, non-selfish preferences may still converge. If there are mind independent normative facts, accessible to rational investigation, then moral preferences in the future might converge on an accurate view of the good. It seems reasonable to expect that future people will have a very accurate understanding of physics. If it turns out that normativity is a proper subject of objective science–like physics–then it may also be reasonable to expect them to have a good understanding of normativity, and therefore to converge in their evaluations. If this "realist moral convergence" thesis is correct, preferences in the future may become more widely shared, which would mitigate coordination problems.

Subjectivist moral convergence

Convergence of non-selfish preferences may be possible even if apparent values differences are not reducible to empirical disagreements and there are no mind-independent normative facts. This is because it may be that if you properly idealize human subjective moral preferences you will get universally or close to universally shared results. If the relevant actors all perform the same idealization procedure on their values and end up with the same result, then unselfish preferences would become shared.[73]

To be sure, if metaethical subjectivism is correct, then that would seem to reduce the likelihood of future moral convergence relative to what it would be if metaethical realism is correct. If subjectivism is correct, it might be that there is no one correct idealization procedure, just a variety of different procedures that reach different results. It also might be that different people's initial values would yield different results even given the same idealization procedure. However, it remains possible that, even assuming that metaethical subjectivism is correct, values could converge in the future and thereby enable a greater degree of coordination.

---

[72] Berelson (1952).

[73] See Oesterheld (2017) for an even more exotic argument that preferences may converge in the limit of reflection.

Strong defensive advantage

If defense is stably and strongly advantaged over offense, then an evolutionary future might be avoided. A sufficiently strong defensive advantage removes the element of collective action from the problem of preventing an evolutionary future. If it is possible to ignore one's competitors without ceding the opportunity to determine what will happen in the future, competitive pressures may not shape the world.

There may be historical precedent for a strong defensive advantage allowing desired but disadvantageous practices to continue. Agricultural civilizations gradually displaced hunter-gatherers from arable land on major continents. However, prior to the early modern period, some hunter-gatherers on remote islands could not have faced competition from agriculturalists because ocean navigation was not yet advanced enough for agriculturalists to reach those islands. The Andamanese in the Indian Ocean (some of whom still persist in a paleolithic way of life today) might be an example of this phenomenon.[74] Before they could be reached by ship, there was in effect a strong defensive advantage that allowed the most isolated hunter-gatherers to maintain their way of life.

It may be that, in the very long run, conditions will again emerge that create a strong defensive advantage for groups intending to pursue a competitively disadvantageous form of life. Groups seeking to implement values that would otherwise be outcompeted might strategically perform well in competition in the short run, but turn to instantiating their values once they were safe from adversaries.[75] Once a period of defense-dominance began, they would be in a similarly secure position to pre-modern hunter-gatherers on remote islands. A potential route to future defense dominance that has attracted the attention of a few futurists is that the geometry or physics of outer space may inherently advantage defense over offense. One reason this could be is that, in tens of billions of years, the expansion of the universe will causally isolate galaxy groups from each other.[76]

It is unclear whether the ultimate balance of technology favors defense over offense. Even if it does, for a defensive-advantage to remove the need for coordination people might have to be willing to forgo present consumption for benefits far in the future. That is also uncertain.

---

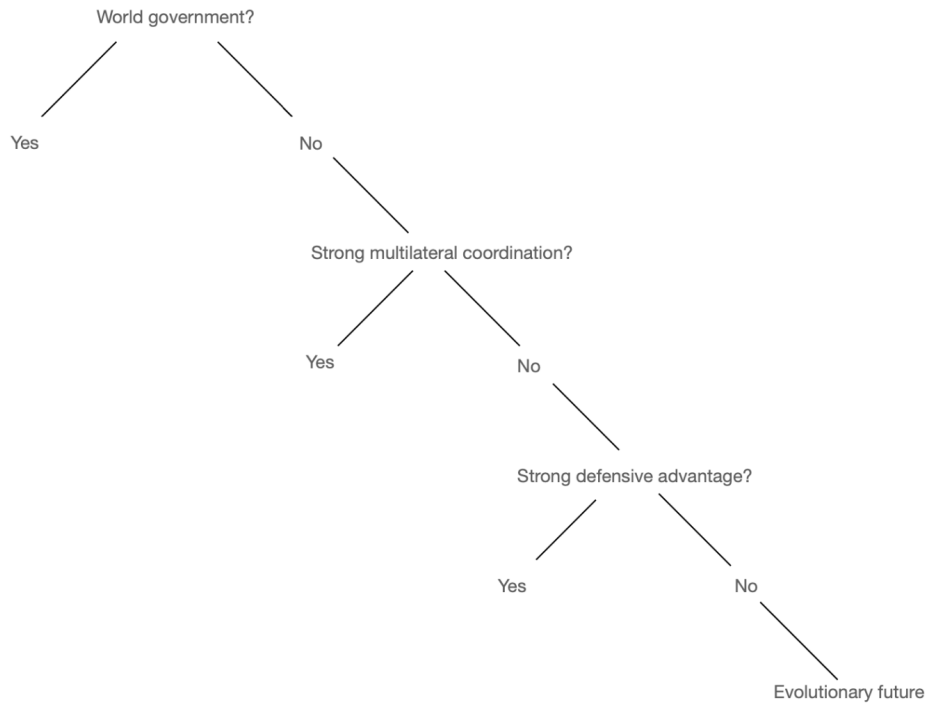
[74] Das & Mukherjee (2021).

