In Defense of Artificial Replacement

Derek Shiller

The pace of technological change is very difficult to predict far in advance, but our current trajectory makes it reasonable to guess that we will have the power create genuine artificial intelligence – artificially created individuals that equal or surpass human beings in all dimensions of cognition, including creativity, power, insight, and wisdom – by the close of this century. Some futurists[1] have worried about our species' continued existence after this development. Such concerns are motivated by the recognition that it may be difficult to predict and control artificial intelligences that are smarter than we are. I think this fear is selfish. The ethically responsible thing for us to do will most likely be to engineer our own extinction.

In this paper, I will present a simple speculative argument for what I will call the *Artificial Replacement Thesis*. The Artificial Replacement Thesis suggests that as a matter of moral beneficence, we should replace our species with artificial beings who are capable of living better lives. After introducing some assumptions and presenting my argument, I will spend the remainder of this paper formulating and replying to objections.

My argument for the Artificial Replacement Thesis relies on two assumptions. I will take these assumptions for granted in the remainder of this paper.

First, I will assume that we will have the power to create intelligent artificial minds that are able mimic natural minds in every morally relevant way we wish. Morally relevant ways might include: consciousness, cognitive flexibility,

[1] See Muller (2015).

emotional capacity, capacity for happiness and unhappiness, ability to engage in interpersonal relationships, creativity, freedom of the will (in whatever sense we have it), and philosophical, religious, or artistic insight. Whatever nature can do with clumps of neurons, we will be able to build computers to do. If this assumption is correct, it means that with the right design, artificial intelligences will be able to fall in love, experience exquisite joy, write novels that probe existential self-doubt, ponder the basic metaphysical structure of reality, and appreciate the beauty of complex mathematical theorems.

Perhaps my most controversial assumption is that it is possible for us to create consciousness out of inorganic matter. While this assumption is integral to my argument, I will not argue for it here.

Second, I will assume that by proper design, it will be comparatively easy to avoid instilling within our artificially intelligent constructs the vicious traits that presently afflict humanity. Human bodies are the results of a variety of evolutionary pressures. These pressures were not conducive to maximizing our ancestors' well-being. Consequently, human life was often nasty, brutish, and short. Though modern technology helps us to fare much better than our ancestors, we share many of their shortcomings.

Some of these shortcomings are physiological. Perhaps the foremost among these is the process of aging. As we age, our bodies deteriorate in a variety of ways: we feel joint pain, we lose our mobility and flexibility, our minds go, we get sick, and we die. We are typically numbed to this inevitable tragedy because we are so accustomed to it, but it is a substantive limit to the well-being we are capable of in our lives.

Other shortcomings are psychological. We suffer from unreasonable and unfulfillable desires. We want to be kinds of people that we cannot be. In pursuit of fleeting temptations, we are disposed to make decisions that go against our own interest. We are aggressive, callous, and cruel to each other.[2] We harbor arbitrary biases against our fellow creatures based on irrelevant characteristics or group membership.

These are not the inevitable vices of any intelligent being. They are part of our species. We need not pass them on to our artificial creations. While there might be some constraints to the kinds of artificial minds that it is possible for us to make, I will assume that we could build artificial constructs that are not afflicted by the same kinds of vices that we are. Our creations could be rational, intelligent, deeply caring, loyal, respectful, wise, good-humored, emotionally stable, and eternally healthy.

The moral core of my argument is the Future Beneficence Principle.

> **Future Beneficence Principle:** Where it is possible to greatly improve the well-being of future generations at a comparatively low cost to ourselves, we should do so, even if doing so will affect the identity of those future beings.

This principle borrows plausibility from its relation to the Future Nonmaleficence Principle.

> **Future Nonmaleficence Principle:** Where it is possible to improve our well-being, at a far greater cost to future generations, we should not do so, even if doing so will affect the identity of those future beings.

Our obligations to the future are notoriously complicated by the fact that our present actions may influence who will come to exist in the future (Parfit 1984, Schwartz 1978). If we choose to conserve resources now, for instance, different people will exist in the future than otherwise would have. As a result of the fact that existence is a precondition for a valuable life, everyone who exists will benefit from our choice, no matter which choice we make.

[2] See Benetar (2006).

