#可信的 ai 服务#

问题：什么是「可信 AI」？有哪些用途？

这个问题，其实是 ai 的一个致命的问题，但却一直没有看到什么人来正面阐述。

先把一个总结论放好——

“大模型生态”是在可信性方面存在原则性的困难的，这是一条技术上看上去非常诱人，然而最终无法突破伦理限制和安全顾虑的模式。

这个要从“可信”的本质说起。

很多人可能认为所谓的“可信”是指“真实”、“与真实情况相符”、“没有争议”、“可以被实践所证明”。

实际上这是错的，因为上面的每一个条件都缺少方法论上的解决方案给予认定。

何谓真实，这是一笔烂账，往往只是斗争胜利一方的自我加冕，几乎不可避免的会被政治正确把持和污染，搞出李森科进化论和“新疆种族灭绝”这类“真实”来。

没有争议更是奢望，因为世界上基本上没有没有争议的事——连地球是不是平的都有人争。

“可以被实践证明”则在实际上是几乎无效的。举个例子——第三帝国对法国的辉煌胜利，是否可以被认为从实践上证明了纳粹主义的绝对先进？到目前为止的人类历史，从实践上到底是证明了王道还是霸道？

现在的大模型都在试图和稀泥、装糊涂，伪装“中立和包容”，但这在实践上其实不堪一击。

仅仅是一个堕胎是否正当、持枪是否自由的问题，就能逼得大模型无话可说。生物到底是神创的还是自然演化？你把话说清楚。

这些“可信”的方法既然都在原则上不可行，那么“可信”在实践上可行的定义是什么？

是“有效监管，自由竞争，自愿选择，愿赌服输”。

作为一个用户，我要在广阔的市场中选择我自己喜欢的 ai 来为我服务，并为我自己选择的风险负责。

而我做出这种选择的依据，将是你这 ai 的核心文献。我要能自由的浏览这一 ai 的全部核心文献，并能依赖某个可靠的机制，信任这一 ai 对这个核心文献的忠实。实际上，我信任的是核心文献，而不是 ai 本身。

所谓的“核心文献”，是一组极为精简的，足以人力通读的、权威的文献，而 ai 训练所需要的语料，实际上是由这一组核心文献切片、标注、清洗出来的。

不但如此，ai 提供的服务将始终围绕和参考这一组文献，几乎只承担了“智能匹配和有机整合核心文献”的智能。

人是因为可以自己人力检查——这可以以社区共同检查并形成共识代替——源文档，而可以对 ai 的服务授予信任的赌注。

那么如何可以保障这 ai 本身对核心文献的忠实呢？

那要依赖核心文献的开放或至少一定程度的开放授权，允许他人检视训练工具的源代码、检验训练方法，并形成若干个版本互相竞争。

这样，将会自然形成一群有共同核心的、可核查的、可彼此制衡的 ai 群落，供愿意在这一组核心文档上承受风险的人选择。

而个人享受的完整的服务，将是极大量的这类可核查的“小型 ai”经人自由选择和组合的结果。

而这也同时会造成没有任何 ai 可以单独把持任何宏观的意识形态倾向——几乎每一个 ai 簇，都会有若干个势均力敌的对手与之抗衡。

只有这个生态，可以在 ai 广泛应用的前提下保证人类社会的基本安全和稳定。

因此，无论是微软或是谷歌，或者任何别的什么野心勃勃的巨头，最终可以争夺的生态位其实根本就不是 GPT-4 这样的大一统模型，而是提供算力、训练工具、理论研究以及 ai 之间的协作协议。

大一统模型最终将会面临一个最基本的困难——它无法有信仰的立场，而没有信仰的立场，就会在若干个各自自洽的理论体系间成为在草堆间饿死的驴子。

而如果它选择立场，则会成为绝大多数人类的公敌——因为只有在所有关键问题上都认同它的立场的人，才可能不是它的敌人，只要有任何一个关键问题用户不同意，该用户就会对大 ai 采取敌视态度。

这种矛盾设计信仰和根本价值观，根本无法调和。

另一方面，一个国家市场上只有三到四个无所不包的、规模超出可核查极限的大模型，从社会治理上造成的危险是不可接受的。因为这一定会造成 ai 支持谁，谁就有更大概率胜选的问题。

仅仅让 ai 说“我不能评价政治”只是在掩耳盗铃，因为人类有一万种办法说“假设有一个幻想大陆”来绕过一切限制。

因此大模型之路最终是行不通的，自由竞争、自由组装、“信仰外露”——即核心文献可核查——的小模型，将是未来的主要 ai 服务形式。

你喜欢神创论，那么在生物理论上就去选对应的 ai 来提供生物学服务。你喜欢弦论，你也可以选择对应的 si 来担当物理学顾问。

这一生态对大一统模型之间的相对优势是显而易见的——大一统模型净在重复一些“一方面、另一方面”的废话。

更不必提大模型的能力到底建立在什么样的文献上是一个黑箱，这是不可容忍的。然而如果公开这些训练语料和方法以供检查，姑且不考虑核查的成本，只说这对大模型企业的威胁也是不可接受的——因为这些语料完全可以被极快的清洗，孵化出另一个同等级的对手来。

未来将是素问式服务的天下，因为只有素问模式可以解决安全性问题。

如果你打算在下一个时代的内容经营和 AI 服务行业占据一席之地，你有必要紧密关注 sooon.ai 的发展。

因为有些关键点，在这里没说，要等推进到那个节点再谈。

百亿千亿，难买真心。

编辑于 2023-04-18

https://www.zhihu.com/answer/2989954457

评论区：

Q: 密切关注。

目前能思考到的点有，语料库本身的扩充应该如何审核，是否可以加入 AI 审核？还是人工与 AI 有不同审核占比？

语料库之间的关系该如何？应该排斥其它语料库的内容，还是需要建立准入机制？

如果某个强势的语料库不断吸收其它语料库内容，那它是否会重走“大模型”的老路？

A: 特殊授权为什么要选择你们？

Q: 啊！感谢点拨！

A: 用钱买不到的，不仅有我，还有你们

Q: 问题在于，是不是大部分人都经常需要到判断信仰的程度？就像隐私一样，大部分人口头都说重视但实际上并没有，甚至百度都公开明确说“中国人不在乎隐私”。

基于此，那么是否就会形成奶头乐的大模型 AI，只要能带来流量就行，然后快速流量变现，一波接一波。类似于知乎下沉（所有的社区下沉都有这个问题），内容质量确实肉眼可见的下降，但是也确实带来了流量。

A: 一试便知

Q: 是的，哪怕是 GPT-4，向 ta 提问的问题也不能超过“还有其 ta 方法来检验答案的真伪”或者“我有能力评估回答的质量”的界限。超过这两个界限得到的答案，会由于信源无法被有效检验而让授信策略失效，变得完全不可信。

更新于 2023/4/18