



人机关系的想象与隐喻——从科幻作品出发

 演讲者 / 陈楸帆

我们来谈谈想象力、叙述和隐喻。

这里列出了一些经典的科幻小说和电影，我的目的是描述一下人机关系在 2100 年会是什么样子。

大家可以看到纵轴是智能水平，横轴是时间。这是一个预测性的模型，不是很具体，只是介绍一下我的想法。

大家可以看到人类智能的基线非常稳定，因为从生物学上讲，从山顶洞人到现在，人类没有很大的变化。无论是神经系统、心态，还是我们吸收知识的效率，和 3 万年前是差不多的，大体上是一个水平的直线。

过电视剧《西部世界》，电影《银翼杀手》，还有游戏《底特律：成为人类》，会发现这是一个非常传统的起义革命模式，这里的机器是实体，智力水平和人类相当，技术也在不断进步。但是到了某些时候，权力体系崩溃了，它们不满足自己在社会当中的地位，所以决定起义和人类打仗，创造一些怪物等等。

这三个科幻作品，都把记忆作为机器反抗的关键因素。这就引出了一个重要问题——我们如何打造一个可以思考的机器，同时让这个机器有记忆。现在我们使用 ChatGPT，关掉窗口后它就忘掉了之前的内容。这是非常重要的。因为人是有记忆的，既使隔代也有记忆，我们具有一致性，我们的意识，身份、知识和经验都是可以传承的。但是在《西部世界》中，机器每一次都会被初始的设计方案把记忆清零。所以机器不能够进行连续的文明迭代，这是非常关键的一点。如何打造一个像人类一样思考的机器，记忆将起到重要的作用。

这三个故事中都有一个坏人——大的技术公司是幕后黑手，所有的系统性问题都是一个大的科技公司设计打造的，这就描述了一种人类本身关于技术的担忧。

可以看到还有一个特别陡的向上的曲线，我把它叫做共生社会。这个可以在最新的动画《万神殿》中看到，改编自刘宇昆的故事。故事讲的是人类找到方法上传我们的意识，所有这些上传的意识智能存在于一个高维的空间，这些意识本身也在进化，也在寻找自己的身份。而科技公司试图杀死这些上传的意识，因为这些意识不受控制。他们设计了机器意识，就像病毒软件一样，去高维空间杀死所有上传的情报。但是与此同时，一个新的智能通过两者的混合产生了。

这个很有意思。就像不同基因一样，它的结构是完全不一样的，会有一些全新的东西从中出现。

我们之前从来没有想象过会出现这样的情况，有点像道家的理念，共生、共存，比如说人类和机器就像合作伙伴一样，可以用各自的优势劣势互补，向上演进。在进化过程中，当生命体想要有一个大的跃迁时，就需要有共生的方式出现。所以我们必须考虑到未来有这种可能性，可能宇宙也会用这样的方式演进。

也许这是我们的出路，仅靠自己是不行的，我们需要一个伙伴帮我们看到认知系统的盲点，就像暗物质，你知道它的存在，但你没办法检测它，也没办法想象它。那我们如何实现这种所谓的“拉康式真实”呢？我们需要另一个接口，非人类智能的接口。

这在伊恩·班克斯的系列小说中也出现过，其中一个人工智能系统叫做 Minds，比人更高明，在社会层面指导人类，创造了和谐的社会体系，调和了人类和其他物种甚至外星人之间的冲突。

所有的这些想象和叙事，都来自我们人性当中隐藏的东西。包括我们对其他物种的恐惧、焦虑、欲望。这就引出我的下一个问题，这里面哪些是关键点，能够帮助我们不仅仅想象或者预测，而且能够影响塑造一个更可持续的、积极的、正向的，能够让大家参与的共生社会。

首先是稀缺性，稀缺性意思是什么？我们所有的资源都是有限的，就像是《三体》所面临的



一些问题，我们需要去掠夺，我们去征服，在宇宙中就像捕食者一样。我们需要在敌人面前隐藏自己。这是一个零和游戏。这种思维意识深深根植于我们的社会文化结构中。如果 AI 学习到这

What's the key factors?

- **Scarcity:** Social Darwinism / Dark Forest hypothesis / Zero Sum Game
- **Relevancy:** Alignmentism / Ecological Niche / Multi-intelligence society
- **Sustainability:** energy / mineral / temperature / carbon emission / space
- **Ethics:** value / rights / equality / decentralization / intersubjectivity
- **Connectivity / Complexity:**
Emergence / Nonlocality / Individual - Collective - Planetary Consciousness

All factors are not standalone but interconnected as a topological structure.