Still, it is widely thought that we have moral reasons to act in ways that will lead to a better future, even if no one in future generations would have existed had we made a different choice. This may be because we have reasons to produce as many people as possible at maximal levels of well-being[3] or because we have reasons to make sure that the people who do exist are as well-off as possible.

While there is sure to be disagreement about the reasons, many will agree that it is wrong to do things will greatly harm future generations in exchange for our own comparatively small benefit. It would be wrong, for example, to continue activities that contribute significantly to global warming for our own economic gain on the grounds that the people who will primarily suffer will also owe their existence to our decision. The Future Nonmaleficence Principle formalizes this idea.

The Future Beneficence Principle suggests that we have similar reasons to provide benefits to future generations when we have the ability to do so. The difference between acting to raise future well-being and abstaining from lowering future well-being is hard to make precise, but it seems unlikely that anything morally significant will turn on it (Bennett 1993). It is difficult enough to draw this distinction in normal contexts. It will be much more difficult when an action's effects are filtered through generational time.

There is also an argument from the consequences of attributing an asymmetry to our obligations.[4] Surely, many of our actions will create both positive and negative effects on the future. If we had special reasons not to cause harm that were not balanced by reasons to do good, then we might be barred

[3] See Sikora (1978), Leslie (1989), Rachels (1998), Huemer (2008), and Gardner (2016).

[4] This bears similarity to an argument made by Sikora (1978). See also McMahan (2009) and Bradley (2013).

from doing things that on the whole lead to a positive situation but which included some significant negatives. It is plausible that the vast majority of choices about policies that have significant future effects are like that. Consequently, the Future Beneficence Principle should appeal to those who are taken by the Future Nonmaleficence Principle.

My particular application of the Future Beneficence Principle involves its implication that, given the choice, we should create beings with greater well-being over beings with substantially less well-being, assuming that it is not too costly to us. This application parallels Savulescu's (Savulescu 2001, Savulescu and Kahane 2009)[5] Principle of Procreative Beneficence, which enjoins us to "select the child, of the possible children [we] could have, who is expected to have the best life, or at least as good a life as the others, based on the relevant, available information." (p. 415)

Savulescu originally presented his principle in the context of choices of medical intervention into normal human reproduction. It implies that we should take steps to avoid having children with traits that are harmful to their well-being, and that we should opt for children with traits that are conducive to their well-being. This idea can be extended to evaluate the question of whether or not we should opt to have normal human progeny or to create artificially intelligent beings. If we can give artificial 'progeny' a better life than biological progeny, Salvulescu's principle seems to suggest that we should forgo natural reproduction on the grounds of beneficence. We should choose to create creatures with the best life. The fact that such creatures are made of silicon and do not emerge directly from our genitals is morally irrelevant.

Here is my argument: human beings live lives that are quite suboptimal. With a good design, we will be able to produce artificial beings whose lives are

[5] See also Harris (2007).

much closer to being optimal. We will then be faced with the choice of continuing to populate the world with human beings, or devoting resources over to creating beings who are capable of much higher levels of well-being.

Our resources are finite and the same resources that might allow human beings to live – effort, land, energy, raw materials – could be more effectively spent on creating and sustaining artificial creatures. When that becomes the case, the beneficent thing to do is to choose that our children be artificial, rather than natural. It will not harm us too much, and it will greatly benefit future generations.

> **Artificial Replacement Thesis:** Once it is possible to design artificial creatures whose lives are significantly better than human lives (and at a more efficient use of our resources), we should engineer the extinction of the human race in order to route available resources to creating and sustaining them.

This proposal should not be read as a justification for forcibly bringing about such a change against the wishes of currently existing people. Nor should it be read as involving the purposeful suicide of anyone. The extinction called for could be achieved by generational replacement, or perhaps a gradual petering out of humanity (where each generation is significantly smaller than the previous).

The Future Beneficence Principle only instructs us to act in the interests of future generations when it is not excessively costly to ourselves. So in order for the thesis to be supported by the principle, it would need to be possible for human extinction to be carried out in relative comfort. Though one might imagine that the last generation of humans would feel anguish, despair, and loneliness, there is no reason why this need be the case. The last humans would have the company of not just each other, but also of their artificial progeny.

