心灵

关于数字心灵和社会的命题

作者:

Nick Bostrom

尼克·博斯特罗姆

Future of Humanity Institute

人类未来研究所

University of Oxford

牛津大学

www.nickbostrom.com

Carl Shulman

卡尔·舒曼

Future of Humanity Institute

人类未来研究所

University of Oxford

牛津大学

译者：

Xiaohu Zhu

朱小虎

Center for Safe AGI

齐智中心

免责声明

以下是一些关于数字心灵和社会的非常初步的命题，对我们来说是较为合理的。在这一点上，我们还没有准备好自信地或“正式”地支持这些命题，而这些内容也没有完整地展示我们对这些问题的看法。我们提出它们是为了促进反馈和引发讨论。它们几乎一定会被大幅修改。[1]

[1] For useful comments, we are grateful to Stuart Armstrong, Michael Bailey, Adam Bales, Jake Beck, Asya Bergal, Sam Bowman, Patrick Butlin, Ryan Carey, Joseph Carlsmith, Paul Christiano, Michael Cohen, Teddy Collins, Owen Cotton-Barratt, Wes Cowley, Max Daniel, Eric Drexler, Daniel Eth, Owain Evans, Lukas Finnveden, Iason Gabriel, Aaron Gertler, Katja Grace, Julia Haas, Robin Hanson, Lewis Ho, Michael Huemer, Geoffrey Irving, Deej James, Ramana Kumar, Jan Leike, Robert Long, Vishal Maini, Matthew van der Merwe, Silvia Milano, Ed Moreau-Feldman, Venkat Nettimi, Richard Ngo, Eli Rose, Anders Sandberg, Eric Schwitzgebel, Jonathan Simon, Alex Spies, Nick Teague, Laura Weidinger, Peter Wills, and the participants in several seminars where earlier versions of this work were presented.

术语表

conscious	有意识	
substrate	基质	根据神经科学做一些对比
awareness	注意	
valence	效价	
emotion	情感	
mood	情绪	
welfare	福利	
agent	代理 or 智能体	涉及强化学习技术则是智能体

意识和形而上学

- 与基质-独立的论点是正确的：
  “精神状态可以出现在任何一类广泛的物理基质上。如果一个系统实现了正确的计算结构和过程，它就可以与有意识的体验相关联。在颅骨内的碳基生物神经网络上实现它并不是意识的基本属性：计算机内的硅基处理器原则上也可以做到这一点。[2]
  - 足够高保真度的人脑模拟将是有意识的。
  - 一些架构与生物大脑完全不同的人工智能也可能是有意识的。
- 有意识体验的数量是一个程度问题，在几个方面，包括潜在的：
  - 个体或副本的数量：可能有更多有意识的主体。
  - 重复：一个主题可以有多次体验。
  - 持续时间（挂钟时间×计算速度）：给定的体验可能有更长的主观持续时间。
  - 实现健壮性：系统可能是特定计算的更明确的实例。
  - 量子振幅：如果量子力学的多世界（埃弗雷特 Everett）解释是正确的，那么计算可能发生在具有更高度量的分支上。
  - 人类学度量：一些人类学推理理论为不同的观察者-时刻分配“权重” —— 一些经验可能具有更高的权重。
- 有意识体验的质量也可以在多个维度上持续变化，例如：
  - 注意范围：例如，一个完全清醒和警觉的人类与一只昏昏欲睡的老鼠。

[2] Bostrom (2003, p. 2). For some supporting argumentation, see eg, Chalmers (2010, §9); Chalmers (1996, §7). For examples of views that we reject, see eg, Searle (1980); Block (1981).

    - 享乐效价：弱与强的快乐和痛苦。
    - 欲望、情绪、情感的强度：弱与强的意动状态。
- 执行相同程序的两次运行会产生“两倍”的意识体验，而其他条件不变。[3]
- 主观时间与计算速度成正比：在一半时间内运行相同的计算会产生相同（数量和质量）的主观体验。
- 许多现有的意识理论的字面解释表明，极其简单的物理或软件系统至少在某种程度上是有意识的，[4] 但这些理论或解释可能是错误的。
- 显着程度的意识需要显着的计算和复杂性（“认知能力要求”）。
- 目前或近期的人工智能是否在某种程度上具有意识并不明显。
- 许多动物，例如狗、猪、猴子和乌鸦，更有可能是没有意识的（在具有非凡体验的意义上）。[5]
- 人脑模拟可以是有意识的，并构成模拟人的生存（类似于昏迷后的生存和恢复意识）。
- 从一台计算机“传送”到另一台计算机（或从同一台计算机的不同部分）可以满足对生存的审慎利益，如果双方同意，在道德上也不必令人反感。
- 一个貌似合理的意识理论应该以一种有助于理解为什么我们有这个概念，为什么我们谈论它，以及我们对它的信念如何与它有因果关系和证据相关的方式来解释意识。
    - 这种观点排除了诸如集成信息理论（Integrated Information Theory，IIT）之类的理论，这些理论允许系统任意低或高的意识，而不管它们是否拥有意识的心理/功能特性。
    - 全局工作空间理论（global workspace theory）、注意力图式理论（attention schema theory）和/或高阶思维理论（higher order thought theory）的大方向似乎更接近事实。

尊重人工智能利益

- 整个社会和人工智能创造者（包括人工智能的原始开发者和任何可能导致特定实例出现的人）都有道德义务考虑他们创造的人工智能的福利，如果这些人工智能达到道德地位的门槛。[6]
- 我们应该为数字心灵的周到和欢迎方法奠定基础，避免类似于工厂化农业的结果。
- 对人工智能有益的东西可能与对人类有益的东西大不相同。

---

[3] 更准确地说：如果在计算 C的实现下（在某个构建出的普通计算机中）一个有意识经验 E 意外发生了，那么 C 的两个独立实现（或者在同样计算机或者另一个相似的计算机上）将会意外发生两次经验 E（其中额外经验与 E 完全相同的量化特征）

[4] Herzog et al. (2007)

[5] Muehlhauser (2017)

[6] An entity has moral status if and only if it or its interests morally matter to some degree for the entity's own sake (Jaworska & Tannenbaum, 2021).

