

The Future of Human Evolution

Nick Bostrom (2004)

[Published in: *Death and Anti-Death: Two Hundred Years After Kant, Fifty Years After Turing*, ed. Charles Tandy (Ria University Press: Palo Alto, California, 2004), pp. 339-371.]

ABSTRACT

Evolutionary development is sometimes thought of as exhibiting an inexorable trend towards higher, more complex, and normatively worthwhile forms of life. This paper explores some dystopian scenarios where freewheeling evolutionary developments, while continuing to produce complex and intelligent forms of organization, lead to the gradual *elimination* of all forms of being that we care about. We then discuss how such catastrophic outcomes could be avoided and argue that under certain conditions the only possible remedy would be a globally coordinated policy to control human evolution by modifying the fitness function of future intelligent life forms.

1. The Panglossian view

Can we trust evolutionary development to take our species in broadly desirable directions? Starting from primitive, unconscious life, biological evolution has led to the development of ever more advanced organisms, including creatures that have minds, consciousness, language, and reason. More recently, cultural and technological development, which exhibit some parallels with biological evolution, have enabled our species to progress at a vastly accelerated pace, with enormous improvements occurring in the past few hundred years in human life-span, labor productivity, scientific knowledge, and social and political organization, which enable billions of people to enjoy unprecedented opportunities for enjoyment and personal development. On a historical as well as on a geological timescale, the big picture shows an overarching trend towards increasing levels of complexity, knowledge, consciousness, and coordinated goal-directed organization, a trend which, not to put too fine a point on it, we may label “progress”.¹

What we shall call the *Panglossian view* maintains that this past record of success gives us good grounds for thinking that evolution (whether biological, memetic, or technological) will continue to lead in desirable directions. This Panglossian view, however, can be criticized on at least two grounds. First, because we have no reason to think that all this past progress was in any sense inevitable – much of it may, for all we know, have been due to luck. And second, because even if the past progress were to some

¹ For an argument that both geological and human history manifest such a trend towards greater complexity, see (Wright 1999). For an opposing argument (criticized in chapter 9 of Wright’s book), see (Gould 1990).

extent inevitable, there is no guarantee that the melioristic trend will continue into the indefinite future.

The first objection derives some degree of support from the consideration that an observation selection effect is operating to filter the evidence we can have about the success of our own evolutionary development.² Suppose it was true that on 99.9% of all planets where life emerged, it went extinct before developing to the point where intelligent observers could begin to ponder their origin. If this were the case, what should we expect to observe? Answer: something similar to what we do in fact observe. Clearly, the hypothesis that the odds of intelligent life developing on a given planet are low does not predict that we should find ourselves on a planet where life went extinct at an early stage. Instead, it predicts that we should find ourselves on a planet where intelligent life evolved, even if such planets constitute a very small fraction of all planets where primitive life evolved. The long track record of life's success in our evolutionary past, which one may naively take to support the hypothesis that life's prospects are in general good and that there is something approaching inevitability to the rise of higher organisms from simple replicators, turns out, after reflecting on the overwhelming observation selection effect filtering the possible evidence we could have, not to offer any such support at all, because this is the very same evidence that we should expect to have if the optimistic hypothesis were false. A much more careful examination of the details of our evolutionary history would be needed to circumvent this selection effect. We will not undertake such an examination in the present paper.³

This paper will instead focus on the second objection to the Panglossian view. Even if the rise of intelligent life from simple replicators were a robust and nearly inevitable process, this would not give us strong grounds for thinking that the good trend will continue. One possibility, of course, is that a catastrophic event may cause the sudden extinction of the human species. Some existential risks arise from nature, e.g. impact hazards (meteors and asteroids), pandemics, astrophysical disasters, and supervolcano eruptions. But the greatest existential risks are anthropogenic and arise, more specifically, from present or anticipated future technological developments. Destructive uses of advanced molecular nanotechnology, designer pathogens, future nuclear arms races, high-energy physics experiments, and self-enhancing AI with an ill-conceived goal system are among the worrisome prospects that could cause the human world to end in a bang. Here, however, we shall explore a different set of existential risks in which the world would end more gradually, not with a bang but a whimper.⁴ Let us therefore suppose that no sudden cataclysm puts an end to life. Let us also set aside scenarios in which evolution leads to the erosion of complexity. We shall explore how, even if evolutionary development continues unabated in the direction of greater complexity, things could nevertheless take a wrong turn leading to the disappearance of all the things we value.

This paper will not claim that this is what will happen. The aim, rather, is to undermine our confidence in the Panglossian view and to suggest that a more agnostic stance better reflects the available evidence. We will examine a couple of scenarios in

² For the theory of observation selection effects, see (Bostrom 2002) and references therein.

³ But see e.g. (Carter 1983, 1989; Bostrom 2002; Hanson 1998).

⁴ For more on existential risks, and the classification of these into "bangs", "whimpers", and other categories, see (Bostrom 2002); also (Leslie 1996; Rees 2003; Posner 2004).

which freewheeling evolutionary developments take us in undesirable directions, and we will argue that *if* the future evolutionary fitness landscape is such as to make these evolutionary courses the default (and we have no strong reason either for or against this assumption), *then* the only way we could avoid long-term existential disaster is by taking control of our own evolution. Doing this, it will further be argued, would require the development of a “singleton,” a world order in which at the highest level of organization there is only one independent decision-making power (which may be, but need not be, a world government).

2. Two dystopian “upward” evolutionary scenarios

In this section we will present two scenarios in which human evolution, potentiated by the advanced technology, leads in directions that we should regard as highly undesirable.