This proposal is likely to be met with a great deal of skepticism. We value our own humanity, and treat the extinction of our species with fear. Though it has been convincingly argued (Lenman 2002) that the end of humanity sooner rather than later is not inherently bad in itself (and this should be especially true if we are replaced with something objectively preferable), we have an understandable attachment to our own species.[6] This attachment manifests itself both in our comparative lack of concern for other animals, and in a repulsion to the Artificial Replacement Thesis. We would like human beings to always be around. Nevertheless, I do not believe that any justification can be found for this preference that carries much moral weight. In the remainder of the paper, I will consider two objections and argue that neither succeeds.

Objection 1: Imperfection is good

According to the first objection, human imperfections actually make our lives better. If we were all extremely intelligent, coolly rational, disease- and disability-free, our lives would lose something of the depth and complexity that makes them valuable (Barnes 2009, Garland-Thomson 2012).

While good in small doses, perhaps too much happiness and self-determination may be detrimental to us. A good life involves making the most with what we have. Perhaps it is valuable to accomplish difficult things (Bradford 2013), and this is not possible without being substantially imperfect.

There are surely some maladies that we are better off without, but perfection would not be worth the loss of depth of experience that our imperfections provide us.

Reply: Artificial intelligences can be imperfect, too.

I grant that lives are better when they are constrained. Overcoming obstacles contributes to the quality of our lives, and if we could get whatever we wanted whenever we wanted it, our lives might be happy but meaningless. It might be a

[6] Though for an opposing viewpoint, see Bennett (1978).

necessary feature of any obstacle worth overcoming that we face some chance of failure to overcome it. And it may be that triumphing over self-imposed limitations does not add as much to the quality of our lives as meeting and exceeding externally-imposed limitations. If so, then we might think that our imperfections actually work in our favor, rather than against us.

This objection only succeeds as an objection to the Artificial Replacement Thesis, however, if either our lives are already fairly close to being optimal, or else if we are less able to convey our artificial creations with the kinds of imperfections that make our lives better. Both ideas are dubious.

First, it seems extremely unlikely that we should have lucked into just the right amounts of misery and difficulty for an optimal life. It is implausible that the majority of detriments are beneficial. It is certainly not something we generally think about our own lives. We strive to improve them by bettering ourselves. We don't consider ourselves lucky to be rash, cruel, or susceptible to cancer. We do our best to prevent our children from having these sorts of imperfections. So we shouldn't think that all of our imperfections really add much value to our lives.

Second, we have a lot of control over the lives of our artificial creations. If we want to make artificial intelligences that die, we can design them to do so. If we want to make artificial intelligences with gambling problems, we can. If we want to make artificial intelligences that strive to find romantic love and fail most of the time, due to their own neurotic insecurities, that is, by assumption, our choice.

If we can actually choose the imperfections of our creations, then we can do so in a way that is maximally beneficial to them. It is unlikely that our particular array of limitations is conducive to maximal well-being. It is unlikely that the optimal life-span happens to be how long human beings live. So we could program our progeny to live as long as it is good for a creature to live, and no longer. Plausibly, it would be better for us to have fewer kinds of pain and suffering, or have them more evenly distributed throughout the population and throughout our lives. Insofar as imperfection is good, artificial lives could be made to have just the perfect amount of imperfection. This should be far preferable to our present state.

Objection 2: Human beings can themselves be perfected.

Medical technology is rapidly improving. We are much better able to treat conditions now than we were a century ago. We may be much better still in another century. Instead of replacing ourselves with artificial creations, we should simply use technology to get rid of the detrimental traits we currently have. We can cure disease and forestall death inevitably through medical advances, and the more persistent psychological problems of humanity might also be curable through genetic manipulation.

Reply: It will most likely be easier to build new lives that avoid our imperfections than it would be to rid ourselves of those imperfections.

Making up for our species' problems requires working within a very specific set of constraints. If we change ourselves too much, we will no longer be human. Our bodies were not created to be easy to fix, so there is no guarantee that we could reach perfection through feasible alterations on this existing design nearly as easily or effectively as it could be achieved starting from scratch.