一点，当奇点接近时，我们不依赖有限的能源或资源，这仍然会变成了一个哲学的问题或者说意识形态的问题。所以我把它放在第一位。

还有相关性，就像尤瓦尔·赫拉利在《智人之上》说的，可能 85% 或者 95% 的人类都是无关的。今天早上我们也谈到了对齐，我们需要的不仅仅是以人类为中心对齐，而是对彼此都有利的对齐。

我分享一个故事：2010 年谷歌退出中国市场时，当时人人都认为它在中国搜索领域最大的竞争对手会是最大的赢家。事实上并不是这样的。可能是因为市场上没有竞争对手了，那家公司失去了自我发展的动力。这个也是很有意思的。当所有的这些物种在生态系统里面共存的时候，我们都有自己独特的地位和价值，彼此之间又是深深地互相联系，互相支持的，所以我们不断地融合成一个共同前进的力量。所以 AI 如果达到了超级智能，人类就要想怎么能够找到在系统里面独特的位置。

另外一个问题是关于环境，当我们谈到可持续性的时候，AI 也需要能源，需要基础设施的建设，也需要降温，所以也会造成一些温室气体的排放，会影响到环境，当然也需要占据空间。可能这些也是 AI 和人类都需要考虑的。因为地球这个星球的承载能力也是有限的，怎么能够以一种可持续的方式共存？这也是很重要的一个话题。

这一切都是不同维度的关键问题。伦理是其中的重要部分，你相信的将改变你的行为方式。

未来 AI 无论是 AGI 还是 ASI，都应该以去中心化的方式去发展，这是我们实现伦理奇点的唯一方式。

因为集中会带来新的不平等，这在历史上反复发生。今天上午赵汀阳老师提到了双重主体性，我认为会是多重主体性，更像《万神殿》——人、机器，还有二者混合出的多种智能体。我们需要去了解整个的生态系统会是怎样的，会有哪些中间体，让我们所有的人能够更加坚强，更加可持续，更加和谐。

最后一点我还没有想得特别清楚。人类都是在个人层面上思考，不会很轻易地彼此分享信息。我们通过对话、语言、分享图片等方式去分享，但是分享效率非常低，有的时候会有很多信息的

损失。AI 可以在纳秒内共享任何数据信息，所以基本上是同步的。这将会完全改变他们作为一个整体的思考和行为方式。可能未来有一些新的种族，比如动物，AI 可能会给它们带来交流能力，不仅仅在同种族之间，还有跨种族之间的交流，这就像一个人能够获得行星级别智能或者意识。这就是涌现的意义，部分的总和远远大于整体。我认为这是很令人着迷的，虽然在数学上还没有得到验证。

所有的这些因素都不是孤立的，他们是彼此以拓扑结构相互关联，一个因素变了之后，另外的因素也会同时改变，所以我们需要共同携手努力，以一种跨学科的方式去研究，而不是从一个单方面给出解决方案。

可能会有一些高维的解决方案，我们还没有看到。想象非常重要，如果不能想象未来是什么样子，我们就不能够打造一个更加美好的未来，这是我们通过叙事达成一致的方式。

Stephen: 问一个小问题，您刚刚说了一些矛盾，特别是机器奴役人类，人类奴役机器，机器起义等等，这是不是一种典型的西方叙事呢？

陈楸帆: 是的，赵汀阳老师说人类是这一星球上最糟糕的物种。有很多的问题不是技术性的问题，是一种文化社会的问题。AI 只是一个表现的界面，暴露出了我们已经有的问题，使得问题放大展现出来。

西方奴隶制是很多科幻小说中类似描写的原型。中国科幻小说最初是向西方学习的，很多地方可以看到西方的影子。但是这几年，中国科幻界开始有自己独特的想法，比如我个人非常喜欢的一个短篇小说，是韩松写的，叫《佛性》，是关于一个机器人和尚的故事，他希望成佛，所以他努力地改变自己的编码，改变他的初始设计，使自己能够成佛。这是一个很好的比喻，东西方并不是矛盾的，更多是关于自我的修养、自我的改善，非常适合今天的讨论。

Stephen: 现在第三位演讲嘉宾就是赵教授，请您给我们介绍一下在 2100 年，人和 AI 之间的关系到底是什么？您对于未来的想象是什么？有请赵汀阳教授。