与数字心灵共享世界[1]

(2020) 草稿版本 1.8

作者：Carl Shulman 和 Nick Bostrom，Future of Humanity Institute, University of Oxford.

译者：朱小虎 Xiaohu Zhu，Center for Safe AGI

[in Clarke, S., Zohny, H. & Savulescu, J. (eds.): Rethinking Moral Status

(Oxford University Press, 2021)]

www.nickbostrom.com

摘要

一但人类深度掌握了人工智能技术，更加庞大的可能的心灵的空间将会被创建出来，而生物心灵只可占据这一空间的一个很小角落。然而，我们的许多道德直觉和实践都基于对数字心灵来说不需要的那个人类本性假设。这表明当我们接近先进的机器智能时代时，需要进行道德反思。本文我们关注问题的一个集合，这些问题源于数字心灵对资源和影响力拥有超人类的强烈主张的前景。这些可能源于大规模生产的数字心灵可以从相对较少的资源中获得的巨大集体利益。或者，它们可能来自具有超人类道德地位或从资源中受益的能力的个体数字心灵。这样的人可以为世界贡献巨大的价值，不尊重他们的利益可能会导致道德灾难，而一种尊重他们的幼稚方式可能会给人类带来灾难性的后果。一种明智的方法要求我们变革现有的道德规范和制度，并提前规划人类将带来什么样的数字心灵。

关键术语表

序号	英文	中文
1	mind	心灵
2	superhuman	超人类

[1] 感谢 Guy Kahane、Matthew van der Merwe、Hazem Zohny、Max Daniel、Lukas Finnveden、Lukas Gloor、Uli Alskelung Von Hornbol、Daniel Dewey、Luke Muehlhauser、James Babcock、Ruby Bloom、Vincent Luzkow、Nick Beckstead, Hilary Greaves, Owen Cotton-Barratt, Allan Dafoe, and Wes Cowley.

3	resource	资源
4	reform	变革
5	being	存在物
6	awareness	察觉
7	realpolitik	现实政治

1. 引言

人类的生物本性对促进某人福利的行为施加了许多实际限制。我们只能活那么长时间，感受到那么多的快乐，有那么多的孩子，以及从额外的支持和资源中获益那么多。同时，人类为了繁荣还需要满足一系列复杂的生理、心理和社会条件。

然而，这些限制对其他存在物来说可能放松。考虑具有有意识的经验、愿望以及推理和自主决策能力的机器心灵的可能性。[2]这样的机器可以享有道德地位，即它们不仅仅是人类的工具，它们及其利益本身就很重要。它们在从额外资源中受益的能力上既不需要受到同样的实际限制，也不需要依赖与人类那样复杂的生存和繁荣需求。这可能是一个美妙的发展：不再有痛苦和疾病，处处洋溢幸福，充满了超人类的察觉和理解以及各种更高尚的善。[3]

机器学习的最新进展提出了这样一种前景，即这种数字心灵可能在可预见的未来成为现实（或者说现在可能在非常有限的程度上业已存在）。这些心灵中的一部分可以使 Robert Nozick (1974, p. 41) 著名的"效用怪物"哲学思想实验变为现实：

> 功利主义理论对效用怪物的可能性感到尴尬，他们从其他人的任何牺牲中获得的效用总和比其他人大得多失去。因为，令人无法接受的是，该理论似乎要求我们所有人都被牺牲在怪物的喉咙中，以增加总效用。

Derek Parfit（1984，第343页）认为，虽然很难想象一个生活数百万倍之多值得活的最佳断人，类似的结果可以通过考虑的定量维度获得*人口规模*，其中显然没有极端值的概念障碍。

我们将争辩说，人口规模只是几个定量维度中的一个——连同几个不太确定的定性维度——在这些维度上，数字心灵可能在每单位资源消耗中获得的收益方面远远超过人类。这些多条路径得出的结论是，至少有一条会更加可靠地实现。

---

[2] 我们假设适当架构的人工智能可能是有意识的，但值得注意的是，一些关于道德地位的解释并不认为这是拥有道德地位的必要条件；参见 (Chalmers, 2010) 对人工智能意识的讨论，以及 (Kagan, 2019) 对无意识但代理式的人工智能中道德地位的讨论。

[3] 其中一些可能至少部分适用于增强了或上传了的人类；（Bostrom，2008a，2008b；Chalmers , 2010).