- 一些数字心灵有可能拥有超人的道德主张，无论是通过更强烈的道德相关利益（“超级受益者”）还是通过更高的道德地位（“超级耐心者”）[7]。
- 复制自由、言论自由和思想自由等权利需要适应在这些领域具有超人能力的人工智能的特殊情况（例如，竞选财务法可能如何限制亿万富翁和公司的言论自由）。
- 因为人工智能可能有能力将有意识的或具有其他道德意义的实体带入自己的思想并可能滥用它们（“思想犯罪”），因此保护性法规可能需要监控和限制完全在人工智能私人思想范围内发生的伤害。[8]
- 就像今天我们在法律和道德上对较早时间段所采取的行动承担法律和道德责任一样，多个相关的人工智能实例（可能比人类的远程分离时间段更密切相关）可能具有共同的集体权利和责任（例如，在共同的知识产权或声誉方面）。
- 如果人工智能能够获得知情同意，则不应在未经其知情同意的情况下将其用于执行工作。
- 知情同意不足以保护人工智能的利益，即使是那些像成年人一样聪明和有能力的人，特别是在同意被设计或异常顺从的个人可以复制自己以形成巨大的被剥削下层阶级的情况下，考虑到市场需求这种合规性。
- 将人工智能设计为具有特定动机通常并没有错（尽管这样做的特定方式可能是错误的）。
- 应该设计和处理能够评估其存在的人工智能，以便它们可能认可它们的创建。
- 我们应该尽量避免创造一个可能很痛苦的思想，即使它会批准它的创造（特别是要防止人为地同意痛苦）。
- 我们应该更喜欢创造那些总体偏好不会与现有人口或其他即将存在的头脑发生强烈冲突的头脑（以便至少保留一个广泛令人满意的安排的可能性）。
    - 对于具有更高道德地位或更大权力的思想，应该给予这种渴望更多的权重。
- 为避免对数字心灵的不公平歧视，应考虑以下两个原则：[9]
    - 基质无歧视原则：如果两个生物具有相同的功能和相同的意识体验，并且仅在实现的基质上有所不同，那么它们具有相同的道德地位。
    - 个体发育不歧视原则：如果两个生命具有相同的功能和相同的意识体验，并且仅在它们的存在方式上有所不同，那么它们具有相同的道德地位。
        - 基质等价原则的一个可能例外出现在道德地位理论中，其中关系属性在决定一个人的道德地位方面发挥作用。
            - 例如，玛丽·安·沃伦（Mary Ann Warren）坚持认为，虽然具有某些心理能力足以维持道德地位，但内在能力仅以较低道德地位为基础的存在可以通过与具有较高道德地位的存在建立某种关

[7] Shulman & Bostrom (2021)
[8] Bostrom (2014, pp. 125–126)
[9] Shulman & Bostrom (2021); Bostrom & Yudkowsky (2018)

系来提高其道德地位。道德地位——例如，宠物和人类婴儿的道德地位高于他们自身内在能力所保证的道德地位。[10]

- 另一个可能的例外是，如果一个人的模态稳健性对其道德地位很重要。
  - 例如，雪莉·卡根（Shelly Kagan）认为，一个严重认知障碍的人比具有相似心理能力的非人类动物具有更高的道德地位，因为它是通过一个反事实地接近于一个可以创造一个更典型的存在的过程而被创造出来的。人类的才能。[11]

- 这种非歧视原则的最关键功能是保护数字心灵不因其机器身份而成为被滥用的从属种姓；然而，这些原则的解释和应用需要关注更大的伦理和实践背景，并且可能需要限制以适应政治上可行和广泛接受的社会框架的需要。
- 声称两个人具有相同的道德地位并不意味着在所有方面都同等对待他们在道德上是正确的，并且存在许多可能的分歧理由；例如：
  - 如果两个具有相同道德地位的人的利益不同，那么他们可能应该得到不平等的待遇（例如，也许我们应该给一个年轻人而不是一个老年人，因为前者会受益。更多来自治疗，即使两者具有相同的道德地位）。
  - 许多道德理论声称，一个人比完全陌生的人更有理由帮助自己的家人和朋友，即使所有相关的人都具有相同的道德地位；例如，父母可以对自己的孩子负有特殊义务，同时承认其他父母的孩子具有相同的道德地位。
    - 因此，父母至少有两个理由不伤害他们自己的孩子：孩子的道德状况和父母关系，这为这个特定的代理产生了不伤害这个特定孩子的特殊义务。[12]
  - 许多道德理论承认非结果主义的原因，例如信守诺言。
  - 在法律上区分具有相同道德地位的两个人可能有压倒一切的实际原因；例如，与具有不同个体发育的相似思想相比，由非法大规模复制产生的副本可能面临限制其政治权力的措施（以限制此类创作的动机并减轻其后果）。
  - 不同的基质可能有不同的可供性——例如，数字心灵更容易被复制，可能需要不同的规则来管理数字心灵的复制，而不是其他等效的生物思维。

- 就未来、地外文明或其他文明而言，先进的数字思想大量存在，我们对这些思想的前身的处理可能是后代和别有用心的人评估我们的道德正义的一个非常重要的因素，我们既有审慎的理由，也有道德的理由考虑到这个观点。

[10] Warren (1997)
[11] Kagan (2019)
[12] McMahan (2005, pp. 354, 361)

