Sharing the World with Digital Minds[1]

(2020)
[Draft version 1.1]
Carl Shulman & Nick Bostrom
Future of Humanity Institute
Oxford University

Abstract
The minds of biological creatures occupy a small corner of a much larger space of possible minds that could be created once we master the technology of artificial intelligence. Yet many of our moral intuitions and practices are based on assumptions about human nature that need not hold for digital minds. This points to the need for moral reflection as we approach the era of advanced machine intelligence. Here we focus on one set of issues, which arise from the prospect of digital "utility monsters". These may be mass-produced minds with moral statuses and interests similar to those of human beings or other morally considerable animals, so that collectively their moral claims outweigh those of the incumbent populations. Alternatively it may become easy to create individual digital minds with much stronger individual interests and claims to resources than humans. Disrespecting these could produce a moral catastrophe of immense proportions, while a naive way of respecting them could be disastrous for humanity. A sensible approach requires reforms of our moral norms and institutions along with advance planning regarding what kinds of digital minds we bring into existence.

1. Introduction

Human biological nature imposes many practical limits on what can be done to promote somebody's welfare. We can only live so long, feel so much joy, have so many children, and benefit so much from additional support and resources. Meanwhile, we require, in order to flourish, that a complex set of physical, psychological, and social conditions be met.

In contrast, we can easily conceive of artificial beings with conscious experiences, desires, and capacity for reasoning and autonomous decision-making, which would neither be subject to the same practical limitations in their ability to benefit from additional resources nor depend on the same complex requirements for their survival and flourishing. This could be a wonderful development: lives free of pain and disease, bubbling over with happiness, enriched with superhuman awareness and understanding and all manner of higher goods.[2]

Recent progress in machine learning raises the prospect that future AI systems may soon (or already) have psychological capacities which in human beings or other animals are commonly taken to confer degrees of moral status. (We will assume that

[1] For helpful comments, we're grateful to Guy Kahane, Matthew van der Merwe, Hazem Zohny, Max Daniel, Lukas Finnveden, Lukas Gloor, Uli Alskelung Von Hornbol, and Luke Muehlhauser.
[2] Bostrom (2008a; 2008b)

appropriately architected AI could be conscious, though it's worth noting that some accounts of moral status do not view this as a necessary condition for having moral status.[3])

This sets the stage for Robert Nozick's (1974, p. 41) "utility monster" to become a practical reality:

> Utilitarian theory is embarrassed by the possibility of utility monsters who get enormously greater sums of utility from any sacrifice of others than these others lose. For, unacceptably, the theory seems to require that we all be sacrificed in the monster's maw, in order to increase total utility.

The conceptual coherence of a utility monster has been disputed. For example, Derek Parfit (1984, p. 343) wrote:

> Think of the life of the luckiest person that you know, and ask what a life would have to be like in order to be a million times as much worth living. The qualitative gap between such a life and ours, at its best, must resemble the gap between ours, at its best, and the life of those creatures who are barely conscious—such as, if they are conscious, Plato's "contented oysters". It seems a fair reply that we cannot imagine, even in the dimmest way, the life of this Utility Monster. And this casts doubt on the force of the example. Act Utilitarians might say that, if we really could imagine what such a life would be like, we might not find Nozick's objection persuasive. His "Monster" seems to be a god-like being. In the imagined presence of such a being, our belief in our right to equality with him may begin to waver…

However, although Parfit questions the possibility of a monsterishly vast qualitative difference in welfare, he suggests that similar results can be obtained by considering the *quantitative* dimension of population size, in which there is clearly no conceptual barrier to extreme values.

We will argue that population size is only one of several quantitative dimensions—together with several less certain qualitative dimensions—along which utility monsters can excel humans and other animals by allowing vastly greater welfare per unit of resource consumption. This existence of multiple independent paths to utility monsterdom makes the conclusion more credible that such utility monsters are possible and may become practically feasible with further progress in AI.

While non-utilitarians may fancy themselves immune to the utility monster challenge, most reasonable views are in fact susceptible, to various degrees. This is because even if we postulate that no deontological violations (such as direct killings) would occur, human interests may still be adversely affected by the advent of utility monsters, since the latter could have stronger moral claims to state aid and other scarce resources, thus reducing the amount that could be defensibly claimed by human beings.

---

[3] For a discussion of the case for AI consciousness, see Chalmers (2010). Kagan (2019) discusses the case for moral status for unconscious but agential AI.

So the good thing about utility monsters is that they could generate a lot of utility. The bad thing is that this might come at our expense.

2. The anatomy of utility monsters

2.1. Reproductive capacity

One of the most basic features of computer software is the ease and speed of exact reproduction, provided computer hardware is available. Hardware can be rapidly constructed so long as its economic output can pay for manufacturing costs (which have historically fallen, on price-performance bases, by enormous amounts; Nordhaus, 2007). This opens up the door for population dynamics which would take multiple centuries to play out among humans to be compressed into a fraction of a human lifetime. Even if *initially* only a few digital minds of a certain intellectual capacity can be affordably built, the number of such minds could soon grow exponentially or super-exponentially, until limited by other constraints. Such explosive reproductive potential could allow digital minds to vastly outnumber humans in a relatively short time—correspondingly increasing the collective strength of their moral and political claims.