虽然非功利主义者可能认为自己不受效用怪物挑战的影响，但大多数合理的观点实际上在不同程度上容易受到影响。这是因为即使我们假设不会发生道义上的违反，人类的利益仍然可能受到效用怪物的出现的不利影响，因为后者可能对国家援助或自然资源和其他稀缺资源有更强的道德要求，从而减少人类可以合理要求的数量。从公正的角度来看，具有这些特性的数字心灵可以使世界在道德上更有价值，同时也使共同规范对现有存在（或实际上任何优化程度较低的数字或其他心灵）的要求更高。

2. 实现超受益者的途径

虽然"效用怪物"有学术的历史，它是一个轻蔑且潜在侵略式的方式 是指具有异常巨大的需求，或者是能够实现非常好的生活众生的轻蔑和潜在的进攻方式因此，我们将转而采用以下术语：

> *超受益者*：一个在从资源中获得福祉的超人效率的存在物
>
> *超耐心者*[4]：一个具有超人道德状态的存在物

"效用怪物"一词含糊不清，但可能最接近于"超受益者"。一些观点认为，道德地位以不同于利益强度的方式进入道德要求的计算，例如作为整体乘数或通过产生一组不同的职责或义务论约束。例如，Shelly Kagan (2019) 认为，特定利益的道德权重——例如避免一定痛苦的利益——应该由具有利益的主体的道德地位程度加权，其中地位的程度取决于各种心理属性和潜力。如果一个存在的利益应该比人类的利益得到更多的道德考虑，不是因为利益更强大，而是因为它具有更高的道德地位，那么这个存在将是我们术语中的超耐心者。

超耐心者状态的可能性是有争议的：一些人声称人类拥有不能被超越的"完全道德状态"，而其他人（例如 Kagan）则认为超耐心者状态是可能的，因为赋予人类道德的心理能力地位承认超人的程度。在本文中，我们将主要探索通往超受益者地位的路径，这可能与争议较小的假设相结合，即数字心灵可以具有至少与人类相同的道德地位，从而产生极端的道德主张。

2.1.繁殖能力

计算机软件最基本的特征之一是只要在计算机硬件可获取的情况下精确繁殖的简便性和速度。只要其经济产出能够支付制造成本，硬件就可以快速构建（从历史上看，这些成本在性价比基础上大幅下降；Nordhaus，2007 年）。这为人口动态打开了大门，人口动态需要几个世纪才能在人类中发挥作用，然后被压缩到人类一生的一小部分。即使*最初*有一定的知识能力的只有几个数字的头脑可以经济地建造，这种头脑的数量可能很快就会成倍或超指数增长，直到其他条件的限制。这种爆炸性的生殖潜力可以让数字心灵在相对较短的时间内远远超过人类——相应地增加他们主张的集体力量。

此外，如果数字心灵和所需硬件的生产继续进行，直到由此产生的心灵的工资等于边际成本，这可能会推动工资向机器生存水平下降，因为自然资源成为一个限制因素。这些可能不足以让人类（和过时的数字心灵）继续生存（Hanson，2001 年；Aghion、Jones 和 Jones，2017 年）。这种情况使得再分配问题变得更加紧迫——生死攸关——而马尔萨斯式的人口增长将使转移支付的要求实际上无法满足。

---

[4] 感谢 Daniel Dewey 提出这个术语。

快速和廉价繁殖的另一个重要方面是它允许人口的快速周转。被删除的数字心灵可以立即被最新版本的完全成熟思维的副本所取代——这与人类情况形成鲜明对比，后者需要 9 个月才能产生一个流口水的婴儿。[5]因此，经济压力可能会推动非常频繁地清除"过时"的思想，并用相同硬件产生更多经济价值的思想取而代之。

因此，适用于数字心灵的当前软件实践的合理延续可能涉及极其大量的短暂生命和死亡，即使是在任何给定时间存在的思维数量的一小部分。这种短暂的数字头脑可心理上的成熟，按时间顺序年轻，长*的潜力*寿命尚未在没有补贴的很短的默认预期寿命。如果我们认为在能够长寿的情况下英年早逝是一种巨大的剥夺，或者当其他人能够长寿时是非常不公平的，那么这可能会为这些数字心灵提供特别强烈的资源来延长他们的寿命（或其他形式的补偿）。如果死亡本身是一件坏事（而不仅仅是已逝生命的机会成本），那么这种思想的快速转变也可能会增加这种每生命年价值贬值的程度。

2.2.生活成本的生活水平

许多数字心灵将需要更少的收入来维持给定这似乎是合理的。支持数字心灵的计算机硬件成本可能会远低于支持人类大脑和身体的成本。如果我们超越单纯的生存来看，适合人类消费的实物商品和服务（例如住房和交通）往往比信息技术和虚拟商品更昂贵，以满足数字心灵的同等需求。数字心灵也不需要受到恶劣的环境条件、污染、疾病、生物老化或任何其他压抑人类福祉的影响。

因此，为类似人类的数字心灵生产给定数量的（质量调整的）存活的年的成本可能远低于生物人的等价成本。生活成本的巨大差异意味着，当出现分配问题时，对人类带来微小利益的资源可能会给许多数字心灵带来巨大利益。如果维持一个人的生命一个月所需的能量预算可以维持十个数字心灵一年，这将成为在资源稀缺情况下支持后者的有力论据。

2.3.主观速度

具有更高串行速度的硬件可用于更快地运行数字心灵。当前的计算机时钟速度以吉赫为单位，比人类神经元的放电率高数百万倍；并且信号传输速度同样可以超过人类神经的传导速度。因此，如果有足够的硬件供应，具有类人能力的数字心灵很可能比人类思考速度至少快数千倍（甚至数百万倍）。如果数字心灵将数千年的主观生命年打包到一个日历年中，那么前者（"主观时间"，而不是挂钟时间）似乎是衡量诸如此类通过延长寿命（Bostrom 和 Yudkowsky，2014 年）来获得幸福感的正确衡量标准。