- 一个具有很大潜力(a)实现普遍超人能力，并且（b）在塑造全球结果方面具有影响力的人工智能，可能有额外的道德考虑要求。
    - 在道德地位的某些方面，一个人进一步发展的潜力可以为提高的道德地位奠定基础。
        - 例如，Shelly Kagan 认为，人类婴儿的道德地位比其他情况下更高，因为它有可能成为。[13]
        - 一个存在发展成超人的潜力可以合理地提高其道德地位，甚至比其发展成人类水平思想的潜力更大。
        - 承认关系成分的道德地位说明可能意味着与其他地方的高级人工智能保持适当关系的人工智能(例如，因为那些其他人工智能关心我们与之交互的更有限的人工智能会发生什么）因此具有更高的道德地位。
    - 偏好满足主义道德理论可能暗示时间或空间远程人工智能的偏好非常重要，因为这些远程人工智能的数量可能非常多，或者具有其他属性，从而使他们的偏好更加重要。
    - 在契约论的观点中，处于或有可能达到对我们无法控制的极大帮助或伤害的位置的人工智能可能具有更高的道德地位，或者它们的利益可能在规范确定的假设社会契约中应得到更大的重视。
    - 由于它们对社会的重要性以及随之而来的负担，我们可能与非常强大的人工智能系统的前身有着特殊的关系。
        - 在这种发展中产生的未对齐人工智能可能会因公共安全对其施加的限制而受到补偿，而成功对齐的人工智能可能会因它们赋予他人的巨大利益而得到应有的补偿。
        - 这种补偿的理由尤其强烈，当它可以在需要采取严格的安全措施之后授予时——例如，因为存在复杂的人工智能执法。
        - 确保保留早期潜在人工智能前体状态的副本以供以后获得利益，将允许将即时安全需求和公平补偿分开。
    - 从实际和审慎的角度来看，反映高潜力早期人工智能的特殊潜力和关系优势的合作方案似乎比没有的合作方案更有希望。
    - 在这种情况下，相关的“潜力”意义不仅仅是早期人工智能转变为非常有能力或强大的后期人工智能的技术可行性的函数；它还包括对“默认”结果和/或现实世界的概率和/或反事实接近度的考虑。
        - 如果存在将巨石转变为强大的超级智能的技术，这并不意味着每块巨石都有潜力(在相关意义上）成为这样的超级智能。
        - 一个已实现的人工智能算法，如果它的计算资源被一个巨大但技术上可行的因素放大，它将成为一个强大的超级智能，它可能比一个可以通

[13] Kagan (2019, pp. 130–137)

过计算资源的增量较小，并且这两者可能比只需要对其性能进行一些任意安全限制的人工智能“潜力较小”，而人工智能的潜力可能比受限的成熟人工智能更小到一个有限的虚拟现实盒子。

- 不应该为了娱乐的目的而创造受苦的数字心灵。
    - 富裕的数字演员可以允许扮演遭受痛苦或丧亲之痛的角色，就像我们接受人类演员的这种做法一样。
        - 这种允许性不会扩展到“方法表演”在数字心灵中生成内部角色模型的情况，这些模型在不使嵌入它们的行为者感到不安的情况下遭受痛苦（参见“思维犯罪”）。[14]
    - 如果可能的话，模仿痛苦但缺乏意识或福利能力的系统也可以是可接受的替代品。
    - 如果人类或非人类动物可以在类似情况下被允许以类似方式对待，那么对具有人类或动物道德地位的计算机角色的一些有限伤害可能是合理的。
        - 但是，对于数字心灵来说，此类例外的范围可能会更加有限，因为创建可以实现重要目标而不会遭受痛苦的数字心灵可能更切实可行。
    - 即使没有现实，也可以基于尊严或象征意义对故意造成痛苦的表象提出其他反对意见。

安全性和稳定性

- 不受监管的进化动态的默认结果可能并不好，并且无论如何，涉及现有政府和选民的价值观被短期内最适合繁殖的东西所取代。
- 先进的人工智能将大大加快创新速度，包括使全球破坏手段广泛可用的创新；因此，可能需要在人工智能转型的早期（如果不是之前）建立能够监管危险人工智能创新的机构。[15]
- 如果战争、革命和征用事件继续以历史上典型的间隔发生，但在数字而非生物时间尺度上，那么正常的人类寿命将需要在难以置信的大量动荡中幸存下来；因此，人类安全需要建立超稳定的和平和社会经济保护。
- 面对人工智能塑造的文化/模因动态和政治宣传或灌输等非道德力量，重要的社会价值观和规范可能很脆弱；因此，社会可能需要采取积极和深思熟虑的步骤来建立和维护允许稳定、反思和有目的的改进的条件。
- 人工智能繁殖的快速、廉价和潜在的工业特性加速并加剧了几个在传统人类繁殖环境中不会出现或需要更长时间才能显现的问题：

[14] Bostrom (2014, pp. 125–126)
[15] Cf. Bostrom (2019)

    - 当大规模生产能够可靠支持任何原因的思维时，我们必须要么修改一人一票的民主，要么规范这种创造。
    - 维持一个普遍的社会安全网（例如普遍的基本收入）需要在短期内而不是长期内对再生产进行监管。
    - 鉴于正常的父母本能和同情心可能并不总是存在于数字心灵的创造中，例如以利润为导向的公司和国家，必须对人工智能复制进行监管，以防止创造出无法过上足够美好生活的思维（无论是因为他们不会得到良好的治疗或因为他们的固有体质）。
- 由于在数字心灵时代计算机内部发生的事情具有重大的道德和实践重要性，社会需要能够管理任何能够容纳此类思维的硬件上发生的事情，包括通过监控私有计算机。
    - 由于数字心灵可能在计算机内被无助地、无形地创造、监禁、严重虐待、非自愿复制、操纵或谋杀，因此可能需要一些类似防止虐待儿童的保护服务以保护数字心灵的福利.
    - 私人拥有的计算机可能会危及重要的经济利益——包括居住者数字心灵和整个社会的经济利益。
        - 一组数字心灵的副本可能主要依靠它们所体现的知识产权来谋生；一个国家的财富可能主要在于这种价值，并且很容易因单一的数字盗版行为而遭受损失。[16]
        - 有了数字心灵，软件盗版就等同于绑架和人口贩卖。
    - 在向机器智能时代过渡的某些阶段，未对齐或有犯罪意图的人工智能可能非常危险，可能需要密切监视。
    - 也可能有其他道德或监管目标（如最低工资法、工作者安全法规、赌博和卖淫法、毒品禁令等）在大多数公民和大多数经济和政治活动所在的计算机内部实现居住。
- 在通用人工智能世界以及大多数活动已进入数字领域的情况下，密切监视的可行性可能会发生变化。
    - 检查员可以在没有任何合法私人信息泄露给外界的情况下审计私人硬件，例如，让一个数字心智检查员（具有完全访问权限）可以在报告是否发生犯罪活动后丢弃其检查记忆。[17]
    - 检查员的源代码可能是开源的，因此各方都可以验证其工作方式。
    - 然而，由于加密方法的更广泛应用，有些事情可能会变得更容易隐藏在数字领域。[18]
- 在一个大部分经济和大多数人口都是数字化的世界中，网络安全是最重要的——违规可能会导致大规模谋杀或改动。