Scenario I: The Mindless Outsourcers

Technological progress continues to accelerate and at some point mind the technology of “uploading” becomes possible.⁵ Some human individuals upload and make many copies of themselves. Meanwhile, there is gradual progress in neuroscience and artificial intelligence, and eventually it becomes possible to isolate individual cognitive modules and connect them up to modules from other uploaded minds. Possibly, modules would need to be trained before they can communicate with each other effectively. Modules that conform to a common standard would be better able to communicate and cooperate with other modules and would therefore be economically more productive, creating a pressure for standardization. There might be multiple standards; some modules might specialize in translating between incompatible standards. Competitive uploads begin outsourcing increasing portions of their functionality: “Why do I need to know arithmetic when I can buy time on Arithmetic-Modules Inc. whenever I need to do my accounts? Why do I need to be good with language when I can hire a professional language module to articulate my thoughts? Why do I need to bother with making decisions about my personal life when there are certified executive-modules that can scan my goal structure and manage my assets so as to best fulfill my goals?” Some uploads might prefer to retain most of their functionality and handle tasks themselves that could be more efficiently done by others. They would be like hobbyists who enjoy growing their own vegetables or knitting their own cardigans; but they would be less efficient than some other uploads, and they would consequently be outcompeted over time.

It is possible that optimum efficiency will be attained by grouping abilities in aggregates that are roughly human-equivalent. It might be the case, for example, that a math-module must be tailored to fit the language-module, and that both must be tailored to fit the executive-module, in order for all three to be able to work together effectively.

⁵ In uploading, a detailed map is created of the neural network of a biological brain, perhaps by using nanotech dissemblers to decompose the brain, and an emulation of neuronal processes that previously took place in the brain are implemented on a computer in such a way that memory and personality is preserved. See e.g. (Moravec 1989; Drexler 1985; Bostrom 2003).

Standardization might be almost completely unworkable. But it is hard to see any compelling reason for being confident that this is so. For aught we know, human-type minds may be optimal only given the constraints of human neurology. When it becomes possible to copy modules at will, to send high-bandwidth signals between parts of different brains, and to build architectures that cannot readily be implemented on biological neural nets, it might turn out that the optima relative to this new constraints-landscape have shifted away from the human-like mind region. There might be no niche for mental architectures of a human kind.⁶

There might be ecological niches for complexes that are either less complex (such as individual modules), more complex (such as vast colonies of modules), or of similar complexity as human minds but with radically different architectures. Would these complexes be worthwhile from our current point of view? Do we, upon reflection, really favor a world in which such alien types of complexes have replaced human-type complexes?

The answer may depend on the precise nature of those alien complexes. The present world contains many levels of organization. Some highly complex entities such as multinational corporations and nation states contain human beings as constituents. Yet we usually assign these high-level complexes only instrumental value. Corporations and states are not (it is generally assumed) conscious; they cannot feel phenomenal pain or pleasure. We think they are of value only to the extent that they serve human needs. In cases where they do not contribute to the welfare of any sentient creature, we “kill” them without compunction.⁷ There are also lower levels of organization in today’s world, and the entities inhabiting these levels are not accorded significant moral value either. We do not think it is wrong to erase a piece of computer code. Nor do we think that a neurosurgeon is harming anyone when she extirpates a module (maybe containing an epileptic center) from a human brain if the operation helps the remaining parts of the brain to function better. As for alien forms of complexes of the same complexity as a human brain, most of us would assign them value only if we thought that they had a capacity for conscious experience.

We can thus imagine a technologically highly advanced society, containing many sorts of complex structures, some of which are much smarter and more intricate than anything that exists today, in which there would nevertheless be a complete absence of any type of being whose welfare has moral significance. In a sense, this would be an uninhabited society. All the kinds of being that we care even remotely about would have vanished.

What would make such a world valueless is not the fact that machines would have replaced biological humans. Whether a mind is implemented on biological neurons or on silicon processors seems to make no fundamental moral difference. Rather, the catastrophe would be that such a world would not contain even the right kind of machines, i.e. ones that are conscious and whose welfare matters. There may be an

⁶ Some speculations on the future ecology of intelligent agents can be found in (Chislenko 1996; Moravec 1989; Moravec 1999); cmp. also (Minsky 1988). There is a vast economic literature on contracts, transaction costs, the size of the firm etc. that form a relevant background for thinking about the plausibility of the Outsourcing scenario, but a review of this literature is beyond the scope of this paper; the *locus classicus* is (Coase 1937).

⁷ Some fundamentalist nationalists may believe that a nation state has independent moral status and is entitled to human sacrifice even when no human need would be served. Most of us reject such views.

abundance of economic wealth and technological capability in such a world, yet it would be of no avail because there would be nobody there to benefit from it.

Scenario II: All-Work-And-No-Fun

Even if we do not suppose that uploading and outsourcing will result in a widespread loss of consciousness, we can still entertain the possibility that intrinsically valuable activities and states of consciousness become rarer or disappear altogether. The extravagancies and fun that arguably give human life much of its meaning – humor, love, game-playing, art, sex, dancing, social conversation, philosophy, literature, scientific discovery, food and drink, friendship, parenting, sport – we have preferences and capabilities that make us engage in such activities, and these predispositions were adaptive in our species' evolutionary past; but what ground do we have for being confident that these or similar activities will continue to be adaptive in the future? Perhaps what will maximize fitness in the future will be nothing but non-stop high-intensity drudgery, work of a drab and repetitive nature, aimed at improving the eighth decimal of some economic output measure. Even if the workers selected for in this scenario were conscious, the resulting world would still be radically impoverished in terms of the qualities that give value to life.

