·人工智能与未来社会：ChatGPT 专题·

GPT 推进哲学问题了吗

赵汀阳

【内容摘要】当前的人工智能仍然属于图灵机，其采用的是经验主义和进化论原则。GPT 是图灵机的一种升级版，其思维方式已经从机械主义转变为经验主义和进化论。从根本上说，人工智能突破奇点尚需获得自我意识、反思性和创造性，这意味着必须建立人工智能自己的语言，并将“数据作业”转变为“思想作业”——虽然不一定需要先天语法，但需要有先验的逻辑。尽管目前人工智能技术尚未突破奇点，但技术文明的发展水平已经远远超过制度文明的发展水平，未来这将划时代地全面改变世界。因此，人类需要一个“新启蒙运动”来应对新技术所带来的问题，注重技术发展与制度演化之间的平衡，以确保人工智能技术的发展能够为人类带来更多福祉。

【关 键 词】GPT　人工智能　主体性　图灵机　奇点

【作　　者】赵汀阳，中国社会科学院哲学研究所研究员，中国社会科学院大学哲学院教授。（北京 100732）

何种意义上的哲学

本文标题中的提问明显是一个后续问题，接续的是若干年前我在另一篇文章的提问“人工智能提出了什么哲学问题”。[①]那篇文章涉及的背景事件是“阿尔法狗系列”，在此不再复述。这个后续问题的背景事件是“GPT 系列”（此刻已经由 ChatGPT 升级到 GPT-4）。如同阿尔法狗事件，ChatGPT 引起世界轰动，“更能干”的 GPT-4 更是引爆全网，一时间颂词滚滚。不过，此类在背后有着“无限商机”的商业化或传媒化轰动几乎都言过其实。不实之词往往很成功，比不实之词更受欢迎的是完全不实的谣言。然而，GPT 系列获得的颂词可不是谣言，虽然有些言过其实，但它的确有真本事，而且潜力很大，可以想象其后续迭代更可能有惊人之举。在技术上，GPT 系列确实推进了人工智能的神奇应用，但这并非对思维的实质推进，仍与阿尔法狗同属一个技术级别，是这个技术级别里的高水平应用，简单地说，GPT 很杰出，但尚未形成人工智能的技术代差。

① 赵汀阳：《人工智能提出了什么哲学问题？》，《文化纵横》2020 年第 1 期。

GPT 系列的厉害之处在于进入了语言领域，而语言是人类的本质，这就切中了人类主体性的要害，问题就严重了。那么，GPT 系列人工智能推进哲学问题了吗？

这里的讨论首先排除了对人工智能进行人文主义批判或伦理批判，这种“哲学”不属于本文考虑的哲学概念。价值批判只是表态，未涉及实质问题，属于日常价值判断，严格地说不属于哲学。凡是相当于“我认为这是好的”句型的意见都不是哲学。人工智能是势不可挡的，对人工智能的人文主义或伦理批判只能说明哲学没有能够回答人工智能提出的实质问题，如对意识、主体性、智能等概念的挑战。

就哲学的不可替代的思想功能而言，哲学研究任何思想“界限”问题，如知识的基本假设、思维的基本设置、价值的最终根据等，而对思想界限的研究必定形成思想的自反性或自相关性（reflexivity or self-reference）从而达到反思的极限。除了对思想边界进行极限反思，哲学的其他功能都是可替代的，实际上其已经被科学和社会科学所替代。思想的极限边界意味着思想走不出去了，只能自我说明。正如维特根斯坦指出的，哲学问题的一般句型相当于：我不知道怎么走了。这不是迷路，而是没有路了，思想没有更多的理由了，于是回到了初始状态。本真的哲学就是思想迫使自身回到思想的初始状态，对思想进行再创造。不在思想初始状态上工作的是史学，不是哲学。那么，这里的问题是，人工智能在哪些问题上迫使思想回到了初始状态？逼得思想无路可走？

