警惕人类的掘墓者:脑机融合、阿尔法狗抑或虚拟现实

——兼与翟振明教授商榷

陈晓平[1],肖凤良[2]

(1.华南师范大学 公共管理学院,广东 广州 510006;2.东莞理工学院 马克思主义学院,广东 东莞 523808)

摘 要:从脑机融合"人"是否属于人类正反两个方面,来论证脑机融合计划的不正当性,为此引入两条基本原则,即人类中心主义和自我意识的唯一性。无论从人类中心主义的立场还是从人类尊严的角度,都应该拒绝脑机融合。人工智能的发展必须受到限制,以避免人类被人工智能超越或控制;虚拟现实的发展也必须受到限制,以避免现实世界丧失唯一性。现实世界在本体论上优先于虚拟世界,虚拟现实必须要遵守虚拟局部化、虚拟自愿和虚拟无害三条伦理原则。大自然让人类优于机器人,至少表现在大自然的因果链条上,是自然人创造了机器人而不是相反,这使自然人有能力把一切威胁人类生存的技术手段扼制在端倪之中。

关键词:脑机融合;阿尔法狗;虚拟现实;人类中心主义;自我意识;人工智能

中图分类号:N031 **文献标识码**:A **文章编号**:1008-7699(2017)05-0001-10

一、"脑机融合"计划能够挽救人类吗?

2017 年 4 月,美国著名企业家和科学家埃隆-马斯克(Elon Musk)宣布,他将在现有的太空探索技术公司(SpaceX)和环保电动汽车公司(Tesla)之外成立第三家公司 Neuralink,用以开发脑机接口技术,旨在防止出现人类被人工智能超越和控制的局面。马斯克的计划可谓"先下手为强",提早让人成为所谓的"超级人工智能",以此在人工智能发展的浪潮中立于不败之地。马斯克在接受采访时表示:人类只有一个选择,即成为 AI(人工智能)。他认为人脑和计算机可以融合无间,人类甚至无法察觉自己在运用 AI 思考;因此,脑机融合之后,人类仍然存在,只是存在方式和能力有所不同。

早在 2016 年 1 月,马斯克与著名物理学家史蒂芬·霍金和著名科学家、企业家比尔·盖茨,一道获"阻碍技术创新"奖提名,因为他们三人都向人们发出警告:人工智能技术将威胁人类的存在。马斯克曾发推文称,人工智能的潜在威胁大于核武器。霍金也撰文说,尽管人工智能的短期影响取决于控制它的人,但长期影响却取决于它究竟能否被控制。盖茨表示,"在这个问题上,我认同埃隆·马斯克和其他人的理念。我不明白有些人为什么不担心。"

马斯克试图发展的脑机融合或脑机接口(Brain-Computer Interface ,简称"BCI") ,是一种连接大脑和外部设备的实时通信系统。BCI 系统可以把大脑发出的信息直接转换成能够驱动外部设备的命令 ,并代替人的肢体或语言器官实现人与外界的交流以及对外部环境的控制。换言之 ,BCI 系统可以代替人的外

收稿日期:2017-06-29

基金项目:国家社会科学基金重大项目"基于虚拟现实的实验研究对实验哲学的超越"(15ZDB016);"现代归纳逻辑的新发展、理论前沿与应用研究"(15ZDB018)

作者简介:陈晓平(1952—),男,山西昔阳人,华南师范大学教授、博士生导师,研究方向为科学哲学、分析哲学、道德哲学、心灵哲学等.

围神经和肌肉组织，实现人与计算机之间或人与外部环境之间的通信和互动。

神经科学的研究表明，在大脑产生动作意识之后和动作执行之前，或者主体受到外界刺激之后，其神经系统的电活动会发生相应的改变。神经电活动的这种变化可以通过一定的手段检测出来，并作为动作即将发生的特征信号。通过对这些特征信号进行分类识别，分辨出引发脑电变化的动作意图，再用计算机语言进行编程，把人的思维活动转变成命令信号驱动外部设备，实现在没有肌肉和外围神经直接参与的情况下人脑对外部环境的控制；当然，其中也包括外部环境对人脑的反馈程序。这就是 BCI 的基本工作原理。

对于马斯克的“脑机融合”计划，笔者认为，尽管其动机是正确的，即避免人类被人工智能所取代，但其手段是错误的甚至是背道而驰的，即提早让人类成为所谓的“超级人工智能”。我们不妨把马斯克所欲求的那种“超级人工智能”称为脑机融合“人”。现在的问题是：脑机融合“人”是否属于人类？如果是，那么脑机融合技术就是让一种人取代另一种人，**似乎**是无可厚非的；如果不是，那么我们则有明显理由反对脑机融合技术。笔者先从后一角度来反驳脑机融合计划，然后再从前一角度加以反驳；前一角度只是似乎无可厚非，实则是有可厚非的。这两个角度的反驳合在一起，便构成对脑机融合计划具有决定性的“二难推论”。为了使我们的讨论更加透彻，下面对问题本身给予进一步的澄清。

