The Future of Human Evolution

Nick Bostrom
Department of Philosophy, Yale University
http://www.nickbostrom.com

May 12, 2001
NOTE: This is an early draft of a work-in-progress

Consider the following scenario: Technological progress continues and even accelerates. At some point in the 21st century, uploading becomes possible. A number of individuals upload and make many copies of themselves. Meanwhile, there is gradual progress in neuroscience and artificial intelligence; at some point, it becomes possible to begin to isolate individual cognitive modules and connect them up to modules from other uploaded minds. Possibly, modules need to be trained for some period of time before they can cooperate effectively. This would lead to a pressure for standardization: modules that conform to a common standard would be better able to cooperate with other modules, and would therefore have a higher economic value. There might be multiple standards, and there might be some modules that specialize in translating between incompatible standards.

What happens now is that competitive uploads begin to outsource an increasing fraction of their functionality: *"Why do I need to know arithmetic when I can buy time on Arithmetic-Modules Inc. whenever I need to perform arithmetic tasks? 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 then manage my assets so as to best fulfill those goals?"* Some uploads might like 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 knit their own cardigans. But they would be less efficient than some other uploads, and they would consequently be outcompeted eventually.

It is possible that optimum efficiency will be attained by grouping abilities in aggregates that are roughly human-equivalent. It could be, for example, that a math-module must be tailor-made to fit the language-module, and that they both must be tailor-made to fit the executive-module, in order for all three to be able to work together effectively. Standardization might be almost completely unworkable. But it is hard to see any compelling reason for why that must be so. For all we know, it may be that human-type minds are 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 be readily implemented on biological neural nets, it might well turn out that the new optimum relative to this new constraints-landscape has shifted away from the human-like mind region. There might not be a niche for complexes that contain precisely the types of components of which human minds are composed.

There might be ecological niches for complexes that are either less complex (say, individual modules), more complex (colonies of modules), or, conceivably, of similar complexity but very differently constituted than human-minds. But would these complexes be worthwhile from our current ethical point of view? When we reflect, do we really prefer a world in which these alien types of complexes have replaced human-type complexes?

I think this depends on the exact nature of those alien complexes. Currently there are levels of organizations – such as multinational corporations and nation states – that are highly complex and whose components include human beings. Yet, we regard these high-level organizations as of merely instrumental value. Corporations and states have no consciousness; they cannot feel pain or pleasure. We think they are good if they serve human (or animal) needs, but in cases where they don't, we have no scruples in "killing" them. Some extreme nationalists might think that a nation state (usually the one they happened to be born in) is a higher moral entity that is entitled to human sacrifices even when they would serve no human need, but most of us reject such views.

There are also lower levels of organization in today's world, but these are not assigned significant moral value either. We don't think it is wrong to erase a piece of computer code, and we wouldn't think that we were harming anyone if we extirpated a module (containing perhaps an epileptic center) from a human brain if that operation helped the remaining parts of the brain to function better. And as for alien forms of complexes of the same level of complexity as a human brain, I suspect that we would assign them moral value only if we thought that they had consciousness.

We can thus imagine a technologically highly advanced world, containing many sorts of complex structures (some of which are much smarter and more intricate than anything that exists today) in which there would nonetheless be a complete absence of any type of being whose welfare has moral significance. In a sense, this would be an uninhabited world, a world in which we and all the beings we care significantly about have gone extinct.

I want to emphasize that it is not the fact that machines have replaced humans that makes such a world undesirable. Whether a mind is implemented on biological neurons or on silicon processors seem to make no difference to me. Rather, it is the fact that there is not even the right type of "machines" in that world – machines of the type whose welfare matters. There may be an abundance of economic wealth and technological capability, but there is nobody there to enjoy it.

Here we must pause to consider a distinction. When we posited that the outsourcing uploads *outcompeted* the "hobbyists", there are two different things we could mean by this. One reading is that the population of outsourcing uploads gradually expands into the domain of resources originally held the hobbyists, so that the latter eventually run out of resources and go extinct. This is the typical outcome in evolution when one type outcompetes another. But we could also conceive of the case where the original group of hobbyists continues to exist indefinitely, and they are outcompeted only in the sense that they comprise a smaller and smaller *fraction* of the total population of agents, and they control an ever-decreasing fraction of the world's total wealth.

