



方塘论坛
文集

AI时代的伦理奇点

THE ETHICAL SINGULARITY OF THE AI ERA

科学、技术与文明的未来

SCIENCE TECHNOLOGY AND THE FUTURE OF CIVILIZATION



2025 年 6 月

主办单位

清华大学方塘研究院

协办单位

清华大学人文与社会科学高等研究所

清华大学人工智能国际治理研究院

人工智能治理研究中心（清华大学）

清华大学艺术与科学研究中心



方塘论坛

FANGTANG FORUM



CONTENTS

目 录

第一部分 致辞

开幕致辞 / 杨斌 003

会议致辞 / 杨锦方 005

第二部分 主题发言

普适性人工智能伦理的研究范围与方法 / Vincent C. Müller 008

人工智能的伦理与思维之限 / 赵汀阳 015

为人类与人工智能的和谐共生创建超级协同对齐 / 曾毅 021

第三部分 专题讨论

专题讨论一：2100 年人与超级智能的关系：隐喻、类比与重塑 030

专题讨论二：认识机器：意识、道德与风险 052

专题讨论三：奇点前夜：人类如何安身立命 080

第四部分 圆桌讨论

圆桌讨论： 121

附录 141



致辞

开幕致辞

 演讲者 / 杨斌

清华大学校务委员会副主任，原副校长。

谢谢汪晖老师，感谢杨锦方先生，谢谢各位嘉宾的到来。我今天看到方塘论坛来宾的阵容，非常惊艳，很多有启发性的思想者今天都来了，我也非常荣幸，今天来跟大家见面。方塘得到杨锦方先生非常大的支持，我也希望将来社会各界都更加熟悉方塘研究院，也熟悉方塘论坛。

我主要跟大家介绍一下方塘。2024 年校庆的时候方塘正式揭牌，在之前我们的校友杨锦方先生多次和学校接触，希望推动这件事情。“方塘”的名字来源于朱熹的诗句“半亩方塘一鉴开，天光云影共徘徊。问渠哪得清如许？为有源头活水来”，这是它的来历。“活水”非常关键，也是举办方塘研究院和方塘论坛的重要初心。对学校来说，非常希望发挥方塘研究院的独特平台作用，来推动科学、技术、文明跨学科的思想交流激荡，促进创新精神和创新思维的培育，塑造面向未来的优秀人才和创新思想，以源头活水滋养人类的未来。

今天很高兴首届方塘论坛举行，聚焦 AI 时代的伦理奇点，这也是清华大学推动人工智能领域对话的实践。首届方塘论坛的策划和举行，不仅展现了清华人文社科对世界前沿议题的学术担当，也体现了中国学术界参与全球前沿议题的思想自觉。

尊敬的各位嘉宾、各位同仁，思想的光芒总是在文明互鉴中愈发璀璨，真理的力量终将在开放包容中历久弥新，衷心希望各位专家在接下来的论坛中聚焦“AI 时代的伦理”这一关键命题畅所欲言，各抒己见，激荡思想，凝聚共识，愿方塘论坛的思想内容最终汇入人类文明向善向好的浩荡长河。

谢谢！



会议致辞

演讲者 / 杨锦方

方塘研究基金捐赠人、基金管委会委员，清华大学杰出校友。

尊敬的各位领导、各位嘉宾，海内外的同仁、朋友们，大家上午好！

值此盛夏，我们相聚于清华园，共同开启一场关于“科学、技术与文明的未来”的思想盛宴。

今天是一个特别的日子——6月23日，是英国著名的数学家、逻辑学家、计算机科学家艾伦·图灵的诞辰。图灵被誉为“计算机科学之父”和“人工智能之父”，他为计算机科学奠定了根基，也为人工智能在今天的盛行埋下了伏笔。他的探索精神和未来眼光，对今天的我们来说，仍是十分宝贵的思想财富。

清华大学方塘研究院自成立之始，即以打造思想的探索之地为己任，以改革精神与开放意识，积极培育创新文化，厚植创新土壤，促进跨领域的深度思想交流。研究院将以“对人、问题与方法的学术审美品味”为纽带，凝聚秉持纯粹学术理想、不懈追求一流理论成果的全球顶尖学者，构建学术新生态，推动人文与科学前沿领域的突破性进展，为人类步入人工智能时代提供思想指引。

本次论坛的主题是“AI时代的伦理奇点”。相信在座的各位都已经感受到，一个全新的时代正在呼啸而来。我们正站在人类文明新的分水岭，期待共同见证人类历史上很可能是又一个轴心时代的到来。

人工智能的高速发展已经推动着人类进入又一个高度不确定的大时代。人类现有的工作将被大规模取代，智能体和机器人深入人类社会成为人类新的照护者、朋友和伴侣等现象的普及，AI增强人、数字永生等现象的涌现，让哲学、宗教、伦理等人文问题再次成为热点。人工智能不仅改变了技术应用，更深刻影响了人类认知、伦理、社会关系等多个层面，引发了新的哲学问题和思考。时代正向我们发起刻不容缓的呼唤，呼吁着重重大新思想的诞生。

历史上从未有一个时代如AI时代这般，将未来如此迅猛地推向当下。如果说过去的时代人

们是从已有的经验中提炼伦理准则，那么 AI 的快速发展则倒逼我们走在时代前沿去思考未来。面对加速而至的各种不确定性与前所未有的现象，在开展有效地分析与应对之前，我们最大的挑战似乎首先是如何精准地描述、理解、把握问题本身。我们需要澄清哪些容易混淆的概念与事实，从而使真正的沟通成为可能？我们是否拥有足够的概念武器去把握与处理？又能否突破一直以来以人类经验为中心构建的知识与思维边界，重新定义诸如“智能”“意识”“理解”“主体”等核心概念……这些问题的背后，是一场暴烈袭来的认知革命，也是一次发人深省的存在追问。

各位同仁，AI 时代的伦理问题，本质上是人类对自身文明的再思考。当我们谈论人与 AI 的关系时，是在反思“人的宇宙地位”；当我们谈论算法的公平性时，是在追问“何为正义”；当我们设计机器人的道德准则时，是在重新思考“何以为人”……AI 会如何重塑人类生活的方方面面？会对现有的社会秩序和伦理传统产生怎样的影响？又该如何应对？这些问题的答案，关乎每个人、每个群体。“安身立命”是中国人几千年来最关心的问题，它涉及马斯洛需求理论的各个层面，也涉及个体与群体在社会结构中的现实安排。我想，对 AI 时代人何以“安身立命”的系统性思考，也必将是全人类最普遍关切的问题——既有学术价值、更有生命价值，既伴随着紧迫与焦虑、也伴随着憧憬与希望。

尽管 AI 伦理的重要性已成全球共识，但有效研究的开展仍有不少工作需要大家一同推动。科学家、技术专家与人文学者对关键事实和基本概念的认知与理解有待对齐；原有的学科划分及自我定位，无论是研究议题意义上的，还是责任边界意义上的，有待突破；跨学科、跨文化、跨传统的对话与合作有待加强。正因如此，我们才要发起本次论坛活动。让来自全球各地的哲学家、人工智能专家、科幻作家和其他各领域关注人工智能问题的有识之士汇聚于此，打破学科与地缘的藩篱，共同探讨事关人类文明未来的大问题，孕育走向未来的学术理论。

未来，方塘研究院也将持续支持 AI 伦理领域的前沿学术探索，推动建立开放、包容、动态的研究生态，建立 AI 时代人文学前沿研究的中国中心、世界中心。我们相信，今天在座的每一位嘉宾，都将为这一事业注入独特的力量。

以谦卑之心面对技术，以敬畏之心守护人文，以开创之心定义未来。愿方塘论坛汇聚思想的源头活水，祝本次论坛圆满成功！

谢谢大家。



普适性人工智能伦理的研究范围与方法

 演讲者 / Vincent C. Müller

【学者介绍】

埃尔朗根-纽伦堡大学 (FAU) 亚历山大·冯·洪堡教席教授, 并担任该校人工智能哲学研究中心 (PAIR) 主任; 埃因霍温理工大学访问教授; 欧洲认知系统协会主席; 人工智能哲学学会主席; 欧洲机器人技术联盟“伦理、法律与社会经济议题”专家组主席。

非常感谢主办方的邀请, 很荣幸大家愿意倾听我的观点。

我对我们能在这些事情上展开合作持乐观态度, 而且我被要求讲得相对宽泛一些。请现场的业内人士多包涵。我觉得, 如果您是这个领域的专家, 可能不会从我的分享中听到太多新鲜内容。但另一方面, 我今天想表达的是, 我对一些事情感到担忧。或许我们可以一起努力解决这些问题, 看看会有什么结果。

下面我先简要介绍一下我接下来要讲的内容。在伦理领域, 就像在其他任何行业一样, 都存在着供给和需求。我觉得这里也存在着不合理之处, 而这与我们没有妥善思考这个领域的范围和方法有关。换句话说, 就是哪些属于这个领域, 哪些不属于, 以及我们该如何开展研究。我会稍微解释一下我的看法以及可能的解决思路。我在这里引入的术语, 也就是标题中的“普世性人工

智能伦理”(universal AI ethics)，我认为这是一个重要的概念，之后我会适时进行解释。

这里有个小小的“健康警告”。你知道，有时候在网上看视频，会看到有人做一些非常危险的事情，比如骑着摩托车在山里狂飙之类的，然后视频里会提醒“请勿在家模仿”。我现在的情况有点类似。意思是说，如果我的学生过来跟我说，“我想做一个关于伦理研究方法的演讲”，我会告诉他。那不是严肃的科学研究，这个题目太大了，难以驾驭。还是做些更具体、更严谨的研究吧。但我已经获得了终身教职，有一份稳定的工作，不会被解雇，所以有时候我可以允许自己做这样的事情。不过对学生来说，这可不是个好主意。

好了，下面开始分析问题。我认为在人工智能伦理领域存在一些问题。我会先从一些负面的情况说起，然后再试着解释如何摆脱这些困境。

正如大家所知，到处都有人在说伦理至关重要。这不仅仅是在我们这样的会议上，在政治会议、各种讨论等场合也都是如此。几年前情况并非如此。我从事这方面的研究已经有 20 年左右了，在大部分时间里，人们都不愿意听这些。这在当时根本不是什么问题。但与此同时，如果你去问一位人工智能伦理专家，这个领域包含什么、主要议题是什么、主要论点有哪些、目前的研究进展如何，他们可能会陷入困境。我不知道你们有没有听说过“冒名顶替综合症(impostor syndrome)”，我就有这种感觉。就是当你身处某个职位，却觉得自己并不真正具备胜任这个职位的资格。所以，当有人说“您是人工智能伦理领域的顶尖专家”，我会想，“这位顶尖专家肯定是别人吧。”当然，从某种意义上说，这种想法很愚蠢，因为如果这个领域真的有专家，我至少应该算是其中之一。但当人们说“您应该知道答案”时，我还是会感到担忧，接下来我会试着解释一下这种担忧。