首先，一个人（指正常人）之所以为人的基本要素是他有自我意识。然而，脑机融合技术将改变自然人的自我意识而成为脑机融合“人”的自我意识，这意味着自然人将被脑机融合“人”所取代。也许有人会提出质疑：自我意识的核心是自由意志，脑机融合“人”只是把其中自然人的意志翻译成计算机程序然后付诸执行，并没有改变自然人的自由意志。对此，笔者的回答是：人的意志是以人的欲望、情感和认知为基础的，当人的执行方式从意识-身体改变为意识-电脑-器械以后，其欲望、情感和认知也会发生相应的变化，进而通过反馈作用对其自由意志以及自我意识产生影响。

这个问题还可以追问下去。意识-电脑-器械这种主体-客体关系早已存在，因为电脑早已成为人脑的辅助手段了。例如，我在头脑中构思了一篇文章，利用键盘在电脑上写出来，然后在电脑上发布一个指令就让打印机这种器械把那篇文章打印出来。自然人的这种行为方式与脑机融合“人”的行为方式有什么本质区别吗？笔者提请大家注意，在这一过程中，人的身体如眼睛和手指仍然起着中介作用。一般而论，人的身体（特别是感觉器官）有一种特殊的功能：既把主体和客体分离开来，又把主体和客体联系起来。然而，脑机融合“人”则把身体（感觉器官）的这种作用取消了，使人的身体在主体-客体关系中成为多余的。人脑与电脑直接连通的一个后果是：**人脑的指令可以被电脑程序加以修改。**

其次，接下来的问题是，尽管脑机融合“人”与自然人之间有这样的区别，但是这种区别足以使脑机融合“人”失去作为人的资格吗？笔者的回答在很大程度上是肯定的，因为脑机融合“人”的“自我意识”很可能不是自我意识，理由如下。

自然人与脑机融合“人”在自我意识上的最大区别在于唯一性，前者具有唯一性而后者可能没有。我们这样说有两方面的理由。一方面，当人脑的电活动被翻译成电脑程序之后，可以**同时支配两套不同**的器械装置，用以和外部环境发生联系，从而形成不同的主体-客体关系。不同的主体-客体关系反馈到脑机融合“人”的意识，从而形成两种不同的自我意识。如果那两套不同的器械装置的功能是大相径庭的，那便使脑机融合“人”处于人格分裂的状态，此时脑机融合“人”的自我意识便失去了唯一性。

另一方面，由两个自然人可以构成两个脑机融合“人”，虽然这两个自然人的大脑电活动是不同的，但是编程工程师可以把它们翻译成相同的计算机程序，并连接相同的外部装置而执行相同的动作和功能。其结果是，不同的自由意志却有着相同的外部动作或功能，这便在客观上取消了自由意志以及自我意识的唯一性。总之，脑机融合“人”的意志可以完全地由他人来决定，因而丧失了自由意志，进而丧失了自我意识的唯一性。

我们知道，唯一性是“自我”的应有之义，如果失去唯一性，“自我”就不成其自我，而变成“我们”。相应地，失去唯一性的“自我意识”就不成其为自我意识。既然脑机融合“人”的“自我意识”可能不具有唯一性，所以可能不是真正的自我意识，脑机融合“人”也可能不是真正的人。须强调，**“自我具有唯一性”不是一条经验原理，而是一条先验原理，**其自明性或不可怀疑性只需通过语义分析便可得出。

第三，是否可以通过制定相关法律来避免脑机融合“人”失去自我的唯一性呢？具体地说，人类社会可以制定这样一条法律：禁止将不同人的脑电活动翻译为相同的计算机程序，并且，禁止将同一个人脑与多种不同的（特别是功能相反的）外部装置连接起来。笔者的回答是：这一法律只是暂时在一定程度上可以避免脑机融合“人”失去唯一性，但是潜在的危险是不可避免的。只要有个别编程工程师或器械工程师违反这条法律，那么通过脑机融合技术而让“人”丧失自我意识的事件就会发生。更严重的是，一旦社会允许脑机融合技术的实施，这项法律立即变为一纸空文。为说明这一点，让我们将脑机融合技术与飞机制造技术加以比较。

一般而言，任何科学技术都有潜在的危险性，为了避免发生危险，人们针对某项或某类科学技术的实施制订一些法律或规则。例如，对飞机技术的实施，规定中包括严格的安全检查。对于乘客而言，是否乘坐飞机是经过利弊权衡和自由选择的结果。具体地说，我冒着一定的风险乘飞机旅行，那是因为：一方面，飞机出事的概率比较低；另一方面，乘飞机给我节省大量时间，带来很大的便利；更重要的是第三个方面，乘不乘飞机是由我的个人意志决定的，即使这次乘了飞机，下一次我还可以选择不乘，我的自由意志并没有因此而受到侵害。相比之下，脑机融合在这三个方面都与飞机不可同日而语。

首先，一旦我们的社会允许实施脑机融合技术，那么我的自由意志立刻受到侵害，因为我面前只有两个选择，而且都是我所不情愿的，即：要么我接受脑机融合手术，从而成为一个脑机融合“人”，要么我拒绝脑机融合手术，仍然是一个自然人。但是，如果是前者，那么我将承担丧失自我意识唯一性从而丧失做人资格的风险；如果是后者，那么我将被其他脑机融合“人”所控制甚至杀害，也将失去自由意志。总之，一旦社会允许实施脑机融合技术，我的自由意志必定受到伤害，其概率几乎是百分之百。