It is questionable whether the hobbyists could in the long run prevent the soup of outsourcing uploads from engulfing and expropriating their property. But suppose, for the sake of the argument, that they can. What we have then is not an extinction scenario, but

a scenario in which there is an immense loss of opportunity. The outsourcing uploads would (to switch metaphor) be a ravaging fire burning up the resources that would otherwise have been used for more meaningful purposes by the sentient uploads. Assuming property rights are enforced, the damage caused by this fire would be limited to the fraction of resources not originally owned by the hobbyists. The hobbyists might even manage to salvage some additional resources by colonizing new matter, although eventually the opportunities for further acquisition would disappear as the outsourcing uploads (being, *ex hypothesi*, more efficient) would have gotten there first. (The dynamics of such a colonization race is described in (Hanson 1998).)

Supposing we can foresee the course of development described in the above paragraphs, what should we do? One option would be to sit back and let things slide. We could bolster our passivity by invoking the greater evolutionary fitness of the outsourcing uploads as a reason for why it is good and desirable that they win out. If they are more fit, are they then not also more worthy possessors of the world's resources? This is obviously a very bad argument, as is plainly seen when it is explicitly articulated, but something like the might-makes-right idea may linger in some dark corners of some people's minds. We should remind ourselves that if some doomsday plague emerged and killed all mammals, that would not entail that it was somehow morally a good thing that the viruses or bacteria "won" although it would mean that they had been proven more fit.

Another attitude would be to lament the outcome but conclude that there is nothing we can do about it. After all, if outsourcing is more efficient, doesn't evolution theory then imply that the outsourcing-trait will spread and squeeze out the hobbyist-trait? Here transhumanists have a reply: Evolution created us, but we now have the moral right and potentially the capability to take control of our evolution (Bostrom et al. 1999; Bostrom 2001). We don't need to sit back and let things slide; we can take an active part in shaping our future destiny to fit our desires and values. We can use evolutionary methods where it suits us, but we can rein in evolution where we see better ways of selecting where to go next; we can substitute *directed evolution* for natural evolution. By so doing, we can climb fitness peaks that are otherwise inaccessible.

However, defeating evolution requires coordination. It is no good that a lot of people choose the preferred pathway if there are others who choose the fitness-maximizing pathway. For the latter will then be the variants that are amplified by selection pressures, and the process is set in motion that inevitably leads to the fittest winning out in the end.

There are only two logically possible ways out of this: either prevent such variants from appearing in the first place, or modify the selection pressure so that it doesn't favor those variants. Let us consider these possibilities in turn.

I assume that it will become possible to prevent random mutations in the narrow sense. Using error-correcting codes, it should be feasible to reduce the probability of errors in the replication process to an arbitrary degree. One must bear in mind, however, that there are more subtle ways in which unanticipated variants can emerge. Transhuman and even posthuman entities will want to change and improve themselves; they will want to acquire new capacities and experiment with new approaches. When installing a new capacity, it may be impossible to guarantee that it will not interact with other modules and other agents in unanticipated ways that result in the "outsourcing-phenotype" –

which then provides the raw material on which selection forces can operate. This leads to the scenario outlined above.

Moreover, even if every replication were error-free and every modification and upgrading produced no radically unanticipated behaviors, this would not prevent the outsourcing-type from spreading unless the initial population were completely free from such individuals. If some people start out with an expansionist mindset, then advanced replication and modification technologies would simply enable them to promulgate their kind more effectively. (Analogously: Contraceptives, no matter how perfect, don't limit population growth if they are not used; they limit the spread only of the genes of those who use them.)

We need therefore consider the second option: modifying selection pressures. Obviously, a very important part of our environment, in terms of determining the differential reproductive success of human gene- or meme-types, is the social realm: the way society, laws and other people's choices define the choices open to us and their effect on our inclusive fitness. Social structures could be set up in a manner that reduce the fitness of the outsourcing-type and enhance the fitness of the hobbyist-type. If these social structures were stable, then evolutionary trajectories would *not* lead to the outsourcing-type gradually outcompeting the hobbyist-type. It would be misleading to characterize this as "society helping the weak and unfit". Rather, the way society is set up partially defines what types are fit and "strong" in the sense of being able to use available means to proliferate. If we want to avoid an evolutionary trajectory which ends up in a region of statespace where the qualities we care for are either completely extinct or at any rate much less widespread than they could have been, then this sort of social sculpting of the conditions for reproductive success must take place (where "reproductive success" is, of course, not limited to sexual biological reproduction but includes also, for instance, copying of uploads, and in general the spread of forms of organizations).