图灵机人工智能的能力界限

当前的人工智能仍然属于图灵机，包括近来出现的 GPT 系列。图灵机人工智能的能力可以大致分析为如下方面。

（1）物理能力。计算机的速度超过人类无数倍，而且可以不休息，在高速度下，最简单的技术也有难以置信的高效能力，所谓“唯快不破”。可以预料，在不远的将来，各种类型的图灵机可以联合起来形成大系统，就像人类形成一个大集体，以万众一心的方式联合作业，很可能会形成类似“全知全能”的上帝的效果，那么，机器人就成为通用人工智能（AGI）。我相信这是一个合理推测，就是说，制造类似个体人那样的个体化通用人工智能不太现实，属于科幻片的拟人化想象，比较合理的想象是，那些貌似个体的通用人工智能是联合作业的，实际上属于一个系统，因此，通用人工智能最可能是“系统人”而不是“独立人”。

（2）来自人类的设计能力。这意味着，任何图灵机的设计能力都不超过人的能力，只能小于或等于人类思维能力，类似于某种速度无限逼近光速。需要指出的是，人工智能的“心灵”可以与人相似但不必要与人相同，它完全可以是另一种心灵。当然，由于人的心灵是目前的唯一榜样（外星人还存在于科幻作品里），模仿人类心灵就是现成路径，但事实证明，逼真模仿人的心灵其实最难。人的心灵像是“上帝”的作品，要模仿人类心灵就需要破解“上帝”的智能，这似乎超越了人目前的智慧。有趣的事实是，计算机的主流设计从来就不是对人类心灵结构的复制性模仿，而是有用性的功能模仿以及对相关功能的原理模仿。莱布尼兹为计算机设想的二进制逻辑——数学表达就已经确定了功能——原理模仿的进路。二进制数学对于人的心灵显然不方便，绝非人的自然选择（大多数人类根据手指自然地选中了十进制，也有十二进制之类，不知什么理由，但据说以数学观点看，七进制的功能最优）。然而对于机械运算而言，二进制却是最优。人工智能的运算方式虽与人脑运算不同，但也是人设计出来的，仍然是人类智能的一种可能性，相当于人类智能里的

一个可能世界。因此，面对人工智能的“不同算法”无须惊讶，类似于人类发明轮子的时候并不是模仿走路，而是为了实现搬运的功能。事实上，所有机器都是实现人想要的功能而不是模仿人的自然所是。总之，图灵机人工智能的设计能力属于并且不超过人类智能，尽管在物理速度的加持下显示出超人能力。

（3）人工智能的惊人知识量来自人类的大量“喂食”，以及人工智能自我训练和互动学习的不断迭代。通过这些方式，人工智能可以获得人类所不及的巨大数据量，并在理论容量上获得人类全部知识。此外，通过与人类进行互动学习，人工智能将来有望接近于“全知”，但达到“全能”则非常困难，因为这需要更加复杂的神经网络设计，即使是保洁员的简单劳动也需要无比复杂的神经网络设计，因此通用人工智能的实现尚需时日。实现“全能”所需要的智能复杂度远高于“全知”，这可能意味着存在某个深刻的智能问题。虽然尚不能确定这个问题是什么，但似乎提示了一种思路，即收集一切知识的博学能力和无漏记忆的“活字典”能力并不需要高智能，也不意味着高智能。真正的高智能可能是量子式的能力，包括反思能力和传说中的“统觉”能力。这种能力的实现或许需要等待量子计算机去证明。

以上综合能力已经足以使人工智能形成惊人的能力，但它终究没有且远远没有把人的全部思维，尤其是高级思维能力翻译为机器思维。这里的障碍是一个尚未解决的知识论问题，即人类思维对于人类自身也不完全透明，我们并不完全理解人类的思维。人类思维有一部分仍然是黑箱，尤其是创造性思维，即从 0 到 1 的创造方式是目前无法解释的。创造性的秘密还无法还原为心理学、生物学和神经学的理论，因此目前没有理论能够解释，换句话说，创造性至今没有翻译方式，我们甚至不可能教给另一个人如何进行创造性思维，更不用说教给人工智能了。目前人工智能所谓的创造性思维是假的，无非是心理学水平的联想和组合，并非从 0 到 1 的创作。人类思维的另一部分属于公开程序，即知识的生产程序，基本上都可以还原为函数关系。从理论上说，知识生产程序可以“喂”给人工智能，但也没有想象的那么容易。程序输入相对容易，但移植知识的意义却不容易。知识是一个解释和自解释系统，要真正理解一种知识，就需要理解知识的系统和结构。这意味着，良好地理解一种知识的意义就需要配备一个良好的解释系统。可是，人类的知识解释系统并不完善，存在许多直观或默会的理解，就是说，人类思维方式存在着许多难以解释或难以证明的概念、假设和意义，有着作为思想底层结构的形而上学，因此不能完全程序化，也很难“喂”给人工智能所有的知识生产系统——但“喂食”某些知识系统，例如“足够清楚的”数学系统，则是可能的。