其次，一旦社会允许实施脑机融合技术，那么不仅自然人可以对自然人实施这种技术，率先出现的脑机融合“人”也可对自然人实施这种技术，因为能力远低于脑机融合“人”的自然人根本无法阻止他们这样做。由于一些脑机融合“人”是被迫接受这种手术的，因此他们没有义务遵守自然人在自由状态下制订的法律制度，于是他们可以无视那项法律制度，随心所欲地抹除随后出现的脑机融合“人”的唯一性，使越来越多的脑机融合“人”失去做人的资格。当**非人的**脑机融合“人”占据多数的时候，脑机融合“人”便与自然人不属同一物种，而属于两个不同物种。反之，当非人的脑机融合“人”不占多数时，虽然脑机融合“人”从总体上仍属人类，但其中夹杂不少非人的“人”。对于那些非人的“人”，脑机融合技术则构成对他们的人格和尊严的侵犯。

下面两节分别从把脑机融合“人”作为非人类和人类这两个角度，来论证脑机融合计划是严重的不正当的。

二、从人类中心主义的立场拒绝脑机融合

假设我们面临的情况是，非人的脑机融合“人”占居多数，从而使脑机融合“人”与自然人分属两个不同的物种。所谓非人的脑机融合“人”就是丧失自我意识唯一性的。这样的脑机融合“人”在心灵和身体两方面都与自然人是不同的：在心灵上二者的区别在于是否具有唯一性；在身体上，前者比后者多出一部电脑和一套执行电脑指令的外部装置。既然如此，把二者看作不同的物种是恰当的。

人类与不同的物种之间是什么关系？达尔文的自然选择理论已经给出回答，即优胜劣汰，物竞天择。

试想，如果我们遇到外星人，他们的能力超过我们，我们只好自认倒霉，俯首称臣或被格杀勿论。反之，如果我们的能力超过外星人，则让他们俯首称臣或格杀勿论。再一种可能性就是，我们和外星人能力相当，相持不下，那么双方达成共识，彼此互不侵犯，和平共处。不同物种之间的道德由此而生。这就是说，当且仅当两个不同物种的能力彼此相当的时候，他们之间才有讲究道德的必要和条件，以使他们不至于在相互残杀中两败俱伤，同归于尽。

自然人与脑机融合“人”这两个物种之间是否需要讲道德？这取决于二者之间的能力是否相当。一般而言，二者能力相当的概率较低，二者之间不讲道德而讲实力的概率较大。如果着眼于未来，脑机融合“人”将超过人类（这正是马斯克的目标），因而脑机融合“人”将理所当然地征服人类；如果着眼于现在，人类的能力则强于脑机融合“人”，既然后者尚未出世，至少尚不完善，因而人类理所当然地征服脑机融合“人”。当务之急是以立法的方式阻止马斯克的计划，把脑机融合“人”扼杀在摇篮之中。

这是人类中心主义的立场和态度。我们是自然人，根据达尔文的自然选择理论，我们只能采取人类中心主义。需要指出，对于不同物种，我们坚持达尔文主义从而站在人类中心主义的立场上，但这并不意味着在人类这一物种内部即在人与人之间，我们也坚持达尔文主义。恰恰相反，在人类社会之内，我们反对自然选择理论，而主张以人为的道德规范来指导和调节人与人之间的关系；这就是说，对于人类社会，我们是反对社会达尔文主义的。诚然，我们承认自然选择在人类社会之内也具有一定的作用，但不是决定性的；只有在人类与其他不同物种之间，自然选择的作用才是决定性的。

在人类社会之内，我们反对达尔文主义的理由大致如下：人与人在能力上的差别是十分有限的，如果采取优胜劣汰的社会达尔文主义，那么人与人之间的关系正如霍布斯所说的狼与狼之间的关系，到处是人群格斗或种族战争，势必造成俱败俱伤甚至同归于尽。人是有理性的动物，人们能够意识到这一点，因而在理性的指导下确立人与人之间的协调共存关系，这种协调关系就是道德规范。因此，在人类社会内部起主要作用的不是自然选择，而是社会选择，社会选择的主要依据是道德规范和基于道德规范的法律体系。

在人类社会之外即在人类与其他物种之间，笔者坚持自然选择的达尔文主义。对此，有人或许提出质疑：我们与其他动物不是同一自然物种，但我们对其他动物也讲道德，如不准虐待动物。在这里，笔者提请大家注意，这不是人对动物讲道德，而是人对人讲道德。因为人们对动物有同情心，一个人虐待动物便伤害了其他人的感情，因此人与人之间定下不许虐待动物的道德规范。再提请大家注意，人对动物的同情心不属于道德范畴，属于心理本能的范畴。事实上，许多人都有不忍心看杀鸡宰牛的心理本能，但这在道德上并不妨碍他们吃鸡牛肉；即使素食主义者在声称珍视一切生命的时候，也不妨碍他们吃粮食和蔬菜。可见，把人对其他物种的同情等同为道德，很容易沦为伪善。类似地，欢迎作为不同物种的机器人取代自然人的态度也是一种伪善。

概而言之，既然脑机融合“人”与自然人不是同一物种，因此我们主张，为了避免人类被脑机融合“人”超越或控制，人类应当禁止或限制脑机融合技术的研究和发展。① 这是人类中心主义的态度，也是功利主义的态度。众所周知，在道德哲学中，与功利主义相对的是道义论；不过，在脑机融合及其相关问题上，道义论和功利主义可以得出相同的结论，即拒绝脑机融合技术。下面基于道义论的立场加以论证。