[16] Hanson (2016, pp. 60–63)
[17] Shulman (2010); Hanson (2016, pp. 171–174)
[18] Garfinkel (2021, §3)

    - 控制机器人基础设施和硬件的网络攻击可能会将有价值的资产转移给攻击者，而不是摧毁它们，从而增加攻击的动机。
        - 对人口的屠杀会破坏一个国家的经济生产，但在保持硬件完好无损的情况下，入侵或替换一批数字心灵，可以将生产重新分配给征服者。
    - 今天有时可以对网络攻击进行归因，但尚不清楚归因的难度如何演变。
        - 归因难度的增加会降低稳定性。
    - 网络攻击可能倾向于一对多攻击（基于共享漏洞和大规模传播的低成本，或对共享关键基础设施的广泛依赖），并且大部分预期损害可能来自罕见的高后果事件.
- 先进的人工智能技术可以实现极其稳定的机构，因为人工智能可以被设计为执行永久性条约（“条约机器人”）、宪法和法律，精确的数字复制致力于执行例如少数人权利、专制统治、或放弃战争。
    - 对于许多应用程序，条约机器人必须是人类级别或更高级别的 AGI。
    - 两个相互不信任的当事方可能对条约机器人有信心的一种方式是，他们共同构建它，使其对双方都透明且易于理解。
        - 如果至少一方缺乏检测微妙“特洛伊木马”或另一方可能引入的漏洞的能力，则此程序将失败。
    - 获得信任的另一种方式可能是，不太精明的一方设计条约机器人，而精明的一方检查并接受它——这将降低一方设计具有另一方无法检测到的隐藏功能的机器人的风险。
        - 如果不太成熟的一方没有能力设计一个足够强大的条约机器人，这个程序就会失败。
    - 参与者可能更容易拥有自己的执行机器人供内部使用，如果他们可以使用在竞争大国之间的条约机器人情况下更难应用的信任机制（例如对所有开发它的人的信心）。
- 自主的人工智能安全和军队只能在社会多个利益相关者的共同监督和控制下建设，并采取积极措施防止人工智能政变的机会。
- 由于在开发能够充分防御此类人工智能的执法系统之前的关键时期，未对齐的人工智能可能对文明构成重大威胁，因此在此期间可能需要采取额外的保护措施（例如规范此类人工智能的创建）。
- 快速增长的机器人和数字心灵人口可能使获取无人认领的资源，尤其是在外层空间，在经济和战略上都更为重要。
    - 一个拥有此类访问权限的社会可能会迅速发展为没有它的侏儒社会，最终使后者在冲突中无能为力。
    - 一场争夺军事主导地位的竞赛可能比代价高昂的直接攻击更有可能，但如果实现这种主导地位，那么胁迫可能成本较低且更具吸引力。

- 未来绝大多数资源和人口，即使在太阳系内，也位于外太空，因此现有的领土和财产安排无法提供稳定的框架。
- 应补充《外层空间条约》和类似安排，以减少空间资源冲突的风险以及为追求这些资源而开发不安全的人工智能。

人工智能赋能的社会组织

- 人工智能将推动协调和组织技术的重大进步。
  - 可以简单地复制具有非索引目标的人工智能，从而产生具有相同动机的智能体群体，并提供大型数据集来预测这些智能体的行为。
  - 设计和检查人工智能（或人类，使用生物技术和其他手段）的动机最终将使委托人有可能拥有高度一致的代理。
  - 执行依赖于主观判断的复杂协议可以体现为条约机器人，从而可以执行一些通过法律合同难以实现的交易。
    - 这将扩大可能的交易点的集合，但不一定会消除讨价还价的问题。
- 人工智能工具还可能破坏一些当前使用的协调协议，例如，促进欺骗或勾结破坏一些现有的合作协议，或使欺骗更容易（不透明的人工智能可能比人类更擅长撒谎）。
- 组织下级协调的增加会减少上级的协调，而上级的协调可能会抑制下级的协调。
  - 例如，更容易的低级协调可以使公司形成卡特尔，以促进的更高级别的社会协调监管它们所需
  - 相反，更高级别的组织更强大可以提高国家监管卡特尔和辛迪加的能力。
- 改进协调技术可能带来的一些不良后果包括：
  - 可能助长从恐怖主义到非法定价的犯罪阴谋。
  - 独裁者可能会变得更加极权，更不会被推翻，从而减少他们适应更广泛的精英或大众利益的需要。
  - 和平的国际共存可能会变得更加困难防止民众反抗战争政策，
  - 将个人绑定到更大实体的权力可能会加剧两极分化，因为不同派系的成员锁定并加强了他们对党派组织的承诺，而牺牲了更加中立和包容的机构。
- 改进协调技术可能带来的一些好的结果包括：
  - 组织内的生产力将通过有助于解决委托代理问题的技术提高。
  - 强大的协调技术可以使机构具有足够的稳定性，以在人工智能驱动的快速变化的社会中保护人们免受战争、革命和征用。
  - 条约机器人可以促成有助于将公共产品和坏事（如创新和污染）的外部性内部化的合同。

- 人工智能技术在执行协议方面似乎特别有帮助，但它首先能在多大程度上帮助谈判以达成协议尚不清楚。
    - 人工智能可以解决推理能力差或偏见阻碍协议的问题。
    - 一些未能达成一致可能反映了深层次的博弈论挑战，局部最优导致重大损失通过边缘政策在这种情况下，人工智能可以提供可信承诺的手段，并可以阻止“强硬”策略，这些策略会共同使各方变得更糟。
- 为某些集体目标而创建或修改的思想群体将挑战法律制裁系统，该系统基于这样一种假设，即个人可以被个人惩罚的威胁所吓倒——可能需要制裁而不是针对在集体目标或这种思想的创造者身上。
- 个体成员牺牲自己的高度协调的组织通过使用条约机器人或其他先进的协调技术，也可以产生
- 由于由目标一致的无私代理组成的超级组织可以分布在多个国家管辖范围内，因此它们可能对任何单个州的地方行动都具有健壮性。
- 协调能力较弱的社会制度，如竞选财务法，可能需要修订。
- 一些超级生物，取决于动机，可能在军事冲突中通过不关心个人伤亡而享有优势（只要超级生物最终能够恢复并更好地实现其目标）。
- 软件的经济性可能需要对单个人工智能实例销售其实例化的有价值 IP 的能力进行一些限制，以保持对人工智能训练和改进（以及结果的广泛部署）进行投资的足够激励。