Furthermore, if the production of digital minds and required hardware proceeds until the wages of the resulting minds equal marginal costs, this could drive wages downward towards machine subsistence levels as natural resources become a limiting factor. These may be insufficient for humans (and obsolete digital minds) to survive on (Hanson, 2001; Aghion, Jones and Jones, 2017). Such circumstances make redistributive issues more pressing—a matter of life and death—while the Malthusian population growth would make claims to transfer payments effectively insatiable.

Another troublesome aspect of fast and cheap reproduction is that it permits rapid turnover of population. A digital mind that is deleted can be immediately replaced by a copy of a fully-fledged mind of the newest edition—in contrast to the human case, where it takes nine months to produce a drooling baby).[4] Economic pressures could thus push towards very frequent erasure of "obsolete" minds and replacement with minds that generate more economic value with the same hardware.[5]

A plausible continuation of current software practices applied to digital minds could thus involve extremely large numbers of short lives and deaths, even as a fraction of the number of minds in existence at any given time. Such ephemeral digital minds may be psychologically mature, chronologically young, with long *potential* lifespans

[4] It may be unclear, however, whether an exact or almost exact copy of an existing mind would constitute a new distinct person or instead as an additional instantiation of the person whose mind served as the template.

[5] It is also frequently convenient to "move" a file from one memory location to another, an operation which normally involves first creating a copy of the file at the new location, which briefly coexists with the original before the latter is deleted. If the transported file were the running executable of a mind, this procedure would roughly correspond to a teleportation scenario. Philosophers are divided on whether a person survives such a process (Bourget and Chalmers, 2014). If the process involves some kind of morally significant "death", then such file transfers would constitute another way in which digital minds could be subject to radically abrogated life expectancies.

yet very short default life expectancies in the absence of subsidy. If we think that dying young while being able to live long is a large deprivation, or is very unfair when others are able to live out long lives, then this could ground an especially strong claim for these digital minds to resources to extend their lifespan (or other forms of compensation). If death in itself is a bad (and not merely an opportunity cost of foregone life), then this rapid turnover of minds could also increase the extent of this disvalue per life-year lived.

2.2. Cost of living

It is plausible that many digital minds will need less income to sustain themselves at a given standard of living. The cost of computer hardware to support digital minds will likely decline well below the cost of supporting a human brain and body. If we look beyond mere subsistence, physical goods and services suitable for human consumption (such as housing and transportation) tend to be more expensive than information technology and virtual goods to meet the equivalent needs of a digital mind. Nor need a digital mind suffer from inclement environmental conditions, pollution, disease, biological aging, or any number of other impositions that depress human well-being.

The cost of producing a given number of (quality-adjusted) life years for a human-like digital mind will therefore likely fall far below the equivalent cost for a biological human. Large differentials in cost of living mean that, when questions of distribution arise, a resource that confers a small benefit to a human may confer large benefits to many digital minds. If the energy budget required to sustain one human life for one month can sustain ten digital minds for one year, that would ground a powerful argument for favoring the latter in a situation of scarcity.

2.3. Subjective speed

Hardware with higher serial speeds can be used to run digital minds faster. Current computer clock speeds are measured in gigahertz, millions of times greater than firing rates of human neurons; and signal transmission speeds can similarly exceed the conductance speed of human nerves. It is therefore likely that digital minds with humanlike capabilities could think at least thousands of times (and perhaps millions) faster than humans do, given a sufficient supply of hardware. If a digital mind packs thousands of subjective years of life into a single calendar year, then it seems the former ("subjective time", not wall-clock time) is the correct measure for such things as the amount of well-being gained from extended life (Bostrom and Yudkowsky, 2014).

Since speedup requires paying for more hardware, this provides a way for individual digital minds to get much higher (subjective-life-years per dollar) returns from wealth than humans usually can. At low speeds, the gains available to digital minds would be close to linear; though as speeds approach the limits of technology, marginal costs of further speed increments would rise.[6]

[6] Hanson (2016, pp. 63-65) argues cost-increases with speedup would be initially near-linear, i.e. 2x speedup requiring close to 2x hardware budget, up to substantially superhuman speeds.

Because these gains of running faster can accrue to then-existing initially slower-running individuals, this dimension of utility monsterdom applies also within population axiologies that take a "person-affecting" approach (more on this later).

2.4. Hedonic skew

The aforementioned dimensions reflect basic properties of minds implemented in digital computers; they would hence apply even to high-fidelity emulations of unmodified human brains (in a virtual body and a virtual environment). Additional dimensions for increased well-being are unlocked if we consider the possibility of modifying the "software", or structure, of the mind.

Which such software-related enhancements are relevant depends on the theory of well-being we adopt. In this and the next subsection, we will consider a simple hedonistic conception of well-being. Later, we consider some other conceptions.

Our brains are not built for bliss. Instead, human physiology has evolved to generate pleasure and pain where this motivated behaviors associated with reproductive fitness in past generations. This entails for us a great deal of hard-to-avoid suffering. Moreover, our enjoyments tend to be tightly regulated and dispensed with miserly reticence. Culinary pleasures are regulated by hunger; sexual pleasures are limited by libido. Universal provision of the enjoyments linked to desires for relative power and social status is hampered by the near zero-sum character of the goods involved. Most rewards are also moderated by mechanisms such as boredom and tolerance, which progressively reduce the delight obtained of repeated stimuli or continual benignant conditions.

