

砖瓦厂早报

Zhuan Chang Zao Bao

2025.03.16

第 469 篇：AI 会失控吗？



前段时间，看到一个 AI 从业者谈到，AI 可能会失控，对其赋予宗教特性，是防止其行为失控的重要办法。

AI 这件事上，讨论起来会有一个麻烦：可能大家聊的，就不是一个东西。

有人可能认为，AI 是一个超级大脑：包含倾向于人的理解能力、自我意识等等。

要是这样的话，事情就会变得很麻烦。

因为，这种定义太宽泛了，没法具体的去对比观点。

但假设：我们把 AI 这个事物，“矮化”一些：

同样作为一个计算硬件，它和非 AI 的区别是——它擅长理解和处理“长文本”的指令。

而非 AI 擅长的是，处理明确的指令。

是不是就清晰多了。

所谓 AI 失控可能指：

1. 理解“长文本”出现偏差，导致输出的指令，出现大偏差。
2. 产生某种“自我意识”，输出挑战“伦理”、“公序良俗”的指令及动作。

那有人给出的解决思路是：给 AI 赋予爱、感恩的特质。

我们试图来理清：

他们理解的问题是什么

他对此想到的解决思路是什么

即如何让 AI 来理解爱和感恩，以及如何评价它是否具有这些特质。

但我们再去看前面给出的定义：AI 和非 AI 的区别，仅仅只是在理解和处理输入文本的长短上。

换句话说：AI 一开始，就是遵循的 0101 的计算机原理。

而非生物学的原理（包括脑科学）。

之所以，有专家认为：需要赋予它爱和感恩。

我认为：是专家可能，压根就没有理解，么子是 AI。

它想象中的 AI，与实际中的 AI，不是一个东东。

毕竟：任何人都无法，仅仅根据一个硬件在处理和理解指令的长短，来定义：人性和非人性。

那豆包啊，deepseek 所表现出来的：深通人意，又是怎样回事呢？

是大型模拟的结果。

这里面的模拟和学习，就需要用到大记忆恢复术，用到 Transformer，用到《Attention is all you need》里提到的算法。

好了，假设专家也认为:ai 的特别之处，就在于牛逼的长文本理解、处理能力。

然后，希望 AI 能够去模拟、学习人类感恩、爱的一面。

是不是也行呢？

去使用 deepseek，会发现它已经比较好的理解了：什么是感恩

而且已经知道，怎样合适的表达出：让别人体会到，它是感恩的——如屏幕上闪烁的谢谢。

把人类的各种行为，做以解析，会发现无论是爱、还是感恩，都是可以拆解为许许多多的行为的：文字、语音、动作的。

这个对于机器而言，学习起来，并不复杂。

且我们目前的 AI，如 deepseek 已经初步具备了这些能力。

那么结论自然是：我们的 AI 已经在向着体现爱，知道感恩的一面，去发展了。

又何谈，什么加入宗教元素呢。

一个硬件，也是可以充满爱和感恩的。

至少：可以让使用它的人，体会到这些特点的。

这些是养宠物，是一个道理：很多人养宠物，是在当爹当妈，全心投入

原因在宠物给出的反馈，可能比一些子女，还要棒。

一个两岁大的边牧，与人沟通，理解人类的意图，响应人的需求，对于很多人而言：

就是正向的反馈。

一个宠物主，之所以，会很深的爱它的宠物，来源于：千千万万的正向反馈。

这与人对人的情感，是类似的。

爱一个人，恨一个人，都产生与：与他（她）之间的信息传递，正向情感和负面情感的占比。

基因中，如血亲这个层面的因素，所能占的比重，可以不是必要调节。

严嵩讲：很多时候，弟子比儿子好，就是这个道理。

假设上面这层逻辑成立：那人机互动，也会遵从这个规律。

一个 AI 主人，是否认为它的 AI 充满人性，充满爱，完完全全取决于：他（她）俩互动的过程中的——信息反馈。

当一个 AI，能理解你的情绪，做好你的家务，协助你的工作，当好你的陪伴者时。

你自然而然，会把你以为的人性中善的一面，加诸于那个硬件之上。

你一定会给它，贴上那些美好的标签。因为：你大概率会赞同，论迹不论心的。

定义它的：是你俩之间的 信息交集

这个道理，在许多消费品身上，同样也是适用。

比如上一个时代，豪车品牌之所以豪华的标签：也是由许许多多多个小标签子集所构成的——发动机功率，缸数，轴距，音响、轮毂、变速箱、地盘、碰撞测试、操控性……

你这些子集，都足够优秀，且价格合适时——品牌的建立，就需要一点时间和运气了。

反过来看，电车的时代，当你的电池，你的智驾，你的碰撞测试成绩，你的内饰等子集，都普普通通时

你身上的豪华标签，自然而然，也会在人类的心智中，慢慢散去。

再具体生活中的许多环节，也有道理的通用性。

比如：用高考成绩这个单一标签，来赛选岗位候选者，通常是靠谱且兼具效率的。

为什么？

因为高考成绩的背后，是下述子集的合集：逻辑能力，

阅读能力，记忆能力，写作能力，细致程度、自驱能力等。

这些能力，又与职场中，特别是工业化生产中的许多职业技能需求，是想通的。

所以，对于许多公司的 HR 而言，单单的按照高考成绩这一个标签来挑选候选者，性价比非常高。

换句话说：事物身上的标签，可以大致的定义这个事物。

打鼻环、泡夜店、蹦（80 后 90 后）老头、不做家务、黑白颠倒、把前男（女）友的名字纹满后背的好女（男）生有没有？

当然有，只是概率低。

而人立于这个世界，人做许多事情的决策成本，是永远无法忽略的。多数人，是没有那么多的精力，去对沟通的对象，做大任务量的考量的。

为了使得决策的效率更高，人自然而然会采用“通过通用标签”来决策。

比如，一些店家，对于某些地区的仅退款行为，愿意花大力气，去起诉。

而对于另外一些地区的仅退款，是采用了信任的决策。

回到最开始：以后随着 AI 技术的不断发展，AI 产品的好与坏，慢慢会被用户所 get 到。

这个过程，就是不同的产品，不断积累各种标签的过程。

感恩性也好，爱也罢，是这些产品中的某个特征。

AI 表达出这些特征，并不困难。

我们以后，对来自于生物的感情，和来自于硅基的感情之区别，会变得越来越模糊。

许多人，深深的爱上一个硅基，一点也不奇怪。

因为，爱与不爱，取决于交互时的表现。