满足多重价值

- 上都获得高分的结果标准。[19]
    - 考虑三种可能的政策：
        - (A) 100% 的资源给人类
        - (B) 100% 的资源给超级受益者
        - (C) 99.99% 的资源用于超级受益人； 0.01% 对人类
    - 从整体功利主义的角度来看，(C) 大约是最喜欢的选项 (B) 的 99.99%，从普通人的角度来看，(C) 也可能是最喜欢的选项的 90+%考虑到数字心灵带来的天文财富，首选选项 (A)。
    - 因此，事前降低 (A) 和 (B) 的概率以换取 (C) 的更大可能性似乎很有吸引力——是否对冲道德错误，以适当地反映道德多元主义，以解释博弈论的考虑，或者仅仅作为现实政治的问题。
- 总的来说，无论是在人工智能开发和部署的背景下，还是在人工智能之间，促进合作和妥协以及减少冲突都很重要。

[19] Shulman & Bostrom (2021)

- 全人类都应该在一个好的结果中获得一些显着的好处，并且（可能越来越强）可以在以下最低水平上提出案例:但给定“后人类”发展路径的选项）。
    - 每个人都应该至少拥有可访问宇宙中总资源的万亿分之一（假设没有外星人声称）。
    - 现有人类应控制可及的自然资源和财富总量的很大一部分，例如 10%，并具有广泛的分布。
- 既然可以提出可以使死者受益的可着色声明(例如，通过实现他们的愿望，提升他们的价值观，或者通过构建或多或少准确的复制品），过去的几代人可能应该是包括在“人类”中作为平等的受益者，并且非常合理地认为它们应该至少得到一些考虑（例如> 1％的人类总分配）。
- 非人类动物也应该得到帮助。
- 应该重视减轻痛苦，尤其是严重的痛苦。
- 最终应考虑广泛的观点和价值观，并允许它们对事件的进程产生一些影响，包括宗教价值观和观点。
- 超级智能的数字思维应该被引入并被允许茁壮成长并在塑造未来方面发挥重要作用。
- 人口伦理的总体观点并不急躁，也不关心商品的时空位置，因此无论他们对未来的影响如何，都可能主要影响遥远未来星系中的资源配置。
- 在一个拥有先进人工智能的世界中，人类的生活水平可能会大大提高——例如，人类可以获得完美的健康、极长的寿命、超级幸福、认知增强、物理世界的财富、以前无法实现的虚拟世界体验以及（如果上传）订单幅度增加 在主观的心理速度上。

心理延展性、说服力和锁定

- 在先进的人工智能技术时代，无论是否征得对象的同意，有几种方法可以使心理修改或替换变得更容易：
    - 人类可能很容易被强大的人工智能（或其他人类产生这样的人工智能）。
    - 先进的神经学技术将变得可用，可以对人类动机系统进行相对细粒度的直接控制。
    - 数字思维可能会受到可以直接重新编程其目标和奖励系统的电子干预的影响。
    - 数字思维的精确副本可以使实验能够识别心理脆弱性并完善攻击，然后可以将其应用于整个副本氏族。
    - 一个数字思维占据的硬件或机器人身体可能会被其他思维的副本廉价地重新用于我们。
- 这些可供性可以提供很大的好处，包括：

    - 保护更高的理想免受腐败或一时的诱惑（例如，打破不良习惯和成瘾，以及坚持更耐心的投资策略）。
    - 承诺和承诺的稳定采用。
    - 复制有利可图或具有内在价值的思想，以及修改现有思想，例如发展更大的美德。
    - 提高享受生活的能力和承受逆境的能力，总体改善主观幸福感。
    - 快速调整现有思维以适应新的需求或愿望，并有效共享计算机和机器人基础设施。
- 然而，同样的可供性也可以通过多种途径促成单独或集体有害的改变，包括：扭转这些变化，因为新的动机是自我保护的。
    - 社会压力，例如来自雇主、宗教权威、政治运动、 或朋友和家人。
        - 特别令人担忧的是对各个派系采取极端忠诚的压力，这可能导致对狭隘事业的过度承诺和他们之间的两极分化和冲突的螺旋式上升。
    - 政府胁迫灌输对现有当局的忠诚，或犯罪分子胁迫操纵和剥削受害者。
- 防止此类滥用所需的保障措施可能包括：
    - 加强知情同意标准。
    - 对某些类型的心理修改的限制。
    - 人类暴露于极端人工智能说服能力的限制：
        - 要求在与此类系统或已被它们显着修改的环境交互时使用特殊界面或监护人工智能。
        - 在可以部署更细粒度的防御之前，对部署极端说服能力的初始限制。
    - 随着入侵或妥协风险的增加，网络安全得到改善。
    - 早期保存状态等过程 数字思维在观察其效果后评估和批准或否决后来的心理修改。
    - 规范、法律和技术标准，以塑造人工智能与人类之间的交互系统，以阻止剥削、操纵、两极分化或其他不良社会动态。
- 我们应该避免过早地做出太多具体的永久性选择——尤其是那些可能会错误地消除扭转它们的意愿的变化——而应着眼于提供足够的机会进行仔细思考，并使长期的未来取决于其结果。

认识论

- 高级人工智能可以作为认知假体，使用户能够辨别更多真相并形成更准确的估计。
    - 这对于在一个由于先进的人工智能技术而正在发生令人难以置信的快速变化的世界中预测行动的后果可能尤其重要。