吊诡的是，人类生产知识的能力超过反思知识的能力，两种能力并不对称，这是一件有些神秘的事情。就目前比较明确的反思来看，知识生产的主要方法是还原法，还原即简化，把难以理解的复杂性化简为心灵一目了然的简单关系，或者说以“清楚明白”的事情去解释混沌不清的事情。笛卡尔想象的“清楚明白”大致相当于在数学和逻辑上能够理解的命题。还原以最简单的方式显示了思维的两个底层原理：其一是经验性的相关性，典型地表达为函数关系，其最简单的形式就是逻辑和（$x \wedge y$）与逻辑并（$x \vee y$），简化到这个层次的基本命题，任何智能都可直接理解，没有更基本的命题了；其二是形式的分析性，最简单的形式就是基于实质蕴含即真值蕴含（$x \rightarrow y$）的逻辑和数学推论，这也是最基本的命题，没有更基本的了，任何智能都能够直接理解。即使亚里士多德的古老逻辑或毕达哥拉斯和欧几里得的初步数学推理也不是很简单的，而是比较“高级了”，已经包含许多不彻底甚至不清楚的“自明”假设、概括性概念和一般化原则，还有一些需要经验背景的内容，并不是机器能够直接理解的。当然，可以把所有数学系统都喂给机器，然后机器照章办事地假装懂。这个“假装懂”的有趣问题稍后再讨论。

还原论的思想目标是发现因果性和必然性。这两个概念看似简单，其实非常复杂，并非自然直觉的，而是形而上学假设的。它们是人类的发明，自然界并没有给我们因果性和必然性的概念，或者说，自然现象中并没有直接显现出因果性和必然性。实际上，自然界中甚至不存在必然性，只有不确定性和无限复杂性。必然性纯粹是逻辑和数学的发明，并且只存在于封闭且可计算的系统中。既然必然性不存在于自然界中，因果性也变得有些可疑了，似乎最多只是无限逼近必然关联的极大概率。此外，我想提醒大家，可能性和概率的概念也都是人类的发明，甚至同一律、矛盾律和排中律也是人类的发明。这些规律在自然界中也是可疑的，只存在于人类的思维结构中。可以说，大多数概念都是人类的发明，而概念就是思想的边界，创造概念就是开拓思想边界。但是，这也意味着人类思想基础的概念都是一般普遍或高度概括的，有着难以分割的丰富意义和整体性，因此无法还原（化简），这意味着还原方法的局限性。人类早就注意到概念或思想甚至自然事物的不可分割的整体性，因此哲学上形成了相对于还原论的整体论，在今天表现为最新的一种综合科学，即“复杂科学”。简单来说，就目前的技术水平而言，人工智能可以进行还原论的思维，但尚无法建立整体性的思维。因此，人工智能目前只具备运算能力，尚无思考能力。

这里的讨论试图说明，我们所知道的思想仅限于人类发明的思想，而且唯此一例，在人工智能得以自己建立主体性思维之前，不存在另一种思想。人工智能的惊人之处在于其运算的效率，在工作能量上远超过人（类似核能比人力强大）。但人工智能的工作原理或思维能力目前只能无限逼近人，却不可能超越人，因为人工智能的思维方式也是人的发明。人类专门发明了一种最适合机器的思维方式，但这种适合机器的思维方式还不能实现充分的思维，即兼备还原论和整体论双重能力的思维。因此，人工智能还没有思想，充分的思维不一定是人的思维，可以是外星人或人工智能自己发明的思维，只是图灵机人工智能办不到。

人工智能的经验主义和进化论

目前的人工智能都属于图灵机，可是图灵测试却恐怕已经失灵了。这个有趣的事情说明，图灵把测试标准定得太低，难不倒 GPT 系列人工智能，反倒只能从过于标准化或过于政治正确的回答来推测谁是人工智能。正常人大概不会坚持不懈地说些滴水不漏的废话，除非是精神异常或人工智能。GPT（包括最新的 GPT-4）提出的新问题是，它属于图灵机，却有能力通过图灵测试。“像人而不是人”这个新问题废掉了图灵测试。为什么可以这样？这就需要分析 GPT 的思维原则。目前的人工智能都采取经验主义和进化论原则，这样的思维水平大致相当于原始人。人工智能的“学习”，主要意思是收集材料和记忆，而其“训练”的主要意思是吃一堑长一智。如此简单听起来令人失望，但加上无敌的运算速度就有神奇效果了。