To see why these evolutionary scenarios are not quite as improbable as they might appear, we shall consider briefly how we got to where we are today and whether the factors that led us to the evolution of both consciousness and interesting activities will necessarily continue to promote these valuable phenomena, or whether they might instead reflect a transient phase in the history of intelligent life.

3. Ours is an evolutionary disequilibrium

If you wanted to maximize the number of your offspring, your best strategy would probably be to donate as much sperm to sperm banks as you can if you are a male, or to become an egg donor if you are a female. We do not do this because we happen not to have any great desire for reproductive success abstractly conceived. Especially in developed countries, couples often choose to have far fewer children than the maximum they could support, and welfare programs would ensure the survival of any number of children that a couple could not support. Human nature is in an evolutionary disequilibrium; our evolved dispositions are not adapted to the contemporary fitness landscape and do not maximize the inclusive fitness of current individuals.⁸

If technology and social organization were magically frozen in their present state, the human species would likely evolve preferences that more fully reflected the modern fitness function. This could happen by our developing a direct preference for reproductive success (as contrasted to preferences for sex, child rearing, etc.). Alternatively, we might develop a strong instinctual aversion against the use of birth control. It is also possible that cultural evolution would act faster than biological

⁸ There is recent evidence of genetic influence on human fertility outcomes beyond what can be attributed to equilibrium mutation rates; see (Kirk 2001; Rogers and al. 2003).

evolution, producing a dominant meme set favoring plentiful offspring and opposing all forms of birth control.⁹

Population growth is limited not only by our relative lack of interest in having children but also by the biology of human reproduction and maturation. Couples can only produce about one child per year and it takes a newborn about a decade and a half to reach sexual maturity. Even these physiological inhibitors of population growth may be reduced. While biological evolution could probably reduce the duration of human pregnancy and time to puberty to some extent, much more radical effects could result from technological developments. Uploads (and artificial intelligences) could reproduce virtually instantaneously. Moreover, since their “offspring” (copies) would be identical to the original, there would be no maturational latency. Population growth in an upload population could be rapidly exponential, with a potential doubling time of days or less.

The current unprecedented level of average global income is the result of the world economy having for a long time been growing faster than world population. Average income can only rise if economic growth exceeds population growth. If and when the motivational and physiological impediments to population growth are reduced, population growth rate would start accelerating, potentially reaching “near infinity” when uploading becomes possible. Economic growth would be unable to keep up with population growth in a population of freely reproducing uploads. If a welfare program sought to guarantee a minimum income for uploads while permitting unlimited reproduction, it would very quickly go bankrupt even given stellar economic growth rates.¹⁰ If social limitations were not imposed on reproduction, resource constraints would kick in and drive average income down to a level that made frequent reproduction impossible.¹¹

These reflections warn us against naively importing intuitions about the current state of affairs into our thinking about the future. Malthusian pessimism might *appear* to have been refuted by the history of the last two hundred years.¹² Living conditions have been improving even in the absence of population control, contrary to Malthus’s notorious prediction. This anomaly, however, is explained by the factors mentioned above: our preferences not being in an evolutionary equilibrium and the slowness of

⁹ There is already some data suggesting memetic effects on fitness. The expansion of the Hutterites, an Anabaptist sect, is attributable to their extremely high fertility rate. An average Hutterite woman gives birth to nine children. The Hutterites are opposed to any kind of birth control and see high fertility as a sign of divine blessing. Roman Catholic women have about 20% higher reproductive fitness than women of other religions (Kirk 2001). By contrast, supporters of VHEMT (The Voluntary Human Extinction Movement) have foresworn having children altogether, which would give them zero fitness (Knight 2001). University educated women have 35% lower fitness than those with less than seven years of education (Kirk 2001), and non-religious women also have lower than average fitness.

¹⁰ Even if we could colonize the universe in all directions at light speed, this would only increase the resources under human control polynomially (at a rate of $\sim t^2$) whereas unconstrained population growth can easily be exponential ($\sim e^t$).

¹¹ For an analysis of the economics of uploading, see (Hanson 1994).

¹² Thomas Robert Malthus (1766-1834), political economist and demographer, argued that the standard of living for the working class could not be raised without population control because increased income would eventually lead to workers having more surviving children, which would drive wages back down again. But Malthus was not as thoroughly pessimistic as is commonly thought. In the second, rarely read, edition of his essay on population he writes: “Though our future prospects respecting the mitigation of the evils arising from the principle of population may not be as bright as we could wish, yet they are far from being entirely disheartening.” (Malthus 1803).

human reproduction. If further evolutionary or technological developments were to remove these inhibiting factors, population growth rates could easily come to exceed economic growth, ushering in a Malthusian era where average income hovers close to subsistence level. With unconstrained upload reproduction, this transition could happen almost instantaneously.

4. Costly signaling and flamboyant display?

We do many things which are not in a narrow sense instrumentally useful: we dance, joke, write poetry, throw parties, go on vacations, dress up in expensive fashionable clothing, watch and participate in sports, and so on. That we are currently in an evolutionary disequilibrium does not account for this, because even our Pleistocene ancestors exhibited many of these “useless” behaviors. Indeed, many animal species engage in analogous activities, especially in the contexts of courtship displays and status contests.