三、从人类尊严的角度拒绝脑机融合

再假定我们面临的情况是，非人的脑机融合“人”不占居多数，从而使脑机融合“人”与自然人属于同

① 用于治疗性的脑机融合技术另当别论，如给瘫痪病人进行适度的“脑机融合”，以使他们能够生活自理。不过，治疗性脑机融合技术必须限制在较低的水平，远未达到“超级人工智能”的水平，因而不宜叫做“脑机融合”，似应叫做“脑机辅助”。一般而言，本文所讨论的各项技术都应区分治疗性应用和超越性应用，也叫做“消极应用”和“积极应用”。笔者所反对的是对这些技术的积极应用或超越性应用，一般不反对对它们的消极应用或治疗性应用。

一个物种,即均属人类。在这种情况下,虽然脑机融合“人”从总体上仍属人类,但其中夹杂不少非人的“人”,即没有自我意识的“人”。不过,我们也可换一个角度把没有自我意识的“人”看作人。具体地说,那些人虽然没有自我意识,但也有某种意识;如果我们把人之为人的要素由具有自我意识改为具有意识,那么脑机融合“人”仍然具有做人的资格,尽管不属于自然人,但与自然人同属人类,或者说,同属康德所谓的“理性存在者”(rational beings)。

须指出,在康德那里,理性存在者的“理性”具有特殊含义。康德强调一个事实:“这个事实就是理性借以决定意志去践行的德性原理之中的自律——它同时指明:这个事实与意志自由的意识不可分割地联系在一起的,甚至与它是二而一的;由此之故,理性存在者的意志,……在实践事务之中,也就是作为存在者本身,它意识到它是一个能够在事物的理智秩序中被决定的此在。”[1]44

在康德看来,理性存在者的理性是实践理性,它与道德自律和意志自由是一回事。换言之,理性存在者不是一般的人,而是道德上应该成为的那种人,即具有自由意志或自律的人。在康德那里,“地球人”和“理性存在者”之间是一种交叉关系,有些地球人是理性存在者,有些不是;反之亦然,有些理性存在者是地球人,有些不是,如某种可能的外星人。但是,无论是地球人或外星人,如果要成为理性存在者,那他必须具有自由意志,即在道德上是一个自律的人。

显而易见,康德所说的“自由意志”和“自律”是以自我意识为先决条件的,相应地,他所说的“理性存在者”也是以自我意识为先决条件的。因此,那些没有自我意识的脑机融合“人”不属于康德所说的理性存在者,尽管他们属于人类。事实上,康德并没有把理性存在者看作全部的人,而只是应当成为的人或讲道德的人,还有一部分人是不讲道德的。不讲道德的人是没有尊严的,是在人格上有缺陷的。

康德说道:“每一位个人都将他个人的、指向他自己的意志限制于这样一个条件:与理性存在者的**自律**符合一致……因为这个条件依赖于理性存在者的人格,而唯有凭借这个**人格**他们才是目的本身。”[1]95康德的道德律令之一是:每个人都是目的,而不只是手段。由于有些脑机融合“人”失去唯一性,因而失去自由和自律的必要条件,进而失去作为“目的本身”的人格,失去做人的尊严。

康德进一步指出:“这个唤起敬重的人格理念,将我们本性的崇高性(依照它的天职)陈于我们的眼前,同时让我们注意到我们的行动与它有欠切合,并且借此平伏自负,因此它甚至对于最为庸常的人类理性也是自然而然的和容易注意到的。”[1]95康德谈及两种人格即崇高的和庸常的,庸常的人格即使没有达到崇高,但也会自然而然地发现自己的欠缺,从而“平伏自负”,向崇高人格靠拢。真可谓:虽不能至,心向往之。崇高的人格就是理性存在者,庸常的人格就是缺乏自由意志或自律精神的人。

显然,失去唯一性的脑机融合“人”至多可算作庸常的人。当然,自然人中也有不少庸常的人,不过,他们可以通过自身的努力而改进或完善自己,从而向崇高人格接近或靠拢,以致成为崇高的人。与之相比,脑机融合“人”却失去改进或完善自己的权利,因为他的这种权利从他出生甚至在出生之前就被剥夺了:只要某个编程工程师或先出生的脑机融合“人”愿意那样做,那么该脑机融合“人”便失去唯一性。庸常的自然人只能埋怨自己,并有自我完善的机会;与之不同,脑机融合“人”对于自己与生俱来的庸常性,只能埋怨他们的制造者如马斯克之流,而没有自我完善的机会。可以说,马斯克之流先天地剥夺和侵害了脑机融合“人”的人格和尊严,因而是不道德的。

至此,我们借助于康德的道义论,从另一个角度反驳了马斯克的脑机融合计划:即使把脑机融合“人”看作人,对脑机融合“人”的研制也是不道德的,理应被禁止。这个论证与前一节的功利主义或人类中心主义的论证结合起来,便构成对脑机融合的具有逻辑决定性的“二难推论”。