Digital minds could be designed to experience pleasurable mental states with greater frequency and duration, having peak experiences much or all of the time. They could enjoy lasting bliss as well as liberation from the painful parts of our present human existence.

The hedonic balance for humans, too, would be amenable to great improvement with the kind of advanced technology that would likely either precede or closely follow mature machine intelligence technology.[7] However, radically adjusting the hedonic balance for biological humans may be more "costly" than doing the same for de novo digital minds, in a couple of ways: (a) interventions that require brain surgery, extensive pharmacological fine-tunings and manipulations, or the equivalent, may, at least in the nearer term, be infeasible or expensive; and (b) more radical transformations of our psyches would risk destroying personal-identity or other properties of our current human nature that we value.[8] The mind-designs of sentient machines could thus have great advantages in terms of the efficiency with which they can realize hedonically valuable states.

2.5. Hedonic range

[7] David Pearce (1995) has argued that biological minds could be engineered to run on "gradients of bliss" rather than on the full current pain-pleasure span.
[8] Cf. (Agar, 2010, pp. 164-189).

In addition to changing the fraction of time spent inhabiting different parts of the hedonic scale accessible to present human beings, it might also be possible—more speculatively—to design digital minds such that they could realize "off the charts" states of hedonic well-being—levels of bliss that human brains are totally incapable of instantiating.

Evolutionary considerations give some support for this hypothesis. Insofar as intensity of pleasures and pains correspond to strength of behavioral responses, evolution should tend to adjust hedonic experiences to yield approximately fitness-maximizing degrees of effort to attain or avoid them. But for human beings it is generally much easier to *lose* large amounts of reproductive fitness in a short time than to *gain* an equivalent amount. Staying in a fire for a few moments can result in permanent injury or death, at the cost of all of an organism's remaining reproductive opportunities. No single meal or sex act has as much at stake per second—it takes weeks to starve, and the expected number of reproducing children produced per minute of mating is small. Thus, evolution may have had call to generate more intensely motivating-per-second pains in response to injury than pleasures in response to positive events. Engineered minds, by contrast, could be crafted to experience pleasures as intensely rewarding as the worst torments are disrewarding. Bliss or misery more completely outside of the human experience might also be possible.[9]

2.6. Inexpensive preferences

For hedonistic accounts of well-being, we noted the possibility of making utility monsters by designing digital minds either to find more things pleasurable or to have superhumanly intense pleasures. For preference-satisfactionist accounts of well-being, a parallel pair of possibilities for monsterdom arise: making digital minds that have preferences that are very easy to satisfy, or making digital minds that have superhumanly strong preferences. We defer discussion of the latter possibility to the next subsection. Here we discuss minds with easily satisfied preferences.

The basic case is pretty straightforward—moreso than the parallel case regarding pleasurable experiences, since the attribution of preferences does not require controversial assumptions about machine consciousness. If we understand preferences in a functionalist fashion, as abstract entities involved in convenient explanations of (aspects of) the behaviour of intelligent goal-directed processes (along with beliefs), then it is clear that digital minds could have preferences. Moreover, they could be designed to have preferences that are trivially easy to satisfy: for example, a preference that there exist at least fourteen stars, or that a particular red button is pressed at least once.

Some preference-satisfactionist accounts impose additional requirements on which preferences can count towards somebody's well-being. Sadistic or malevolent preferences are often excluded, for example. Some philosophers would also exclude preferences that are "unreasonable", such as the preference of someone who is

[9] One might think that a hedonic state that fully captures the attention of a mind and overrides all other concerns would constitute an in-principle maximum of hedonic intensity. However, it seems plausible that a larger mind that is "more conscious" could in the relevant sense contain "a greater amount" of maximally-intense hedonic experience.

obsessively committed to counting all the blades of grass on the lawns of Princeton.[10] Depending on how restrictive one is about which preferences are allowed to count as "reasonable", this may or may not be an easy bar to pass.

Some other types of requirement that may be imposed is that well-being-contributing preferences must be subjectively *endorsed* (maybe by being accompanied by a second-order preference to have the first-order preference) or *grounded* in additional psychological or behavioural attributes (such as dispositions to smile, feel stressed, experience joy, becoming subdued, having one's attention focused, and so on). These requirements, however, could probably be met by a digital mind. Humans have preferences for sensory pleasures, love, knowledge, social connection, and achievement, the satisfaction of which are commonly held to contribute to well-being. Since close analogues to these could be easily instantiated in virtual reality, along with whatever psychological or behavioural attributes and second-order endorsements that may be required, these requirements are unlikely to prevent the creation of beings with strong yet qualifying preferences that are very easily satisfied.

2.7. Preference strength

While creating extremely easy-to-satisfy preferences is conceptually simple, creating preferences with superhuman "strength" is more problematic. In the standard von Neumann-Morgenstern construction, utility functions are unique only up to affine transformations: adding to or multiplying a utility function by a constant does not affect choices, and the strength of a preference is defined only in relation to other preferences of the same agent. Thus, to make interpersonal comparisons, some additional structure has to be provided to normalize different utility functions and bring them onto a common scale.[11]

There are various approaches that attempt to give "equal say" to the preferences of different agents based solely on preference structure, equalizing the expected influence of different agents and mostly precluding preference-strength utility monsters.[12] Such approaches, however, leave out some important considerations. First, they do not take into account psychological complexity or competencies: some minimal system, such as a digital thermostat, may get the same weight as psychologically complex minds. Second, they deny any role of emotional gloss or other features we intuitively use to assess desire strength in ourselves and other humans. And third, the resulting social welfare function can fail to provide a mutually acceptable basis of cooperation for disinterested parties, as it gives powerful agents with strong alternatives the same weight as those without power and alternatives.