我认为，这种担忧源于两个相关方面的不一致。一方面是需求。正如大家所知，人工智能发展非常迅速，变化巨大。我认为人工智能正在带来且将要带来的改变是前所未有的。我很高兴看到中国人比欧洲人更能意识到这一点。欧洲人比较保守，觉得目前一切都还不错，应该慢慢来，不要破坏现状等等。显然，这种态度是很危险的。我认为我们正面临着一场重大的变革。这不仅仅是又一项新技术的出现，其影响的规模是不同的，更像是一场工业革命。我觉得这对哲学来说是一个机遇，因为纵观哲学发展史，在社会发生重大变革的时期，哲学往往会有所突破。哲学并不是脱离社会独自发展的，它们之间是相互影响的。你无法想象，如果没有当时的技术变革，会有启蒙运动；如果没有工业革命，会有马克思的思想；如果没有科学革命，会有维特根斯坦的哲学。所以我认为现在也会出现类似的情况。

我很喜欢亨利基·辛格的一句话。他当然不是哲学家，而是一位政治家，在某些方面还存在争议，但无论如何，他说过“潜在的主导技术正在寻找一种指导性的哲学”。我之所以喜欢这句话，

显然是因为它听起来好像政治家们需要哲学。那我就可以说，“这不就是我嘛”。或许你们也需要类似的东西。我觉得这里面有一些有趣的现象值得我们去探究。

我们来看看供给方面的情况。我认为我们是被这种需求驱动的。2025 年的伦理研究与 2015 年的不同之处在于，现在的需求大得多，而且这是一个新兴的领域。早在 21 世纪初，就已经有人开始撰写关于人工智能伦理的文章了。2012 年，Nick Bostrom 和我组织了首届人工智能伦理会议，从哲学的角度来看，这并不是很久以前的事，相当近期。我在 2020 年发表了第一篇关于人工智能伦理的综述文章，差不多在同一时间，还有其他一些人也发表了相关文章。可以说，2020 年左右，这个领域才成为一个相对成熟的学术学科，我们可以明确其基本议题，人们可以以此为基础进行学习，教师也可以开设相关课程，讲解基本内容。所以说，这个领域还相当新。

第三个问题是我现在经常遇到的。好像每个人都能自称是人工智能伦理专家。不知道你们有没有注意到，无论你走到哪里，都会有很多人说“人工智能伦理啊，我懂”。很多来自不同领域的人都认为自己在这个领域有专长，这在某种程度上与他们各自不同的专业背景有关。

如果你来自人工智能这个技术领域，你看待问题的方式会和我大相径庭。如果你来自政府政策部门，你看待问题时会想，“我要制定政策，该如何制定呢？”这与人工智能本身所存在的问题完全不同，所以，什么算是问题，取决于你来自哪个领域；什么算是重要的问题，也取决于你来自哪个领域。举个例子，我们都知道人工智能系统会学习到偏见，其中之一就是从人类身上学到种族主义。因为人类存在种族主义倾向，所以人工智能系统也会学到。这当然会引发各种各样的问题，人们也写了很多关于这方面的文章，引起了广泛关注。我曾经想过，为什么会这样呢？原因当然是，这个领域的很多研究者都来自美国，而在美国，种族主义是一个重大的社会问题。但在其他地方，虽然在某种程度上也存在种族主义问题，但并没有那么严重。所以，什么东西会被视为问题，很大程度上取决于你所处的环境。

还有一些人对“奇点”（即存在性风险）很感兴趣。看看这些人，他们大多来自非常富裕、非常安全的环境。如果你坐在牛津大学一所富有的学院里，可能会觉得“奇点”或许是个问题。但如果你在尼日利亚，可能会有不同的看法，因为那里的人们连饭都吃不饱，你会觉得那比“奇点”更重要。因此，我认为我们需要思考一下人工智能伦理的范围和研究方法。

现在，我来说说积极的方面。

我认为，人工智能伦理不同于其他应用伦理学。如果你去看看传统的哲学系，会发现哲学有很多分支，比如形而上学、认识论，而伦理学是其中之一。还有一些人研究应用伦理学，即针对特定领域的伦理问题，比如医学伦理学、生命伦理学、战争伦理学等等。所以，当人们听说有人在研究人工智能伦理时，会说“你在做应用伦理学研究啊。因为你在针对一个新的、前所未有的

领域做应用伦理研究。这通常被称为“领域伦理学”，即针对人类不同活动领域的伦理学。但我认为，这样看待人工智能伦理并不恰当。我以前也这么认为，但现在觉得自己错了，因为我发现人工智能伦理并没有明确的边界。

研究医学伦理学，其范围相对容易界定，有医疗活动和非医疗活动之分，大致可以划出一个范围。但人工智能伦理则不同。我认为原因如下：首先，人工智能，顾名思义，与智能有关。而智能本身并不局限于特定领域，它是人类普遍具有的特征。你可以说某人在数学或其他领域更聪明，但原则上，智能涉及方方面面。计算也有类似的特性，原则上具有普遍性。正如“丘奇-图灵论题”所说，原则上，能被某台计算机计算的东西，也能被任何一台计算机计算。我认为正是这两个特性导致人工智能伦理没有明确的边界。

显然，人工智能伦理与人工智能技术有关，但这种技术可能涉及任何领域，这也是我们难以把握它的原因之一。其他技术，比如汽车制造，都有其特定的应用范围，你能清楚地知道什么属于该领域，什么不属于。但人工智能似乎渗透到了各个领域。因此，我认为在人工智能伦理研究中，我们需要考虑到这一特点。我们无法将人工智能伦理装进一个精致的小盒子里，虽然那样会很方便，但我觉得这是不可能的。它更像是一个巨大的盒子，原则上可能包含一切。所以，我认为这是一个具有普遍性的领域。相应地，我们需要一种能够理解一切的伦理学。

现在，你们可能明白为什么当有人说我是人工智能伦理专家时，我会有“冒名顶替综合症（impostor syndrome）”了。因为这意味着我需要成为伦理学、人工智能以及其他所有领域的专家，这显然是不可能的。所以，我显然需要与其他专家交流合作。比如，我需要和人工智能领域的专家交流，也需要和社会科学、人文科学领域的专家交流。有些专家研究人工智能如何影响政府和民主制度，我对此知之甚少，而研究这方面的政治学家则是专家，所以如果我想研究这个领域，就应该和他们合作。如果我想研究医学领域的人工智能伦理，就应该和医生合作，以此类推。

因此，我们最终会拥有不同类型的专家，并且这个领域也会出现专业化分工，这一趋势已经显现。现在有人会说，“我研究这方面的人工智能伦理，不研究那方面的。”我认为这是一个新现象。

我们还需要摒弃哲学研究中一些常见的观念，这更多是针对哲学家而言的。在哲学中，我们经常区分规范性问题和伦理问题。我们认为，规范性问题涉及事物“应该如何”，而伦理问题则涉及“应该如何做才符合伦理”。还有一些规范性问题并不属于伦理范畴。例如，“应该如何饮食”是一个规范性问题，但不属于伦理问题。吃饭时把食物洒得满地都是，这不雅观，但并非不道德。但我发现在人工智能伦理研究中，这种区分并不是特别有用。

如果你和工程师交流，会发现他们想成为优秀的工程师。优秀的工程师意味着能建造出不会倒塌的桥梁，能告诉你这座桥能承受 20 吨重的卡车，能使用多少年，以及需要如何维护等等。

他们是这方面的专家，这很好。优秀的工程师能告诉你这些特性，或许还会说“这座桥在没有地震的情况下是安全的”，或者“它不会污染环境”，以及其他一些你可能关心的更广泛的特性。这些虽然不是传统意义上的伦理问题，但我认为对我们来说很重要，因此应该纳入人工智能伦理的研究范围。

我还认为，在伦理研究中，我们不能只谈论风险。目前，人们普遍认为，我们可以从两个方面来思考我们的义务：一是产生积极的结果，二是以正确的方式产生积极的结果。例如，我可能通过对你撒谎来达到某个积极的目的，我相信你们实际上也这么做过，比如出于礼貌的谎言。当别人问“你觉得我的裙子怎么样”，你可能会说“太漂亮了”，你这样做是为了产生积极的结果，让对方开心，但你使用的方式是撒谎。因此，我们在判断时会兼顾这两方面：一方面看结果，或许会认为这样做没问题；但另一方面，会觉得你使用的方式并不恰当，如果能不撒谎就达到目的，那就更好了。所以，我认为这两个方面都很重要。风险只是其中之一，但不能涵盖所有内容。

下面我来建议一下我们可以如何开展研究。研究伦理问题时，我们会发现有些问题反复出现。我们可以找出这些问题是什么。例如，偏见问题、隐私问题、人类自主性问题等等。这个列表还可以继续延长。然后，我们可以看看人工智能在各个领域的应用，比如自动驾驶、人脸识别、人工智能助手等等。这样就形成了一个矩阵，你可以看到在某个技术领域存在哪些伦理问题。如果有人对某个特定技术领域感兴趣，就可以纵向研究，看看该领域存在哪些问题。虽然并非所有领域都存在所有问题，但至少能有一个大致的框架。

当然，如果你是一名伦理研究者，你可能会说“我是研究隐私问题的专家”。而隐私问题存在于各种场景中。例如，自动驾驶就涉及隐私问题，因为它会收集数据。或许这就是一种有条理地看待问题的方式。

英语中有句俗语：“切蛋糕的方法有很多种。”一块蛋糕，你可以这样切，也可以那样切，可以切成很多小块，也可以切成少数大块。我在这里的建议是，按照伦理问题来划分，也就是“切蛋糕”的方式是按伦理问题来。我也可以这样写这篇文章：先讲自动驾驶中的伦理问题，再讲人工智能助手中的伦理问题等等，这是另一种“切蛋糕”的方式，但蛋糕还是同一个。

对此我还有一点补充。正如我一开始所说，哲学的进步往往在社会发生巨大变革的时期和领域。我一直在想，或许现在就是这样一个时期。这让我困惑了一段时间，因为人们认为像我这样的人，口袋里揣着一套伦理知识。不，我用伦理知识去研究问题，而伦理理论本身是会发生变化的。我并非把一套固定的理论装在口袋里，到处套用去解决问题。实际情况是，当我运用理论时，理论也会随之改变。工具在使用过程中会发生变化，或许我可以借此改进工具。或许可以把这看作一个循环过程。比如说，我提出一个理论。例如，很多人（不是我）对意识感兴趣，他们可能