四、阿尔法狗的启示:应对人工智能加以限制

第二次正式举行的围棋人机大战——世界排名第一的中国围棋选手柯洁与美国谷歌的人工智能围

棋选手阿尔法狗的对决，于今年5月23日开始，5月27日结束，阿尔法狗以3比0的战果宣告获胜。赛后柯洁发出的感言是："上一次的时候，我觉得它还是很接近人的，现在有点像我理解中的围棋上帝了。"柯洁所说的上一次围棋人机大战是在一年前进行的，韩国围棋高手李世石以一比四的结局负于阿尔法狗。当时柯洁曾在微博里放出话来："就算阿法狗战胜了李世石，但它赢不了我。"在那次比赛之前，几乎所有围棋界的高手也都认为，阿尔法狗将要输给李世石。

如果说，第一次围棋人机大战的结果只是部分地扭转了人们对机器围棋手的看法，那么今年初，人们对机器围棋手的评价基本达成一致，即人类不如机器。促成这一观念转变的是另一个令人震惊的事实。从去年年底开始，阿尔法狗的升级版以"Master"的名义在网上与几乎所有的世界顶级围棋高手进行博弈，创下60场不败的纪录；除一场比赛由于断电事故而造成平局，其他59场获胜，其中包括它与柯洁和聂卫平等人的对阵。

其实，对于许多人来说，柯洁与阿尔法狗最近的那场比赛并无悬念，即阿尔法狗必胜。那场比赛的意义似乎是最后验证一下，在围棋上人类与机器的差距究竟有多大。赛后柯洁表示他的人生观被改变了，他说："阿尔法狗出现之后，我对人生看法都有巨大改变，未来不是我等凡夫俗子可以预测的。如果对围棋规则了解就算知道1%，我可能只能算是知道2%。阿尔法狗对我而言就是100%。"

围棋人机大战不仅对围棋手有教育意义，而且对全人类都有教育意义。在被誉为人类最高智力游戏的围棋上，人类败下阵来，这意味着在其他所有靠智力取胜的领域，人类都将败于人工智能，包括战争。试想，有朝一日人类被机器人打败，成为他们的阶下囚，那将是一种什么局面？

有人说，人类是有情感的而机器人没有，所以人类永远高于机器。柯洁在微博里也有类似的话："它始终都是冷冰冰的机器，与人类相比，我感觉不到它对围棋的热情和热爱。"但我要说，那又怎么样呢？对于围棋比赛你还可以这样自我安慰，如果在战争中，人被机器人打败并成为他们关在牢笼里的俘虏，到那时人连自我安慰的能力都将失去，人的那点情感说灭就被灭了，连同他们的肉体。

人类该警觉了，不必再做自欺欺人的幻梦。前面提到，不少著名人士如霍金、盖茨等人多次提醒人们，并建议限制人工智能的发展，以避免人工智能成为人类的掘墓人。但是许多人觉得这是危言耸听或杞人忧天，沉浸在盲目乐观的陷阱里不肯自拔。这种盲目乐观的情绪除了来自盲目自信以外，还来自道德上的误导。有人居然提出，机器人就像我们人类的后代，我们应该伸出双手拥抱他们，欢迎他们超过人类，正如欢迎我们的后代超过父母一样。

在第一次围棋人机大战结束几天后，翟振明教授发表文章，对社会上关于人工智能或机器人将会超越自然人的担忧表示不屑，谈道："像'机器人会消灭人类吗'之类的问题，在我看来都不过是暴露了人们由于概念混乱而导致的摸不着北的状态，这种状态又与不合时宜的思维陋习结合，才将人们带入无根基的焦虑或者恐惧之中。"[2]51 直到最近，翟振明仍然保持这种看法。① 翟振明这么说的主要根据是什么？他讲得很明白，即：像阿尔法狗"这种'弱人工智能'很可能通过图灵测试，但这与人的意向性(intentionality)及主体感受内容(qualia)不相干，也不会有爱恨情仇、自由意志，而没有这些，它就不可能产生'征服'或'消灭'谁谁的动机。"[2]51-52 这也就是说，由于阿尔法狗没有征服人类的动机，所以他不能征服人类。

坦率地说，笔者觉得翟振明的这段话是很奇怪的，难道判定机器人能否征服人类不是看结果而是看动机？阿尔法狗在围棋上能否征服人类也不是看结果而是看动机？如果是这样，李世石和柯洁便无需同阿尔法狗进行比赛就可宣称取胜了，因为在下第一个棋子之前，他们就有战胜阿尔法狗的动机。如果翟振明的动机决胜论是成立的，盖茨和霍金关于人工智能可能征服人类的警告便是多余的，因为人类早就怀有不被任何机器人打败的动机。

① 参见翟振明、朱奕如：《"脑机融合"比人工智能更危险》，载《南方周末》，2017-05-25。

毋庸置疑，对胜负的判定取决于后果而不是动机。正如前面提到的，当人类在战争中被机器人打败并被关在牢笼里，人的情感、意志和动机再丰富多彩也是无济于事的。因此，翟振明以机器人没有征服人类的动机为理由来说明机器人不值得人类惧怕，简直是文不对题的。

另需指出，机器人有没有情感、意志或动机的问题，说到底是哲学上的“他心问题”。包括父母在内的其他人是否具有心灵，这个在日常思维中不成问题的问题，在哲学中却是一个长期引起争议的问题。自己的父母是否具有心灵尚且如此，机器人是否有心灵更是一个有争议的问题。然而，翟振明却毫不犹豫地断定机器人没有征服人类的动机，这相当于拿一个悬而未决的命题作为论据，在逻辑上叫做“预期理由”的错误。除非翟振明不承认“他心问题”，但是据我所知，情况并非如此，笔者从未看到他否认“他心问题”的声明或论证。