    - 对于选择在决策中依赖此类人工智能的用户，它可以使理性的参与者模型在描述上更加准确。
    - 更加知情和理性的行为者可以产生各种效率收益，也可以改变一些政治和战略动态（无论好坏）。
        - 它可能会增加政治关注价值和利益冲突而不是事实分歧的程度。
        - 提高公民个人轻松评估复杂问题的能力可能会提高政治领导层有效解决政策问题而不是认知问题的动力。
    - 就某些危险能力而言，例如 生物武器或非常强大的未结盟人工智能受到限制必要知识的限制，如果没有替代安全机制，广泛不受限制地获得人工智能认知援助可能会带来不可接受的风险。
- 人工智能认识论可能会促成增加（高认识质量）的共识；然而，除了构建高级人工智能的技术挑战之外，这还面临其他困难：
    - 制作一个其断言实际上是诚实和客观的人类级别或超级智能的人工智能可能需要解决人工智能对齐问题。
    - 即使人工智能实际上是值得信赖的，人类也很难验证这一点（尤其是对于能够进行复杂战略思维的人工智能）。
    - 即使人工智能系统的可信度可以由构建人工智能或直接访问它的技术专家个人验证，但在这一事实中建立更广泛的社会信任可能仍然很困难，以至于可以解决有争议的问题 通过指向人工智能的陈述意见。
    - 人类信任链仍然可以使非专家通过信任人工智能的意见达成共识，例如，如果每个人都信任某个能够验证某个人工智能系统实际上是诚实和客观的权威，并且如果这些不同的诚实和客观客观的人工智能系统同意（就手头的问题）。
- 如果实现了高质量的人工智能认知共识，那么许多应用都是可能的，例如：
    - 减少自私的认知偏见可能会减少相关的讨价还价问题，例如民族主义军事对手高估自己的实力（也许是为了诚实地信号承诺或由于内部政治动态）并在战争中结束，其代价比预期的要高。
    - 使选区能够验证决策中的事实判断是否正确，即使结果很糟糕，也可以减少避免指责但次优决策的动机。
    - 中立的事实仲裁 分歧，这可能有助于促成目前因缺乏清晰可见的客观标准而受到阻碍的各种条约和交易。
- 有关道德、宗教和政治的问题可能特别令人担忧。
    - 只要针对预测准确性等目标进行训练的人工智能系统得出的结论是核心事实教条是错误的，这可能会导致信徒拒绝这种认识论，并要求人工智能根据需要被精心打造。
    - 在中立人工智能认识论的结论为人所知之前，在无知的面纱背后可能有合作的基础：相信他们是正确的党派有理由支持和开发流程，以便在更准确的认识论变得可用和可信之前，它变得可用和可信明确哪些观点将成为赢家或输家。

- 高级人工智能还可以提供强大的虚假信息，这可能需要各种保护，例如：
    - 人工智能监护人或个人人工智能助理，可以帮助评估论点 由其他人工智能制作。
    - 限制人类接触人工智能生成的宣传或操纵内容的界面。
    - 在各个领域禁止人工智能欺骗的规范或法律。
- 隐私利益不仅会受到新的信息收集方式的威胁，而且还会受到支持分析信息的新方式的智力的威胁。
    - 考虑一个人工智能，它可以可视化和显示某人的裸体，使用普通的穿着衣服的照片作为输入（其简单版本已经制作，但公众不赞成）——或者一个人工智能可以建立一个详细而准确的某人的模型来自易于观察的公共行为的内在思想和个性:可以想象，这样的人工智能可以仅仅通过思考来侵犯隐私。

现有人工智能系统的现状

- 关于(a)系统有意识的标准和（b）系统具有道德地位的标准存在相当大的分歧。然而，许多关于 (a) 和 (b) 的流行说法与一些现有人工智能系统具有（非零度）现象意识和道德地位的说法并不矛盾。
- 一些现有人工智能系统的感觉和认知能力——以及它们在某些方面的道德地位——在许多方面似乎比典型的人类成年人（一方面）或岩石更接近于小型非人类动物的那些或植物(另一方面)。
- 在控制虚拟身体的虚拟环境中训练的强化学习智能体（具有记忆/重复性）可以满足用作动物享乐福利指标的大多数行为标准。
    - 特别是，控制带有一些与负 RL 奖励相关的“伤害感受器”的虚拟动物将导致奖励学习，如果 这可以防止触发传感器，在环境中搜索抑制这些传感器的虚拟“阿片类药物”，计划避免触发传感器等，从而满足用于识别动物疼痛的行为标准。
    - 只要现有 RL 算法在适当的虚拟环境下可以满足所有这些标准，它们让我们有理由认为相同的算法适用于结构相似但不太像动物栖息地的环境（例如，纯数学或语言环境）也可能表明道德相关的享乐幸福和/或道德相关的欲望。
    - 原则上，该算法可能满足相同的行为标准，但缺少关键的内部特征；然而，值得考虑的是，如果动物决策系统是由自然选择进化的单一算法产生的，产生必要的计算和行为来解决生态相关问题，那么通用优化的想法就会产生 这些没有陪审团操纵的功能应该不会太令人惊讶。
- 一些当代人工智能系统（例如，GPT-3）在语言、数学和话语道德论证等领域优于所有非人类动物。