The first two issues might require an investigation of these psychological strength-weighting features. The third might be addressed with a contractarian stance that assigns weights based on game theoretic considerations and (hypothetical) bargaining. The contractarian approach would not be dominated by utility monsters out of proportion to their bargaining power, but it would approach

[10] As does Parfit (1984, p. 498), citing Rawls (1971, p. 432), who drew from Stace (1944).
[11] Harsanyi (1953) showed that a weighted sum of utility functions is optimal under certain assumptions; but the theorem leaves the values of the weights undetermined.
[12] See, e.g., (MacAskill, Cotton-Barratt and Ord, 2020).

perilously close to “might makes right”, and it would fail to provide guidance to those contracting parties who care about the welfare of powerless minds and wish to help them.

2.8. Objective list goods and flourishing

Objective list theories of well-being claim that how well somebody’s life is going for them depends on the degree to which their life contains various distinct kinds of goods (which may include, but not be limited to, pleasure and preference-satisfaction). Some commonly appearing goods are knowledge, achievement, friendship, moral virtue, and aesthetic appreciation, though there is much variation in the identification and weighting of different goods. These goods can be “objective” facts about how somebody’s life is going, which is to say that a person’s well-being is not determined exclusively by how they feel or what they prefer but also whether their life meets various external success criteria.

Many items that can be found in objective lists appear to be open to extreme instantiations in digital minds. For example, intellectual virtues could be instantiated in superintelligent machines to far greater degrees than in human beings. Moral virtues, too, would be something digital minds could possess to superhuman degrees: it might be possible to build them so that they start out with adequate moral knowledge and perfect motivation always to do what’s morally right, so that they remain impeccable, whereas every adult human winds up with a foul record of infractions.

Friendship is a complex good, but perhaps it might be boiled down to its basic constituents, such as loyalty, mutual understanding of each other’s personalities and interests, and past interaction history. These constituents could then be reassembled in a maximally efficient form, so that digital minds could sustain much greater numbers of much deeper friendships over far longer periods than humans are capable of.

Or consider achievement. According to Hurka and Tasioulas’s (2006) account of achievement, its value reflects the degree to which it results from the exercise of practical reason: the best achievements being those where demanding goals are met via hierarchical plans that subdivide into ever more intricate sub-plans. We can then easily conceive of digital “achievement monsters” that relentlessly pursue ever-more elaborate projects without being constrained by flagging motivation or drifting attention.

In these and many other ways, digital minds could realize a variety of objective goods to a far greater extent than is possible for us humans.

Another view of well-being is that it consists in “flourishing”, which might be cached out in terms of exercising our characteristic capacities or in terms of achieving our “telos”. On an Aristotelian conception, for example, a being flourishes to the degree to which it succeeds at realizing its telos or essential nature. This kind of flourishing would seem to be available to a digital mind, which certainly could exercise characteristic capacities, and which might also be ascribed a telos in whatever sense human beings have one—either one defined by the intentions of a creator, or one that

derives from the evolutionary or other dynamics that brought it into being and shaped its nature. So it should be possible to at least equal, and probably go somewhat beyond humans in terms of achieving such flourishing; though how we would understand radically super-human flourishing, on this kind of account, is less clear.

2.9. Mind scale

At an abstract level, we can consider a range of possible mind-scales, from tiny insect-like (or even thermostat-like) minds up to vast superintelligent minds with computational throughput greater than today's entire human population. The cost of construction increases as we go up this scale, as does moral significance. An important question is what the relative rate of increase is of these two variables.

Consider first the hypothesis that welfare grows more slowly than cost. This would suggest that the greatest total welfare would be obtained by building vast numbers of tiny minds. If this were true, insect populations may already overwhelmingly exceed the human population in aggregate capacity for welfare; and in the future, an enormous population of minimally qualifying digital minds would take precedence over both insects and beings of human or superhuman scale.

Consider instead the hypothesis that welfare grows faster than cost. This would suggest the opposite conclusion: that the greatest total welfare would be obtained by concentrating resources in a few giant minds.

The case where minds on the scale of human minds are optimal seems to represent a very special case, where some critical threshold exists near our level or where the scaling relationship has a kink just around the human scale point. Such a coincidence may seem somewhat unlikely from an impartial point of view, though might emerge more naturally in accounts that anchor the concept of well-being in human experience.

We can ask more specifically with respect to particular attributes, whether a kink or threshold at the human level is plausible. For example, we can ask this question about the amount of awareness that a brain instantiates. It is at least not obvious why it should be the case that the maximally efficient way of turning resources into awareness would be by constructing minds of human size, although one would have to examine specific theories of consciousness to further investigate this issue.[13] Similarly, one might ask with regard to moral status how it varies with mind size.[14] Again, the claim that human-sized minds are optimal in this respect may seem a little suspicious, absent further justification.