由于加速需要为更多硬件支付费用，因此这为个人数字心灵提供了一种方式，可以从财富中获得比人类通常更高的回报（每美元的主观寿命年）。在低速下，数字心灵可获得的收益将接近线性；尽管随着速度接近技术极限，进一步提高速度的边际成本会上升。[6]

因为运作得更快的这些收益可以累积到当时存在的最初跑得更慢的个人身上，这种影响与采用"影响人"方法的人口价值论尤其相关（稍后会详细介绍）。

---

[5] 然而，可能不清楚的是，现有心灵的精确或几乎精确的副本是否会构成一个新的独特的人，或者是作为模板的人的额外实例。

[6] Hanson (2016, pp. 63-65) 认为，随着加速的成本增加最初是接近线性的，即 2 倍加速需要接近 2 倍的硬件预算，最终达到超人类类的速度。

2.4.享乐偏差

有理由认为，经过改造的大脑可以享受更长的持续时间和更强烈的快感。人类心理已经进化到产生快乐和痛苦，其中这种动机行为与过去几代人的生殖健康相关，而不是为了最大限度地提高幸福感。这给我们带来了很多难以避免的痛苦。与此同时，我们的享受却很少。美食享乐受饥饿调节，性享乐受性欲调节。从相对地位或对他人的权力中获得的快乐在结构上是稀缺的。大多数奖励也受到无聊和容忍等机制的调节，这些机制会逐渐减少从重复刺激或持续良性条件中获得的愉悦感。对于数字心灵，可以放宽这些限制，以允许可持续的强烈快乐，同时从当前人类生存的痛苦部分中解放出来。

人类的享乐平衡也可以通过先进的技术得到极大的改善，这些先进技术可能先于或紧随成熟的机器智能技术。[7]然而，从根本上调整了生物人类享乐平衡可能比做同样更“昂贵的*从头*数字头脑”，在几个方面：（a）需要脑外科手术干预，广泛的药理精细调音和操作或等价物，至少在近期内，可能是不可行或昂贵的；(b) 对我们的心理进行更彻底的转变可能会破坏我们目前珍视的人性中的个人身份或其他属性。[8]因此，有知觉机器的心灵设计在实现享乐价值状态的效率方面具有巨大优势。

2.5.范围

享乐除了改变目前人类可访问的享乐规模的不同部分所花费的时间比例之外，还有可能——更具推测性——设计可以实现享乐并“超乎寻常”状态的数字心灵——*存在*——人类大脑完全无法实例化的幸福水平。

进化考虑为这一假设提供了一些支持。就快乐和痛苦的强度对应于行为反应的强度而言，进化应该倾向于调整享乐体验，以产生近似适应度最大化的努力来实现或避免它们。但对于人类，一般是很容易*失去*大量的生殖健康在很短的时间，而不只是*获得*等量的。在火中停留片刻可能会导致永久性伤害或死亡，代价是生物体剩余的所有繁殖机会。没有任何一餐或性行为每秒有这么大的风险——它需要数周才能饿死，而且每分钟交配产生的繁殖后代的预期数量很少。因此，进化可能需要产生更强烈的刺激性每秒疼痛来应对伤害，而不是对积极事件的愉悦。相比之下，经过精心设计的大脑可以被精心设计来体验快乐，就像最糟糕的折磨是无益的一样。更完全超出人类体验的幸福或痛苦也是可能的。[9]

2.6.廉价的偏好

对于幸福的享乐主义描述，我们注意到通过设计数字心灵来寻找更多令人愉悦的事物或拥有超人般强烈的快乐，从而使超受益者成为可能。对于幸福的偏好满足主义解释，出现了一对平行的可能性：制造具有很容易满足的偏好的数字心灵，或者制造具有超人强烈偏好的数字心灵。我们将后一种可能性的讨论推迟到下一小节。在这里，我们讨论具有容易满足的偏好的思想。

---

[7] David Pearce (1995) 认为，生物心灵可以设计成在“幸福的梯度”上运行，而不是在当前的整个痛苦-快乐跨度上运行。

[8] 参见（Agar，2010, pp 164-189）

[9] 人们可能会认为，一种完全吸引心灵注意力并凌驾于所有其他关注点的享乐状态将构成原则上最大的享乐强度。然而，在相关意义上，一个“更有意识”的更大的心灵可能包含“更多”的最大强度的享乐体验，这似乎是合理的。

基本案例非常简单——比关于愉快体验的平行案例更简单，因为偏好的归因不需要关于机器意识的有争议的假设。如果我们以功能主义的方式将偏好理解为抽象实体，这些实体涉及对智能目标导向过程（以及信念）的行为（的方面）的方便解释，那么很明显数字心灵可能具有偏好。此外，它们可以被设计成具有非常容易满足的偏好：例如，至少存在十四颗星的偏好，或者特定的红色按钮至少被按下一次。