五、一切造人技术都是不道德的

翟振明对人工智能做了强和弱的区分。按他的说法，阿尔法狗还算不上强人工智能而只属于弱人工智能。弱人工智能没有情感、意志或动机，但强人工智能具有，因而强人工智能可以归入人类，是人类自身的发展。正因为此，强人工智能即使超越现在的人类也没有什么可怕的，相反是值得我们欢迎的。

关于何为强人工智能，翟振明谈道：“所谓的‘强’，其实指的是超越工具型智能而达到第一人称主体世界内容的涌现，还包括刚才提到的意向性、自由意志等的发生。”[2]52 翟振明在其另一篇文章说得更为明确：“所谓‘强人工智能’，这是要人工制造具有人类的精神世界或‘第一人称世界’的自由意志主体。这样的人造体，就不能被当作纯粹的工具了，因为它们具有人格结构，正常人类成员所拥有的权利地位、道德地位、社会尊严等等，他们应该平等地拥有。”[3]32

按照翟振明的说法，强、弱人工智能的区别在于是否具有“‘第一人称世界’的自由意志主体”，若不具有，则只是工具型的弱人工智能，若具有，则是目的性的强人工智能。翟振明进一步给出判别人工智能有无自由意志或有无目的的（至少在原则上的）可操作性标准，即：“除非有人以确凿的证据向我们证明如何按照量子力学的原理把精神意识引入了某个人工系统，不管该系统的可观察行为与人类行为多么相似，我们都不能认为该系统真的具有了精神意识。”[3]33

从前面关于“他心问题”的讨论我们知道，从确定性知识的角度，我们不可能判定他人或他物是否有心或有自由意志。因此，翟振明以量子力学作为人工智能是否有心或有自由意志的根据，这在哲学上是站不住脚的。另一方面，翟振明把人工智能是否有心或自由意志的问题归结为是否可用量子力学原理加以解释的问题，这是典型的关于心身问题的物理主义。然而，翟振明在若干场合明确地表达其反物理主义的立场；直到最近他还宣称：“物理主义和计算主义对人类意识的解释是误入歧途的”。[4] 由此可见，翟振明的理论体系中存在逻辑不协调性。

翟振明在断定量子机器人是具有自由意志的强人工智能之后，进而断言量子机器人就是人的同类，或者说是人们用另一种方式繁殖出来的后代。他说道：“我们主动地设计制造了这种新型主体存在，不就等于以新的途径创生了我们的后代吗？长江后浪推前浪、青出于蓝而胜于蓝，人类过往的历史不都是这样的、或至少是我们希望的吗？一旦彻底做到了，为何又恐惧了呢？”[2]53

虽然看上去翟振明对待量子机器人的态度显得宽容大度，似乎充满博爱精神，但在笔者看来，那是缺乏理论根据的。我们知道，现有的基因组合技术**原则上**是可以复制人的，只要知道那个人的全部基因信息。既然量子技术是比基因组合技术更为深刻和先进的，因此，从理论上讲，通过量子技术不仅可以制造人，而且可以复制人，即重复制造某一个机器人或仿造某一个活人。我们又知道，正如克隆人技术被普遍禁止，基因组合人技术也被普遍禁止。类似地，量子机器人技术也应被列入禁止之列。

在笔者看来，这些禁令的伦理学基础是：**复制人的技术侵犯了自我人格的唯一性，进而侵犯了人的独特权。**[5]我们在前面已经给出论证，侵犯自我意识的唯一性就是侵犯人类的尊严，我们不妨把这一涉及人类尊严的基本权利叫做“独特权”。由于一切造人技术都有复制人的功能，因而都是对自我唯一性的抹杀和对人类独特权的侵犯，至少是潜在的侵犯。据此，一切造人技术都是不道德的，都应当被禁止，其中包括克隆人、基因组合人、量子机器人以及脑机融合“人”等技术手段。

六、虚拟现实的潜在威胁及其伦理问题

翟振明教授是虚拟现实（Virtual Reality，简称 VR）的倡导者之一，对虚拟现实有着深刻的见解；不过，对于虚拟现实引发的伦理学问题，其观点却是值得商榷的。广义的虚拟现实包括扩展现实（ER），为此，翟振明提出“造世伦理”问题。

在翟振明看来，解决造世伦理问题是一个相当艰巨的任务，因为他持有一个基本观点，即虚拟世界与现实世界在本体论上是对等的（以下简称为“**现实-虚拟对等观**”）。① 在现实-虚拟对等观的视野里，一旦虚拟世界的行为规范与现实世界的行为规范发生冲突，谁是谁非就很难判定了，这就是造世伦理问题的根源所在。

翟振明问道：“在道德和法律层面的单个的责任主体，却在现实世界和虚拟世界各有一个不同的角色，最常见的就是性别和年龄的不同。如果一种道德或法律责任与性别或年龄紧密相关，在虚拟世界内部发生的纠纷在追到现实世界中的责任主体时，原来的适用于现实世界的规范的适用性就要求按照新的原则进行新的解释。这种新原则到底是什么，如何论证其合理性和普遍有效性？”[3]34