Social shaping of the conditions for reproductive success is not a controversial proposal; it is fact of life obtaining to every life form that lives in societies. Before continuing our discussion about the future, let's spend a moment to contemplate some aspects of human society as it is today. I suggest that current society is, in a certain sense, in an anomalous state regarding reproduction and evolution. To be more specific, it is not in an evolutionary equilibrium (Kirk 2001): our preferences and inherited dispositions are not the ones that would maximize our fitness. 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 don't do this, because we happen not to have any great desire for reproductive success in the abstract sense. If we imagine current society frozen in its present form, then eventually humans would presumably evolve to want to fitness in an abstract sense (and perhaps to have a strong instinctual aversion against the use of contraceptives and birth control methods). Cultural evolution would possibly act faster, producing dominant memes according to which contraceptives are bad.[1] So one way in which the current

[1] 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 blessing. By contrast, supporters of VHEMT (The Voluntary Human Extinction Movement) (Knight 2001) have foresworn having children altogether. In passing, we may note that the scenario where we all one day become illuminated and decide to commit suicide would

scheme of things is anomalous (I'm not putting any negative evaluative connotation into that word!) is that our preferences aren't tuned to the present technological conditions for reproduction.

Another way in which the current regime is unstable is that it depends on economic growth rates keeping up with and exceeding population growth rates (in most places – unfortunately not in some African countries). Average income can only continue to rise as long as economic growth exceeds population growth. Population growth is limited, first of all, by facts about human biological reproduction: couples can only have so many children per year, and it takes a long time before a child reaches reproductive age. It is also limited by the fact, referred to in the previous paragraph, that our preferences are not fine-tuned to maximize the number of offspring under current conditions. Especially in developed countries, couples often choose to have many fewer children than they could sustain (and welfare programs would take care of any number of children they couldn't sustain). Both of these limitations can disappear. Lack of enthusiasm for having lots of kids could evolve away eventually. More drastically, the rate at which those individuals who do want to have many kids can achieve their goals would "go to infinity" when uploading is feasible. There is no way for economic growth to keep up with population growth in a population of freely reproducing uploads. If a welfare program seeks to guarantee a minimum income for uploads while allowing unlimited reproduction, it would very quickly go bankrupt even given stellar economic growth rates.[2]

These reflections serve to remind us that we shouldn't naively import intuitions about the current state of affairs into our thinking about the future. Malthus's teachings[3] don't seem to describe the world we see (where living conditions have been improving, without population control, contrary to his prediction), but this is explained by the two factors referred to above (preferences not in equilibrium and the slowness of human reproduction). In the upload world, by contrast, reproduction is virtually instantaneous, and this will lead to a rapid evolution of reproductive preferences that optimize fitness under whatever the socio-economic conditions are at that point.

To return to the earlier thread, we are thus asking how future socio-economic conditions would have be constituted in order that they define a fitness landscape that does not channel evolution in a direction that leads to the extinction of beings of the sort we care about. In other words, how can we shape conditions so that they favor the hobbyist-type?

---

not count as a whimper because if it really were better not to exist (as Silenus told king Midas in the Greek myth) then no desirable form of posthumanity had failed to be realized. *Erroneous* collective suicide is an existential risk albeit one whose probability seems extremely slight.

[2] 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 ($\sim e^t$).

[3] Thomas Robert Malthus (1766-1834), political economist and demographer, argued that the standard of living for the working class could not raised without population control because increased income would eventually lead to workers having more surviving children, which would drive wages back down again. Malthus was not as thoroughly pessimistic as is commonly thought, however. 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).

One simple but crude method would be ban outsourcing behavior. This would be very costly, since we are assuming that there are many tasks that are most efficiently done through outsourcing. A better approach would be to tax outsourcing and subsidize hobbyist behavior. That way, outsourcing would be used for the tasks where it brings the greatest cost-savings, while the level of taxation could be set so as to ensure that the hobbyist-type continues to thrive.

But is it really necessary to tax outsourcing in order to ensure that the hobbyist-type survives? Note that the only reason why we would ever institute such a tax is because we happen to like hobbyist-type beings and their activities. So to the extent that we have that preference, we would be willing to spend our own money on hobbies. Consequently, so long as we are around (maybe as uploads) and so long as we own all resources, then we will choose to devote a certain fraction of our income on hobbies. If we write "consumption" for hobbyist activities and "investment" for outsourcing behavior, then it seems that even absent "sin-taxes" on outsourcing, we would voluntarily choose to spend an optimal fraction of our resources on hobbies, and what we invest in outsourcing could be seen as delayed consumption of hobbies. The situation wouldn't be any different from the world today, where we see no need to encourage present consumption by taxing investment.