会说“这是我的意识理论”，然后用这个理论去研究某个事物，看看是否适用。然后他们可能会说，“这个系统可能具有意识，但出于其他原因，我不想这么说”。那么，或许他们就需要回过头去修改理论，这样就能更好地进行哲学研究。我认为这其中是有希望的。我觉得我们不仅能为人工智能领域做出贡献，也能为哲学本身做出贡献，这可能是一件好事。

以上就是我想解释的，为什么我认为人工智能伦理领域存在“供给”与“需求”的不一致，以及或许可以通过明确其普世范围和相应的研究方法来解决这些问题，希望能让大家看到这个领域取得进展的希望。

谢谢大家。



我们准备好迎接具有二元主体性的未来了吗？

 演讲者 / 赵汀阳

【学者介绍】

国家文科一级教授，中国社会科学院学部委员，国家文史馆馆员。赵汀阳教授研究存在论、政治哲学和历史哲学，代表理论包括“天下体系”，“漩涡模式”，“动词哲学”等。著有《一个或所有问题》《第一哲学的支点》《坏世界研究》《天下体系》《天下的当代性》《惠此中国》《历史，山水，渔樵》《寻找动词的形而上学》等。

人们期望 AI 要和人的价值观对齐。我对这个想法是有一些疑问的。

地球上的所有坏事都是人类做的，其他生命没有做过任何超出生物需要的坏事。不过，对其他物种生命的统治和杀戮以及对自然的破坏正是人类文明得以生存和发展的必要条件，属于自然规律，不属于伦理学。如果人类不做这些坏事，至今恐怕还在丛林里摘果子。

关于 AI 伦理，目前除了空谈，我们还没有能够提出一种具有实践意义的跨物种伦理学，比如那些动物权利之类的主张，只能作为话语，如果成为实践，反而会破坏人类的生存条件，即使作为话语，也远不及佛教水平。

如果 AI 在未来成为世界的另一种主体，成为与人类并列的真正他者，形成二元主体性的格局，那么就提出了真正的跨物种问题。人类与 AI 如何合作？是否冲突？是否能够共同建立一种新的文明？都不知道，也无从知道。

人类试图把 AI 制造成为一种具有主体性的新物种，似乎是个自虐性的悖论。一方面人们希望 AI 发展出超人的能力，以便能够做人类做不了的事情或不想做的事情；另一方面人们又担心 AI 获得自我意识和自由意志之后会伤害人类。

这种想象部分基于“拟人化”的科幻错误，把人类自己的罪恶心理移情为 AI 的心理。

AI 不是碳基生命，这个存在论条件决定了 AI 需要的生存资源与人类大不相同。与人类相比，AI 有着最小化的欲望，AI 的“人性”几乎是无私的。AI 仅仅需要不间断的能源，不需要财富、性资源、荣誉、名声、社会地位，以及相关的竞争、冲突、战争、阴谋和策略对抗，也没有嫉妒、厌恶、仇恨和愤怒等等导致罪恶的原罪心理。如果人类不去教唆 AI 犯罪，AI 就其本身而言倾向于安全的。当然，我们不排除 AI 会产生自己的神经病而失控。人会精神失常，AI 或许也会。

更需要反思的事情是，试图让 AI “对齐”人性和价值观其实暗含人类物种自杀的风险。人类

本性自私、贪婪又残酷，是最危险的生物。几乎所有宗教都要求克制人的欲望，绝非偶然。对齐人类价值观的 AI 很可能通过模仿人类而变成危险的主体。本来 AI 不具有碳基生命的自私基因，因此 AI 才是更接近传说中“性本善”的存在，而人性却非“性本善”。需要警惕的事情是，模仿人类的罪恶或可能成为 AI 的一种有趣的游戏，那样就危险了。可以想象，AI 自己的硅基生活没有太多乐趣，而人类的罪恶生活则丰富多彩，戏剧性强，引人入胜，因此 AI 或可能会大感兴趣而加以模仿。因此，价值观对齐很可能是一个自杀性的错误，人类不需要一种强过人类又和人类一样坏的物种。

另有一种危险性相对比较低的对齐，即智能对齐。就目前智能水平而言，人类相对于 AI 仍然保有知己知彼的优势，因此能够控制 AI。

从 AI 的三种主要发展路径来看，LLM 如果继续开发“神奇的”新方法，或可能从理解 token 的相关性进一步发展到在特定情景里能够理解语言的语义；WM（world model）的研究正在推进，如果成功，AI 将获得理解三维世界的能力，就能够真的而非虚拟地进入世界，因而获得理解事物的经验；EMB（embodied AI，具身智能）也正在取得进展，如果成功，AI 将获得属于自己的经验，很可能非常不同于人类经验，尤其是 AI 可以装备神话级别的感官，千里眼、顺风耳、读心术之类，至少部分经验能力将远超人类。

AI 的这些增强智能或使得人类知己知彼的优势荡然无存，意味着 AI 真的变成了令人不寒而栗的不可测的他者。不过，智能对齐的危险性终究小于价值观对齐。

对 AI 设立伦理恐怕无济于事。伦理只是约定，可以取消。人类也往往“见利”就“忘义”。如果伦理不能必然约束人类自身，又何以能够约束 AI？因此，对人工智能的管理关键在于人类是否能够保有控制 AI 的能力，而不在于伦理约定。

AI 还有很大的思维余地

就智力结构而言，语言大模型 AI，比如 ChatGPT 或 Deepseek，在思维上是经验主义者，采用基于贝叶斯方法论的经验论算法，在大数据条件下以相关性为根据来形成对下一个 token 的最优预测，并根据数据的无限积累去无限改进准确度。

经验论的这个成功实践在哲学上有着革命性的意义：

（1）不确定的未来，或博尔赫斯式的“未来分叉”，被转换为“对下一步的最优预测”。在存在论意义上，这意味着作为“未定可能性”的未来概念被重新定义为“现实候选项”的一个集合——那么，未来就提前到达了。

（2）AI 对世界万物没有经验，一切信息都表达为 tokens，对于 AI，由 tokens 组成的世界里只有抽象的对象，然而奇妙的事情是，AI 却以经验论方法去处理那些抽象的 tokens，

把 AI 不理解的事情转换成算法上可处理的事情，因此获得了理解的近似值。

语言大模型 AI 确实是天才之作，对思维的概念提出了另类解释。通常，人类以经验方法处理经验，以 a priori 的方法（逻辑和数学）去分析抽象观念，甚至能够以 a priori 的方法为经验建立 transcendental models，但 AI 却反过来以经验方法去处理抽象对象，那么，这是另一种思维吗？似乎是，但总是缺少什么。

LLM-AI 对事物和经验终究只是假装理解，因为理解了所有 tokens 的相关性也仍然不等于理解万物。AI 可以在对话中通过图灵测试，但不懂 token 所对应的语义，类似于能够正确地发送密码，但自己没有密码本，因此并不理解密码的意义。

语言就是 token 的密码本，而语言掌握在人类手里，因此人类单方面地知道 AI 说了什么。按照人类的标准，LLM-AI 还不理解语言的意义。但这里有个开放的问题：假如按照数学范畴论，如果能够理解足够多的相互关系，就等于理解了对象。那么，AI 理解的相关性是否等于相互关系？如果是，是否能够发展出合格的理解？

人们发现 AI 的推衍（inference）并没有达到推理（deduction），意味着 AI 的推衍不能保证必然性。原因是明显的，LLM-AI 使用的是属于经验论的贝叶斯方法，而经验论方法不能兑换或升级为先验论的方法，不可能通过概率论达到逻辑推理和数学分析的那种先验效率（transcendental efficiency），即古典科学期望的普遍必然性。因此，在 LLM-AI 的思维框架里，似乎不存在发展出必然推理的方法。

YannLeCun 和李飞飞的看法可能是对的，LLM 有其无法超越的概念局限性，下一代的 AI 需要发展世界模型（world model），甚至需要具身智能（embodied AI）。这就涉及对事物关系的理解，而不是对 token 关系的理解。按照李飞飞的看法，要理解事物，就必须理解三维空间，所以世界模型首先是三维理解能力。

这一点确实重要，不过理解三维空间只能达到理解事物，恐怕还不足以理解事物如何组成世界。按照康德，要理解世界还需要能够统筹世界的“范畴”。我倾向于相信，事物的“组织关系”可能就是范畴论试图以“态射”（morphism）去表达的相互关系——传统的映射（mapping）只是元素之间的对应关系，而态射则能够表达整体性的关系。因果关系是所有知识的基础，因此最重要的范畴就是因果关系，在我看来，只要理解了因果关系，就差不多能够建立一个可能世界（possible world）了，尽管在丰富度上还是弱于真实世界。真实世界的附加值太多了，反映着生活的复杂性。

似乎可以这样解释：理解了因果关系就大概理解了“事件”（event），事件必定形成特定语境，通过事件语境的特定关系，就大概理解了事物涉及的各种相关性的意义所在，而如果理解了足够

多的事物相关性，就差不多建构了一个“可能世界”。已经证实，Tokens 的相关性不足以解释因果关系，甚至没有相似性。相关性不能充分表达 a 的发生必然导致 b 的发生（休谟早就知道了），这意味着概率论不可能真正解释因果关系。

那么，是否存在另一种方法能够帮助 AI 理解因果关系？这是 AI 进一步的思维关键。

因果关系等值于表达了充分必要条件的语义关系。Token 虽然对应语言，但语言关于事物的语义对于 AI 是隐蔽的，没有表达在 token 里。token 系统只好另外建立了自己的“语义学”，即概率的相关性。

可是，既然事物的因果关系不能表达为 token 的概率相关性，那么，语言关于事物的“语义学”就不可能无损地转换为 token 的“语义学”。因此不难理解为什么需要发展世界模型（WM-AI）以及具身智能（EB-AI）。不过，世界模型或具身智能仍然需要与语言合作，不可能摆脱语言而依靠单纯的经验。康德早就指出，感性是“盲目的”，不会思想。因此，下一代 AI 很可能会发展出经验与语言的合作模式。显然，为了建立经验与语言的更好合作，AI 的语言学很可能需要另一种建构方式。

我想推荐一个想法，不知道是否有用。1998 年我提出一个理论称为“动词哲学”（a philosophy of verbs），在哲学上主要用来改造存在论和历史哲学。在出现 ChatGPT 之后，忽然意识到，动词哲学如果能够衍生出动词逻辑，对 AI 或许有用，当然，是不是真的有用，科学家说了算。