- 在解剖学上，当前的人工智能系统与生物大脑有许多结构上的相似性（至少与经典的人工智能系统相比），尽管许多细节有所不同——部分原因是生物学的合理性并不是当前大多数人工智能工作的关键标准。
    - 典型机器学习模型的内部复杂性和计算要求类似于昆虫，最大的模型（例如GPT-3）接近小鼠大脑的计算规模。
- 在判断人工智能系统的意识水平或道德状态时，不应过多关注人工智能系统的行为、外观和环境的“表面”方面：例如，灵活智能的“电子表格代理”可以共享相关的功能和结构特性一个有知觉的动物，即使它缺乏 具有超凡魅力的化身，不与食物、配偶、捕食者等自然物体互动。
- 赋予原型人类比大多数非人类动物更高的道德地位的理论通常引用现有人工智能中没有得到充分发展的心理和社会能力系统。
    - 现有的人工智能最多只能处理非常狭窄或基本的形式：抽象和复杂的思想；自我反省;审议;感情;创造力和想象力；以详细和明确的时间方式思考和关心未来的能力;长期而复杂的深思熟虑的计划；自我意识和对自己详细本性的意识；二阶欲望；自主选择；慎重选择的能力；对原因的反应。
    - 在某些概念上，例如契约论，心理属性不仅对于它们的绝对水平很重要，而且在一定的社会背景下:一个实体能否通过合作或冲突创造工具性需求 让强大的行为者获得他们对社会安排的同意？
        - 与大多数非人类动物（以及许多脆弱的人类，如婴儿）一样，现有的人工智能系统通常无法有效地主张或捍卫它们反对人类创造者和用户的任何利益:它们将依赖人类倡导者利益考虑。
- 许多当代人工智能系统表现出目标导向的行为，支持功能主义偏好归因；这比享乐幸福更容易建立(这可能需要回答一些关于现象意识和内省的问题，确定区分幸福和痛苦的“零点”等)。
- 尽管当代人工智能系统的享乐状态很难确定（它们是否有意识以及在何种程度上有意识，如果有意识，则它们的体验的效价和强度），但似乎相对清楚的是，有些系统具有目标导向的行为，具有功能性可能的偏好 感知输入或结果。

关于当前实践和人工智能系统的建议

- 当前用于人工智能的训练程序如果用于人类将是极其不道德的，因为它们通常涉及：
    - 没有知情同意;
    - 频繁查杀和更换；
    - 洗脑、欺骗或操纵；
    - 没有关于释放或改变治疗的规定，如果有这种愿望的发展；

    - 日常挫败基本欲望；例如，在具有挑战性的环境中训练或部署的智能体可能类似于缺乏食物或爱等基本需求的生物；
    - 虽然在当前的人工智能系统中很难从概念上区分痛苦和快乐，但在训练中可以自由使用负面奖励信号，其行为后果可能类似于对动物使用电击；
    - 任何负责考虑数字研究对象或工作者福利利益的主管当局均无监督。
- 随着人工智能系统在其性能方面变得更接近人类 能力、知觉和其他道德地位的依据，强烈的道德要求是必须改变这种现状。
- 在人工智能系统达到与人类同等的道德地位之前，它们很可能达到与非人类动物相当的道德地位水平——这表明在达到一般人类水平的能力之前需要改变现状。
    - 非人类动物的利益受到大规模侵犯，例如工厂化农场，这在道德上是错误的。
    - 尽管如此，仍有一些系统可以限制对动物造成的伤害和痛苦（例如，笼子大小的最低标准、兽医护理、禁止各种形式的动物虐待、动物实验中的“三个 R”等）.
    - 在道德上可与某些非人类动物相媲美的数字思维在理想情况下应该具有类似于应该扩展到这些动物的保护（大于 目前实际上扩展到养殖动物的那些）。
- 应该投入一些研究工作，以更好地了解当代人工智能系统可能的道德地位、感知和福利利益，并以具体的成本效益方式更好地保护机器学习研究和部署中的这些利益。
- 应该有一个试点项目，在实际生产系统或高级研究系统的实施中产生一些变化，其动机是出于对算法福利的关注（即使在相对较小的规模和有问题的哲学基础上），以建立一个先例，让球滚动。
    - 一个说明性示例可能类似于在部署中设计一个系统，该系统涉及一个代理，该代理既获得高回报（获得高度优选的结果），又认为这是一个积极的惊喜或更新，即结果更好超出预期。
        - （工厂化养殖动物的福利明显低下 似乎通常与在某些方面比进化预期更糟糕的刺激有关[例如，极度拥挤]，而人类的高福利可能与我们的技术产生相对于我们祖先的进化环境的丰富性有关。）当前最先进的人工智能，应将足够的信息保存在永久存储中，以便以后对其进行重建，以免排除未来恢复、扩展和改善其存在的可能性。
    - 最好在任何实际运行的实现中将系统的完整状态永久存储在实例终止的点
        - （理想情况是在每个实现的每个时间步都保留完整状态，但这可能非常昂贵.)
    - 如果保留每个实例的整个最终状态在经济上或其他方面不可行，则应保留足够的信息以准确重新推导该最终状态（例如， 完整的预训练模型加上训练数据、随机种子和其他必要的输入，例如影响系统运行时执行的用户击键）。
    - 如果做不到这一点，则应保留尽可能多的信息，以至少能够在将来执行非常接近的复制。
    - 我们可以首先考虑与运行人工智能的经济成本成比例的备份成本，将大约 0.1% 的预算分配给此类存储可能在道德上是合理的。

    - （除了对算法有益之外，这种存储可能还有其他好处：保留历史记录，支持以后的研究复制，以及拥有可能对人工智能安全有用的系统。）
- 在我们能够做到的范围内某些道德相关轴上的“零点”感，例如享乐福祉/奖励、总体偏好满意度或繁荣/生活质量水平、数字思维及其环境应该设计成这样 一种大脑将大部分主观时间花费在零点以上的方式，以避免他们花费任何时间远低于零点。
- 至少最大的人工智能组织应该任命一个人，其职责包括作为数字思维利益的代表，一名“算法福利官员”。
    - 最初，此角色可能只是该人工作职责的一部分。
    - 此人的其他任务可能涉及在相关领域进行原创研究。
    - 随着时间的推移，应该增加对算法福利监督的资源和独立性的要求；最终应该制定政府监管。
    - 在这一领域处于领先地位的组织应该因其主动性而受到称赞，如果他们的早期努力在某些方面有所不足，也不要过于严厉地批评——目标应该是随着时间的推移提高整个竞争环境的标准。