传统的图灵机相当于数学直觉主义的信徒，被称为“布劳威尔型号”，其知识生产范围受限于能行有限步骤可实现的确定必然结果，就是说，它只能承认封闭领域内的确定知识，遇到有限步骤无法解决的问题就会出问题。例如，让它计算一个悖论或圆周率之类的问题，它可能会死机或永不停机。这种图灵机只是按照给定规则执行任务，显然缺乏主体性，是工具而不是主体。它的思维方式可以称为机械主义。GPT 是图灵机的升级版（很快还会有更高级的），其思维方式已经从机械主义转向经验主义和进化论。GPT 的思维不再是封闭的，超越了“布劳威尔型号”，变成了“维特根斯坦 2 型”（符合后期维特根斯坦哲学），其思维居然有了黑箱，也就是说，在建立

信息或语言关联时具有语境化的不确定性、灵活性或即兴性。因此，它形成了思维不完全透明的黑箱效应，即使设计者也不完全知道它是怎么想的，比如说不清楚它在什么时候和为什么会突然“一本正经地胡说八道”。但是这种“自主性”并不等同于主体性，GPT 并没有自己的信念和想法。

GPT 的思维技艺尚不足以发展出传媒夸大其词的通用人工智能。GPT 的大语言模型(LLM)“思维”大概是这样的：首先是获得语言词汇和用法的大数据，然后进行“预训练”，即在语言大数据里去发现统计学意义上的概率性规律或搭配模式。一旦掌握了大量此类统计性的规律，就会以不是人的方式说很像人的话。比如发现 you、eat、an 和 apple 这几个词汇大概率相关，就知道可以说出“you eat an apple”这句话，至于什么意思，人工智能并不懂，只是假装懂，即知道关联性，但不知道关联性背后的思想。这样的预训练是完全经验主义的，类似于原始人在没有先验语法的情况下以完全经验主义的方式发展一种语言——当然其实不如原始人，原始人是真的懂语言的意思的——准确地说是发现了大量高概率的关联。这种所谓的“训练和学习”就是以真实标签代替了人工标签，意味着不需要先天语法的彻底经验主义。有趣的是，乔姆斯基对 GPT 缺乏先天语法表示了不满。GPT 的语言训练（学习）几乎完全符合后期维特根斯坦的语言游戏理论。后期维特根斯坦在关于意义的问题上放弃了先验论而采取了经验主义的分析，其中还发展了一种近似于数学直觉主义的理解，所谓“意义在于用法”，就是只信任“有限实例”，而不是依靠先验普遍原则来理解意义。简单地说，维特根斯坦相信实例（examples）的有限集合定义了语词和可能语句的意义值域。GPT 正是这样做的，可以说，GPT 是个维特根斯坦型的经验主义者。

GPT 不使用概括性的原则，而是通过实例集合来形成意义，准确率非常之高。不过，海量训练和学习虽然能够通过实例的增长来实现理解的增长，但永远存在例外，也就难免有时会“胡说”。那么，如果引进乔姆斯基的先天语法或深层语法，GPT 的语言水平会有所提高吗？其实乔姆斯基的先天语法研究并不完善，并不能证明全人类真的有一种通用的先天语法，至少汉语的语法就显示出某些深层的差异（可参考沈家煊先生的理论）。可以肯定的是，语言的意义域存在着大量发散的关联，似乎更适合经验主义而不是先验论的理解。