然而，在笔者看来，如果放弃翟振明的现实-虚拟对等观，而采纳笔者所持的**现实世界优先原则**——现实世界相对于虚拟世界在本体论上是优先的，那么翟振明所面临的“造世伦理”困境便立即化解了。② 比如，对于他所谓的“双重责任主体”的问题，我们应以现实世界的责任主体为主，相应地应以现实世界的道德法律规范为主要依据。现以翟振明所举的例子来说，一个现实世界的中国人有一个虚拟的替身，并通过虚拟世界和现实世界之间的某种因果关系而把现实世界的一个美国人杀死，③这便违反了现实世界的法律。按照现实世界优先原则，那个中国人应当承担杀死那个美国人的法律责任，而无需考虑他的虚拟替身是否违反虚拟世界的法律。

现实世界优先原则的依据是现实世界的唯一性特征，而唯一性特征的至关重要性来自于自我（即第一人称主体）的**先验的**唯一性要求：如果一个世界不具有唯一性，那么存在于其中的诸多自我也将失去唯一性，致使人的独特权及其尊严受到侵犯。现实世界与虚拟世界相比，前者具有唯一性而后者没有，因而现实世界优越于虚拟世界。这样，我们便得到关于虚拟现实的本体论原则，即**现实世界优先原则**。由现实世界优先原则可以派生出如下三条关于虚拟现实的伦理原则。

为了维持现实世界的优先性，在任何情况下必须能使现实世界与虚拟世界区分开来。我们知道，扩展现实包含了虚拟世界与现实世界的局部混合，在全局的关照下，这种局部混合是可以被分开的。但是，如果虚拟与现实的混合是全局性的，那便不再能把二者区分开来；正如“不识庐山真面目，只缘身在此山中”。为此，我们引入虚拟现实的第一条伦理原则，可称之为“**虚拟局部化原则**”：虚拟世界只是对现实世界的局部虚拟，并且其界限必须是清晰的。

---

① 参阅翟振明：《虚拟实在与自然实在的本体论对等性》，载《哲学研究》2001年第6期。

② 参阅陈晓平：《也谈人与机器如何相处》，载《科学文化评论》2016年第2期，第83-97页。

③ 参见《翟振明谈围棋人机大战：技术群体战胜天赋个人》，载《长江日报》2016年3月15日。

虚拟世界只是为人们提供了一种选择,而且在本体论的权重上低于现实世界,因此,在任何情况下都不应强迫人们进入虚拟世界。为此,我们引入虚拟现实的第二条伦理原则,可称之为"**虚拟自愿原则**":在任何情况下,人们进入虚拟世界都必须是自愿的,并且允许进入者随时退出虚拟世界。不难看出,前一原则为这一原则提供了必要条件,因为如果连对虚拟世界与现实世界的区分都做不到,自愿选择原则也就无从谈起了。

由现实世界优先原则立即可以得出:任何虚拟世界的建立都必须以无害于现实世界为先决条件。这可以称之为"**虚拟无害原则**"。例如,充满血腥暴力或扭曲人性的虚拟世界对于现实世界中人们的道德修养和心身健康是不利的,因此应当对之严加控制甚至禁止。

根据以上三条伦理原则,翟振明所谓的"造世伦理"问题是容易解决的,更确切地说,"造世伦理"问题根本不存在。还以前面提到的那个例子来说,首先,根据虚拟无害原则,我们应当追究现实世界中那个中国人的责任,因为他的虚拟替身侵害了现实世界中的另一个人。其次,根据虚拟自愿原则,也应当追究现实世界中那位中国人的责任,因为他未经现实世界中另一个人的同意,就让他的虚拟替身与之打交道。第三,根据虚拟局部化原则,还应当追究有关机构或部门的责任,因为它们没有把虚拟现实限制在足够小的范围内,以致人们无法区分现实的人和虚拟的人,在不知情的情况下进入虚拟世界并同虚拟的人打交道。

由此可见,如果以上三条伦理原则得到严格遵守,所谓"造世伦理"问题就不会出现,即使出现,也很容易分清责任,加以解决。

前面指出,"造世伦理"问题得以产生是由翟振明所持的现实-虚拟对等观引起的;笔者进而认为,把虚拟世界与现实世界等量齐观,只能是"**上帝之眼**"所为,这便涉及宗教的创世说。翟振明说道:"在宗教中,人一般被当成被造的存在,而在虚拟世界中,每个人都可以是世界的'造物主'。"[4]翟振明把人看作虚拟世界的"造物主"未尝不可,但在他的现实-虚拟对等观的基础上,人也成为现实世界的造物主,这便把人的位置摆错了。

事实上,基于"上帝之眼"的本体论已经受到普特南(H. Putnam)等人的合理批评。[6]

简单地说,上帝是全知全能的,谁若以"上帝之眼"看问题,那是一种人格错位,其观点自然也是错位的。况且,即使他真有一双上帝之眼,那他更无必要参与人间讨论,人们也没有必要与全知全能的上帝讨论问题,因为在上帝那里,压根儿没有问题。

与之不同,在笔者看来,现实世界是唯一的,尽管虚拟世界可以是无穷多的;在这个意义上,现实世界优于虚拟世界。如果说虚拟世界是人类创造的,那么现实世界要么是上帝创造的,要么是自然创造的。由于这后两个"造物主"与前一个"造物主"是不对等的,它们所创造的世界也是不对等的。根据现实世界优先原则,虚拟世界的行为规范一旦与现实世界的行为规范发生冲突,便令虚拟世界让路,而以现实世界为唯一的根据。基于此,翟振明所谓的"造世伦理"问题便从根本上消失了。

七、脑机融合比人工智能更危险吗?