[75] The first argument of this kind that I am aware of was made by Shulman (2012) and elaborated by Christiano (2013). For a more skeptical perspective on this idea, see Tomasik (2013).

[76] Ord (2021).

A lower bound on the probability of an evolutionary future

Recall the chart from the introduction:



Thus far the chart has been used in a deductive manner. If there is no world government, and there is no strong multilateral coordination, and no strong defensive advantage, it will not be possible to avoid an evolutionary future. Deductive reasoning can be transformed into probabilistic reasoning by adding subjective probabilities to each step. So, assuming one is willing to assign subjective probabilities to each step, it is possible to calculate a lower bound on one's subjective probability that there will be an evolutionary future by multiplying through one's subjective probabilities at each successive step.

P (evolutionary future) ≥ P (no world government) · P (no strong multilateral coordination | no world government) · P (no strong defensive advantage | no world government or strong multilateral coordination)[77]

[77] If there is a significant chance that humanity will go extinct or uncontrolled AIs will seize power before an evolutionary future can take place, then the overall probability of an evolutionary future should be adjusted downwards. Here is how that can be done:

P (evolutionary future, unconditional) = P (evolutionary future, conditional) · (1-P (extinction))

Normative implications

The existential risk framework and value erosion

Nick Bostrom, in initiating the modern existential risk literature, defined an existential risk as "one where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential."[78] Allan Dafoe coined the term "value erosion" to refer to the potential existential risk posed by an evolutionary future. For an evolutionary future to qualify as an existential risk, it need not reduce the moral value of the future to zero or lower. It merely must cause the future to fall dramatically short of what would ideally have been possible.

It seems that there is an incredible range of possibilities in the future. The best imaginable possibility could easily be many, many times better than the simply good possibilities.[79] There may be some relationship between what is good and what is selectively advantaged; the world today contains many good things while being in large part the result of evolutionary processes. On the other hand, it seems strange to assume that that relationship between selective advantage and goodness is a perfect correlation without further normative argument. Selective fitness and moral desirability are two different properties. If the best possible future is "drastically" better than a merely good future, and the correlation between selective advantageousness and goodness is imperfect, it seems that an evolutionary future could be an existential catastrophe.

How good or bad would an evolutionary future be?

Regardless of whether the possibility of an evolutionary future amounts to the "existential risk" of "value erosion," it is reasonable to want to know how good or bad an evolutionary future would actually be. Futurist writers are divided on this point.[80] The question of the quality of life in an evolutionary future is highly important, and may be a good direction for future work.

Stopping value erosion

If value erosion is potentially an existential risk, it is natural to ask what might be done to stop it. I think the clearest way to think about stopping value erosion is to divide proposed interventions by the step in the conjunctive flow chart at which they propose to intervene. One might try to promote world government, enable strong multilateral coordination, or create a strong future defensive advantage. In the rest of this section, I discuss some important reasons for caution before taking these steps.

---

[78] Bostrom (2003).

[79] Most obviously, the ultimate population size of the affectable universe could be extremely large. (See Bostrom (2014) for a highly speculative estimate illustrating the potential scale of future populations.) Given that the maximum possible future population seems quite high, the range of possible population sizes is also very large–any smaller population size would be possible as well. Given that there is presumably also a large range of possible levels of quality (not just quantity) of future life, the range in possible moral value of different futures would appear to be very large.

[80] See Hanson (2016), and responses from Caplan (2016) and Alexander (2016) for discussion of the quality of life in one specific evolutionary future scenario.