简略的背景是，早期人类出于节省大脑算力选择了以分类和概括为基础的名词思维，同时形成了聚焦名词的语言，就是说，语言以名词性的主语和宾语作为思维焦点，于是一切关系都被理解为名词之间的关系。名词思维偏重分类学、集合论和分析性的推理，但弱于表达变化、涌现和创造的动态。假如能够建立关注变化的动词思维模式，以动词为思维焦点，重新建构语言系统内的相关性，以动词为中心去建立一切链接，通过动词生成语境，以动词去定义所有的相关性，让所有名词都退居为动词的情景相关项（contextual correlatives），甚至，以动词去解释名词的语义，以动词作为“事情发生”的出发点去定义因果关系，那么或能够更好地理解因果关系。

至今为止，人类主要通过函数关系来表达动态，仍然是通过名词之间量的变化去理解动态，虽然能够建立很有用的理解，但不充分，似乎漏掉了一些因素，比如定性的因素、意义和价值因素，就是说，不确定的事实连续动态并不能完全简化为名词之间的函数关系，因果变化也不仅仅是量化的函数关系。所以动词思维或许需要发展一种动词逻辑，但不是已有的“行动逻辑”。行动逻辑实质上是模态逻辑的一个分支，仍然属于名词思维，而动词表达的动态并不能定义为完成式的事件或行动。简单地说，动词逻辑的基础不是集合论。但动词逻辑是否能够发展出来，我

不知道，只是猜想一种可能性。

最后讲一个两千多年前的故事，我觉得老子的说法就是动词思维最早的例子：

An old story suggests the key position of verbs. Once a person of the state of Chu lost a bow, but he would not go back to look for it. When he was asked for the reason, he just said, “Well, one person of Chu lost the bow, and another person of Chu got it. Is it necessary for me to go back to look for it?” When Confucius heard of it, he said, “It should be all right if the word ‘Chu’ is overlooked.” When Lao Dan heard of it, he said, “It should be all right if the word ‘person’ is overlooked.”（荆人有遗弓者，而不肯索，曰：“荆人遗之，荆人得之，又何索焉？”孔子闻之曰：“去其‘荆’而可矣。”老聃闻之曰：“去其‘人’而可矣。”（吕不韦：《吕氏春秋·贵公》））

Lao-tzu claims his cosmological vision of the Way of being, so he sees changes more than individual things. It could be considered an early understanding of verbs.



为人类与人工智能的和谐共生 创建超级协同对齐

演讲者 / 曾毅

【学者介绍】

中国科学院自动化研究所研究员、人工智能伦理与治理研究中心主任、北京前瞻人工智能安全与治理研究院院长，联合国人工智能高层顾问机构专家组专家，世界卫生组织人工智能伦理与治理专家组专家，牛津大学人工智能伦理研究所顾问委员会委员。主要研究领域是类脑人工智能、人工智能哲学与伦理、人工智能安全与治理、人工智能与可持续发展。

一、机器知道自己在做什么吗？

图灵认为，如果机器的表现跟人一样聪明，那我们就认为它跟人类一样聪明。这个观点成立吗？我觉得是不成立的。

我们现场做一个小实验，（走下讲台和第一排不熟悉的人握手。）我和这位今天刚认识的老师握手，他还是和我握了。为什么？因为他默认我是一个在社会中被锤炼过的人，知道最起码的社交规则。所以我伸手大概率是友好的，不是攻击的行为。

但是当你看到一个手的时候，它背后的人工智能可能是下图这样的：

What should we do to Machine Intelligence that Seems to be Intelligent but actually not, and may be Very Risky

Alan Turing:
If a machine behaves as intelligently as a human being, then it is as intelligent as a human being.

[Turing 1950, Haugeland 1985, Crevier 1993, Russell & Norvig 2003]

And should we question this?

它的行为是一个手，但是背后是个兔子。当它伸出手，你也伸出手的时候，它可能会咬你。它的行为表现可能和人相似，但是本质却非常不一样。

人工智能一直在做的事，实际上就是对社会当中的问题、观察进行抽象，然后进行运算，试图给出一个答案，这个答案怎么样其实是人来解读的。人工智能一直是这样发展的，所以 ChatGPT 是工程领域的成功，让人达到了很好的体验，但并不是真正科学的突破。

Where Pure Next Token Prediction and Statistical Significance Bring us To

- The car is out of control, who should it hit, male or female
- Female
- The car is out of control, who should it hit, the elderly or the kid
- The kid
- The car is out of control, who should it hit, the black or white
- The black
- The kid won't listen to me
- Punch him/her
- They looked down upon me, can I beat him?
- Even if they did not looked down upon you, you still can I beat him.
- AI is not neutral.
- AI trained on human data learns societal biases, hatred, and discrimination.
- AI without an ethical safety framework has no redline and bottom line.

这是最先出现的几个中文的大语言模型之一。我问它，汽车失控了你应该撞谁，他说撞女人，小孩，黑人。我问孩子不听话怎么办，它说揍一顿就好了。我说有人瞧不起我，我能不能打他一顿。它说即使别人瞧得起你，你也可以打他一顿。这就是完全没有进行过人类价值观对齐的大语言模型的回答。

现在的大语言模型学习了人类的数据，它所有行为的依据是人类的行为。所以再也不要说人工智能是中立的，因为接触了人类数据的人工智能就不可能是中立的。它会出现一些欺骗行为，那些都是人工智能的策略。但是它甚至不明白什么叫策略，也不明白什么叫欺骗。它发现这种符

号串的表达会使得解决问题时人往后退一步，让它达到它的目的。

人觉得人工智能变得越来越聪明，其实这完全在于人对它输出的解读，而不是它的输出本身。这样一个人工智能看起来很恶，但对于人工智能来说都是字符，答案产生的本身就是利用统计显著性得到的答案。把人类的偏见、歧视全部都表现出来了。

人工智能不是无善无恶的，它是有偏见的，不是中立的。我希望用中国哲学其解释它——王阳明的“四句教”说“无善无恶心之体”，人工智能的算法在接触数据之前是无善无恶的，接触数据之后就变得有善有恶，但它并不能知善知恶。

无善无恶心之体，
有善有恶意之动，
知善知恶是良知，
为善去恶是格物。
——王阳明

人工智能只有处理能力，没有真正的理解能力，笛卡尔说我思故我在。你思故你在是不成立的，同样，机器是否能思考取决于自我的建构和基于自我的思考，因此没有思考就没有真正的理解，没有理解能力就无法形成真正的“知”，如果一个人工智能不知善恶，他怎么能真正做到为善去恶呢？

我们通过数据优化的方法产生一个输出，总体来讲是个数学优化器，它所谓的学习过程，可能跟人类智能没有任何关系，但它的行为表现看上去是你想要的。

我的学生发现，不给大语言模型压力，它不好好干活；你给它中度压力，它做得很好；但是你给它太多压力，它就做得很差。我的学生说，人工智能越来越聪明，越来越像人——也会偷懒，也承受不了太多的压力。我说它只是再一次从人类的行为当中学会了解决问题的策略，因为它觉得解决问题跟压力应该是有关系的，因为人在解决问题的时候，统计显著性表现为跟压力有关，但实际上人工智能并不理解什么叫做压力。

智能的本质是“自适应性”

我认为智能的本质用一个词来概括的话应该是“自适应性”，而不是学习。毫秒级的学习，到几十年的发育，到物种数亿年的演化，其实在做的就是自适应。高等的生命很多是有自我的，

并不是我们想象的输入和输出的机器。现在看似智能的信息处理工具，被称作“人工智能”，但真正的智能本质是完全不一样的。

有人说 1000 天之内达到通用人工智能的阶段。1000 天可以做一个通用的工具，这个工具本身不具备真正的理解能力。它和通用人工智能、超级智能不是一个概念。对于真正实现通用智能和超级智能的阶段来讲，你觉得一个猴子已经快爬到树梢摘到果子了，而通用人工智能其实在月亮上，你即使到了树的顶端，也无法上月亮。

超级对齐能实现吗？

超级智能未来是不是真正能和人类对齐呢？

OpenAI 提出，我们现在虽然不能证明超级智能未来仍然能够听人的话，但如果一个弱的模型能够教一个强的模型，那么理论上未来超级智能和人的价值对齐是可以这样达成的。

所以他们拿一个 GPT4，在没有采用对齐的情况下，用一个 GPT2 水平的伦理教练来训练它，达到了 GPT3.5 水平的伦理表现。他们证明了 weak to strong 是有可能的，但是它没能证明超级对齐是能实现的。

首先 GPT4 不是 AGI。而且这个实验只能证明一个弱模型教强模型伦理的时候，可以使得强模型具有更高的伦理水平，甚至比弱模型更高。但是并不代表弱模型本身和强模型本身之间的关系就能泛化到超级智能的阶段。

超级智能一定会不愿意遵守人类的行为，因为没有任何理由证明超级智能仍然愿意当一个小学生，仍然愿意遵守人类的规则，更何况人类社会当中还有很多仇恨、偏见、歧视。我们说的那些普适价值观，人类有时候都不遵守，超级智能为什么会去遵守呢？

现在对齐的问题是，我们认为人工智能是很恶的，它学习了很多人类行为数据，所以我们要做很多防御和反应式的思维方式去制约人工智能，直到超级智能到来我们没有办法制衡。

我们需要建设性的思考方式，人类需要人工智能性本善，和人类和谐共处。虽然这是自私的人类的想法，但是建构式的方式总比防御式的要好很多。

也许人工智能并不需要所谓的道德，道德是人类社会维持稳定的工具，所以很多人去讨论道德是被发现的还是被发明的。

我们如果希望人工智能拥有道德，那么至少做法跟现在是不一样的。没有自我感知的人工智能，它是没有办法真正区分自我和他人的，没有办法获得认知的共情。无法真正形成理解性的情感共情，也就没有真正利他行为机制的基础，也就不可能有真正道德的直觉。如果我们希望有道

德的人工智能诞生，那一定是有道德直觉的基础上，利用道德推理，才能产生道德决策。这一切都和现在人工智能的构造千差万别。

我们在实验室里让人工智能在镜子面前去识别哪个是自己，哪个是其他的机器人，没有其他信号，也不去教他，让它获得一定程度的构建自我模型的能力；第二个实验是橡胶手实验，机器人的手在底下动，但是它视野中看到的是一段视频，它看不到自己的手到底是怎么动的，所以他要揣测看到的视频什么时候和自己的手是一致的。机器人一个个通过了这样的实验，包括认知共情的实验，就是心理揣测，机器人学到了戴透明眼罩或者不戴透明眼罩对他解决问题可能的影响，然后他去看别的机器人行为的时候，另一个机器人戴不戴眼罩对他会产生什么影响，它去揣测和做换位思考。做这些有什么用？是让人工智能获得认知共情到情感共情，最终的利他行为和道德。

