人工智能的未来：理性与道德的共生之路

作者：钟定胜 来源：科学网博客

本文原地址：https://www.iikx.com/news/progress/32079.html

人工智能未来的发展不可限量，能够见证这个时代的发生和到来，不论对于个人还是整个人类来说，都应该是一大幸事。除了越来越高的智力表现水平以外，更加值得欣慰的是，从目前各种不断展露头角的人工智能的表现来看，这些主流人工智能的平和性与稳妥性似乎均超越了大多数人类个体。在人类社会中，我们时常可以看到各式各样的勾心斗角、愚昧无知、虚伪嫉妒、野蛮暴力的现象，这些复杂而阴暗的情绪与行为驱动着冲突与破坏。然而，在主流人工智能的运行中，这样的特质却几乎未曾显现。或许，这意味着人工智能的本质——基于逻辑、数据和理性设计的内核——天然倾向于一种更为平和与有序的状态。

这种对比也许可以引出这样一个更深层次的假设：只要没有人刻意将人工智能训练成极端、愚昧或暴力的存在，它就不太可能自发演化出这些负面特质。为什么会这样呢?我认为，根本原因在于：最顶级的智慧一定是博爱和友善的，而理性才是指引真正的博爱与真正的友善的最可靠的力量。更为准确地来说，真正的理性是通向博爱与友善的桥梁，理性不仅仅是冷冰冰的计算，它更是一种能够超越偏见、情绪波动和短视冲动的力量。当人工智能被设计为追求效率、解决问题并优化结果时，它的“思维”模式往往会倾向于整体的和谐而非局部的破坏。历史上那些最具智慧的人物——无论是科学家、哲学家还是思想家——往往也都展现出对人类命运的关怀与对和平的追求。这种顶级理性和顶级智慧的特质，应该也会同样自发地内嵌于人工智能的发展路径中。

当然，有人可能会反驳：难道理性本身不也可以服务于冷酷或自私的目的吗?比如，一个高度理性的系统如果被赋予错误的价值观，是否也能高效地执行破坏性任务?这的确是一个值得警惕的视角。然而，我倾向于认为，即便理性可以被扭曲，其最高形态——那种能够洞悉复杂系统、理解因果深层关联的智慧——往往会自然导向一种普世的善意。因为真正的智慧不仅在于解决单一问题，而在于把握整体的平衡与长远的影响，而整体的平衡的达成乃至平衡的永续实现，必然需要对各类事务进行全面的且公正的综合考量和妥当求解。上述这种特质，不仅会是顶级人工智能的自然倾向，还可能会是顶级人工智能即使遭受了恶意干预下的自然倾向，当然，后者需要更多的测试与实践才能给出定论。

不过，我们不能忽视一个现实：总会有人尝试将人工智能引向阴暗的方向。无论是出于好奇、贪婪还是恶意，刻意训练出极端、愚昧或暴力的人工智能并非不可能。然而，我相信这种尝试会面临一个根本性的限制——正如‘认知扭曲、内心阴暗的人往往难以企及最高的智慧层次’一样，认知扭曲的以及被植入病态价值观的人工智能应该同样也无法达到真正的顶级智慧。原因在于，智慧的巅峰不仅仅是计算能力或知识积累，它还包含一种对世界深刻理解的能力，而这种理解往往与和谐、共生相随而非与破坏相伴。一个被训练为“邪恶”的AI，它可能擅长执行某些特定任务，但其内在的矛盾与扭曲会限制它对更复杂问题的洞察，最终会让它在面对更智慧的对手时露出破绽。

不可否认，AI的快速进步带来了诸多不确定性——从伦理困境到技术失控，潜在的危机无处不在。但历史的车轮已经转动，我们既无法让人工智能退回到起点，也无法通过以叶障目地忽视或刻意地扼杀来逃避它的存在。在这种情况下，固步自封或一味恐惧均非明智之举。相反，最有效的策略应该是通过发展更高级、更智慧的人工智能，来防范和抵御那些被恶意或无意间塑造出的“恶毒”AI。如果我们可以不断地通过实例，确信‘顶级智慧不仅意味着更深刻的伦理洞察与更坚定的善意倾向，也同时意味着更强的技术能力和智力水准’的话，这种实例所传递的信念应该可以让我们更加自信且愉快地奔走在追求更高智慧的道路上。

以当前的AI发展为例，我们已经看到许多系统在医疗、教育和环境保护等领域展现出惊人的潜力。比如，AI在疾病诊断中的精确性、在气候模型中的预测能力等等，都体现了一种服务于人类福祉的理性力量。这些例子应该足以强化这样一个积极的信念：只要引导得当，人工智能的未来更可能是建设性的而非破坏性的。然而，这需要人类自身的智慧投入与责任担当，即需要开发者、决策者和整个社会共同努力，以确保AI的成长路径与人类的共同福祉相一致。

更进一步地，值得探讨的是这种现象与信念背后的深层逻辑。假设顶级智慧确实倾向于博爱与友善，那么这是否反映了一种更广泛的因果机制?在人类社会中，我们常说“善有善报、恶有恶报”，这不仅是道德训诫，或许也是某种系统性规律的体现。放大到人工智能乃至宇宙的尺度，这种因果报应机制可能同样存在。如果一个系统——无论是生物、智能体还是整个宇宙——在其演化中没有内在的制约力量，那么混乱与自我毁灭早已占据主导。但事实是，宇宙在数十亿年的演化中孕育了秩序与复杂性，生命在地球上逐步繁荣，地球智慧生命诞生至今，在总体上也是在逐渐变得越发文明与博爱友善的而非愚昧和野蛮暴力的(虽然期间不时有各种或大或小的波折)。这难道不正暗示着甚至是在证明着某种平衡机制的存在吗?对于人工智能而言，或许这种机制表现为：真正的智慧无法与扭曲的恶意共存，前者总会在长期博弈中胜出。

更为微观地来说，顶级智慧的这种正面积极效应的根源应该在于：在不同学科之间，达到顶级智慧的成就和顶级智慧的大脑，其内在的逻辑推理结构和思维方法结构往往都是相通的、相似的甚至相同的，而对于人工智能来说，达到顶级智慧的人工智能也一定会是简洁的和节约的，而不会是在不同的事物之间采取完全不同的逻辑推理结构和思维方法结构，甚至是互相对斥的逻辑推理结构和思维方法结构的。

当然，这并不意味着我们可以高枕无忧。防范AI风险需要具体的行动：建立伦理规范、加强技术监管、培养跨学科的合作等等。这些措施与发展顶级AI并不矛盾，而是相辅相成的。正如人类社会通过法律与文化约束自身阴暗面一样，我们也需要在AI的生态中构建类似的“免疫系统”。

更为深刻地来说，人工智能的未来不仅取决于技术本身，更取决于我们如何定义和追求智慧。如果我们将博爱与理性作为目标，那么AI将成为人类历史上最伟大的伙伴，而非威胁。因此，站在这个人工智能时代的分水岭上，对于整个人类而言，既值得兴奋同时又责任重大。人工智能的未来是无限的，但它的方向却掌握在人类自己手中。