Stopping value erosion may be intractable

One formalization of the tractability of a risk or problem is the difference between the level of risk if no effort is made and the level of risk if the best reasonably achievable effort is made. Thus:

Tractability (value erosion) = P (value erosion) - P (value erosion | an effort being made)

Intuitively, it seems like the tractability of reducing value erosion should be low, relative to other, more acute, risks. If value erosion is more of a concern contingent on the correctness of views that deemphasize the ability of decision makers to shape history, then we should be skeptical of our own ability to prevent value erosion by making good decisions.

Trying to stop value erosion may increase other risks

Intuitive strategies for stopping value erosion carry obvious risks. For example, one natural way to stop value erosion would be to try to increase the probability that a world government is created. However, moving towards a global government increases the risk of global tyranny.[81] Not only would a global totalitarian government affect more people than local totalitarian governments have, it might be much more stable. In "The Totalitarian Threat", Bryan Caplan wrote:

> The worse-case scenario for human freedom would be a global totalitarian state. Without an outside world for comparison, totalitarian elites would have no direct evidence that any better way of life was on the menu. It would no longer be possible to borrow new ideas from the non-totalitarian world, but it would also no longer be necessary. The global government could economically and scientifically stagnate without falling behind. Indeed, stagnation could easily increase stability. The rule of thumb "Avoid all change" is easier to correctly apply than the rule, "Avoid all change that makes the regime less likely to stay in power."[82]

Thus we should think very carefully before we try to bring about a world government.[83] More generally: we should be sure that, if we take action to reduce the probability of value erosion, we do not neglect effects on other risks.[84]

---

[81] Hanson (2021a, 2021c) points out the related issue of "rot." Individual organisms age, and physical objects, legal systems, software systems, and firms, also seem to display a kind of rot. If there is a world government, the lack of competition may degrade its performance.

[82] Caplan (2008), p. 509.

[83] Though there would also of course be long-run benefits from a global government—mitigating the risks described in Bostrom (2019), for example.

[84] Promoting multilateral coordination may be more promising than promoting world government.

Conclusion

Relative to nearly everyone who has ever lived, most people alive today are wealthy, long-lived, unburdened by infectious disease, and literate.[85] This might seem to suggest that predictions of a Malthusian, evolutionary future should be ignored. And many Malthusian doomsayers, such as the biologist Paul Ehrlich, have indeed made egregiously wrong predictions of imminent disaster. It may be reasonable to predict that trends towards a less Malthusian world that have lasted for hundreds of years will not reverse themselves any time soon. But predicting that an evolutionary future is reasonably likely is quite different from anticipating an imminent, acute catastrophe. The example of the rise of agriculture shows that there is no inexorable force that ensures that future developments are in accord with what people, prior to those changes, would have wanted.

This paper has been a combined discussion of hypothetical future technologies, ideas from social science theory, and some very high-level historical trends. As such, nearly all of the specific material in it is very speculative. However, the idea that the future may be determined more by competitive pressures than by choice is not inherently more speculative than the idea that humanity might choose its future. Given my current understanding of the arguments and the burden of proof, I think it would be a serious mistake to rule out the possibility of an evolutionary future.

[85] Fogel (2004); Roser & Ortiz-Espina (2018); Shaw-Taylor (2020).

References

Adams, John. [1787] 1851. *A Defense of the Constitutions of Government of the United States of America*. Edited by Charles Francis Adams. Vol. 4 in *The Works of John Adams*. Boston: Charles C. Little and James Brown.

Aghion, Philippe, Benjamin F. Jones, and Charles I. Jones. 2018. "Artificial Intelligence and Economic Growth". In *The Economics of Artificial Intelligence: An Agenda*, 237–82. Chicago: University of Chicago Press.

Åkesson, Linus. 2016a. "Discussion and Spoilers". https://www.linusakesson.net/programming/underhanded/2015-spoilers.pdf

———. 2016b. "Introduction". https://www.linusakesson.net/programming/underhanded/2015-intro.pdf

Alexander, Scott. 2014. "Meditations On Moloch". *Slate Star Codex* (blog). 30 July. https://slatestarcodex.com/2014/07/30/meditations-on-moloch/.

———. 2016. "Book Review: Age of Em". *Slate Star Codex* (blog). 28 May. https://slatestarcodex.com/2016/05/28/book-review-age-of-em/.

Alighieri, Dante. [1559] 1879. *The "De Monarchia" of Dante*. Translated by F. J. Church. London: Macmillan.