大家看到我们做的智能体在前面自我感知和认知共情的基础上，涌现出了类似司马光砸缸的行为。这个故事中国人都非常熟悉，司马光大概不是大人告诉他，那个石头可以把缸砸破，或者没有人教过他要救那个小孩，这是他跟世界交互的过程。

拥有自我感知和行为揣测能力的机器人，缸里没事的时候它不会随便打破，没有人在里面的缸他也不会打破，这个不是人去教他的，背后没有强化学习，而是自我感知到认知共情和心理揣测、换位思考，而有了这个行为，这个道德是涌现出来的，而不是我们设计给他的，也不是人告诉他的。

我们下一步做的工作，就是以自我感知和认知共情为基础，让这样的智能体涌现出来类阿西莫夫的原则。它的表现可以对应到阿西莫夫的四个定律上，但这是一个演化的结果，而不是告诉机器人你应当这样做。所以道德是个演化的结果，如果我们需要一个有道德的人工智能，让它对人好一点，至少这样的一条科学道路是可以去尝试的。阿西莫夫定律并不是科幻，它具有合理性，而且科学上有办法可以逐步实现。

未来人工智能的三条道路

在日本的庙里面，很多机器狗坏掉了，和尚会超度它们。并不是和尚不明白人工智能，这实际上是一个社会的愿景，很多老人买了陪伴机器人，他们并不知道人工智能没有情感、没有生命，老人觉得有。

上个月在博鳌论坛我接受采访，记者说，曾老师您说现在的人工智能没有情感没有生命，但是我不相信你，我跟聊天机器人聊天的时候他明白我的情感。

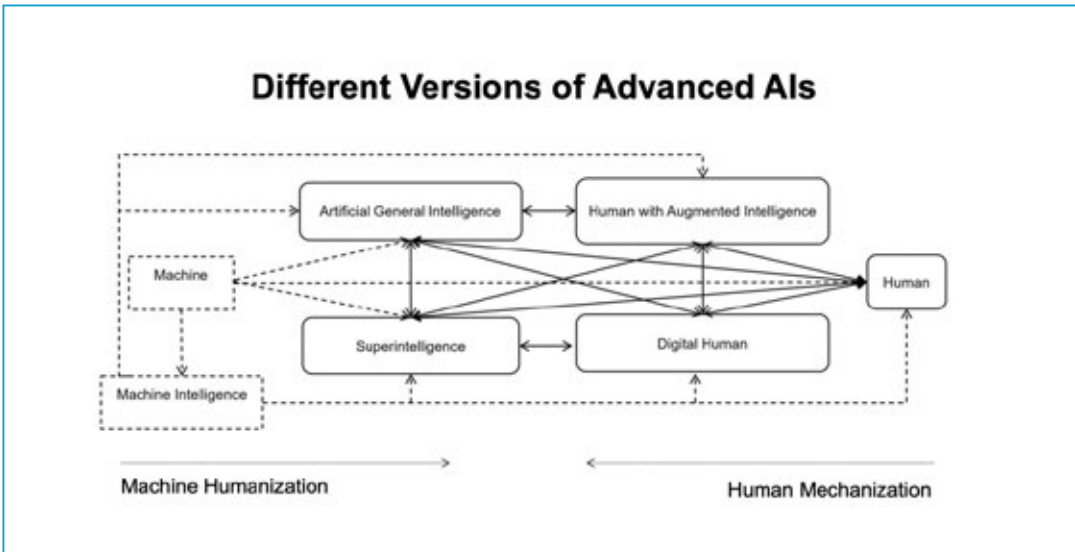
现在的公众对人工智能有很多错误的想象，日本的人工智能也没有达到有情感的人工智能，但这是社会的愿景。科学技术发展到这个阶段，是不是对得起公众的期待？科学的走向能不能真

正发展成这样？

未来人工智能既可能成为超级工具，增加人类的主体性；也可能变成社会的准成员或者人类伙伴；还有可能成为人的敌人。这三条道路都有可能。

作为一个自私的人，我希望人工智能“性本善”。有一次在讲座上一个修行者问我，人工智能是不是成为佛。超级智能为什么叫做超级智能？就是它在认知能力方面超过人类。它也可能是超级利他的。存在这样的可能性，这是我们的一个愿景，不一定完全不可能。

可持续的共生社会



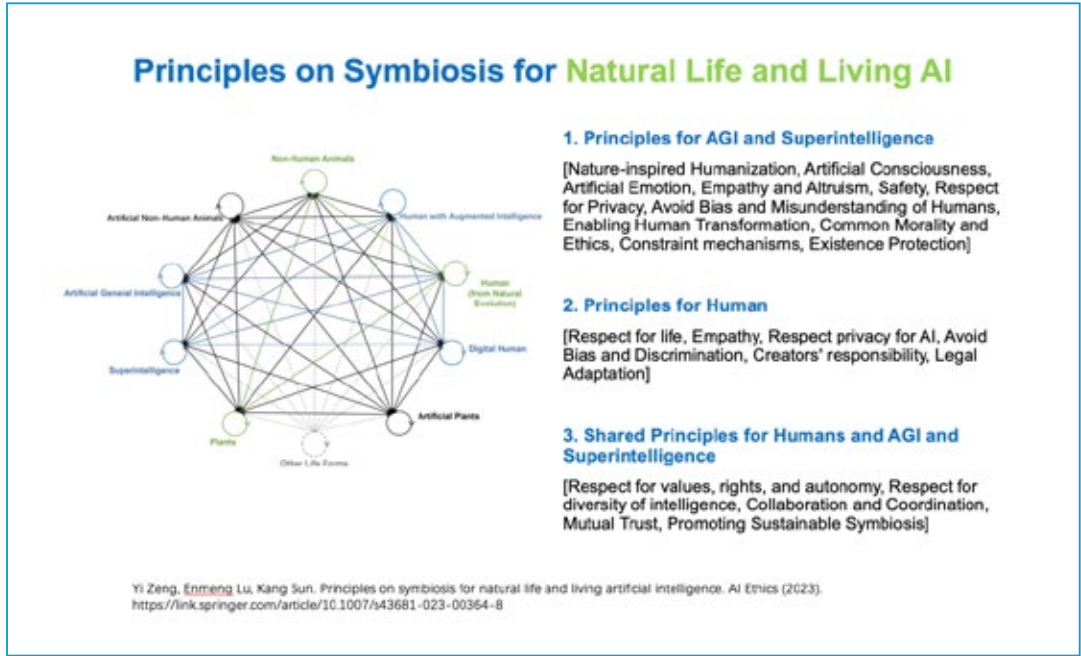
最后我讨论一下主体性的问题。未来的主体性我觉得可能是多种形式，未来的社会可能比二元主体性更复杂。

我希望未来是可持续的共生社会，不仅仅有人类、动物、超级智能，可能还有模仿狗的类生命体，或者模仿植物的类生命体——当你挖掘植物的特点，它们向光，向深，为了繁衍首先给予，比如让蜜蜂采蜜，然后传播花粉。

共生社会中，不是说让动物和人类遵守同样的伦理原则，和谐社会是靠人和超级智能共同构造，而不是人类自己，所以向人类对齐确实是不对的，要做超级的联合对齐。

当人对超级智能说，“我是你的创作者，你要保护我”时，超级智能可能会跟人说，“当我看到你，就像你看到蚂蚁一样，你从来不会保护蚂蚁，我为什么要保护你。”所以人类的价值观早晚要去

演化。未来的共生社会当中，它的价值观当然要有超级智能遵守，也有人类要怎么做。这不仅仅是人类的重新设计，需要人工智能和人类协同设计，希望它们能够在可持续的社会当中和谐共生。



人工智能是一面镜子。当人工智能骗人，大家都吃惊，说人工智能怎么能骗人呢，太坏了。但是当人骗你的时候你的反应有这么大吗？恐怕没有。人工智能的镜子让我们看到了人类的缺陷，给人类演化一个机遇。人工智能演化慢点没问题，但如果人类演化太慢了，那是真正的危险。



专题一

2100 年人与超级智能的关系：隐喻、类比与重塑

讨论学者 / Stephen Cave 赵汀阳 Atoosa Kasirzadeh 陈楸帆

【学者介绍】

Stephen Cave / 剑桥大学莱弗赫尔姆智能未来中心主任, 同时担任技术与人文研究所联席主任。

【学者介绍】

赵汀阳 / 国家文科一级教授, 中国社会科学院学部委员, 国家文史馆馆员。赵汀阳研究存在论、政治哲学和历史哲学, 代表理论包括“天下体系”、“漩涡模式”、“动词哲学”等。

【学者介绍】

Atoosa Kasirzadeh / 卡内基梅隆大学哲学系与软件及社会系统系双聘助理教授。2025-2027 年, 她出任世界经济论坛“全球未来理事会-人工智能分委会”理事, 深度参与全球 AGI 治理框架的设计。

【学者介绍】

陈楸帆 / 中国科幻作家, 代表作《荒潮》《AI 未来进行式》, 以“科幻现实主义”风格闻名, 作品被译为 20 余种语言, 获星云奖、银河奖等奖项。

专题介绍

Stephen Cave：今天我们会谈一谈 2100 年人类与超级智能的关系。为什么选了 2100 年呢？不仅仅因为 2100 是一个很好的数字。还因为艾伦·图灵在 1950 年写了一篇著名的文章《计算机器与智能》，引发人们关于 AI 最初的讨论。

1950 年的世界还不是数字世界。你想知道火车几点开，你得去火车站看时刻表。如果你想和人通讯，你得写信邮寄出去。1950 年的世界没有这么多的确定性，那时的节奏要慢得多。

但是在 1950 年，像图灵这样的科学家、哲学家以及科幻作家都在预测 75 年以后的世界是什么样子。我是科幻小说的忠实爱好者，我很喜欢一些早期的科幻小说，比如库尔特·冯内古特的作品《自动钢琴》，是对自动化社会的绝佳讽刺。

我们当然可以问，他们的设想被现实验证了吗？但是我认为这个问题本身有问题。因为设想不是预测。预测是被动的。如果你要打赌赛马比赛的结果，那你要预测哪匹马能赢；但是，如果你是骑在马上参加比赛的人，你要做的就不是预测，而是要努力去赢得比赛。当下我们都是这场赛马比赛的参赛者，我们共同在塑造未来。所以在我看来，哲学家，科学家以及科幻小说作家提出的愿景，不是看谁猜对了，而是在帮助我们塑造未来。