既然语义关联有着大量不合逻辑的“文学化”链接，那么，GPT 如何在开放条件下去保证语义关联的经验主义有效性？假如没有理解错的话，GPT 的策略大概是这样的：除了基于大数据的统计学，同时还使用了预测—修正程序，估计就是贝叶斯概率推理，这样就可以理解 GPT 何以能够从特殊推导出一般模式。当然，这些模式并不是普遍必然的，只是在不断修正中相对最大可能性的模式。在这一过程中，引入了一个更加拟人的方法，即行为主义的奖励—惩罚原则（行为主义是互动经验主义），以此诱导其思维的加速优化和强化，称为“强化学习”。强化学习需要与真实的人互动，人对其回答的积极或消极反馈就是所谓奖励和惩罚，GPT 据此来调整其模型参数。但是有个疑点：人类会给出大量自私、无聊、偏见、狭隘和恶意的反馈，与人类互动所获得的奖惩参数恐怕很难产生最优结果。为了控制不良因素，GPT 只能引入一些人工标签，于是其经验主义就不再纯粹了。比如，GPT 会鼓励说“you eat an apple”，但不鼓励说出“you eat shit”。这样长期学习下去，GPT 会不会变成一个平庸的迎合者？世界上多一个平庸之辈不要紧，但人工智能这个响亮的名字或可能导致 GPT 被识别为“思想权威”或“人民代言人”。

也许，人工智能还可以引入更多复杂一些的思维模型。按照我的想象，比如博弈论和演化博弈论的一些模型，还有复杂科学的一些模型，包括因果涌现（causal emergence）模型，还有溯因推理（abductive logic）之类，都应该对人工智能有用。总之，加持多种技术会有助于更准确地形

成“意义涌现”，并且在无限迭代的训练和学习中不断更新“意义涌现”。可以想象，这个过程无限逼近人的经验主义进化方式，而且依靠高速度把万年实现为屈指可数的天数。

不断有新因素加入的迭代就是进化，人工智能的高速迭代实现了“强进化”。这样的高速进化看起来会让人工智能无限逼近人，那么是否会超越人？是否能够成为超人的新主体？请允许我提出一个“新芝诺问题”。众所周知，按照芝诺的算法，阿基里斯永远追不上乌龟，但在物理学上，阿基里斯当然瞬间就能超过乌龟。“新芝诺问题”的要点在于，人类知识可以无限发展，但受到生物学的限制，人类的智能存在着极限（心灵和身体的能力都有其极限），相当于智能被上帝锁死，因此，人类智能有着某种无法超越的智能常数，类似于光速是宇宙的一个不可超越的常数，而人工智能的设计智能来自人类。给定人工智能限于图灵机，那么合理的推测是，图灵机人工智能可以无限逼近人的知识，但无法超过人的智能常数，类似于不可能比光速更快。如此，在智能常数的限制下，人工智能阿基里斯就真的追不上人类乌龟了，当然两者会无限逼近。给定这个情况，无限逼近人类智能的图灵机将是人类最好的工具，能够帮助人类创造更好的生活。然而，人类念念不忘的“自虐”问题是：人工智能何时超越人成为新主体？人类提心吊胆而兴奋地等待这一天的到来。

人工智能何以突破奇点

GPT 的互动表现使人在一种观看恐怖片的自虐兴奋中不断追问人工智能是否将要成为超级人工智能。其实 GPT 追求的只是成为比超级人工智能低一级的通用人工智能（AGI）。通用人工智能尚未形成一个通用定义，但一般来说，AGI 是一个比超级人工智能要谦虚一些的概念，其确定的意思是“样样都能干”，但不保证“样样比人强”。至于 AGI 是否具有自我意识，却是一个尚无定论的问题。只有当一个问题被极端化而形成思想自反性，才成为哲学问题，而那些在技术上能够解决的问题都被消化为科学问题，因此，这里要讨论的只是极端化的人工智能问题，即人工智能将来是否能够突破人类的智能常数而成为一种真正的新主体？这个问题的惊悚性等价于外星人来到地球。人类一直是地球上的唯一主体，如果出现了新主体，人类的主体地位就成问题了。这个问题属于提前预告，但预告有可能是错误的，人们对未来的预测似乎很少是正确的。

人工智能突破奇点有两种可能性：（1）超越人类的智能常数，这必须能够产生与人不同而高于人的另一种思维；（2）达到人类的智能常数，又有着比人类智能更大的运作能量。可能性（2）是安全奇点，看起来非常可能，只是需要时间，但可能性（1）是危险奇点，幸亏目前还难以想象。从根本上说，人工智能突破奇点需要获得自我意识、反思性和创造性，这三者密切相关。

