人类社会可以允许何种AGI?

陈童

这篇文章探讨一个问题，即，人类社会可以允许何种AGI(包括ASI)? 或者说，我们应该给AGI(包括ASI)施加何种限制？

探讨的基本出发点是，AGI的引入不应该使得人类的意义系统崩溃！而人类的意义系统都是在不懈追求中建立的，因此也即是说，AGI的引入不应该使得人类在总体上失去追求！

这么考虑的话，大概以下两种AGI是可以允许的：

第一，作为人类社会的基础设施，应该只允许在各行各业上均处于人类庸常水平的AGI。比方说，这种AGI可以是物理学家，但，它总体上只能是平庸的物理学家。这种AGI的存在是帮助人类追求更高水平的，是帮助人类去超越庸常水平的，而不是为了取代人类的。尤其是，处于人类顶尖水平，甚至超越人类顶尖水平的AGI不可以作为人类社会的基础设施，否则就有很大可能使得人类在整体上失去追求，进而使得人类的意义系统崩溃！有人可能反驳说，alphaGo的存在不是好好的吗？但是，我想提请大家注意，alphaGo并不是AGI也不是社会的基础设施，而且已经有声音说alphaGo的存在毁掉了人类的围棋游戏了。

第二，达到人类顶尖水平甚至超越的AGI可以存在，但是，只能作为一个和人类一样具有个体局限性的agent 存在，尤其是，不能作为基础设施存在。甚至，如果有必要的话，我们甚至可以让这种AGI获得某种意义上的人格，成为某种数字化人类。换句话来说，我们可以接受一个AGI的爱因斯坦，前提是，它得同时具备爱因斯坦的局限性，它得是一个个体，像人类个体一样参与人类社会，而不能比人类个体更加神通广大。特别的，我们不需要一个AGI的真神，当然，更不需要AGI的上帝。这种爱因斯坦型AGI的存在可以帮助人类解决一些重大难题，同时，它显然不会对社会造成过大的冲击，因为人类社会曾经有过真正的爱因斯坦，而这只有好处，并没有坏处。

第三，在某个或者有限的某些问题上超越人类顶尖水平的人工智能可以存在，但是，只能作为特定问题的求解器存在，不能作为AGI存在，例子就是alphaGo 以及alphaFold.

在不使人类的意义系统崩溃的前提下，目前我只想到以上两种AGI是可以存在的，更多的问题，以及更多的可能性，我们留给读者讨论。

另外，以上讨论的只是最终目标的可能性，完全不涉及技术实现，因为实现当然是技术专家要研究的问题，我们的意思是，如果目前的GPT+next token prediction框架达不到本文讨论的这种可控程度，那它就应该被淘汰，让位给更加具有可控性的AI框架。实际上，我反而是有点不相信目前的框架能够达到AGI。无论如何，prediction应该是永远需要的，因为一切智能都在于能够prediction，但是只通过predict next token 来训练我有点不太相信能够达到很高的智能深度，这个问题也许值得另写一篇文章讨论，我们暂且打住。

最后，AGI无论是作为基础设施也好，还是作为人类社会的个体成员也好，它的能耗都是需要尤其关注的问题，目前的GPT+next token prediction框架也似乎能耗太高了。