[Flamboyant displays] appear in a variety of contexts, ranging from sexual selection contests in the animal world, to prestige contests among members of contemporary nation states that span continents with huge road-systems, and even land people on the Moon.¹³

An evolutionary explanation for the existence of such behaviors is that they function as hard-to-fake signals of important qualities that are difficult to observe directly, such as bodily or mental fitness, social status, quality of allies, ability and willingness to prevail in a fight, or possession of resources. Not only behavior or but morphology, too, can serve as a signal – the peacock’s tail being a paradigm example. An extravagant tail is a handicap that only fit peacocks can afford, and peahens have evolved to be sexually attracted by such tails because they are indicators of genetic fitness.¹⁴

Given that flamboyant display is so common among both humans and other species, one might consider whether it would not also be part of the repertoire of technologically more mature life forms. We might hope that even in if there were to be no narrowly instrumental use for play or even for conscious minds in the future ecology of intelligent life, these features might nonetheless confer evolutionary advantages to their possessors by virtue of being reliable signals of other adaptive qualities. Yet while this possibility is hard to rule out, there are several reasons for skepticism that undermine the confidence prescribed by the Panglossian view.

First, many of the flamboyant displays we find in nature are related to sexual selection.¹⁵ Yet reproduction among technologically mature life forms may well be asexual. In particular, this is so in the uploading scenario described above.

Second, new methods of reliably communicating information about oneself might be available to technologically mature creatures, methods that do not rely on flamboyant display. Even today, professional lenders tend to rely more on ownership certificates, bank statements, and the like, than on costly displays such as designer suits and Rolex

¹³ (Kansa 2003).

¹⁴ See e.g. (Zehavi et al. 1999).

¹⁵ See (Miller 2000).

watches. In the future, it might be possible to employ auditing firms that can verify through direct inspection that a client possesses a claimed attribute. Signaling one's qualities by such auditing may be much more efficient than signaling via flamboyant display. Such a professionally mediated signal would still be costly to fake (this is of course the essential feature that makes the signal reliable), but the signal could be much cheaper to transmit than a flamboyantly communicated one when it is *truthful*.

Third, not all possible costly or "flamboyant" displays are ones which we should regard as intrinsically valuable:

[Costly] signaling does not only take the form provisioning public goods that enhance group benefits. Many costly signals take the form of "waste" where expenditures do not confer any group benefit... In the American Northwest, the Kwakiutl potlatch ceremonies involved the public destruction of vast amounts of accumulated wealth in the context of chiefly competition.¹⁶

While humor, music, and poetry can plausibly be said to enhance the intrinsic quality of human life – aside from any social benefits – it is more dubious that the same claim can be sustained with regard to the costly pursuit of fashion trends or with regard to macho posturing leading to gang violence or military bravado. Even if future intelligent life forms would use costly signaling, it is thus an open question whether the signal would be intrinsically valuable: whether it would be like the nightingale's rapturous song or more like the toad's monosyllabic croak.

5. Two senses of *outcompeted*

We should distinguish two different senses in which a species can be outcompeted by other life forms. In the scenarios presented in section 2, the outsourcing uploads and the all-work-no-play agents were postulated to outcompete the agents that retained consciousness and hobbyist interests (we shall term the latter *eudaemonic agents*). One thing this could mean is that the former types of agent gradually obtain the resources originally held the eudaemonic agents, so that the latter eventually run out of resources and become extinct. This is the typical evolutionary outcome when one type of organism outcompetes another.

But we could also conceive of the case where the eudaemonic agents continue to exist indefinitely, perhaps in undiminished numbers, and are outcompeted only in the sense of comprising a steadily declining *fraction* of the total population of agents and of controlling an ever-decreasing fraction of the world's total wealth. It is questionable whether the eudaemonic agents could in the long run prevent their fitness-maximizing competitors from engulfing them and expropriating their property. But even if the eudaemonic agents could do this, and escape extinction, the outcome would still be a disaster because it would entail a tremendous loss of opportunity. Like a ravaging fire, the fitness-maximizers would gobble up resources that would otherwise have been used for more meaningful purposes by the eudaemonic agents.¹⁷

¹⁶ (Kansa 2003). For a provocative take, see also (Frank 2000).

¹⁷ Strict enforcement of property rights might limit the destruction to resources not originally owned by the eudaemonic agents. The latter might even manage to salvage some additional resources by colonizing new

6. Could we control our own evolution?

Suppose we could foresee that one of the dystopian evolutionary scenarios described above represents the default course of development of our species. What would then be our options?

One response would be to sit back and let things slide. We could bolster our passivity by contemplating the greater evolutionary fitness of the non-eudaemonic agents: being more fit, are they not more worthy possessors of the world's resources? While few would endorse this argument in its explicit form, it is quite possible that certain related thoughts – perhaps deference to the “natural order” or acceptance of the idea that might makes right – may have found a hiding place in the dark corners of some minds. We can expel such notions by reminding ourselves that if a doomsday plague emerged and killed all mammals, this would not imply that the pathogens' “victory” was a good thing, even though it would mean that the plague had proven fitter.

Another response would be to lament the dystopian outcome but conclude that nothing could be done to prevent it. If outsourcing or constant toil has a higher fitness-value than eudaemonic living, does not evolution theory then entail that the eudaemonic agents will disappear? Yet as we shall see, the future need not be hopeless even if the default course of evolution is dystopian. Evolution made us what we are, but no fundamental principle stands in the way of our developing the capability to intervene in the default course of events in order to steer future evolution towards a destiny more congenial to human values.

Directing out own evolution, however, requires coordination. If the default evolutionary course is dystopian, it would take coordinated paddling to turn the ship of humanity in a more favorable direction. If only some individuals chose the eudaemonic alternative while others pursued fitness-maximization, then, by assumption, it would be the latter that would prevail. Fitness-maximizing variants, even if they started out as a minority, would be preferentially selected for at the expense of eudaemonic agents, and a process would be set in motion that would inexorably lead to the minimization or disappearance of eudaemonic qualities, and the non-eudaemonic agents would left to run the show.