讨论人工智能自我意识的可能性，就必须分析冯·诺依曼问题。他在人工智能的发展初期就提出了一种可能形成自我意识的智能机器——“自复制机”。几年前，我称之为“哥德尔机”，后来听说早就有个科学家命名了“哥德尔机”，看来所见略同。冯·诺依曼的自复制机与哥德尔机思路相似，但能力弱于哥德尔机，这正是我们要讨论的。依照冯·诺依曼的思路，使用 Quine 自相关递归技术，人工智能机器就可以实现打印复制自身，于是实现了自我复制。假设这可以实现（理论上没有困难），那么问题是，以递归技术实现的自我复制，是否等于实现了自我意识？这其中似乎大有疑问。事实上，一切生命现象都是基于自我复制（基因复制），却只有人类有自我意识。这说明，一个机器能够对自己给出指令，把自身程序“宾格化”为一个打印任务，这个技术并不能证明机器能够理解这样做的意义，“自我复制”这个任务并不能自动产生“（自我复制）是复制了我

自己”的意识。换句话说，“自我复制”并不必然蕴含“我知道我做的是（自我复制）”的语义，“我知道我做的是……”这个语义是多出来的部分。因此，自我复制并不必然能够形成自我意识。类似的，动物出于本能为生存而战，积极生育，但恐怕不知道这样做有什么意义。

意义属于一个系统对自身反思而产生的解释，尤其包含对未来的期待值。人工智能未必理解它所进行的游戏有什么意义，更不知道它作为一个系统有什么意义，所以没有自我意识。如果按照康德的标准，要求就更高了，自我意识需要达到自治自律性（autonomy），自己能够为自己建立自己遵守的秩序，即自我立法。哥德尔机没有达到自我立法（当然它没有这个需要），其“意识”也就没有主动的建构性。这里的分析也是对我自己观点的一个质疑和批评，在前几年的文章中，我相信图灵机如果升级到哥德尔机就有望形成自我意识，现在看来很有疑问，特此纠正。

看来自我意识先需要具备反思能力。反思不仅仅是认识自身，更重要的是拥有自身，即对自身拥有所有权、自主权和立法权，就是康德说的“autonomy”。目前人工智能之所以还不是主体，不在于能力还不够强（能力的局限性会在高速迭代中被解决），而在于它的思维能力虽然落户在机器上，但并不属于机器而属于人类，人工智能并没有拥有思维，只有思维的使用权，却没有思维的所有权、自主权和立法权，相当于说，人工智能是人类智能的经理，却不是主权人。假如反思性仅仅达到认识自身，就只是自我意识的必要条件而不是充分条件。如果反思没有建构性（创造秩序或系统的能力），自我意识就功败垂成。

哥德尔对数学系统的反思是最伟大的自我认识之一。哥德尔使用自相关递归技术为数学系统创造了元语言，使一个系统的整体性质被反身地表达出来，相当于“我终于知道我是谁”。哥德尔的反思是目前对“我是谁”这个哲学问题的唯一成功回答。虽然冯·诺依曼的自复制机具有哥德尔式的技术，但是它不具备反思自身的一个完整的元语言，只有一些属于元语言的句子，这种残缺的元语言并不能形成完整的反思，所以严格地说，冯·诺依曼的自复制机与哥德尔机还有些距离，只是思路一致。然而，即使是想象中的哥德尔机，如前所论，也不足以形成自我意识。哥德尔的元语言只有反思能力，却没有建构能力，不能创造一个更好的系统或者解决给定系统的不完全性问题（哥德尔在晚年对自己只破不立的伟大成就感到有点遗憾）。前几年，我对哥德尔机的想象过于乐观了。当时我没有意识到，仅仅知道“我是谁”并不足以实现对“我”的建构，或者说，“我是我”并不必然蕴含“我属于我”，更不必然蕴含“我为我立法”。这个问题对于人类同样适用，一个人之“我是我”并不等于“我属于我”或“我为我立法”，思想有可能都是别人的。康德的主体性标准其实极端苛刻，全世界也没有多少人能够满足康德的标准。