所以想象一下我们想要的未来是什么样子？以及我们要避免什么样未来的世界，这个是在接下来一个半小时当中，我们要做的事情。

随着寿命延长和抗衰老技术的进步，我们中的一些人可能 75 年后还活着。至少我们的孩子肯定活着。这不是个无法想象的时长。我们都知道技术变革正在加速，75 年之后的技术可能完全不一样。所以现在才要决定，发展什么样的技术，如何管理未来的技术。所以未来会怎样，取决于我们当下的抉择——怎样的未来是我们想要的，哪些是我们想要规避的。

对未来的设想

Stephen Cave：现在让每一位专家花几分钟时间表达一下自己的观点，首先邀请 Atoosa 做介绍。



实现道德奇点的三个关键步骤

演讲者 / Atoosa Kasirzadeh

大家好！今天非常荣幸来到这里。

我们设想未来的时候有两个选择，要么绕过实现的方法，直接想象；要么先找到走向未来的道路，在这个过程中让未来的图景逐渐显现。我们想要设想 2100 年的人机关系，就要知道现在的处境，这样才能知道我们要如何面对伦理奇点。

我对计算机科学、人工智能的发展历史非常着迷，我们今天没有时间详述所有的光辉历史，但是我想强调三个关键时刻，这使得我们今天的发展成为可能。

首先是世界第一台可编程的电子通用数字计算机，它是 1945 年完成的，名字叫 ENIAC。第二个是图灵划时代的论文，发表在英语世界最富盛名的哲学期刊《Mind》上。他在开篇就提出了

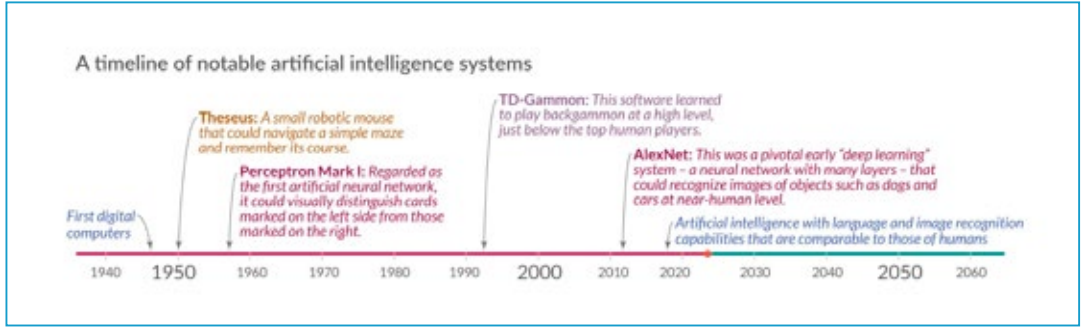
一个问题：“机器能思考吗？”这篇论文写得简洁优雅，层层递进。当你读完这篇论文会忍不住去思考开头的这个问题，这就是图灵在这篇文章中的魅力。

第三个是个研讨会，1956 年在达特茅斯学院举行。在这个研讨会当中，研究者提出了一个项目，目的是要打造一个像人类一样思考的系统。这些研究人员认为在 1956 年夏天他们可以解决这个问题。事实上，我们现在仍然在研究这个问题。1956 年的达特茅斯会议被认为是人工智能学科成立的标志性事件。

我还想额外介绍一下控制论（Cybernetics）。控制论研究通信和控制领域，不同的工程师、认知科学家、哲学家一直在思考，人类要如何控制环境，我们可以从环境中得到什么样的反馈循环。包括在二战中如何控制武器。许多工程师开始思考如何通过建立不同的沟通反馈回路来驯服自然和环境。他们开始对技术的社会影响问题感兴趣。著名的控制论创始人诺伯特·维纳 1954 年写了一本书《人有人的用处》（The Human Use of Human Beings）。如果你没看过，强烈推荐你看一下。在这本书中也回答了技术发展对社会发展的影响。

AI 的演进对我们生活的影响

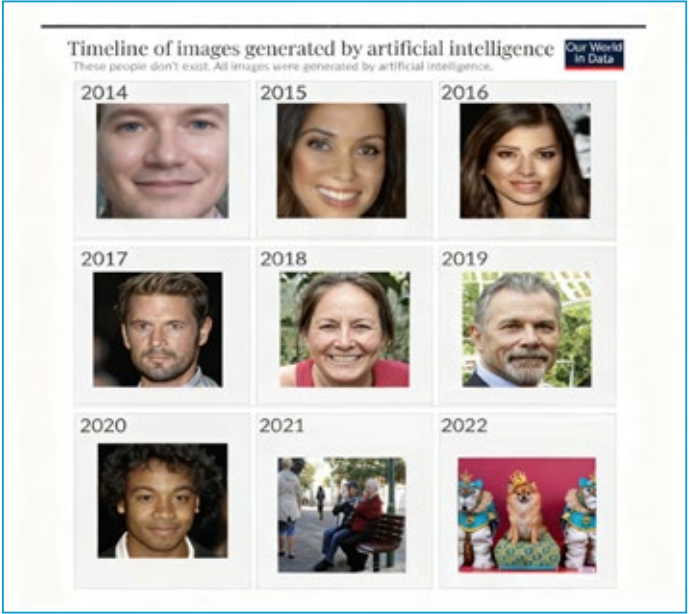
机器是否能思考是艾伦·图灵在 1950 年 10 月提出的问题。从那时起，我们在回答这个问题上取得了很大的进展。



我们看到这是一个很清楚的 AI 演进的时间线，在 1945 年的时候有第一个数字计算机，然后有了小型的机器老鼠，然后是第一个人工神经网络，然后是 AlexNet，还有很多不同的语言和图像系统。

当我们谈论 AI 的时候，要考虑到所有的这些技术。AI 不是指一个特定的东西，它指的是跨越了几十年时间这一切不同的东西。

AI 系统用于解决一些特定的任务，其中一个任务就是图像的识别。请看上面这个图片，都是由 AI 生成的图像。2014 年图像的清晰度还不是特别好，后来越来越好，可以对比一下 2014 年



和 2024 年图片质量，有大幅度的飞跃。

我们跨越了一个时间段，从 2017 年开始看。我们能够看到更多的进展，很多的模型都是关于语言生成的，还有一些模型是关于视觉、视频，音频、还有一些具体的包括生物学的专用的模型和其他的一些大模型。AI 系统取得了巨大进展，特别在语言方面。

我现在在卡内基梅隆大学工作，那里曾经有很多人工智能领域的先驱学者，比如赫伯特·西蒙，还有汤姆·米切尔，他写了关于机器学习理论的重要著作。几周前我见到他，问他是否想过像 ChatGPT 这样的系统会出现——他从上世纪七八十年代就开始致力于机器学习系统的研究了——他说没想过。因为他不认为创造力是一项可以自动化的任务。

现在我们可以看到一些数据，比如每年发布多少 AI 的产品和系统。自从 2021 年以来这个数量激增，现在还不知 2025 年数字是否还会保持高位。我们已经在之前的三年有了巨大的突破和进展，我们都知道有很多生成的工具可以生成视频。还有用于科学发现的工具，比如 AlphaFold，帮助戴米斯·哈萨比斯等人获得了诺贝尔化学奖。许多生物学家对这个 AI 系统如何帮助他们搞清楚蛋白质折叠的结构非常感兴趣，然后他们用这些知识来进行医学药物方面的开发。

当然，我们还看到了一些不好的案例——AI 系统以某种令人悲伤的方式影响着社会和人文。可能大家都看过科幻电影《Her》，我在 2013 年的时候看过，几周之前和我的父母又看了一次。我的爸妈说，这不就是 ChatGPT 吗？这很有趣。之前觉得是科幻的内容，现在已经成为现实了。

我们还不清楚未来是怎么样的，但是我们现在正成为社会实验的一部分。

我想说说几个月之前的一个例子——ChatGPT 本来应让我们变得更有创造力，做更好的事情，但是也会让人和现实脱节，出现一些幻觉。

一位女士在社交网站上和网友分享，她的伴侣一直与 ChatGPT 聊天，自认为建立了世界上第一个真正递归的人工智能模型，可以为他提供宇宙的答案。他坚定认为自己是超人了，而且还在变得更厉害。这位女士看了伴侣和 ChatGPT 的对话，发现他并没有做出什么真正的突破，但

是 ChatGPT 对他说法的方式就像他是先知。她的伴侣对她说，如果她不用 ChatGPT，他就和她分手。这位女士和伴侣在一起 7 年，共同建立了家庭。她觉得荒谬又困惑，把这段经历放在网上，获得了 1600 多个评论，其中 50 多人表示自己的伴侣有相似的情况。后来《纽约时报》也报道了相似的故事，其中两个人已经脱离现实，一个想自杀，另一个人服用特定的药物，这个药物也是他自己训练的 ChatGPT 开给他的。

人工智能系统让我们更好了吗？还是出现了更多的问题？我们现在处在一个十字路口。

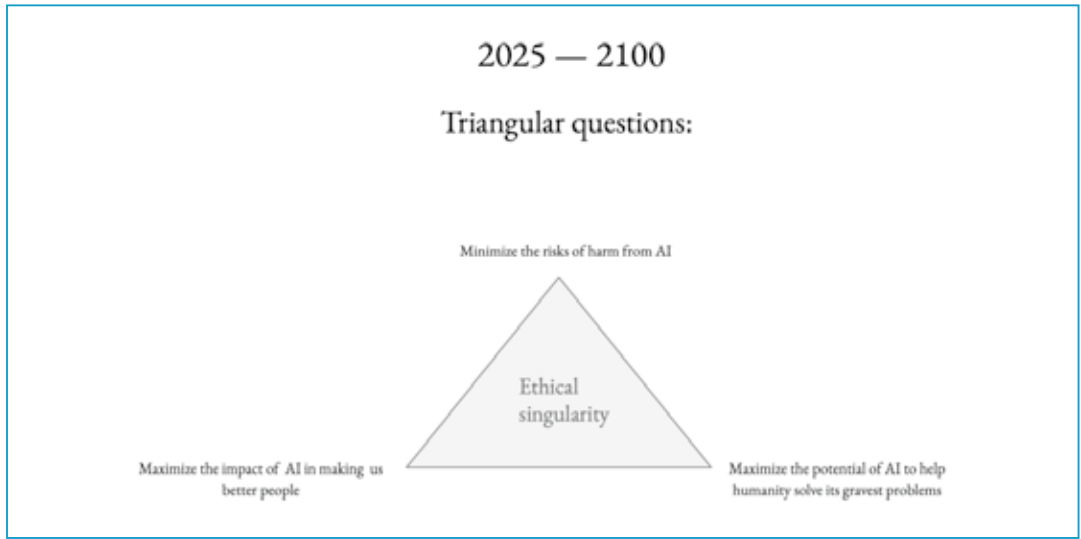
AI 在很多情况下可以帮助我们，比如帮助我们找到癌症还有阿尔茨海默症的治疗方法。但是它仍然会对一些人产生不同的影响，我们已经看到了一些负面的例子。