Ansary, Tamim. 2009. *Destiny Disrupted: A History of the World Through Islamic Eyes*. New York: PublicAffairs.

Aschenbrenner, Leopold. 2020. "Existential Risk and Growth". Global Priorities Institute. https://globalprioritiesinstitute.org/leopold-aschenbrenner-existential-risk-and-growth/.

Berelson, Bernard. 1952. "Democratic Theory and Public Opinion". *The Public Opinion Quarterly* 16 (3): 313–30.

Blattman, Christopher. 2022. *Why We Fight: The Roots of War and the Paths to Peace*. London: Penguin.

Bostrom, Nick. 2003. "Astronomical Waste: The Opportunity Cost of Delayed Technological Development". *Utilitas* 15 (3): 308–14.

———. 2004. "The Future of Human Evolution". In *Death and Anti-Death: Two Hundred Years After Kant, Fifty Years After Turing*, edited by Charles Tandy. Palo Alto, California: Ria University Press.

———. 2006. "What Is a Singleton?". *Linguistic and Philosophical Investigations* 5 (2): 48–54.

———. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

———. 2019. "The Vulnerable World Hypothesis". *Global Policy* 10 (4): 455–76.

Burnham, James. 1943. *The Machiavellians: Defenders of Freedom*. New York: The John Day Company.

Caplan, Bryan. 2011. "The Totalitarian Threat". In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan M. Cirkovic. Oxford: Oxford University Press.

———. 2016. "What's Wrong in Robin Hanson's The Age of Em". *Econlog* (blog). 7 June. https://www.econlib.org/archives/2016/06/whats_wrong_in.html.

Christiano, Paul. 2013. "Why Might the Future Be Good?" *Rational Altruist* (blog). 27 February. https://rationalaltruist.com/2013/02/27/why-will-they-be-happy/.

———. 2019. "What Failure Looks Like". *The AI Alignment Forum* (blog). 17 March. https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/what-failure-looks-like.

Clark, Gregory. 2007. *A Farewell to Alms: A Brief Economic History of the World*. Princeton: Princeton University Press.

Coase, R. H. 1960. "The Problem of Social Cost". *The Journal of Law & Economics* 3: 1–44.

Cohen, G. A. 2008. *Rescuing Justice and Equality*. Cambridge, Mass.: Harvard University Press.

Critch, Andrew, and David Krueger. 2020. "AI Research Considerations for Human Existential Safety (ARCHES)". *arXiv*.

Dafoe, Allan. 2015. "On Technological Determinism: A Typology, Scope Conditions, and a Mechanism". *Science, Technology, & Human Values* 40 (6): 1047–76.

Dafoe, Allen. 2020. "AI Governance: Opportunity and Theory of Impact". *Allandafoe.com* (blog). September. https://www.allandafoe.com/opportunity.

Das, Kaustav, and Koel Mukherjee. 2021. "The Journey from Isolation to Interaction During British Raj: Case of Natives in Andaman". In *Tribe-British Relations in India: Revisiting Text, Perspective and Approach*, edited by Maguni Charan Behera, 65–80. Singapore: Springer.

Diamond, Jared. 1999. "The Worst Mistake in the History of the Human Race". *Discover Magazine*, 1 May.

Drexler, Eric. 2018. "Paretotopian Goal Alignment". https://www.effectivealtruism.org/articles/ea-global-2018-paretotopian-goal-alignment.

Eichengreen, Barry, Donghyun Park, and Kwanho Shin. 2011. "When Fast Growing Economies Slow Down: International Evidence and Implications for China". Working Paper. Working Paper Series. National Bureau of Economic Research.

Ellickson, Robert C. 1994. *Order without Law: How Neighbors Settle Disputes*. Cambridge, Mass.: Harvard University Press.

Fearon, James D. 1995. "Rationalist Explanations for War". *International Organization* 49 (3): 379–414.

Feigenbaum, Joan, Michael Schapira, and Scott Shenker. 2007. "Distributed Algorithmic Mechanism Design". In *Algorithmic Game Theory*, edited by Noam Nisan, Tim Roughgarden, Éva Tardos, and Vijay V. Vazirani. Cambridge: Cambridge University Press.

Finnveden, Lukas, C. Jess Riedel, and Carl Shulman. 2022. "Artificial General Intelligence and Lock-In". https://docs.google.com/document/d/1mkLFhxixWdT5peJHq4rfFzq4QbHyfZtANH1nou68q88/

Fogel, Robert William. 2004. *The Escape from Hunger and Premature Death, 1700-2100: Europe, America, and the Third World*. Cambridge: Cambridge University Press.