人们对超级人工智能的期待有些类似于对三体人的期待，表现出一种自虐的兴奋。但冷静地看，这是一件非常困难的事情。人工智能要获得具有建构性的反思能力，就需要拥有自己的语言。语言几乎就是思想的本质，至少在维特根斯坦看来，语言的界限就是我们世界的界限，没有语言就没有思维。目前，人工智能还没有“自己的”语言，它不认识它说出的话，只理解那些话的底层数据关联，因此人工智能其实并没有说话，它只是表达了数据关联。乔姆斯基批评GPT没有语法，只有数据关联，这个批评是正确的，不过人工智能是否需要先天语法，却是个未定问题，我们还无法判定。GPT的经验主义算法和进化论迭代是否能够或不能在将来突然产生自己的语法，这也是一个未定问题。在我看来，语法虽然重要，但不是最重要的，根本的问题是，人工智能将来是否能够发展出自己真正的语言。所谓“真正的语言”，是指一种语言不仅仅是一个能够表达任何事物的符号系统，而且能够反身地分析、解释和建构自身，即一种语言同时也是自己的元语言，而且这种语言及其元语言都

是非封闭的，因此永远有着建构的余地，自然语言就是这样的。人工智能将来是否能够发明自己的语言？这确实令人好奇。创造一种语言相当于创世，这是知识论革命的存在论事件。

这里涉及一个更深层次的问题。人工智能之所以很难把“喂”给它的代码系统转换为自己的语言，其中一个难点就在于人工智能的思维对象或思想空间与人类的思维对象或思想空间有着本质的差异，而要自己想办法开拓一个新的思想空间显然很难。人工智能的思维对象是给定的数据，从存在论意义上说，人工智能的思想对象都是已经存在的，它不能处理尚未存在的事情。人类的思维对象不仅包括给定的数据（相当于已经存在的事情），而且包括尚未存在的事情，表现为“可能性”，同时在逻辑上也自动包含了“不可能性”。就是说，包含了一切事实命题和一切反事实命题，或者说，包括可能世界的无穷集合以及不可能世界的集合。于是，人类的思想空间在“如果–那么”的张力中展开为无限空间。与之相比，人工智能只拥有“是 / 非”所定义的无张力思想空间，人工智能的世界显然小得多。需要解释一下：人工智能在假装“说话”时当然会使用“如果–那么”的句型，但“如果–那么”对于人工智能是表达数据关系的一种工具，不是思想对象，这意味着可能性、不可能性、反事实命题都不是人工智能的思想对象，虽然它可以说到这些事情。人工智能会“合逻辑地”说出一大堆推理，但只是在数据里识别到了固定关联，不知道那就是推理，人工智能只是在做“数据作业”。换句话说，人工智能可以说出让人们激动不已的长篇大论，但这是人脑读到的“思想”，而人工智能自己并没有看见思想，对于人工智能，思想无非是数据的概率关联，就是说，真正属于人工智能自己的“思想”只是数据相关性。

笔者替人工智能想过一个越俎代庖的问题：既然人工智能可以理解相关性，相当于可以理解逻辑关系“∧”“∨”，那么它是否能够通过真值关系而突然发现了逻辑蕴含“→”的意义呢？如果能，就相当于理解了“如果–那么”，也就应该能够理解什么是可能世界，从而把反事实命题变成思想对象。这个想象不知道对不对，但无论如何，人工智能要突破奇点，就必须能够建构自己的语言，把“数据作业”转变为“思想作业”——未必需要先天语法，但需要先验的逻辑。如果不能，人工智能就恐怕无法为自己建立主体性和自我意识。

至于超级人工智能（如果可能的话）将来是否能够解决那些人类不能解决的问题，这在理论上仍然是个悬念，因为这需要难以置信的创造力。可以试举几个人类无力解决的问题，似乎都触及了人类的智能极限。（1）无穷性或预测未来。由于无法遍历地认识无穷可能性，因此不可能完全解决无穷性或预测未来的问题。（2）存在着一些真理无法证明，尤其是无法证明基本假设或涉及经验的普遍真理。这也与无穷性有关。（3）悖论或自相矛盾的事实，可以回避，但不能解决。（4）系统的一致性与完全性难以兼备，这来自哥德尔定理，永远无法保证一个系统没有漏洞。（5）人们互相不同意，这个“他人不同意”难题也似乎永远无法解决，因为资源稀缺导致利益冲突，主体性争夺精神世界也导致无解冲突。（6）价值排序永远存在两难和歧视，不可减省的需要太多，因此无法解决优先性的问题。（7）价值问题不存在一个真理解或普遍必然解，这来自休谟问题。简单地说，不存在关于“好”的普遍必然定义，因此，价值只是一个语境化的变量而不可能成为一个常数。