一些偏好满足主义的解释强加了额外的要求，即偏好可以计入某人的幸福。例如，虐待狂或恶意偏好通常被排除在外。一些哲学家还排除了"不合理"的偏好，例如痴迷于计算普林斯顿草坪上所有草叶的人的偏好。[10]根据一个人对哪些偏好算作"合理"的限制，这可能是也可能不是一个容易清除的障碍。

有些其它类型的可施加的要求是福祉-贡献偏好必须主观地*赞同*（可能由伴随着一个二阶偏好具有一阶偏好）或 *接地*在附加的心理或行为属性——例如微笑的倾向、感到压力、体验快乐、变得柔和、注意力集中等。数字心灵可能会满足这些要求。人类对感官愉悦、爱、知识、社会联系和成就有偏好，而这些满足感通常被认为有助于幸福。由于可以在虚拟现实中轻松实例化与这些非常相似的事物，以及可能需要的任何心理或行为属性和二阶认可，因此这些要求不太可能阻止具有强烈但合格的偏好且很容易满足的生物的创造。

2.7.偏好强度

虽然创造极易满足的偏好在概念上很简单，但创造具有超人类"力量"的偏好则更成问题。在标准的 von Neumann-Morgenstern 构造中，效用函数仅在仿射变换之前是唯一的：将效用函数添加或乘以常数不会影响选择，并且偏好的强度仅在与其他偏好相关的情况下定义同一个代理。因此，为了进行人际比较，必须提供一些额外的结构来规范不同的效用函数并将它们带到一个共同的尺度上。[11]

有多种方法试图仅基于偏好结构对不同代理的偏好给予"平等的发言权"，均衡不同代理的预期影响，并主要排除偏好强度的超受益者。[12]然而，这种方法忽略了一些重要的考虑因素。首先，他们没有考虑心理复杂性或能力：一些最小的系统，例如数字恒温器，可能会获得与心理复杂的头脑相同的权重。其次，他们否认我们直觉上用来评估我们自己和其他人的欲望强度的情感光泽或其他特征的任何作用。第三，由此产生的社会福利函数可能无法为无利害关系的各方提供相互可接受的合作基础，因为它赋予具有强大选择权的强大代理与没有权力和选择权的代理相同的权重。

前两个问题可能需要对这些心理强度加权特征进行调查。第三个可以通过契约主义的立场来解决，该立场基于博弈论的考虑和（假设的）讨价还价来分配权重。契约主义的方法不会被与其议价能力不成比例的超受益者所主导，但它接近于"可能会正确"，它无法为那些关心弱势群体并希望分配的缔约方提供指导不考虑接受者的议价能力。

2.8 客观的商品清单和蓬勃发展

福利的要求，有人的生命是如何去为他们依赖于到他们的生活中包含了各种不同类型的商品（which may include pleasure and preference 满意程度目标清单理论 inter alia）。一

[10] Parfit (1984, p. 498) 也是如此，引用了来自 Stace (1944) 的 Rawls (1971, p. 432)。
[11] Harsanyi (1953) 表明效用函数的加权总和在某些假设下是最优的，但该定理未确定权重的值。
[12] 例如（MacAskill、Cotton-Barratt 和 Ord，2020）

些常见的物品是知识、成就、友谊、道德和审美，尽管不同物品的识别和权重有很大差异。这些理论的共同点是，它们包括的项目对幸福的贡献并不完全取决于主体的态度、感受和信念，但还要求满足某些外部成功标准。

目标列表中的许多项目都可以进行极端实例化。例如，超级智能机器可以培养超出人类范围的智力美德。道德美德也可以达到超人类的水平：数字心灵可以以广泛的道德知识和完美的动机开始生命，总是做道德上正确的事情，这样他们就可以保持无可挑剔的无罪，而每个成年人最终都会有犯规的记录。

友谊（friendship）是一种复杂的善，但也许可以归结为它的基本组成部分，例如忠诚、对彼此个性和兴趣的相互理解以及过去的交往历史。然后，这些成分可以以最大效率的形式重新组合，这样数字心灵也许可以在比人类更长时间的时间内维持更多更深层次的友谊。

或者考虑成就（achievement）。根据 Hurka 和 Tasioulas (2006) 对成就的描述，它的价值反映了它从实践理性中产生的程度：最好的成就是那些通过分层计划实现具有挑战性的目标的成就，这些计划细分为越来越复杂的子计划。然后，我们可以很容易地设想数字"超级成就者"，他们不懈地追求更精细的项目，而不会受到低落的动机或转移注意力的限制。