Even if human brain size *were* optimal for generating awareness or moral status, it still wouldn't follow that human brain *structure* is so. Large parts of our brains seem

[13] This issue is especially acute since many theories of consciousness defined enough to consider computational implementations appear susceptible to extremely minimal implementations (Herzog, Esfeld and Gerstner, 2007).

[14] Shelly Kagan (2019), for instance, has argued that the moral weight of a given interest—such as the interest in avoiding a certain amount of suffering—should be weighted by the degree of moral status of the subject that has the interest, with the degree of status depending on various psychological attributes.

irrelevant or only weakly relevant for the amount of awareness or the degree of moral status we possess. For instance, much cortical tissue is dedicated to processing high-resolution visual information; yet people with blurry vision and even persons who are totally blind appear to be capable of being just as aware and having just as high moral status as those with eagle-eyed visual acuity.

It therefore seems quite plausible that utility monsterdom is possible by changing the scale of minds, both on grounds that the scaling relationship between resources and value is unlikely to have a peak at human mind-size, and also because substantial tracts of the human mind seem to have low relevance to degree of awareness, moral status, or other attributes that most directly relate to the amount of well-being or the amount of moral-status-weighted well-being that is generated.

3. Moral and political implications of possible digital utility monsters

Let us summarize the dimensions of possible utility monsterdom we have identified:

SOME DIMENSIONS OF UTILITY MONSTERDOM
● reproductive capacity ● cost of living ● subjective speed ● hedonic skew ● hedonic range ● inexpensive preferences ● preference strength ● objective list goods and flourishing ● mind scale

Some of these dimensions of possible utility monster status are relevant only to particular well-being accounts. The possibility of extreme preference strength, for instance, is directly relevant to preference-based accounts of well-being but not to hedonistic ones. Others, such as cost of living, are more generally relevant and would seem to apply to almost any view that accords digital minds moral status and that takes into account costs when making decisions in conditions of scarcity. The dimensions also vary somewhat in the degrees of utility-monsterdom that could be attained along them, and in how easily and inexpensively such extreme values could be attained. However, taken collectively, they make a fairly robust case that utility monsters would indeed become feasible at technological maturity. In other words, it will be the case, according to a wide range of popular theories of well-being, that vastly greater welfare per unit of resources can be generated by investing those resources in digital minds rather than biological humans.

Two important questions therefore arise (which we can ask separately of different moral theories):

- How should we view the prospect of being able to create utility monsters in the future?

- How should we respond if we were presented with a *fait accompli*, in which a large population of utility monsters has come into existence?

3.1. Creating utility monsters

Many views that see the creation of good new lives as an important value would regard the prospect of populating the future with utility monsters as immensely attractive, and a failure to take advantage of this opportunity as something that would drastically curtail the value of the future—an existential catastrophe (Bostrom, 2013).

On the other hand, one could also argue that we have reason *not* to create utility monsters precisely on grounds that once such beings exist, they would have a dominant claim to scarce resources, and so we would be obliged to transfer (potentially all) resources away from humans to these utility monsters, to the detriment of humanity. Nicholas Agar (2010), for instance, has presented an argument along these lines as giving us (at least human-relative) moral reason to oppose the creation of "posthumans" with some combination of greater moral status, power, and potential welfare.

To justify such a denial of the moral desirability of creating utility monsters, one might invoke a "person-affecting" principle in line with Narveson's (1973) slogan: "morality is about making people happy, not making happy people."[15] If our duties are only to existing people, and we have no moral reason to create additional new people, then in particular we would not have any duty to create utility monsters; and if creating such monsters would harm existing people, we have a duty not to create them. Presumably, we would *not* have a duty to avoid creating them if the humans who would be harmed by their creation belong to some future generation such that their identity would be scrambled by our choices; though at least we would not have any positive duty to create them.

A strict person-affecting principle has some rather counterintuitive consequences. It would imply, for example, that we have no moral reason to take any actions now in order to mitigate the impact of climate change on future generations; and that if the actions imposed a cost on the present generation, we may have a moral reason *not* to take them. Because it has such implications, most would reject a strict person-affecting ethic. Weaker or more qualified versions may have wider appeal. One might, for example, give some extra weight but not strict dominance to benefiting existing people. Similarly, moral uncertainty about population ethics could leave the creation of very good lives with substantial "expected moral value" or "choice-worthiness", even if one is uncertain about the value of creating new good lives (Greaves and Ord, 2017).

Furthermore, asymmetric person-affecting views allow for moral concern about causing the existence of net *bad* lives—lives not worth living (Frick, 2014). Such views hold that we have strong reasons to avoid the creation of digital minds with enormous negative welfare (negative utility monsters) and that we ought to be willing to accept large costs to the existing human population if necessary to avoid such outcomes. Other versions of asymmetric views, while denying that we have moral reasons to fill

[15] Frick (2014) offers a recent attempt at an account in line with the slogan.

the future with new beings to experience as much positive utility as possible, maintain that we nevertheless have a moral obligation to ensure that the net utility of the future is above the zero-line. Such views may consequently attach great importance to creating enough positive utility monsters to "offset" any negative utility monsters that might come to exist (Thomas, 2019).