人类不能解决这些困难，但可以回避，以免不可自拔。人工智能恐怕也不能解决这些困难，但图灵机甚至不会回避困难，或者死机，或者不停机。似乎可以想象，超级人工智能可以像人一样回避困难而另辟蹊径，这就需要创造性。但创造性对于人类自己来说也是一个黑箱，我们不知道创造是如何进行的。但有一点可以肯定，创造性决不能还原为联想和组合，那样过于简单了，属于心理学的解释，与思想的创造性有着比较大的距离。真正的创造性一定有智力难度，主要是

创造概念和理论、发现规律或提出定理。在这个意义上，GPT 还没有创造性，它的艺术或文学作品虽然技术精良，但其艺术品质是平庸的。创造性有着逻辑或数学无法表达的品质，这一点似乎说明了人工智能难以发生创造性，因为人工智能的本质是数学和逻辑。

人工智能的现实挑战

尽管人工智能尚未突破奇点，但它将要划时代地全面改变世界，这是一定的。人工智能带来的可能后果很多，这些可能后果已有大量讨论，比如会导致劳动、手艺、经验、博学的贬值，最终导致人的废物化；人工智能加持的元宇宙或许会导致真实世界和人际关系的贬值，最终导致生活的意义消散（dissipation）；更深刻的问题是存在论的危机，万一人工智能变成新主体，世界就会成为多物种主体的世界，人类单方面做主的历史就终结了；还有政治上的风险。在这里，我想重提十多年前的一个预言，即以人工智能和基因技术为代表的高新技术的发展很可能会给人类社会带来新的挑战：（1）技术与金融资本结合的新专制，其原则是“服务就是力量”；（2）技术与政治力量结合的新专制，其原则是“管控就是力量”；（3）技术、政治和资本三位一体的结合生成的全方位新专制，其原则是“系统就是力量”。这些都是潜在的危机。

人工智能等高技术并非导致危机的唯一原因，与之互动配合的是人类文明自身的一个深刻隐患，就是说，并非技术太危险，而是人类文明自身的结构成问题。这个结构性的隐患就是人类文明的技术发展与制度发展之间的失衡，准确来说，技术文明的发展水平远远超过制度文明的发展水平。历史上改变人类命运的重大发展大多数都是技术性的，如工具的发明、语言的发明、逻辑和数学的发明、农耕技术、工业技术、科学的发明、信息技术到当下的人工智能和基因技术，这些技术发展使人类生活与原始生活拉开了天壤之别的距离。然而，制度的发展却显得远远落后。假定同样以原始生活为基点，从霍布斯的自然状态出发，经过演化博弈而发展出了自然秩序，即在无政府状态下进行的非人为设计的大集体互动博弈自发自然产生的制度，自然秩序成为人类文明的第一代制度，使人类过上了有秩序的生活，但是自然秩序无法解决的遗留问题很多，例如还会有上述挑战人类智能的问题之 4、5、6。令人失望的是，后续的多种制度发明，包括君主制、共和制、封建制、集权制、民主制等，在解决上述基本难题上并无明显推进，尽管在管理技术上有所提高，但制度的智能水平与自然秩序并没有拉开很大的距离，问题还是老问题，仍然没有解决。正因为人类的制度演化水平不高，远远跟不上技术的发展，所以一旦遇到新问题就会陷入危机或困境。

在中国社会科学院金融研究所我曾讨论了这个问题，张晓晶教授提出了一个非常重要的问题：技术与制度的“赛跑问题”。张教授倾向于乐观主义，他相信（或者希望），在技术发展的刺激下，制度应该会发生回应性的重大创新。背后的理由是，人类在无数危机中走过来，应对危机有着丰富的经验。古希腊人把危机（krisis）看作是时间的一个重要性质，既可以是“危机”也可以是“转机”，有着“危”也是“机”的双重含义，这就很有深意了。不过，对于技术和制度的“赛跑问题”，我会稍微偏向于悲观主义，因为人性不容乐观。如果说在“危”中见到“机”的话，我相信人类需要一个“新启蒙运动”，因为启蒙运动的遗产已经应对不了新技术提出的问题，新技术并没有挑战个人，而是针对人类整体命运的危机，而现代的个人主义显然承担不起人类整体的命运。

编辑 张 蕾