This happy outcome depends on a number of assumptions. First, we need to assume that property rights will always be nearly perfectly enforced. If the outsourcing types develop and eventually become immensely powerful, they might rob the hobbyists (either collectively in one big coup, or in a series of smaller assaults between groups or individuals). Second, we assume that the values of the initial hobbyist population are forever preserved. This would either require that the initial hobbyists don't die and that their values don't change over the eons, or else that they choose to reproduce almost exclusively in ways that cause their offspring to share their parents' hobbyist inclinations to an undiminished degree. And third, that the existence of a thriving hobbyist population is not a public good. Let's consider these three assumptions in more detail, starting with the last one.

You and I and a million other people might all desire that there be hobbyists in the world a long time from now. But each of us may also know that our individual actions will have a negligible effect on that outcome; so we each spend our resources on other goals, and the end result is that there are no hobbyists in the distant future, although we all agree that we would have been better off if some fraction of our resources had been set aside in a hobbyist conservation fund. The free-rider problem prevented us from contracting to bring about that superior outcome.

Objection: "Surely most hobbyists themselves want to survive indefinitely, and hence (supposing the other two assumptions were satisfied) there would be no need for an agreement between them to make sure that the hobbyist type continues to flourish a long time from now." – This is only partially true. Yes, most hobbyists might like to live indefinitely, but this is not their only desire. The may also have an impersonal ethical preference that there be lots of hobbyists in the future. This latter preference is for a public good (it's non-rivalrous and non-excludable).

Objection: "But wouldn't this public good be produced as a side effect of the hobbyists going about pursuing their personal survival? After all, if each hobbyist wants (a) the public good of many hobbyists existing in the future, and (b) that they themselves

exist in the future, and they know that they can't significantly influence (a) then they should direct all their resources towards (b); and if all hobbyists do that, the net effect is that all hobbyist resources are devoted (unintentionally) to the goal that there be many hobbyists in the future." – Again, this is only partially true, since it presupposes that the hobbyists have no other preferences than (a) and (b). But realistically, their preference function will be influenced by other factors. For instance, their preference for (b) is presumably modulated by a time discount factor, making (many) hobbyists (at least slightly) prefer to have personal hobbyist consumption now rather than in the distant future. They would therefore invest less than an optimizing outsourcing type, and so in the long run the fraction of all resources owned by hobbyists would dwindle even if property rights over the whole cosmos were assigned at the outset and assumptions (1) and (2) were completely satisfied.

Assumption (2) might at first blush look implausible, but as we noted above, error-correction codes should make reproduction arbitrarily reliable; uploads don't suffer biological aging and the habit of keeping backup copies at dispersed locations should reduce the risk of accidental death to an arbitrary level as well. There remains the risk of value drift as a result of installation of new capacities. It is not completely clear at this time to what extent individual prudence could reduce that risk.[4]

Assumption (1), that property rights will be enforced, is crucial if an "invisible hand" is to select a favorable outcome. If outsourcers are capable of and willing to simply expropriate hobbyist capital, then the latter will of course become extinct unless they in turn manage to expropriate some of the outsourcers's resources. If each tries to expropriate from the other, and there is no superior agency that prevents them from doing so, then they are in a state of war. There is no reason to think that this war would necessarily divide neatly into two warring sides, Outsourcers vs. Hobbyists. Instead, the situation is one where various shifting alliances could form between individual outsourcers and hobbyists. As in all wars, there would be a cost in terms of lost lives or property and wasted opportunities for collaboration. But assuming no "doomsday weapons" are deployed (big assumption), would it lead, in the long run, to the hobbyists becoming extinct and the outsourcers taking over?

It seems that this would be a case of two groups of organisms competing for the same resources, one group being more efficient than the other. One would expect the less efficient group to die out. This means that the hobbyists would either go extinct, or lose their hobbyist inclinations, or be forced to forever keep those inclinations in check (so

[4] It would be relevant in this context to consider the idea of safety pacts. – You sign an agreement with other individuals you trust to reverse some implementation if they find that it has changed your nature in ways that you would not have liked. You could also store a copy of your earlier self and allow it a time period during which it can remold its brainchild. In fact, it would be interesting to explore in more detail the various strategies that could be use to pursue safe development. But the relevant issue here is whether the safety measures that can be taken by voluntary safety pacts are as effective as those that could be taken by democratic society. It seems likely that they can.

It could also be worth to relate these issues of personal development to the literature on second-order desires and dispositional theories of value. This could possibly form the basis of a transhumanist "dynamic ethics", which would deal with ways in which value systems and ethical systems could (ethically) change over time. Also related to the problem of defining "improvement", and to the problem of the ethics of new persons (Glover 1984; Parfit 1984).

that they don't manifest themselves in hobbyist behavior, and assuming the resources it takes to encode the inert hobbyist inclinations are negligible).