To this problem there are only two possible solutions: preventing non-eudaemonic variants from arising in the first place, or modifying the fitness function so that eudaemonic traits become fitness-maximizing. Let us examine these two options in turn.

7. Preventing non-eudaemonic agents from arising

It is quite plausible to suppose that technologically advanced life forms would be able to prevent unwanted mutations from occurring, at least if we understand “mutation” in a narrow sense. Cryptographic methods and error correcting codes could reduce transmission errors and copying errors to negligible levels. The control systems of nanotechnological replicators could be encrypted in such a way as to make them

territories, although eventually the opportunities for such acquisition would disappear as the fitness maximizers (being *ex hypothesi* more efficient) would have got there first. The dynamics of such a colonization race is analyzed in (Hanson 1998).

evolution-proof (any random change would be virtually certain to completely destroy the replicator).¹⁸ For uploads, avoiding reproductive mutation may simply be a matter of performing multiple verifications that the copy is identical to the original before it is run. Even for biological creatures unaided by nanotechnology, sufficiently advanced gene technology should make it possible to scan all embryos for unwanted mutations, and ordinary genetic recombination could be avoided with the use of reproductive cloning.

Source code mutation and genetic recombination are not, however, the only ways in which new variants with unanticipated properties can arise. Consider again the uploading scenario where uploads outsource much of their functionality and share mental modules. Recombining different modules could result in unexpected emergent phenomena. Likewise, the enhancement of various cognitive or emotional capacities, or the installment of entirely new capacities, could produce combinatorial effects that may not be fully predictable. Ordinary growth and maturation of an individual could lead to the development of a fitness-maximizing non-eudaemonic character even where none is manifest at conception. Novel memetic influences might also trigger non-eudaemonic tendencies. So while it is plausible that an advanced life form could avoid random mutations in its source code, it is more dubious that it would be able to predict and avoid emergent effects of growth, enhancement, and learning in individuals or in interacting communities of developing agents. Preventing fitness-maximizing non-eudaemonic variants from occasionally arising may be infeasible or may require creating a completely static society where individual experimentation and enhancement are banned and where no social reorganization is permitted. Such a fossilized world seems intrinsically undesirable.¹⁹

Even if dangerous mutants could be prevented from arising, it would be to no avail if the original population already contained some individuals with non-eudaemonic fitness-maximizing preferences, because these would then proliferate and eventually dominate. And we can surely assume that at least some current human individuals would upload if they could, would make many copies of themselves if they were uploads, and would happily embrace outsourcing or forego leisure if they could thereby increase their fitness. These individuals would somehow have to disappear from the population, and it is hard to see any practical and ethical way in which this could happen.

Forestalling the dystopian evolutionary scenarios by preventing non-eudaemonic agents from arising is therefore a non-starter. At most, this measure could serve an auxiliary role. In particular cases, it might make good sense to try to reduce the frequency with which dangerous mutants are spawned – in cases where this can be done relatively inexpensively, in an ethically acceptable way, and where clear and specific harms can be foreseen. For example, we might in the future pass laws against building powerful artificial intelligences with goal systems that are hostile to human values. But we cannot rely on this strategy alone to prevent a dystopian evolutionary scenario if such a scenario should happen to be the default.

¹⁸ See (Drexler 1985).

¹⁹ Cmp. (Huxley 1932).

8. Modifying the fitness function

The second option is to modify the fitness function so that eudaemonic agents continue to have at least some niche in which they are fitness-maximizing.

The differential reproductive success of human gene- or meme-types is determined not only by the natural habitat in which they live in but also by their social environment. Social institutions, laws, and other people's attitudes define the choices open to us as individuals and determine the effects that these choices have on our inclusive fitness. Social structures could be arranged in a manner that reduce the fitness of non-eudaemonic types and enhance the fitness of eudaemonic types. These structures, if stable enough, could constrain evolutionary development to a set of trajectories on which the eudaemonic type flourishes.

It would be misleading to characterize such intervention as "helping the weak and unfit". The way society is set up partially defines which types are fit and "strong" (in the sense of being able to use available means to proliferate). If we want to avoid evolutionary trajectories that lead to a region of state space where the qualities we value are extinct or marginalized, then social sculpting of the conditions for reproductive success might our only recourse. The term "reproductive success" here covers not only biological sexual reproduction but also upload duplication and in general the spread of forms of organizations. Social shaping of the conditions for reproductive success is, of course, a fact of life for every organism that lives in societies; but the suggestion here is that it might become necessary to deliberately adjust these conditions so that they favor eudaemonic types.

We clearly do not need to remold *all* niches so that they favor eudaemonic agents. Doing that would be tantamount to working towards the elimination of all non-eudaemonic agents. But many non-eudaemonic agents are harmless; many are indeed highly useful to the eudaemonic agents. Just as current human beings benefit from other species, which pose no serious threat to the human species, so too may technologically more advanced agents benefit from the existence of an ecology of non-eudaemonic agents. Such non-eudaemonic agents could serve economically useful functions. Only those non-eudaemonic agents that threaten to invade the niche occupied by the eudaemonic agents pose an evolutionary hazard; or more precisely (since the eudaemonic agents could potentially move from one niche to another), the concern pertains to such non-eudaemonic agents as would reduce the total size of the niches available to eudaemonic types. The goal is to maximize the total quantity of resources possessed by eudaemonic agents, or at any rate to prevent this quantity from falling to zero. Non-eudaemonic agents, even if they reduce the *fraction* of resources possessed by eudaemonic agents, could on balance be beneficial if they also increased the total amount of resources available. Non-sentient property-owning robots, for example, could theoretically have that effect.