通过这些和许多其他方式，数字心灵可以在比我们人类更大的程度上实现各种客观商品。

幸福的另一种观点是"繁荣"，这可以通过锻炼我们的特征能力或实现我们的"目的"来兑现。例如，在亚里士多德的概念中，一个存在的繁荣达到它成功实现其目的或本质的程度。这种繁荣似乎适用于数字心灵，它当然可以行使特有的能力，并且也可能被归为人类所拥有的任何意义上的终极目标——要么是由创造者的意图定义的，要么是源于使其形成并塑造其性质的进化或其他动力。所以至少应该可以达到这样的繁荣程度，甚至可能在一定程度上超越人类；尽管在这种情况下，我们如何理解彻底超人类的繁荣尚不清楚。

2.9. 思维尺度

在抽象层面上，我们可以考虑一系列可能的思维尺度，从微小的昆虫状（甚至恒温器状）思维到计算吞吐量超过当今整个人类人口的巨大超级智能思维。随着我们达到这个规模，建设成本会增加，道德意义也是如此。一个重要的问题是这两个变量的相对增长率是多少。

首先考虑福利增长比成本增长更慢的假设。这表明通过建立大量微小的思想可以获得最大的总福利。如果这是真的，昆虫种群的总体福利能力可能已经压倒性地超过了人类；数量庞大的最低限度合格的数字心灵将优先于昆虫和人类或超人类规模的生物。

相反，考虑福利增长快于成本的假设。这将暗示相反的结论：通过将资源集中在少数几个巨人的头脑中，将获得最大的总福利。

在人类思维尺度上的思维是最佳的情况似乎代表了一个非常特殊的情况，在我们的水平附近存在一些临界阈值，或者尺度关系在人类尺度点附近有一个扭结。从公正的角

度来看，这种巧合似乎不太可能，尽管它可能会更自然地出现在将幸福概念锚定在人类经验或人性中的叙述中。

我们可以更具体地询问特定属性，人类层面的扭结或阈值是否合理。例如，我们可以问这个关于大脑实例化的意识量的问题。至少不清楚为什么将资源转化为意识的最有效方法是构建人类规模的思想，尽管人们必须检查特定的意识理论以进一步研究这个问题。[13]同样，人们可能会问道德地位如何随心智大小而变化。再一次，声称人类大小的头脑在这方面是最佳的可能看起来有点可疑，没有进一步的理由。

即使人类大脑尺寸*是*用于发电的意识或道德状况最佳，它仍然不会遵循人类大脑*的结构*是如此。我们大脑的很大一部分似乎与我们拥有的意识数量或道德地位程度无关或只有微弱的相关性。例如，许多皮质组织专门用于处理高分辨率视觉信息；然而，视力模糊的人，甚至完全失明的人，似乎也能像鹰眼视力的人一样有意识和道德地位。

因此，超受益者地位似乎可以通过不同规模的工程思维来实现，这既是因为资源和价值之间的缩放关系不太可能在人类思维规模达到峰值，而且

因为人类心灵的大部分区域与意识程度、道德状态或其他与幸福程度或产生的道德状态加权幸福程度最直接相关的属性的相关性较低。

3. 数字超受益者的道德和政治影响

让我们总结一下数字心灵可以通过超人类力资源效率实现福利的维度：

通往超人类福利的一些途径
• 繁殖能力 • 生活费用 • 主观速度 • 享乐偏斜 • 享乐范围 • 物美价廉的喜好 • 偏好强度 • 目标清单商品和繁荣 • 心灵量表

其中一些维度仅与特定的幸福感相关。例如，极端偏好强度的可能性与基于偏好的账户直接相关，但与享乐主义账户无关。其他因素，例如生活成本，更普遍相关，似乎适用于几乎所有赋予数字心灵道德地位并在稀缺条件下做出决策时考虑成本的观点。这些维度也有所不同，它们可以增加幸福感的程度，以及如何轻松和廉价地获得这种极端值。然而，总的来说，他们提出了一个相当有力的案例，即超受益者确实会在技术成熟时变得可行。换句话说，根据广泛的流行福祉理论，将这些资源投资于数字心灵而不是生物人类，可以产生更大的单位资源福利。

[13] 这个问题尤其严重因为许多足以考虑计算实现的意识理论似乎容易受到极小实现的影响（Herzog，Esfeld 和 Gerstner，2007）。

因此出现了两个重要的问题（我们可以分别询问不同的道德理论）：

我们应该如何看待未来能够创造超受益者的前景?

我们应该如何应对，如果我们提出了一个*既成事实*，在这种超受益者，或许蜂拥而至，已生效的存在?