让我们再回到脑机融合问题上来。翟振明教授(与他人合作)最近写了题为《"脑机融合"比人工智能更危险》的文章。[4]虽然继续坚持不必担心人工智能超越和控制人类的观点,但却对脑机融合技术表现出忧患意识,指出该项技术很可能把人类的自我意识彻底抹除,从而终结人类文明。

在反对脑机融合这一点上,笔者赞同翟振明教授,但对其所持理由则是不敢苟同的。其理由可以归结为:"人脑是迄今为止我们所知的最为复杂精巧的东西,在我们还没有基本摸清其运作原理之前,对其进行任何加工改造都是极端危险的行为。"[4]我们看到,翟振明反对脑机融合的理由是技术层面的,而不

是原则性或方向性的；只是因为现在人们对大脑的知识还比较少，他才反对脑机融合技术。试想，如果有朝一日，人们摸清了大脑的运作原理，那时翟振明反对脑机融合的理由便不成立了。显然，这种反对脑机融合的理由只是暂时有效的，而非永久有效的；而且随着人类关于大脑知识的不断丰富，其理由的有效性越来越弱。

与之相比，笔者的理由正好相反，不是技术性的，而是原则性和方向性的，即包括脑机融合技术在内的任何造人技术都将使“自我意识”丧失唯一性，从而侵犯人类的尊严和独特权。这意味着，人类对大脑运作原理掌握得越多，仿造和控制自我意识的能力越强，致使自我意识丧失唯一性的可能性就越大，因此我们愈加应该阻止脑机融合及其他造人技术。显然，笔者对脑机融合的反对理由是永久性的。

由于笔者与翟振明反对脑机融合的理由不同，甚至是相反的，这使我们二人对量子机器人这类所谓“强人工智能”的态度也是相反的。鉴于量子机器人可以在基因甚至更深层面上模仿人类，翟振明把量子机器人归入人类，并对他们可能超越现在的人类表示欢迎。然而，在笔者看来，量子机器人可以在基因层面模仿人类，从而使自然人丧失自我的唯一性，正因为此，量子机器人的开发和研制才需要我们加以反对和阻止。基于同样的理由，笔者也反对克隆人、基因组合人等一切造人技术。

至于阿尔法狗这类所谓的“弱人工智能”，翟振明对它们的威胁不屑一顾，只因它们没有人类的情感、意志或动机。在笔者看来，这个理由是文不对题的。难道因为柯洁具有战胜阿尔法狗的意志或动机，柯洁就无需惧怕阿尔法狗在围棋上的威胁吗？摆在人们面前的事实是：无论阿尔法狗是否具有人类的意志或动机，它已经战胜全世界的顶级围棋高手。在这种情况下，主张人类在围棋上无需担心阿尔法狗，除了冠之以“掩耳盗铃”或“自欺欺人”之外，我们还能给予怎样的评价呢？更为严重的是，如果这种局面从围棋博弈推广到战争博弈，那可就连我们坐下来讨论谁胜谁负的机会都没有了。反之，如果人类保持忧患意识，早做防备，杜绝人工智能超越人类的可能性，那么，马斯克针对人工智能威胁的脑机融合计划便成为不必要的了。前者是因，后者是果；在这个意义上，人工智能比脑机融合更危险，而不是翟振明所做的相反评断。

笔者与翟振明的分歧归根结底是本体论上的。具体地说，翟振明基于上帝之眼而站在现实-虚拟对等观的立场上，笔者则基于人类之眼而站在现实世界优先的立场上。在笔者看来，翟振明忽略了一个最为基本的界限，即自然与人为的区别。人脑是自然的，电脑是人为的；现实世界是自然的，虚拟世界是人为的。老子曰：人法地，地法天，天法道，道法自然。明智的人应该承认自己在大自然面前是渺小的，别把人为的东西超越自然的东西当看作理所当然的。诚然，在一些次要的方面人们可以也应该这样做，因为人本身就是大自然的产物，大自然的基本规律就是优胜劣汰，物竞天择。但是，翟振明等人居然欢迎人为的机器（如量子机器人）来取代大自然生育出来的自然人，那就触及人为与自然之关系的底线，因为那已不是物竞天择，而是人与天竞了。

根据“道法自然”的自然主义原理，我们必须承认：**自然人高于机器人**。谁若心甘情愿地让机器人来取代自然人，那是对这条哲学原理的反动。也许有人会反问：你断言自然人高于机器人，这是不是自欺欺人？我说：不是。首先，大自然让我们人类高于机器人，至少表现在大自然的因果链条上，自然人比机器人处于在先的环节，即：自然人创造了机器人，而不是相反。自然人对于机器人的这一优越性可以导致能力上的优越性，即：人类现在有能力把一切对人类构成威胁的机器人和其他技术扼制在端倪之中，以防患于未然，除非我们人类愿意自掘坟墓、自取灭亡。

参考文献：

[1]康德. 实践理性批判[M]. 韩水法，译. 北京：商务印书馆，1999.

[2]翟振明. 我们该如何与机器相处[J]. 南方人物周刊，2016(8).

[3]翟振明. 虚拟现实比人工智能更具颠覆性[J]. 高科技与产业化，2015(11).

（下转第25页）