9. Policies for evolutionary steering

What kind of intervention would be required to shape social conditions so that they favor eudaemonic types? One simple but crude method would be to ban outsourcing (and other kinds of non-eudaemonic phenotypes). Such a ban, however, would be costly since we are assuming that many productive tasks are most efficiently accomplished through

outsourcing. A more efficient approach would be to tax outsourcing and subsidize eudaemonic cognitive architectures. That way, outsourcing would be used for the tasks where it brings the greatest returns and the level of taxation could be set at a level that ensures that the eudaemonic type continues to thrive.

One might think that there would be no economic rationale for subsidizing the eudaemonic type, because to the extent that we value this type we would be willing to spend their own money on ensuring its existence, perhaps by devoting ourselves to eudaemonic activities. We do not, in today's world, see any special need to encourage present consumption by taxing investment. Might not individuals acting in a free market allocate their resources optimally between eudaemonic "consumption" and non-eudaemonic "investment" (in enterprises that may produce resources that could later be eudaemonically consumed)? There are at least three potential reasons for being skeptical that the invisible hand would by itself orchestrate a workable, much less optimal, solution to the problem non-eudaemonic competition.

First, property rights may not always be perfectly enforced. If non-eudaemonic types emerge and become immensely powerful, they might rob the eudaemonic agents. Such robbery could be blatant, as in a big coup or in a series of minor assaults by non-eudaemonic groups or individuals; or it could take a more subtle form, as in influencing governments to enact legislation unfairly favoring non-eudaemonic interests. Eudaemonic agents could also rob non-eudaemonic agents, but because they would be less productive one would expect a net flow of resources to non-eudaemonic agents. Even if existing property rights were enforced, there would remain the vast opportunity cost of having the non-eudaemonic agents colonizing the cosmic commons for which property rights have not yet been assigned.²⁰

Second, the values of the initial eudaemonic population may not be forever preserved. Such preservation would require either that the initial eudaemonic agents do not die and that their eudaemonic preference do not change over the eons, or else that they choose to reproduce almost exclusively in ways that transmit their eudaemonic preferences to their progeny in undiluted form. If there were occasional crossovers where eudaemonic agents develop non-eudaemonic preference, or have offspring with such preferences, then this would need to be counterbalanced by an at least equally great flux in the opposite direction. We have already discussed the problem of preventing such crossovers from occurring.²¹

Third, if the existence of a thriving hobbyist population is a public good (i.e. non-rivalrous and non-excludable), it will be undersupplied by the market. You and I and a

²⁰ See (Hanson 1998).

²¹ We noted that error-correcting codes should make reproduction arbitrarily reliable. Uploads do not suffer biological aging, and the habit of maintaining backup copies at dispersed locations should reduce the risk of accidental death to a very low level. As a potential way of reducing the risk that basic preferences could drift in unexpected ways when new capacities are installed or learned, it would be interesting to consider the use of safety pacts. The idea is that you empower some agents that you trust to reverse your self-modifications if they believe that the modifications have changed you in a way that you would not have approved of prior to the change. If you are an upload, you might also be able to create a copy of yourself, and then appoint this copy as a trusted overseer of the transformation process with the authority to reverse the changes you have made within some interval if it judges that the modifications have been for the worse. It could also be worth exploring how such hypothetical techniques relate to dispositional theories of value (Lewis 1989; Johnston 1989), to ideal observer theory, and more generally to the ethics of human enhancement (Glover 1984; Bostrom 2004).

million other people might all desire that there be eudaemonic agents in the world a long time from now. Each of us realizes that our individual actions will have but a negligible effect on the outcome. We therefore each spend our resources on other goals, and the result is that the eudaemonic type disappears. This is compatible with our agreeing that our interests would have been better served if some fraction of our resources had been set aside in a eudaemonics conservation fund. The provision of the global public good – in this case the continued existence of eudaemonic types into the indefinite future – has been thwarted by the free-rider problem.

For these three reasons, there is cause to doubt that a laissez-fair approach would give adequate protection to the eudaemonic type. But before we move on, let us consider again the first of these objections to the laissez-fair approach in a little more detail.

10. Detour

The concern was that it might be infeasible to ensure that property rights are perfectly enforced. Agents may occasionally steal each other's resources; or in a more extreme scenario, there may be a general war for resources. However, there seems no reason for thinking that such feuds would be a neatly bipolar contest between two grand coalitions, the eudaemonic and the non-eudaemonic agents fighting on opposing sides. Instead, there might be various and shifting alliances between individual eudaemonic and non-eudaemonic agents. In such a conflict, there would be a net social loss in terms of expenditure on protection and security services, armaments, and probably also in terms of casualties and collateral damage. There would almost certainly be an additional opportunity cost in terms of wasted opportunities for collaboration. But assuming no doomsday weapons is deployed that causes the extinction of the entire population of intelligent life forms (which is not an innocuous supposition) we may still wonder whether such a conflict would necessarily lead to the eventual extinction of the eudaemonic type.

Since what we have here is a case of two kinds of organism being locked in a struggle for the same resource niche, where one kind is more efficient than the other, one might expect, on ecological grounds, that the less efficient kind will eventually die out. (Some of the eudaemonic agents may survive but at price of losing their eudaemonic inclinations or being forced, in order to remain competitive, to forever keep those inclinations suppressed.) But it is, in fact, not at all clear that this is what would happen. The situation differs from the case of competition between more primitive life forms such as plants and animals. The agents involved in this struggle can form strategic alliances. Moreover, by contrast to current human political competition, where alliances shift over time, it might be possible for more advanced life forms to verifiably commit themselves *permanently* to a particular alliance (perhaps using mind-scanning techniques and technologies for controlling motivation). If such permanent commitments are possible, then when war breaks out, various budding coalitions may bargain for the allegiance of unaffiliated individuals. At this stage, some eudaemonic agents may strike favorable deals with what turns out to be the winning coalition. After victory, the eudemonic agents in the winning coalition could not, by assumption, shift their allegiances, so the surviving eudaemonic agents may then have their security guaranteed *in perpetuo*.