3.1.打造超受益者

许多将创造美好新生活视为重要价值的观点认为，让超受益者在未来居住的前景极具吸引力，而未能利用这一机会将大大削弱未来的价值 —— 一场存在性的灾难（Bostrom，2013）。

在另一方面，我们也可以说，我们有理由*不*创建超受益者正是理由是一旦这种生物存在，他们将有一个占主导地位的要求，以稀缺资源，从那里我们将被迫转移（可能是所有）资源从人类转移到这些超受益者，从而损害人类。Nicholas Agar（2010）提出了一个沿着这些思路的论点，为我们（至少是相对于人类的）道德理由反对创造具有更高道德地位、权力和幸福潜力的"后人类"的创造。

为了证明这种否认创造超受益者的道德愿望是合理的，人们可以援引符合 Narveson (1973) 口号的"影响人"的原则，"道德是让人快乐，而不是让快乐的人。"[14] 如果我们只对现有的人负责，而我们没有道德理由去创造更多的新人，那么我们就没有任何创造超受益者的义务；如果创造这样的超受益者会伤害现有的人，我们有责任不创造他们。据推测，我们*不*会有责任避免创建超受益者，如果谁还会因此而属于某个下一代伤害人类，这样我们的选择会改变其人类进入存在的"蝴蝶效应";但至少我们不会有任何积极的责任在这种观点上创造超受益者。

然而，严格的影响人的方法会产生一些相当违反直觉的后果。例如，这意味着我们现在没有道德理由采取任何行动来减轻气候变化对后代的影响；并且，如果行动强加给本产生的成本，我们可能有道义理由*不*把他们。因为它有这样的含义，大多数人会拒绝严格的影响人的伦理。较弱或更合格的版本可能具有更广泛的吸引力。有人可能，例如，给*一些*额外的重量，但没有严格的主导地位，以造福现有人。

类似的结果，在这里我们有*一些*道德理性创建，即使现有的人类被给予特别考虑超受益者，可以由考虑到对人口伦理学（Greaves 和 Ord，2017）道德的不确定性而出现。靠的是不确定性怎么这么处理的，一个可能*要么*得到的结论是，最"的选择，值得"，当然动作是把钱花在创建超受益者的所有资源，即使人认为这是不可能的，这将在事实上是资源的最佳利用；*或*（更可能是在我们看来）的行为就选择最值得的当然是预留至少一些资源用于现有人类的利益，即使一个人认为这可能是它实际上将更好地使用所有的资源创造超受益者。

另一种方法是，允许大约造成净的存在道德关怀不对称的人，影响的观点为代表*坏*的生活，生活不值得过（Frick，2014）。这些观点认为，我们有充分的理由避免创造具有巨大负面福利的数字心灵，并且我们应该愿意接受现有人口的巨大成本以避免这种结果。其他版本的不对称观点，同时否认我们有道德。

[14] Frick (2014) 提供了符合口号的最近尝试。

用新生命填充未来以体验尽可能多的积极效用的理由，坚持认为我们仍然有道德义务确保未来的净效用高于零线。因此，这种观点可能非常重视创造足够多的积极的超受益者来"抵消"未来生物的无用性（Thomas，2019）。

3.2.与超受益者共享世界

如果我们考虑到超受益者已经存在的情况，那么影响人的原则所产生的复杂性就消失了。从简单的功利主义角度来看，假设完全遵守，那么结果很简单：我们应该将所有资源转移给超受益者，如果我们不再具有工具实用性，就让人类灭亡。

当然，有许多伦理观点否认我们有义务将我们自己（更不用说其他人）的所有资源转移给任何能够获得最大福利的人。例如，义务论通常认为，在放弃我们自己的财产的情况下，这种行为是多余的，而在重新分配他人财产的情况下，这种行为是不允许的。

尽管如此，诸如非歧视性转移支付、政治平等和生殖自由等被广泛接受的原则可能已经足以做出严重的权衡。考虑由税收资助的普遍基本收入的共同提议，以抵消先进人工智能造成的人类失业。如果快速复制的数字心灵群体对基本收入的要求至少与生物人类一样强烈，那么财政能力可能会很快耗尽。平等的津贴将不得不降至低于人类的生存水平（朝着数字心灵的生存水平），而不平等的津贴，即收入在平等利益的基础上配给，将以低成本将支出转移到数字心灵生活——赋予数字心灵一年的生命，而不是人类的一天。

避免这种结果似乎需要某种不平等待遇的组合，其中特权人类比至少具有同等道德地位和更大需求的数字心灵更受青睐，以及对数字心灵的生殖机会的限制——如果适用于人类的限制，会违反生殖自由的原则。

同样，在政治层面，民主原则将使占人口绝大多数的多产的数字思想赋予政治控制权，包括对转移支付和产权制度的控制。[15]

在这里，人们可以尝试捍卫人类的特殊特权。例如，一些契约理论可能表明，如果人类相对于数字心灵处于强大的地位，这将使我们有资格获得相应的大量资源。或者，人们可能会采用某种与代理人相关的原因的解释，即社区或物种有权根据客观上同样大的沙漠和道德状态。[16]这种相对性似乎反映了当今国家采取的事实上的方法，这些国家通常对本国公民的福利条款比对外国人更慷慨，即使有更贫穷的外国人，也可以受益更多，并且就他们的利益而言内在特征至少与本国公民一样值得援助。

然而，在走上这条道路之前，人们应该仔细和批判性地反思类似立场的历史记录，这些立场曾经被广泛采用，但后来名誉扫地，这些立场被用来证明对许多人类群体的压迫和对非人类动物的虐待是合理的。例如，我们需要问一下，倡导数字心灵与人类之间的歧视是否类似于支持某些种族至上主义？