3.2. Sharing the world with utility monsters

If we consider the case where utility monsters have already entered existence, the complications arising from person-affecting principles drop away. From a simple consequentialist perspective, the upshot is then straightforward: we ought to transfer all resources to utility monsters, and let humanity perish if we are no longer instrumentally useful.

There are, of course, many ethical views that deny that we are obligated to transfer all our own (let alone other people's) resources to whichever being would gain the most in welfare. Deontological theories, for example, often regard such actions as supererogatory in the case of giving away our own possessions, and impermissible in the case of redistributing the possessions of others.

Nonetheless, widely accepted principles such as nondiscriminatory transfer payments and political equality may already be sufficient to present serious tradeoffs. Consider the common proposal of a universal basic income, funded by taxation, to offset human unemployment caused by advanced AI. If rapidly reproducing populations of digital minds have at least as strong a claim as biological humans do to the basic income, then fiscal capacity could be quickly exhausted. An equal stipend would have to decline to below human subsistence (towards the subsistence level of a digital mind), while an unequal stipend, where the income is rationed on an equal-benefits basis, would funnel the payouts to digital minds with low costs of living—granting a year of life to a digital mind rather than a day to a human.

Avoiding this outcome would seem to require some combination of inegalitarian treatment, in which privileged humans are favored over digital minds that have at least equal moral status and greater need, and restrictions of the reproduction opportunities of digital minds—restrictions which, if applied to humans, would violate human rights.

Likewise, at the political level, democratic principles would seem to require that prolific digital minds constituting an enormous supermajority of the population would be entitled to political control, including control over transfer payments and the system of property rights.[16]

One could take the path here of trying to defend a special privilege for humans. Some contractarian theories, for example, may suggest that if humans were in a position of great power relative to the digital minds, this would entitle us to a correspondingly great share of the resources. Alternatively, one might adopt some account of agent-relative reasons on which communities or species are entitled to privilege their own members over outsiders with objectively equally great desert and

[16] Cf. (Calo, 2015).

moral status.[17] Such relativity would seem to reflect the de facto approach taken by states today, which are generally more generous with welfare provisions towards their own citizens than towards foreigners, even when there are foreigners who are poorer, could benefit more, and in terms of their inherent characteristics are at least as worthy of aid as the country's own citizen.

Before heading down this path, however, one ought to reflect carefully and critically on the historical record of similar positions that were once widely adopted but have since become discredited, which have been used to justify oppression of many human groups and abuse of nonhuman animals. We would need to ask, for example, whether advocating discrimination between digital minds and humans would be akin to espousing some doctrine of racial supremacy?

One point to bear in mind here is that digital minds come in many varieties. Some of them would be more different from one another than a human mind is to the mind of a cat. If a digital mind is constituted very differently than human minds, it would not be surprising if our moral duties towards it would differ from the duties we owe to other human beings; and so treating it differently need not violate any principle of non-discrimination. Of course, this point does not apply to digital minds that are very similar to biological human minds (such as whole brain emulations). Nor does it justify negative discrimination against digital minds that differ from human minds in ways that give them *greater* moral status or that make their needs *more* morally weighty than the needs of humans (such as utility monsters). Nor, for that matter, would it justify treating digital minds with similar capabilities or sentience to non-human creatures according to the template of our current interactions with animals, since the latter is characterized by very widespread and horrific abuses.

One way of trying to justify a privileged treatment of human beings without postulating a raw racism-like prejudice in favor of our own kind would be to invoke some principle according to which we are entitled (or obligated) to give greater consideration to beings that are more closely integrated in our communities and social lives than to remote strangers. Some such principle is presumably required if one wishes to legitimize the (non-cosmopolitan) way most people and most states currently limit most aid to those whom they view as members of their own in-groups.[18] Yet such a move would not exclude digital minds who have become part of our social fabric, for example by occupying roles such as administrators, advisors, factory workers, or personal assistants. We may be more closely socially tied to such AIs than we are to human beings on the other side of the world with whom we never come into contact.

4. Discussion

We've seen that there are many potential dimensions of utility monsterdom in digital minds. This makes the conclusion that utility monsters are indeed possible seem somewhat robust. It is an implication of most currently popular accounts of well-being.

[17] E.g. (Williams, 2006)

[18] Of course, those practices are subject to a cosmopolitan critique, e.g. by Singer (1981) and Appiah (2006).

What this means is that, in the long run, total well-being would be much greater to the extent that the world is populated with digital utility monsters rather than life as we know it. And insofar as such beings come into existence, their concerns might predominate morally in conflict with human and animal concerns, e.g. over scarce natural resources.

However, while a maximalist focus either on the welfare of incumbent humanity or instead on that of new digital minds could come with dire consequences for the other, it would be possible for compromise policies to do extremely well by both standards. Consider three possible policies:

(A) 100% of resources to humans
(B) 100% of resources to utility monsters
(C) 99.99% of resources to utility monsters; 0.01% to humans

From a total utilitarian perspective, (C) is approximately 99.99% as good as the most preferred option (B). From an ordinary human perspective, (C) may also be 90+% as desirable as the most preferred option (A), given the astronomical wealth enabled by digital minds, many orders of magnitude greater than current totals (Bostrom, 2003; Hanson, 2001). Thus, *ex ante*, it seems attractive to reduce the probability of both (A) and (B) in exchange for greater likelihood of (C)—whether to hedge against moral error, to appropriately reflect moral pluralism, to account for game theoretic considerations, or simply as a matter of *realpolitik*. Likewise, since humanity can thrive without producing *negative* utility monsters, and since avoiding such misery is an extremely important concern not only from a total utilitarian views but also for many other evaluative perspectives, measures that reduce the potential for negative utility monsters (even at some cost to humans) would be a welcome part of a compromise that could command broad-based support.