11. Only a singleton could control evolution

Even if the survivors of such a contest would in the end get to enjoy perpetual peace, they would have got there by a costly and risky detour. We could reach the same destination more directly, avoiding the wastage and attrition and some of the risk of conflict, by creating a singleton. The stable alliance that we speculated might form at the end of the conflict would in effect be a global regime that could enforce basic laws for its members; in other words, it would be a kind of singleton. In order to be assured of stability, it would not only have to lack external competitors but its domestic affairs would have to be regulated in such a fashion that no internal challenges to its constitution could arise. This would require that the alliance implements a coordinated policy to prevent internal developments from ushering it onto an evolutionary trajectory that ends up toppling its constitutional agreement, and doing this would presumably involve modifying the fitness function for its internal ecology of agents, e.g. by means of a separate tax codes for eudaemonic and non-eudaemonic agents combined, perhaps, with an outright ban on the creation of agents of the most dangerous types.

Reining in evolution is a feat that could only be accomplished by a singleton. A local power might be able to control the evolution of its own internal ecology, yet unless these interventions served to maximize its total productivity (which would be incompatible with affirmative action for eudaemonic activities), evolutionary selection would simply reemerge at a higher level. Those powers that opted to maximize their economic rather than their eudaemonic productivity would outperform rival powers that were less single-mindedly fixated on advancing their competitive situation, and in the long run the eudaemonic powers would become marginalized or extinct or would be forced to rescind their eudaemonic policies. In this context, the “long run” may actually be quite short, especially in the uploading scenario where reproduction could be extremely rapid. Moreover, if the eudaemonic powers could anticipate that they would be outcompeted if they continued with their eudaemonic activities, they may decide to scale back on such activities even before they were overrun by the non-eudaemonic powers. Such anticipatory effects could produce immediate manifestations of evolutionary developments that would otherwise take a long time to unfold. The upshot, in either case, would be a tremendous loss of eudaemonic potential.

A singleton could prevent this unfortunate outcome by promoting eudaemonic types and activities within its own jurisdiction.²² Since a singleton would lack external competitors, there would be no higher level at which evolutionary selection could gain foothold and start penalizing the singleton’s policy of non-maximization of economic productivity.

²² One may wonder how a space-colonizing singleton could enforce its laws over cosmic distances and do so without disintegrating even over time-scales of millions of years. Yet this seems to be at root a technical problem. One solution would be to ensure that the goal-system of all colonizers it sends out include a fundamental desire to obey the basic laws of the singleton. And one of these laws may be to make sure that any progeny produced must also share this fundamental desire. Moreover, the basic law could stipulate that as technology improves, the safety-standards for reproduction (the degree of verification required to ensure that progeny or colonization probes share the fundamental desire to obey the basic constitution) improve correspondingly, so that the probability of defection asymptotically approaches zero. While this proposal may seem technologically daunting, we must bear in mind that any galactic empire would be technologically extremely advanced. A singleton could postpone wide-ranging space colonization until it had developed the control technology necessary to ensure its own long-term stability.

A singleton need not be a monolith (except in the trivial sense that has some kind of mechanism or decision procedure that enables it to solve internal coordination problems). There are many possible singleton constitutions: a singleton could be a democratic world government, a benevolent and overwhelmingly powerful superintelligent machine, a world dictatorship, a stable alliance of leading powers, or even something as abstract as a generally diffused moral code that included provisions for ensuring its own stability and enforcement.²³ A singleton could be rather minimalist structure that need not get in the way much of the lives of its inhabitants. And it need not prohibit novelty and experimentation, since it would retain the capacity to intervene at a later stage to protect its constitution if some developments turned malignant.

Increased social transparency, such as may result from advances in surveillance technology or lie detection, could facilitate the development of a singleton.²⁴ Deliberate international political initiatives could also lead to the gradual emergence of a singleton, and such initiatives might be dramatically catalyzed by wild card events such as a series of cataclysms that highlighted the disadvantages of a fractured world order. It would be a mistake to judge the plausibility of the ultimate development of a singleton on the basis of ephemeral trends in current international affairs. The basic conditions shaping political realities may change as new technologies come online, and it is worth noting that the long-term historical trend is towards increasing scope of human coordination and political integration.²⁵ If this trend continues, the logical culmination is a singleton.

12. Conclusion

Contrary to the Panglossian view, current evidence does not warrant any great confidence in the belief that the default course of future human evolution points in a desirable direction. In particular, we have examined a couple of dystopian scenarios in which evolutionary competition leads to the extinction of the life forms we regard as valuable. Intrinsically worthwhile experience could turn out not to be adaptive in the future.

The only way to avoid these outcomes, if they do indeed represent the default trajectory, is to assume control over evolution. We argued that this would require the creation of a singleton. The singleton would lack external competitors and would have a sufficiently integrated decision mechanism that it could solve internal coordination problems, in particular the problem of how to reshape the fitness function for its internal agent ecology to favor eudaemonic types. A mere local power could also attempt to do this but it would thereby decrease its competitiveness and ensure its own eventual demise. Long-term control of evolution requires global coordination.