It also remains to be seen how effective genetic tampering, social engineering, or psychiatric treatment can be. Genes don't influence phenotypical traits neatly. Who we are is largely determined by a set of extremely complex interactions between our genes. All this means that many of our imperfections may be extremely difficult or impossible to fully remove without creating larger problems elsewhere. Psychology is presently far from solving our emotional maladies, and we are making much more progress toward artificial intelligence than we are to completely resolving our negative psychological traits.

Even if we could create human beings with optimal lives in the same time frame in which we could create optimal artificial lives, we cannot get around the fact that our lives take a lot of resources to sustain. Humans have bodies of particular sizes that come along with particular caloric needs. Those needs have so far led us to convert a large percentage of the Earth's surface to agriculture. Our ideal lifestyle requires a lot of space and property for each individual. Sustaining a large human population has had catastrophic effects on our environment that an artificial population might avoid. If we could produce vastly more efficient artificial creatures that did not require organic material to live, then by allowing humanity to go extinct, we could make our present resources go a lot farther.

Conclusion

The assumptions made in the course of the argument are likely to be rather contentious. Nevertheless, we cannot ignore the possibility that we will be able to create artificial creatures with lives of optimal well-being in the not-too-distant future. If we can do that, a genuine utopia on Earth may be within our grasp. We must merely step out of the way to let it happen.

Bibliography

Barnes, Elizabeth (2009). Disability, Minority, and Difference. *Journal of Applied Philosophy* 26 (4):337-355.

Barnes, Elizabeth (2014). Valuing Disability, Causing Disability. *Ethics* 125 (1):88-113.

Benatar, David (2006). *Better Never to Have Been: The Harm of Coming Into Existence*. New York; Oxford University Press.

Bennett, Jonathan (1978). On Maximizing Happiness. In Richard I. Sikora & Brian M. Barry (eds.), *Obligations to Future Generations*. White Horse Press 61--73.

Bennett, Jonathan (1993). Negation and Abstention: Two Theories of Allowing. *Ethics*104 (1):75-96.

Bradford, Gwen (2013). The Value of Achievements. *Pacific Philosophical Quarterly* 94 (2):204-224.

Bradley, Ben (2013). Asymmetries in Benefiting, Harming and Creating. *Journal of Ethics* 17 (1-2):37-49.

Gardner, Molly (2016). Beneficence and Procreation. *Philosophical Studies* 173 (2):321-336.

Garland-Thomson, Rosemarie (2012). The Case for Conserving Disability. *Journal of Bioethical Inquiry* 9 (3):339-355.

Harris, John (2007). *Enhancing Evolution: The Ethical Case for Making Better People*. Princeton University Press.

Huemer, Michael (2008). In Defence of Repugnance. *Mind* 117 (468):899-933.

Hurka, Thomas (1983). Value and Population Size. *Ethics* 93 (3):496-507.

Lenman, James (2002). On Becoming Extinct. *Pacific Philosophical Quarterly* 83 (3):253–269.

Leslie, John (1989). The Need to Generate Happy People. *Philosophia* 19 (1):29-33.

McMahan, Jeff (2009). Asymmetries in the Morality of Causing People to Exist. In David Wasserman & Melinda Roberts (eds.), *Harming Future Persons*. Springer 49--68.

Müller, Vincent C. (ed.) (2015). *Risks of Artificial Intelligence*. CRC Press - Chapman & Hall.

Parfit, Derek (1984). *Reasons and Persons*. Oxford University Press.

Rachels, Stuart (1998). Is It Good to Make Happy People? *Bioethics* 12 (2):93-110.

Savulescu, Julian (2001). Procreative Beneficence: Why We Should Select the Best Children. *Bioethics* 15 (5-6):413-426.

Savulescu, Julian & Kahane, Guy (2009). The Moral Obligation to Create Children with the Best Chance of the Best Life. *Bioethics* 23 (5):274-290.

Schwartz, Thomas (1978). Obligations to Posterity. In Richard I. Sikora & Brian M. Barry (eds.), *Obligations to Future Generations*. White Horse Press 3—3.

Sikora, Richard (1978). Is It Wrong to Prevent the Existence of Future generations. In Richard I. Sikora & Brian M. Barry (eds.), *Obligations to Future Generations*. White Horse Press 112--166.

Sparrow, Robert (2015). Imposing Genetic Diversity. *American Journal of Bioethics* 15 (6):2-10.