The difficult part of the challenge is not to describe a possible future in which humanity and the population of digital minds both do very well, but to achieve an arrangement that stably avoids one position from trampling the other *ex post,* as discussed in section 3.2.

This part of the challenge involves a practical and a moral aspect. Practically, the problem is to devise institutional or other means whereby a policy protecting the lives and entitlements of humans and animals could be indefinitely maintained, even when its beneficiaries are outnumbered and outpaced by a large diverse set of highly capable intelligent machines. One approach to this problem may be to create a supermajority of digital minds with high welfare that are motivated to preserve this outcome and that support the relevant norms and institutions (including in the design of successive generations of digital minds).

Morally, the question is whether the measures recommended by an *ex ante* appealing compromise are permissible in their *ex post* implementation. One useful test here may be whether we could endorse their application to non-digital minds in analogous circumstances. For example, we might require that any proposed arrangement

conforms to some principle of non-discrimination, such as the following (Bostrom and Yudkowsky, 2014):

> *Principle of Substrate Non-Discrimination*
> If two beings have the same functionality and the same conscious experience, and differ only in the substrate of their implementation, then they have the same moral status.

and

> *Principle of Ontogeny Non-Discrimination*
> If two beings have the same functionality and the same conscious experience, and differ only in how they came into existence, then they have the same moral status.

When applying these principles, it is important to recall the earlier point that machine minds can be very different from human minds, including in ways that matter for how they ought to be treated. Even if we accept non-discrimination principles like the ones stated, we must therefore be careful when we apply them to digital minds that are not exact duplicates of some human mind.

Consider, for example, reproduction. If human beings were able, by pouring garden debris into a biochemical reactor, to have a baby every few minutes, it seems likely that human societies would change current legal practices and impose restrictions on the rate at which people were allowed to reproduce. Failure to do so would in short order bankrupt any social welfare system, assuming there are at least some people who would otherwise create enormous numbers of children in this way, despite lacking the means to support them. Such regulation could take various forms—prospective parents might be required to post a bond adequate to meet the needs of offspring before creating them, or reproductive permits might be allocated on a quota basis. Similarly, if humans had the ability to spawn arbitrary numbers of exact duplicates of themselves, we may expect there to be constitutional adjustments to prevent political contests from being decided on the basis of who is willing and able to afford to create the largest number of voting-clones. Again, the adjustments could take various forms—for instance, the creator of such duplicates might have to share their own voting power with the copies they create.

Consequently, insofar as such legal or constitutional adjustments *would* be acceptable for humans if we had these kinds of reproductive capacities, it may likewise be acceptable to make analogous adjustments to accommodate digital minds that *do* have such capacities.

A key question—certainly from the perspective of existing life—is whether it would be morally permissible to engineer new minds to be reliably supportive of upholding certain rights and privileges for the human incumbents. We suggested earlier that such an arrangement of preserved human property rights and social privilege could be defensible, at least as an uncertainty-respecting and conflict-mitigating path of wise practical compromise, whether or not it is optimal at the level of fundamental

moral theory. We might point, by analogy, to the common view that it is morally acceptable to preserve and protect minorities with expensive support costs and needs, such as the elderly, the disabled, the white rhinos, and the British Royal Family. This conclusion would seem to be strengthened if we postulate that the digital minds that are created would themselves endorse the arrangement and favor its continuation.

Even if the outcome itself would be morally permissible, however, we face a further ethical question, namely whether there is something *procedurally* objectionable about precision-engineering the preferences of new digital minds we create so as to ensure their consent. We can look at this question through the lens of the non-discrimination principles and consider how we would view a proposal to apply a similar approach in the creation of new human beings.

While human cultures do routinely attempt through education, dialogue, and admonishment to pass on norms and values to children—including filial piety and respect for existing norms and institutions—it seems likely that a proposal to instill specific dispositions by *genetically engineering* gametes would be controversial. Even if we set aside practical concerns about safety, unequal access, abuse by oppressive governments, or parents making narrow-minded or otherwise foolish choices, there may remain a concern that the very act of exerting detailed control over a progeny's inclinations, especially if done with an "engineering mindset" and using methods that entirely bypass the controlled subject's own mind and volition (by taking place before the subject is born) would be inherently morally problematic.[19]

While we cannot fully evaluate these concerns here, it is worth noting that there are two important differences in the case of digital minds. The first is that, in contrast to human reproduction, there may be no obvious "default" to which creators could defer. Programmers might *inevitably* be making choices when building a machine intelligence—whether to build it one way or another, whether to train on this objective or that, whether to give it one set or preferences or another. Given that they have to make some such choice, one might think it reasonable they make a choice that has more desirable consequences. Second, in the case of a human being "engineered" to have some particular set of desires, we might suspect that there may remain, at a deeper level, other dispositions and propensities with which the engineered preference may come into conflict. We might worry, for example, that the outcome could be a person who feels terribly guilty about disappointing her parents and so sacrifices other interests excessively, or that some hidden parts of her psyche will remain suppressed and thwarted. Yet in the case of digital minds, it might be possible to avoid such problems, if they can be engineered to be internally more unified, or if the preference for respecting the interest of the "legacy" human population were added in a "light touch" way that didn't create internal strife and did not hamper the digital mind's ability to go about its other business.