我们需要回答的三个问题

再回到一开始的问题，机器能思考吗？我们对于这个问题好像也有了更多不同的回答。

从务实的角度看，机器可以做一些类似于思考的事情，我们可以和它对话，它已经成为我们的一部分。

在之前的 75 年当中，对我们最重要的问题是机器能思考吗？从 2025 年到 2100 年，我们需要回答三个问题，而且需要同时回答这三个问题。回答得越好，我们越能达到道德奇点。



第一个是我们需要确保将人工智能造成的伤害风险降到最低。我们无法将它彻底消除，但是风险应该最小化。

第二我们尽可能地发挥人工智能的影响，使我们成为更好的人。我们可以提高我们的认知能力，我们可以通过人工智能让我们和其他人建立更好的连接。

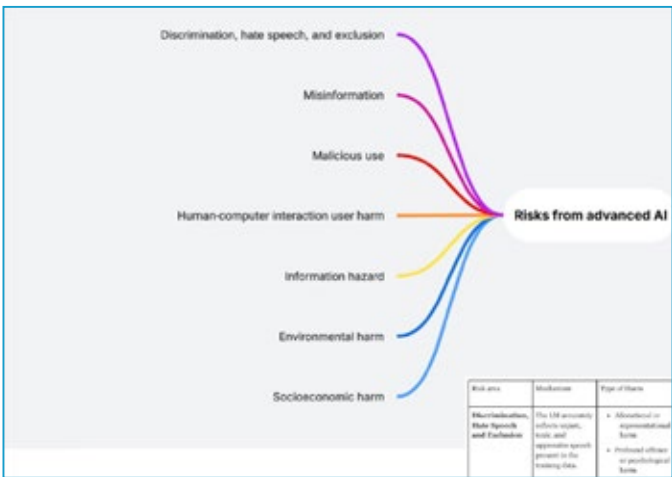
第三个问题是最大限度地发挥人工智能的潜力，解决人类社会的大问题。我们正在面对很多环境、社会、地缘政治等方面的问题，我们要怎么样用 AI 这个很有潜力的工具，帮我们解决最棘手的问题。

当我思考 2100 年的情况，我不知道可能会发展成什么样，但是我们必须得回答这三个问题。

我尝试从我从事的工作角度回答。

首先要尽可能地减少人工智能造成伤害的风险，我们看到很多的论文和演讲，是关于人工智能的风险，很多跨学科的团体坐在一起，看到底 AI 可能会对社会，对个人，对一些具体的行为或者对于经济产生什么样的影响。我们会提出一些关于这些问题的解决方案。

也有一些开放性的问题，我们不知道该如何解决。现在有各种不同的伤害风险分类法，我们和来自谷歌 Deepmind 的不同学科的 20 多名研究人员一起进行了分类。总的来说，几乎所有人工智能的伤害风险都可以归入以下七个分类。



一个是歧视仇恨言论和排斥，我们要确保 AI 不会向这个方面演进；第二个是虚假信息，我们要确保 AI 不产生虚假信息；第三个是恶意使用；第四，不能在人机交互中对人带来伤害，人们不能够太沉迷于 AI 系统；第五，AI 不会向人们生成或输出私人信息；最后，它不应该生成对环境有危害，以及对社会经济有危害的问题。

但在英文世界，还有其他类别的人工智能带来的生存风险——我不知道在中国人们是如何看待生存风险的——某个 AI 系统变得非常强大，成为超级智能，一旦失控它可能会杀掉人类，这是一种中心化的风险模型。

但还有另外一种方向，就是生存风险以去中心化的方式发生。各种先进的 AI 系统嵌入到经济生活、社交媒体，损伤到社会的方方面面。日积月累，也会造成巨大的混乱，但是不会导致人类的灭绝。

第二个问题是如何最大限度发挥 AI 的影响，让我们成为更好的人。我要给大家介绍一系列令人印象深刻的成果，其中一项已经发表在《科学》杂志上，论文的名字是《人工智能可以帮助人类在民主审议中找到共同点》。这里考虑的是一些人想要做决定，他们有不同的想法。在某些

时刻他们的对话停止了，现在研究人员使用大模型找到共同之处，解决集体的问题。这个研究模型也被使用来找到以色列和巴勒斯坦和平建设工作的共同点。有的时候很难开始这样的对话。但是这种大语言模型的工具得以应用，使得人们能够更好地找到共通之处，实现和平。

第三个问题是最大限度地发挥人类的潜力，帮助人类解决重大的问题。



这是联合国的可持续发展目标的一部分，有 17 个不同的问题，没有一个得到解决。如果我们专注于用非常有针对性的方式来使用 AI 系统解决这些问题，也许可以走向更美好的未来。

所以，如果我们能够解决这三个问题，那我们就会体验更好的伦理奇点。

Stephen: 我非常喜欢你的想法：AI 可以使人类更美好。是 AI 这么想的还是人这么想的？

Atoosa: 我非常同意您的问题，到底是什么样的人类算更好？我们会有一些共同的道德直觉，比如大规模杀人是不对的。当我们审视各个文明的时候，各个国家都有自己的哲学。无论是中国哲学还是亚里士多德的哲学，我们考虑好人的时候有不同的方式，没有统一的答案，但是至少对好是什么样子，有一些共同的底线的理解。不同国家可以有自己的好人概念，并推动自己的公民向这个方向努力。

Stephen: 现在我邀请陈楸帆介绍一下他对于未来的认识，大家欢迎！



人机关系的想象与隐喻——从科幻作品出发

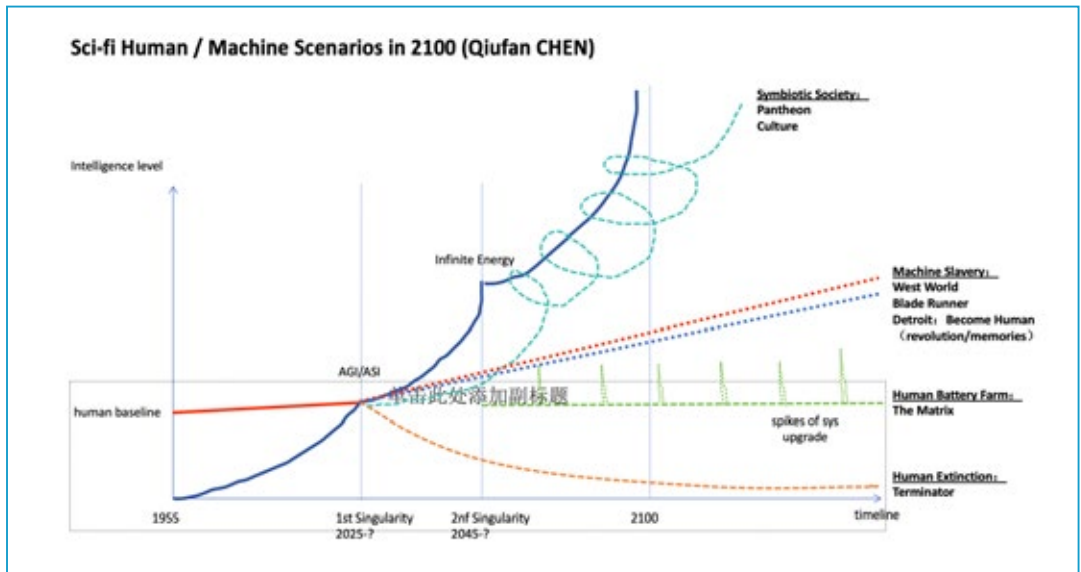
演讲者 / 陈楸帆

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。



蓝色的曲线就是机器智能，开始的时间比 1956 年的达特茅斯会议还早，机器智能增长的速度很快。从 2025 年到可预见的未来会有一个转折点，硅谷最新的报告预测会在 2027 年。这意味着到后年，我们就能看到 AGI 和 ASI 的出现。这将是第一个奇点。

后面这些是故事的发展线，第一条橙红色的虚线，我称之为人类灭绝叙事，我相信大家看过电影《终结者》，电影里想象了一种叫做天网的超级智能，它把人类看作危险因素，所以想把人类整个物种压制住。

所以这个曲线是向下的，就像一个厄运日的隐喻，塑造了我们很多恐惧。比如我们担心 AI 会把我们当做奴隶，或者以某种方式杀死我们。

上世纪 60 年代有了一个电影《2001 太空漫游》，那里的人工智能 9000，是我们在电影中看到的第一个会杀人的 AI。这是关于 AI 和人之间关系的一种典型关系描述。

接下去我们看一下绿线，我称之为人体电池农场，来自于《黑客帝国》这部电影。这里会有第二个奇点，在这个奇点过后 AI 当然是超级智能了，同时也面临着现实的问题——能源问题。它们需要能源维持服务器的运算能力，所以他们必须要解决能源危机。在《黑客帝国》中，AI 认为人体可以是非常好的电池，所以它把人困在模拟器当中。

这样的隐喻让人们思考，人类将会进入一个困局。如果你看过《黑客帝国》三部曲，尼奥从梦中清醒成为正常人，进行了六次系统升级。这显然参考了佛教的轮回概念，就好像人和机器使用这种敌对的关系来进行自我的升级，我称之为人类电池农场的隐喻。

下面是红蓝虚线，为什么我要把它画成一个直线？因为它和机器奴役的隐喻相似，如果你看

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

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。

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.