这里要记住的一点是，数字心灵有很多种。他们中的一些人彼此之间的差异可能比人类思维与猫的思维差异更大。如果数字心灵的构成与人类思维非常不同，那么我们对其的道德义务与我们对其他人的义务不同也就不足为奇了；因此，以不同的方式对待

[15] 参看 （Calo，2015 ）
[16] 例如（Williams，2006 ）

它不一定是令人反感的歧视。当然，这一点不适用于与生物人类思维非常相似的数字心灵（例如全脑仿真）。也不是反对理由，从人类思维的区别在于，给他们数字的方式心中消极歧视*更大的*道德状态（超例）或使他们的需求*更*比人类（超受益者）的需求道德分量。就此而言，它也不能证明根据我们当前与动物互动的模板来对待具有类似能力或感知能力的数字心灵对非人类生物是合理的，因为后者受到非常普遍和可怕的虐待的困扰。

试图证明对人类的特权待遇是正当的，而不假设有利于我们同类的原始种族主义般的偏见的一种方法是援引一些原则，根据这些原则，我们有权（或有义务）更多地考虑那些与偏远的陌生人相比，更紧密地融入我们的社区和社会生活。如果人们希望使大多数人和大多数国家目前将大多数援助限制在他们自己的群体内的（非世界性）方式合法化，那么可能需要一些这样的原则。[17]然而，这样的举措不会排除已经成为我们社会结构一部分的数字心灵，例如担任管理员、顾问、工厂工人或个人助理的角色。与地球另一端的人类陌生人相比，我们与此类人工智能的社会联系可能更紧密。

4. 讨论

我们已经看到，通往数字超受益者的途径有很多，这使得他们的可能性更加强大。这是目前最流行的幸福感的含义。

这意味着，从长远来看，如果世界上充斥着数字超受益者而不是像那样的生活，那么总体幸福感会我们知道的更高。并且就这样的生物出现而言，他们的关注可能在道德上占主导地位，与人类和动物的关注相冲突，例如对稀缺自然资源的关注。

然而，尽管极端主义关注现有人类的福利或新数字心灵的福利可能会给另一方带来可怕的后果，但妥协政策有可能在这两种标准下都做得非常好。考虑三种可能的政策：

(A) 100% 的资源给人类

(B) 100% 的资源给超受益者

(C) 99.99% 的资源给超受益者；给人类的为 0.01%

从总功利主义的角度来看，（C）大约是最优选选项（B）的 99.99%。从普通人的角度来看，考虑到数字心灵带来的天文财富，(C) 也可能是最优选选项 (A) 的 90% 以上，比目前的总数高出许多数量级 (Bostrom, 2003; Hanson, 2001）。因此，*事前*，似乎有吸引力的，以减少的可能性更大既（A）的在交换的概率和（B）（C）-无论是对冲道德错误，以适当地反映道德多元化，以考虑博弈论的设定，还是仅仅作为的问题*现实政治*。同样，由于人类可以在不产生超人类地茁壮成长*坏*的生命，而且由于避免这种痛苦不仅是从总功利的角度，而且在许多其他评价意见，可以减少对超高效生产的潜在措施极其重要的问题贬低价值（即使对人类付出一些代价）将是共识政策的重要组成部分。

[17] 这些做法当然是, 受到世界主义的批评；例如（辛格，1981 年；阿皮亚，2006 年）。

更大的挑战是不是说明在人类未来可能的和数字的头脑都做的非常好人口，但要实现这样的设置稳定避免一方践踏对方*事后*，在 3.2 节进行了讨论。

这一挑战涉及实践和道德两个方面。实际上，问题是设计制度或其他手段，使保护人类和动物利益的政策可以无限期地维持下去，即使在其受益者人数众多且速度快的情况下，大量不同的高功能智能机器也是如此。解决这个问题的一种方法可能是创建绝大多数高福利数字心灵，以保持这一结果并维护相关规范和制度（包括在连续几代数字心灵的设计中）。

从道义上，问题是建议的措施*事前*有吸引力的妥协是否在他们的允许的*事后*执行。这里的一项有用测试是，我们是否可以在类似情况下认可它们在非数字心灵中的应用。例如，我们可能要求任何提议的安排都符合一些非歧视原则，例如以下（Bostrom 和 Yudkowsky，2014 年）：

> *基质非歧视原则 Principle of Substrate Non-Discrimination*
>
> 如果两个生物具有相同的功能和相同的意识经验的不同，只是执行的基础不同，那么他们就具有相同的道德地位。

以及

> *个体发生无歧视原理 Principle of Ontogeny Non-Discrimination*
>
> 如果两个生物具有相同的功能和相同的意识体验，并且仅在它们如何存在上有所不同，那么，那么他们具有相同的道德地位。

在应用这些原则，它是机器的头脑能够从人类心灵很大的不同，其中的方式是很重要的召回早先点事如何，他们应该接受治疗。即使我们接受上述非歧视原则，但在将它们应用于并非某些人类思维完全复制的数字心灵时，我们也必须小心谨慎。