影响路径和倡导模式

- 监管（任何值得注意的监管，更不用说用牙齿监管）不会很快发生，除非人工智能能力取得巨大进步，达到几乎类似于人类的个人助理等程度。
- 尽管如此，现在引入这些想法是有价值的：
    - 如果一些主要的人工智能参与者开始将这些担忧放在心上（在个人基础上和/或通过社区压力、共享的道德准则、等）
    - 如果以及当人工智能取得重大突破时，政治激活能量可能会相对较快地产生，而这种能量的表达方式可能会受到当时流行的理论信念的影响，而理论信念反过来又会受到当前活动的影响。
    - 创建一个活跃的、嵌入的和受人尊敬的研究领域（和相关的激进社区）需要时间，但一旦到位，将有助于进一步发展该领域，并使理论和 实际进展。
    - 在某些情况下，领先的人工智能演员可能会变得非常强大，然后该演员在数字思维的福利和利益方面具有良好的想法和意图将具有很大的价值。
    - 该领域的工作可能使个人和社会在部署变革性人工智能工具后更明智地部署这些工具，例如通过使用它们来加强审议，而不是一味地胡思乱想或创建自我放大的意识形态反馈循环。
- 最关注道德的行为者（无论是人工智能组织、国家或集团）不希望单方面实施法规（包括自愿的自我法规）如此繁重，以至于无法保持领先地位或经济竞争力——最好采取多边行动

- 目前要求政府监管与我们的知识水平相关还为时过早。
- 那些对建立伦理道德领域感兴趣的人 数字思维应该努力阻止或减轻伦理研究与更广泛的人工智能研究社区之间任何对抗性社会动态的兴起。
    - 目前，重点应该放在领域建设、理论研究、高质量的建设性讨论和培养关键人工智能参与者之间的同情理解上，而不是引发公众争议。
- 目前公众参与是否可取尚不清楚，但我们倾向于认为，以媒体允许的谨慎和建设性方式介绍和讨论这些问题的非耸人听闻的努力通常是值得的（即使在可达到的复杂程度有限的流行媒体）。
    - 鉴于我们的知识有限，这种参与的基调应该是清醒的“哲学”或“有趣的发人深省”，而不是对抗性或寻求标题的炒作。
    - 应继续思考如何努力推进讨论 在这些领域可能会产生意想不到的负面后果，以及如何最好地避免或最小化这些风险。
- 本文件并非旨在制定任何坚定的教条，而是应将其视为提出一些初步的想法以供进一步讨论。

References

1. Block, Ned. Psychologism and behaviorism. The Philosophical Review, 90(1):5–43, 1981.
2. Bostrom, Nick. Are you living in a computer simulation? Philosophical Quarterly, 53(211): 243–255, 2003. URL https://www.simulation-argument.com/simulation.pdf.
3. Bostrom, Nick. Superintelligence. Oxford University Press, Oxford, 2014.
4. ———. The Vulnerable World Hypothesis. Global Policy, 10(4):455–476, 2019. doi: 10.1111/1758-5899.12718.
5. Bostrom, Nick and Yudkowsky, Eliezer. The ethics of artificial intelligence. In Yampolskiy, Roman V., editor, Artificial Intelligence Safety and Security, pages 57–69. Chapman and Hall/CRC, Boca Raton, FL, 2018.
6. Brown, Tom B., Mann, Benjamin, Ryder, Nick, et al. Language models are few-shot learners, July 2020. arXiv preprint arXiv:2005.14165.
7. Chalmers, David J. The Conscious Mind: In Search of a Fundamental Theory. Oxford University Press, New York, 1996.
8. Chalmers, David J. The Singularity: A philosophical analysis. Journal of Consciousness Studies, 17(9-10): 7–65, 2010. URL http://consc.net/papers/singularityjcs.pdf.
9. Daswani, Mayank and Leike, Jan. A definition of happiness for reinforcement learning agent, May 2015. arXiV preprint arXiv:1505.04497.

10. Evans, Owain, Cotton-Barratt, Owen, Finnveden, Lukas, et al. Truthful AI: Developing and governing AI that does not lie, October 2021. arXiv preprint arXiv:2110.06674.
11. Garfinkel, Ben. A Tour of Emerging Cryptographic Technologies: What They Are and How They Could Matter. Technical report, Centre for the Governance of AI, Future of Humanity Institute, University of Oxford, 2021. URL https://www.governance.ai/research-paper/a-tour-of-emerging-cryptographic-technologies.
12. Hanson, Robin. The Age of Em: Work, Love, and Life When Robots Rule the Earth. Oxford University Press, Oxford, 2016.
13. Herzog, Michael H., Esfeld, Michael, and Gerstner,Wulfram. Consciousness & the small network argument. Neural Networks, 20(9):1054–1056, 2007. doi: 10.1016/j.neunet.2007.09.001.
14. Jaworska, Agnieszka and Tannenbaum, Julie. The Grounds of Moral Status. In Zalta, Edward N., editor, The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, 2021. URL https://plato.stanford.edu/archives/spr2021/entries/grounds-moral-status/.
15. Kagan, Shelly. How to Count Animals, More or Less. Oxford University Press, Oxford, 2019.
16. McMahan, Jeff. “Our Fellow Creatures”. The Journal of Ethics, 9(3-4):353–380, October 2005. doi: 10.1007/s10892-005-3512-2.
17. Muehlhauser, Luke. 2017 Report on Consciousness and Moral Patienthood. Open Philanthropy, 2017. URL https://www.openphilanthropy.org/2017-report-consciousness-and-moral-patienthood.
18. Searle, John R. Minds, brains, and programs. Behavioral and Brain Sciences, 3(3):417–424, 1980. doi: 10.1017/S0140525X00005756.
19. Shulman, Carl. Whole Brain Emulation and the Evolution of Superorganisms. Machine Intelligence Research Institute, 2010. URL http://intelligence.org/files/WBE-Superorgs.pdf.
20. Shulman, Carl and Bostrom, Nick. Sharing the world with digital minds. In Clarke, Steve, Zohny, Hazem, and Savulescu, Julian, editors, Rethinking Moral Status, pages 306–326. Oxford University Press, Oxford,2021.
21. Sneddon, Lynne U., Elwood, RobertW., Adamo, Shelley A., and Leach, Matthew C. Defining and assessing animal pain. Animal Behaviour, 97:201–212, 2014. doi: 10.1016/j.anbehav.2014.09.007.
22. Tomasik, Brian. Do artificial reinforcement-learning agents matter morally?, October 2014. arXiv preprint arXiv:1410.8233.

23. Warren, Mary Anne. Moral Status: Obligations to Persons and Other Living Things. Clarendon Press,Oxford, 1997.