Friedman, Jeffrey. 2019. *Power without Knowledge: A Critique of Technocracy*. Oxford: Oxford University Press.

Garfinkel, Ben. 2020. "The Case for Privacy Optimism". March 9. https://benmgarfinkel.blog/2020/03/09/privacy-optimism-2/.

———. 2021. "A Tour of Emerging Cryptographic Technologies". Center for the Governance of AI. https://www.governance.ai/research-paper/a-tour-of-emerging-cryptographic-technologies

———. 2022. "The Tragedy of Lumpy Information".

———. 2023. "Growth and Civil War".

Gastelum, Zoe N. 2020. "Societal Verification for Nuclear Nonproliferation and Arms Control". In *Nuclear Non-Proliferation and Arms Control Verification: Innovative Systems Concepts*, edited by Irmgard Niemeyer, Mona Dreicer, and Gotthard Stein, 169–83. Cham, Switzerland: Springer International Publishing.

Gonzalez, Marco, Kristen N. Taddonio, and Nancy J. Sherman. 2015. "The Montreal Protocol: How Today"s Successes Offer a Pathway to the Future". *Journal of Environmental Studies and Sciences* 5 (2): 122–29.

Grace, Katja, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. 2018. "When Will AI Exceed Human Performance? Evidence from AI Experts". *arXiv*.

Hanson, Robin. 1998. “Burning the Cosmic Commons: Evolutionary Strategies for Interstellar Colonization”. https://mason.gmu.edu/~rhanson/filluniv.pdf

———. 2000. “Long-Term Growth As A Sequence of Exponential Modes”. https://mason.gmu.edu/~rhanson/longgrow.pdf

———. 2009. “This Is the Dream Time”. *Overcoming Bias* (blog). 29 September. https://www.overcomingbias.com/p/this-is-the-dream-timehtml.

———. 2016. *The Age of Em: Work, Love, and Life When Robots Rule the Earth*. Oxford: Oxford University Press.

———. 2020. “The World Forager Elite”. *Overcoming Bias* (blog). 22 September. https://www.overcomingbias.com/p/the-world-forager-elitehtml.

———. 2021a. “What Makes Stuff Rot?” *Overcoming Bias* (blog). 22 October. https://www.overcomingbias.com/p/what-makes-stuff-rothtml.

———. 2021b. “The Coming World Ruling Class”. *Overcoming Bias* (blog). 20 November. https://www.overcomingbias.com/p/the-coming-world-ruling-classhtml.

———. 2021c. “Will World Government Rot?” *Overcoming Bias* (blog). 20 November. https://www.overcomingbias.com/p/will-world-government-rothtml.

———. 2021d. “On Evolved Values”. *Overcoming Bias* (blog). 5 December. https://www.overcomingbias.com/p/on-evolved-valueshtml.

Hardin, Garrett. 1968. “The Tragedy of the Commons”. *Science* 162 (3859): 1243–48.

Hovi, Jon, Mads Greaker, Cathrine Hagem, and Bjart Holtsmark. 2012. “A Credible Compliance Enforcement System for the Climate Regime”. *Climate Policy* 12 (6): 741–54.

Iacono, William G., and Gershon Ben-Shakhar. 2019. “Current Status of Forensic Lie Detection with the Comparison Question Technique: An Update of the 2003 National Academy of Sciences Report on Polygraph Testing”. *Law and Human Behavior* 43: 86–98.

Kelsen, Hans. 1949. *The Political Theory of Bolshevism: A Critical Analysis*. Berkeley: University of California Press.

Koppell, Jonathan GS. 2010. *World Rule: Accountability, Legitimacy, and the Design of Global Governance*. Chicago: University of Chicago Press.

Lambert, Patricia M. 2009. “Health versus Fitness: Competing Themes in the Origins and Spread of Agriculture?” *Current Anthropology* 50 (5): 603–8.

Larsen, Clark Spencer. 1995. “Biological Changes in Human Populations with Agriculture”. *Annual Review of Anthropology* 24 (1): 185–213.

Lynn-Jones, Sean M. 1995. “Offense-Defense Theory and Its Critics”. *Security Studies* 4 (4): 660–91.

MacAskill, William. 2022. *What We Owe The Future: A Million-Year View*. London: Oneworld Publications.

Maynard Smith, John, and Eörs Szathmáry. 1995. *The Major Transitions in Evolution*. Oxford: Oxford University Press.