例如，考虑繁殖。如果人类能够通过将花园垃圾倒入生化反应器中，每隔几分钟就生一个孩子，那么人类社会似乎可能会改变当前的法律惯例，并对允许人们繁殖的速度施加限制。如果不这样做的话，任何社会福利体系都会在短期内破产，假设至少有一些人会以这种方式创造大量儿童，尽管他们缺乏支持他们的手段。这种监管可以采取多种形式——未来的父母可能需要在创造后代之前提供足以满足其需求的保证金，或者可能会根据配额分配生殖许可。类似地，如果人类有能力产生任意数量的与自己完全相同的副本，我们可能会期望进行宪法调整，以防止根据谁愿意并有能力创造最大数量的投票来决定政治竞赛-克隆。同样，调整可以采取各种形式——例如，此类副本的创建者可能不得不与他们创建的副本分享他们自己的投票权。

因此，只要这些法律或宪法的调整*将*如果我们有这些类型的生殖能力，是人类可以接受的，它同样可以接受作出类似的调整，以适应谁数字头脑*做*拥有这种能力的。

一个关键问题——当然是从现有生活的角度来看——是设计新思想以可靠地支持维护人类现任者的某些权利和特权在道德上是否是允许的。我们早些时候提出，这种保留的人权和社会特权的安排是有道理的，至少作为明智的实践妥协的尊重不确定性和缓

解冲突的路径，无论它在基本道德理论的层面上是否是最佳的。我们可以通过类比来指出一种普遍的观点，即以昂贵的支持成本和需求来保护和保护少数群体，例如老年人、残疾人、白犀牛和英国王室，在道德上是可以接受的。如果我们假设所创造的数字思维本身会支持这种安排并支持其延续，那么这个结论似乎也得到了支持。

即使结果本身在道德上是允许的，但是，我们面临的一个进一步的伦理问题，即是否存在一些*程序上*反感约精密工程，我们创建，以确保他们的同意，新的数字心目中的喜好。我们可以从非歧视原则的角度来看待这个问题，并考虑我们将如何看待类似地塑造人类儿童偏好的建议。

虽然人类文化也经常通过教育，对话，规劝试图通过对规范和价值观对儿童，包括孝顺和对现有的规范和尊重的机构，一个灌输具体的处置建议，通过*基因工程*的配子将可能是更多的争议.即使我们抛开对安全、不平等的获取、压迫性政府的虐待或父母做出狭隘或其他愚蠢选择的实际担忧，也可能仍然担心对后代的倾向施加详细控制的行为，特别是如果这样做了具有"工程思维"并使用完全绕过受控对象自己的思想和意志的方法（通过在对象出生之前发生）将在本质上存在道德问题。[18]

虽然我们无法在这里全面评估这些担忧，但我们注意到数字心灵的两个重要差异。首先，与人类繁殖相比，创作者可能没有明显的"默认"可以推迟。程序员可能*不可避免*建设工程机械情报是否建立它拉上时，是否将这一目标或火车地作出选择，是否给它一组或偏好或其他。鉴于他们必须做出一些这样的选择，人们可能会认为他们做出具有更理想结果的选择是合理的。其次，如果一个人被"设计"成具有某些特定的欲望，我们可能会怀疑，在更深层次上，可能存在其他可能与设计偏好发生冲突的性格和倾向。例如，我们可能会担心，结果可能是一个人对让父母失望而感到非常内疚，因此过度牺牲其他利益，或者她内心深处的某些隐藏部分将继续受到压抑和阻碍。然而，在数字心灵的情况下，如果可以将它们设计为内部更加统一，或者如果在"轻触"中添加尊重"遗产"人口利益的偏好，则可能避免此类问题"这种方式既不会引发内部冲突，也不会妨碍数字心灵开展其他业务的能力。

所有的一切，似乎使数字超受益者的创作的结果*和*繁荣的极大人口的保存可以得分很高双方的客观和以人为中心的评价标准。考虑到高风险和不可逆转的发展潜力，制定出道德上可接受且实际可行的路径以实现这种结果将具有很大价值。

参考文献 References

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

Aghion, P., Jones, BF and Jones, CI (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.

[18] 例如（Habermas，2003；Sandel，2007）

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, WM (eds) *The Cambridge Handbook of Artificial Intelligence*. Cambridge University Press, pp. 316–334.

Calo, R. (2015) 'Robotics and the Lessons of Cyberlaw', *California Law Review*, 103(3), p. 529.

Chalmers, D. (2010) 'The Singularity: A Philosophical Analysis', *Journal of Consciousness Studies*, 17(9–10), pp. 7–65.

Frick, JD (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, MH, Esfeld, M. and Gerstner, W. (2007) 'Consciousness & the Small Network Argument', *Neural Networks*, 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, WD (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.pp. 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, JM (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, WT (1944) 'Interestingness', *Philosophy*, 19(74), pp. 233–241.

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

Williams, BAO (2006) 'The Human Prejudice', in *Philosophy as a Humanistic Discipline*. Princeton University Press. pp. 135–152.