One complication that should be noted is that the "shifting alliances" condition may not be satisfied. Maybe there will be techniques that enable parties to reliably commit themselves forever (e.g. mind-reading techniques?). Then when the war breaks out, various budding coalitions would bargain for the allegiance of those individuals who have not yet committed themselves. At this stage, some hobbyists may strike favorable deals with what turns out to be the winning coalition. After the victory, the outsourcers in this coalition could not then, on this hypothesis, shift their alliances again; the surviving hobbyists would have their security guaranteed forever.

This possibility represents nothing but a dangerous and costly detour, however. The end result of such a war between fixed alliances, assuming intelligent life doesn't go extinct in the process, would be that a new order is established. In the best case, hobbyists would be a part of this order and would have their property rights secured; and the hobbyist-type may then survive indefinitely. But its survival depends on natural evolution being reeled in by the winning pact forming a *singleton*, that is: a global regime that can pass and enforce laws over all people. Only a singleton can succeed in arranging the forces of selection so that they drive evolution in a desirable direction. A world order with many competing powers will be subject to evolutionary pressures that are outside anybody's control: each power may direct evolution within its own jurisdiction, but the various powers would still be in competition with one another; so evolution would still take place, only the entities on which it would operate would be "powers", complexes of individuals, rather than particular creatures. Evolution would favor complexes that passed laws internally that resulted in the power's resources being used for the outsourcing-type behavior rather than the hobbyist-type behavior. In the long run, complexes containing the hobbyist-type would be driven out of existence.

In practice, the "long run" may be quite short, depending on just how great the efficiency-advantage that the outsourcers have over the hobbyists. Since hobbyists can think ahead, they may choose to limit their expenditure on hobbies and behave more like outsourcers. This could buy them time, but at the expense of cutting back the activities that make them hobbyists. In either case, the days of fun would be severely diminished by the existence of the outsourcers if the latter were in direct ecological competition with the hobbyists.

Let us summarize the conclusions of this section so far. We have noted that natural evolutionary pressures in a posthuman world may well direct development in a direction that leads to the extinction of the sort of beings that we regard as valuable (the world ending with a whimper). The fact that the forms that may result from natural evolution would be the fittest is no reason to favor that outcome. But this outcome can be avoided if we can direct evolution, by shaping the fitness function by which a being's reproductive success is determined so that it favors the hobbyist-type beings that we prefer should exist (and among whose numbers we ourselves may hope to be). In order to do this, however, it is not sufficient that some local power or group decides to set up internal conditions so that they favor the hobbyist type. For evolution would then simply occur on a higher level, where more efficiency-oriented groups would eventually win out. What is needed is that on the highest level there is only a single power: a singleton that has the ability to regulate the conditions under which lower entities compete.

A singleton need not be a monolith: it could harbor within itself an abundant ecology of diverse types of beings and values. A singleton could, for example, be a democratic world-government, or a friendly superintelligence (Yudkowsky 2001). In a minimal version it could, in principle at least, be set up so that its role is limited to ensuring that the property rights of its inhabitants cannot be violated. This would be sufficient to ensure that the hobbyist is preserved, since some members would choose to use their resources to live a hobbyist life of indefinite duration. The outsourcing type would still out-compete the hobbyist type in the sense of possessing an ever-increasing fraction of the singleton's resources. But the absolute amount of resources possessed by the hobbyists need not decrease, and could in fact increase, as they gain a return on what proportion of their capital they are saving and as the singleton's total resources keep growing. More activist kinds of singletons are of course also imaginable.

References

Bostrom, N. (2001). "Transhumanist Values." *Manuscript*. http://www.nickbostrom.com

Bostrom, N., et al. (1999). *The Transhumanist FAQ*. http://www.transhumanism.org/resources/faq.html

Glover, J. (1984). *What Sort of People Should There Be?*, Pelican.

Hanson, R. (1998). "Burning the Cosmic Commons: Evolutionary Strategies for Interstellar Colonization." *Working paper*. http://hanson.gmu.edu/workingpapers.html

Kirk, K. M. (2001). "Natural Selection and Quantitative Genetics of Life-History Traits in Western Women: A Twin Study." *Evolution* **55**(2): 432-5.

Knight, L. U. (2001). *The Voluntary Human Extinction Movement*. http://www.vhemt.org/

Malthus, T. R. (1803). *An Essay on the Principle of Population*. London, J. Johnson.

Parfit, D. (1984). *Reasons and Persons*. Oxford, Clarendon Press.

Yudkowsky, E. (2001). *Friendly AI 0.9*. http://singinst.org/CaTAI/friendly/contents.html