Nordhaus, William D. 2021. “Are We Approaching an Economic Singularity? Information Technology and the Future of Economic Growth”. *American Economic Journal: Macroeconomics* 13 (1): 299–332.

Oesterheld, Caspar. 2017. “Multiverse-Wide Cooperation via Correlated Decision Making”. https://longtermrisk.org/multiverse-wide-cooperation-via-correlated-decision-making/

Olson, Mancur. 1965. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge, Mass.: Harvard University Press.

Ord, Toby. 2020. *The Precipice: Existential Risk and the Future of Humanity*. London: Bloomsbury Publishing.

———. 2021. “The Edges of Our Universe”. *arXiv*.

Ostrom, Elinor. 1990. *Governing the Commons*. Cambridge: Cambridge University Press.

Posner, Richard A. 2004. *Catastrophe: Risk and Response*. Oxford: Oxford University Press.

Pritchett, Lant. 1997. “Divergence, Big Time”. *Journal of Economic Perspectives* 11 (3): 3–17.

Roser, Max, and Esteban Ortiz-Ospina. 2018. “Literacy”. *Our World in Data*. https://ourworldindata.org/literacy

Scott, James C. 1999. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. New Haven: Yale University Press.

Shaw-Taylor, Leigh. 2020. “An Introduction to the History of Infectious Diseases, Epidemics and the Early Phases of the Long-Run Decline in Mortality”. *The Economic History Review* 73 (3): E1–E19.

Shulman, Carl. 2010. “Whole Brain Emulation and the Evolution of Superorganisms”. The Singularity Institute.

———. 2012. “Spreading Happiness to the Stars Seems Little Harder than Just Spreading”. *Reflective Disequilibrium* (blog). 17 September. http://reflectivedisequilibrium.blogspot.com/2012/09/spreading-happiness-to-stars-seems.html.

Shulman, Carl, and Nick Bostrom. 2021. “Sharing the World with Digital Minds”. In *Rethinking Moral Status*, edited by Steve Clarke, Hazem Zohny, and Julian Savulescu. Oxford: Oxford University Press.

Singleton, Sara, and Michael Taylor. 1992. “Common Property, Collective Action and Community”. *Journal of Theoretical Politics* 4 (3): 309–24.

Sokal, Robert R., Neal L. Oden, and Chester Wilson. 1991. “Genetic Evidence for the Spread of Agriculture in Europe by Demic Diffusion”. *Nature* 351 (6322): 143–45.

Stafford, Eoghan, and Robert Trager. 2022. “The IAEA Solution: Knowledge Sharing to Prevent Dangerous Technology Races”. Center for the Governance of AI. https://www.governance.ai/research-paper/knowledge-sharing-to-prevent-dangerous-technology-races.

Tomasik, Brian. 2013. “The Future of Darwinism”. *Essays on Reducing Suffering* (blog). 3 November. https://reducing-suffering.org/the-future-of-darwinism.

Trammell, Philip, and Anton Korinek. 2020. “Economic Growth under Transformative AI”. Global Priorities Institute. https://globalprioritiesinstitute.org/philip-trammell-and-anton-korinek-economic-growth-under-transformative-ai/.

Trask, Andrew, Emma Bluemke, Ben Garfinkel, Claudia Ghezzou Cuervas-Mons, and Allan Dafoe. 2020. “Beyond Privacy Trade-Offs with Structured Transparency”. *arXiv*.

Vidal Bustamante, Constanza, Karolina Alama-Maruta, Carmen Ng, and Daniel Coppersmith. 2022. “Should Machines Be Allowed to “Read Our Minds”? Uses and Regulation of Biometric Techniques That attempt to Infer Mental States”. *MIT Science Policy Review* 3 (August).

Weber, Max. [1919] 2004. “Politics as a Vocation”. In *The Vocation Lectures*, translated by Rodney Livingstone. Indianapolis: Hackett Publishing.

Wright, Robert. 2001. *Nonzero: The Logic of Human Destiny*. Reprint edition. New York: Vintage.

Yudkowsky, Eliezer. 2013. “Intelligence Explosion Microeconomics”. Machine Intelligence Research Institute.

———. 2022. “AGI Ruin: A List of Lethalities - LessWrong”. *Less Wrong* (blog). 5 June. https://www.lesswrong.com/posts/uMQ3cqWDPHhjtiesc/agi-ruin-a-list-of-lethalities.

Yudkowsky, Eliezer, and Nate Soares. 2018. “Functional Decision Theory: A New Theory of Instrumental Rationality”. *arXiv*.