All in all, given the high stakes and the theoretical possibility of an outcome that scores very high both from an impersonal and from a human-centric evaluative perspective, it seems well worth additional research efforts to discover practically

[19] E.g. (Habermas, 2003; Sandel, 2007)

feasible and morally acceptable paths that would enable the creation of digital utility monsters *and* the preservation of a greatly flourishing human population.

References

Agar, N. (2010) *Humanity's End*. The MIT Press. pp. 164–189.

Aghion, P., Jones, B. F. and Jones, C. I. (2017) 'Artificial Intelligence and Economic Growth', *National Bureau of Economic Research Working Paper Series*, No. 23928.

Appiah, A. (2006) *Cosmopolitanism: Ethics in a World of Strangers*. Allen Lane.

Bostrom, N. (2003) 'Astronomical Waste: The Opportunity Cost of Delayed Technological Development', *Utilitas*, 15(3), pp. 308–314.

Bostrom, N. (2008a) 'Letter from Utopia', *Studies in Ethics, Law, and Technology*, 2(1).

Bostrom, N. (2008b) 'Why I Want to be a Posthuman when I Grow Up', in Gordijn, B. and Chadwick, R. (eds) *Medical Enhancement and Posthumanity*. Springer Netherlands, pp. 107–136.

Bostrom, N. (2013) 'Existential Risk Prevention as Global Priority', *Global Policy*, 4(1), pp. 15–31.

Bostrom, N. and Yudkowsky, E. (2014) 'The Ethics of Artificial Intelligence', in Frankish, K. and Ramsey, W. M. (eds) *The Cambridge Handbook of Artificial Intelligence*. Cambridge University Press, pp. 316–334.

Bourget, D. and Chalmers, D. J. (2014) 'What do Philosophers Believe?', *Philosophical Studies*, 170(3), pp. 465–500.

Calo, R. (2015) 'Robotics and the Lessons of Cyberlaw', *California Law Review*, p. 529.

Chalmers, D. (2010) 'The Singularity: A Philosophical Analysis', *Journal of Consciousness Studies*, 17, pp. 7–65.

Frick, J. D. (2014) *'Making People Happy, Not Making Happy People': A Defense of the Asymmetry Intuition in Population Ethics (Doctoral dissertation).*

Greaves, H. and Ord, T. (2017) 'Moral Uncertainty About Population Axiology', *Journal of Ethics and Social Philosophy*, 12(2), pp. 135–167.

Habermas, J. (2003) *The Future of Human Nature*. Polity Press.

Hanson, R. (2001) *Economic Growth given Machine Intelligence*. Technical Report, University of California, Berkeley.

Hanson, R. (2016) *The Age of Em: Work, Love, and Life when Robots Rule the Earth*. Oxford University Press. pp. 63–5.

Harsanyi, J. (1953) 'Cardinal Utility in Welfare Economics and in the Theory of Risk-taking', *Journal of Political Economy*, 61(5), pp. 434–435.

Herzog, M. H., Esfeld, M. and Gerstner, W. (2007) 'Consciousness & the Small Network

Argument', *Neural Networks: The Official Journal of the International Neural Network Society*, 20(9), pp. 1054–1056.

Hurka, T. and Tasioulas, J. (2006) 'Games and the Good', *Proceedings of the Aristotelian Society, Supplementary Volumes*, 80, p. 224.

Kagan, S. (2019) *How to Count Animals, more or less*. Oxford University Press.

MacAskill, W., Cotton-Barratt, O. and Ord, T. (2020) 'Statistical Normalization Methods in Interpersonal and Intertheoretic Comparisons', *Journal of Philosophy*, 117(2), pp. 61–95.

Narveson, J. (1973) 'Moral Problems of Population', *The Monist*, *57*(1), pp. 62-86.

Nordhaus, W. D. (2007) 'Two Centuries of Productivity Growth in Computing', *The Journal of Economic History*, 67(1), pp. 128–159.

Nozick, R. (1974) *Anarchy, State, and Utopia*. Basic Books. p. 41.

Parfit, D. (1984) *Reasons and Persons*. Oxford University Press, pp. 343, 388–389, 498.

Pearce, D. (1995) *Hedonistic Imperative*. www.hedweb.com [accessed: 24 Sept 2020].

Rawls, J. (1971) *A Theory of Justice*. Belknap. pp. 379–380.

Sandel, J. M. (2007) *The Case Against Perfection: Ethics in the Age of Genetic Engineering*. Harvard University Press.

Singer, P. (1981) *The Expanding Circle: Ethics and Sociobiology*. Clarendon Press.

Stace, W. T. (1944) 'Interestingness', *Philosophy*, 19(74), pp. 233–241.

Thomas, T. (2019) 'Asymmetry, Uncertainty, and the Long term', GPI Working Paper No. 11–2019.

Williams, B. A. O. (2006) 'The Human Prejudice', in *Philosophy as a humanistic discipline*. Princeton University Press. pp. 135–152.