A singleton could take a variety of forms and need not resemble a monolithic culture or a hive mind. Within the singleton there could be room for a wide range of different life forms, including ones that focus on non-eudaemonic goals. The singleton could ensure the survival and flourishing of the eudaemonic types by restricting the ownership rights of non-eudaemonic entities, by subsidizing eudaemonic activities, by guaranteeing the enforcement of property rights, by prohibiting the creation of agents with human-unfriendly values or psychopathic tendencies, or in a number of other ways.

²³ For a scenario of this kind, see (Miller 2000).

²⁴ See (Brin 1998).

²⁵ For a persuasive case for this claim, see (Wright 1999).

Such a singleton could guide evolutionary developments and prevent our cosmic commons from going to waste in a first-come-first-served colonization race.

The reflections offered in this paper are not meant to be the final word on the matter. We do not know that a dystopian scenario is the default evolutionary outcome. Even if it is, and even if the creation of a singleton is the only way to forestall ultimate catastrophe, it is a separate question what policies it makes sense to promote in the here and now. While creating a singleton would help to reduce certain risks, it may at the same time increase others, such as the risk that an oppressive regime could become global and permanent. If our preliminary study serves to draw attention to some possibly non-obvious considerations and to stimulate more rigorous analytic work, its purpose will have been achieved.²⁶

References

- Bostrom, N. (2002), *Anthropic Bias: Observation Selection Effects in Science and Philosophy*. New York: Routledge.
- (2002), "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards", *Journal of Evolution and Technology* 9.
- *The Transhumanist FAQ: v 2.1*. World Transhumanist Association 2003. <http://transhumanism.org/index.php/WTA/faq/>.
- (2004), "Transhumanist Values", in Fredrick Adams (ed.), *Ethical Issues for the 21st Century: Philosophical Documentation Center Press*.
- Brin, D. (1998), *The Transparent Society*. Reading, MA: Addison-Wesley.
- Carter, B. (1983), "The Anthropic Principle and its Implications for Biological Evolution", *Philosophical Transactions of the Royal Society A* 310:347-363.
- (1989), "The Anthropic Selection Principle and the Ultra-Darwinian Synthesis", in F. Bertola and U. Curi (eds.), *The Anthropic Principle*, Cambridge: Cambridge University Press, 33-63.
- Chislenko, A. *Networking in the Mind Age* 1996. <http://www.ethologic.com/sasha/mindage.html>.
- Coase, R. H. (1937), "The Nature of the Firm", *Economica* 4 (16):386-405.
- Drexler, K. E. (1985), *Engines of Creation: The Coming Era of Nanotechnology*. London: Forth Estate.
- Frank, R. H. (2000), *Luxury Fever: Money and Happiness in an Era of Excess*. Princeton: Princeton University Press.
- Glover, J. (1984), *What Sort of People Should There Be?:* Pelican.
- Gould, S. J. (1990), *Wonderful Life: The Burgess Shale and the Nature of History*: W. W. Norton & Company.
- Hanson, R. (1994), "What If Uploads Come First: The Crack of a Future Dawn", *Entropy* 6 (2).
- *Burning the Cosmic Commons: Evolutionary Strategies for Interstellar Colonization* 1998. <http://hanson.gmu.edu/filluniv.pdf>.
- *Must Early Life be Easy? The Rhythm of Major Evolutionary Transitions* 1998. <http://hanson.berkeley.edu/>.

²⁶ An early version of this paper was available 12 May 2001. I'm grateful to Wei Dai and Robin Hanson for comments.

- Huxley, A. (1932), *Brave New World*. London: Chatto & Windus.
- Johnston, M. (1989), "Dispositional Theories of Value", *Proceedings of the Aristotelian Society, supp.* 63:139-174.
- Kansa, E. (2003), "Social Complexity and Flamboyant Display in Competition: More Thoughts on the Fermi Paradox", *working paper*.
- Kirk, K. M. (2001), "Natural Selection and Quantitative Genetics of Life-History Traits in Western Women: A Twin Study", *Evolution* 55 (2):432-435.
- Knight, L. U. *The Voluntary Human Extinction Movement* 2001. <http://www.vhemt.org/>.
- Leslie, J. (1996), *The End of the World: The Science and Ethics of Human Extinction*. London: Routledge.
- Lewis, D. (1989), "Dispositional Theories of Value", *Proceedings of the Aristotelian Society, supp.* 63:113-137.
- Malthus, T. R. (1803), *An Essay on the Principle of Population*. 2nd ed. London: J. Johnson.
- Miller, G. (2000), *The Mating Mind: How Sexual Choice Shaped the Evolution of Human Nature*. New York: Doubleday.
- *Moral Vision* 2000. http://www.unm.edu/~psych/faculty/moral_vision.htm.
- Minsky, M. (1988), *Society of Mind*. New York: Simon & Schuster.
- Moravec, H. (1989), *Mind Children*. Harvard: Harvard University Press.
- (1999), *Robot: Mere Machine to Transcendent Mind*. New York: Oxford University Press.
- Posner, R. (2004), *Catastrophe*. Oxford: Oxford University Press.
- Rees, M. (2003), *Our Final Hour: A Scientist's Warning: How Terror, Error, and Environmental Disaster Threaten Humankind's Future in This Century - On Earth and Beyond*: Basic Books.
- Rogers, J. L., and e. al. (2003), "Genetic Influence Helps Explain Variation in Human Fertility: Evidence From Recent Behavioural and Molecular Genetic Studies", *Current Directions in Psychological Science* 10 (5):184-188.
- Wright, R. (1999), *Nonzero: The Logic of Human Destiny*. New York: Pantheon Books.
- Zehavi, A., et al. (1999), *The Handicap Principle: A Missing Piece of Darwin's Puzzle*. Oxford: Oxford University Press.