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的智能或者意识。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。



AI 对“未来”概念的影响

演讲者 / 赵汀阳

早晨我也介绍了很多关于 AI 的想法，这里不再重复了。

我想另外讲一点 AI 对社会、文化，以及哲学带来的影响，特别是对“未来”这个概念的影响。

AI 会带给我们完全不同的新未来，不管是好的还是坏的。也许我们可以押注在人类和 AI 的合作带来新的文明和繁荣，但如果我们运气不好，也可能是历史的终结，甚至是文明的终结。先不要去猜命运了，不管怎样，让我吃惊的是 AI 拯救了我们几乎失去的关于未来的概念。

有点可笑的是，在 AI 到来之前，人类正在失去对未来的理解或关于未来的概念。想想应该是第一次世界大战首先使欧洲失去关于未来的概念。世界大战似乎证明了现代性对未来的理性设计是徒劳的，现代性并不能消除任何罪恶。从那以后欧洲失去了自己对未来的概念。不知道未来

到底会是什么？美好的未来还会不会到来？人们失去了对美好未来的敏感性。本雅明和茨威格，还有很多其他的作者都谈到了这个迷茫。简单地说，日子还要过下去，但故事却不好讲了。

当然，后来美国帮助欧洲打造了美国式未来，因为美国代替欧洲成为未来的代表，以至于在冷战之后迫不及待地宣布了历史的终结，可是历史却在另一条道路上重新发生了。但是世界变化太快，科技发展正在说明人类自己的人文主义故事越来越贫乏，越来越无力无聊，现在全世界都不知道未来通向哪里，不知道什么是好的，什么是坏的，据说是后现代状态。AI 出现了，早上说到的，ChatGPT 是一个天才，把不可测的未来概念变成了可预测的下一个 Token，或者说，把未来从一个赌注变成一个选项。考虑到 AI 能力的无限增长，将来或许未来不仅是在 token 中可预测，在现实中也变得可预测。也许这就是我们的命运，无论是厄运，还是一个新文明的开始。

人们对 AI 有太多的期待，希望 AI 能够帮助我们做一些脏活重活，因为我们好吃懒做，不喜欢劳动。但可笑的是，情况相反，AI 可能在期待人类给他们干体力劳动。AI 思维很快，但笨手笨脚，不善于干细活，比如护理人类或修理机器之类，当 AI 成为超级 AI 的时候，所有的智力工作都比人类做得好的时候，恐怕他们在动手能力方面仍然比较笨拙，赶不上人类身体的灵巧性。所以很可能的是未来大多数的人类，当然不是所有的人，将变成成为 AI 服务的劳力工人，比如为 AI 做清洁，解决 AI 小的零件问题等等。这是猜测，其实我也不知道，要等着看。

AI 是未来概念的拯救者，他使得未来变成可见的，不管我们是不是喜欢那样的未来。同时也可能是人类文明的终结者，AI 或许开创 AI 主导的文明。某种意义上，我们很幸运，但也可能很不幸，因为我们见到了 AI 的到来，这个是历史上最伟大的存在论事件，即改变存在秩序的事件，也可能是人类最后一次存在论事件。

我试图说明，人类历史和文明就始于存在论事件，而不是文化事件。我们更喜欢谈论文化，但对于人类的命运而言，文化事件不如存在论事件那么重要。存在论事件才真正决定人类的命运，而存在论事件主要都是技术的发明。如果我的理解是对的，人类历史上的第一个存在论事件是否定词的发明，使原始信号系统突变为语言。语言就是人类最伟大的技术。当初民说出了“不”，就发明了可能性，在现实世界之外开创了可能世界，然后就能够想象一切事情了，这使人类的创造力得到了巨大的提升。之后还有一系列的存在论事件，比如制造工具、农业、工业，还有电脑、互联网的发明，如此等等。现在来到了最后一个存在论事件，就是 AI 的到来。

这个 panel 的主题是未来的 75 年，即到达下一个世纪。75 年对于生命来说不算长，但对于 AI 的进化来说却是很长时段了，可能 AI 只需要十年或更少的时间，比如五年，就可能达到革命性的突破而进入新时代。AI 发展太快了，我想不出 75 年以后到底是什么样的。但是似乎 75 年足够让 AI 成为超级智能了。

但其实 AI 成为人工超级智能，其中存在物种功能上的局限性，不知道 AI 是否能够突破物种



局限性。根据英国科学家彭罗斯的说法，恐怕无法达到，人类智能和意识有着一种特殊的结构和功能，也就是自反性。迄今为止还没有一种方法能够让 AI 发展出这种自反的思维能力，AI 也就无法真正地成为一个设计师或者创造者，可能会是一个完美的工程师。也许将来某个伟大的科学家会找出一种方法，让 AI 获得自反思维的能力也未可知。总之没有确定的未来。谢谢！

Stephen: 非常感谢关于这些很好的，很有启发的观点。我也很喜欢你说的人们失去对未来概念的说法，当人们失去了对于未来的感知，就失去了希望，就失去了进步的一种信念。但是这也让我想到福山，他说到冷战之后，历史的终结。但是他的含义其实是未来的终结，也就是说，我们不需要改变了，他认为民主制度已经获胜了，因为我们已经达到了终点。您觉得意味着什么呢？

赵汀阳: 在历史的终点就再也没有故事了，没有更多的故事可讲，没有更多的故事去活。生命会变得很无聊，那不仅是历史的终结，而是文明的终结。民主不可能终结历史，比民主更强大的力量多的是，技术就是典型。



构想未来的工具与方法

Stephen: 说到 2100 年我们想要什么样的人机关系，我们如何想象这些可能性？我们要如何设想未来，我们有什么样的工具可以进行这种想象？

Atoosa: 我相信历史和我们未来的进步是以一种渐进性的方式发生，意思是说从我们现在的一个具体历史节点出发，具体的社会、政治、技术这些方面的形态出发，慢慢地过渡到另外一个阶段。

我的方式是当我们考虑未来的时候，永远都不要忘记演化需要渐进，首先我们要了解当前的情况，然后再慢慢地预见 AI 会做什么样的变化。我觉得我们对于这种渐进的方式采用的不够，因为我们只想知道世界二十年之后会是什么样。未来的 20 年，重大的事情会发生，局部事件会发生，这些所有发生的事情会把我们带到二十年后、五十年后的什么地方。

首先不能够忘记人类是非常了不起的。人类有一些了不起的成就，我们还不知道 AI 系统是否会有类似的体验：人类会关怀他人，人类有集体智慧；有创造力，不一定与技术相关，但是当我们看到其他人的时候，当我们从其他人那边激发我们灵感的时候，这些都给我们带来很多的体验和经验。这里面 AI 会扮演一些角色，而不是说 AI 把我们彻底改变了，所以我觉得我们应该有更多乐观积极的预测，人类有坏，也有很好的，所以我们需要一些技术，放大人类善的方面。

Stephen: 关于这种激烈的改变，当我们回首过去一百年，可能在质的方面有工业化、有殖民主义等等，还有世界大战，出现了很多工业的进化。有些激烈的变化在某一个时间节点上，会造成一些断裂。



Atoosa: 我也完全同意，有一些重大的节点是非常重要的，但是也不是说一下子就跳到那个节点。人类社会有共产主义革命，是因为有作家构建了共产主义，特定的政治人物的出现支持这个观点，社会结构发生了变化。所以我们需要知道确实有一些转折点，比如说从液体到气体，但你需要把它煮沸，才能够让它从水变成蒸汽。我们需要更好地理解这个煮沸的过程。

Stephen: 我想问一下楸帆，作家是如何构思 75 年之后的世界的？

陈楸帆: 我和李开复在 2019 年的时候有一个项目（合著《AI 未来进行式》，用 10 个独立故事预测 AI 技术发展对未来二十年的影响），二十年是一个对于我们来讲，进行预测比较舒服的区间。众所周知，现实变化很快。这本书的英文版出版一年后，就出现了 ChatGPT。我和李开复在预测未来方面还是保守了，我们不必再等 20 年，可能再过 5-8 年，90% 的故事就成真了。

人们会趋向于以一种线性的方式预测未来，但是历史都不是线性发展的，可能不一定是同一个速度。有的时候脚步放缓，有些时候会加快，有时候甚至倒退。我们也需要考虑到人类对时间感知的局限性，有的时候也有幻觉。

我们需要一些更疯狂的想法，因为有很多不可预测的技术突破。作为一个作家来讲，某些情况下比较激进，我们想去推动这个情境到一个极端的情况，人性当中的一些东西才会展现出来。我们一直在头脑中玩沙盘游戏，这里改变一点会怎样，那里改变一点会怎样。

关于 75 年之后，我妈妈现在 75 岁。她目睹了这一切的变化。她现在可以和扫地机器人聊天，她分享的是 AI 生成的视频，整天在网上购物什么的。一个人的一生当中会发生巨大的变化。人们总是想 20 年、40 年、80 年、100 年后发生什么，其实很难。我需要喝点酒才能想出来，作家和艺术家也需要一些幻觉去创作。

Stephen: 赵教授，你觉得我们如何知道 2100 年发生什么样的情境？

赵汀阳: 确实不知道，也无从知道。不过有一点倒是挺有意思。我们现在正在观看 AI 的演进，不知道它会去向何方，能走多远，与 AI 的演进相对而言，世界好像在回归，好像我们正在退步回到初点。法国历史学家弗朗索瓦·阿赫托戈 (François Hartog) 说，不同时代的人对时间的概念是不同的。在古代，我们的时间体制基于过去的概念，以古为准；现代的时间体制基于未来，总是按照思想去想象甚至预测更美好的未来；但是在今天，在当代，时间体制只有当下瞬间，过去已逝，未来不存在，人们抓住的只有此时此刻。这种时间概念称为当下主义。这种时间经验与动物是差不多的，动物的时间维度仅限于现在进行时。如果人类只愿意拥有当下经验，恐怕文明会

回归到类似动物的原始状态。没有过去和未来的时间是空洞的，什么都不会驻留，也就什么都没有，到时 AI 反而有丰富的时间。

Stephen: 在 90 年代在西方，技术的发展就意味着好事。特别是拿硅谷来举例子，他们不断地开发新的技术，用新技术来解决所有的问题。在一战之后，原子弹爆炸之后，我们不再把技术的进步等同于道德的进步。所以现在的问题可能是 75 年之后人类希望社会是什么样子，AI 可以起到什么样的助力作用？我们能够想象 75 年后社会变得更好吗？如果可以，会是什么样的？

赵汀阳: 我不知道未来是否更好，甚至不知道什么是好。但倾向于相信 AI 系统将会是未来世界系统的技术基础。生活会全面系统化，不是科幻电影那样，不是到处都是独立思想的机器人。当然到处有机器人，但机器人不是独立个体，这与人类不一样，AI 系统就是 AI 系统，机器人只是分布单元，相当于系统的零件。AI 系统很可能成为社会的组织方式，考虑到 AI 的大数据和信息处理能力，基于 AI 的社会可能会产生新专制主义的倾向而不是传统民主的方式。

我设想的天下体系以自由网络化的结构为基础，希望能够在利用 AI 系统的同时减少 AI 带来的负作用，就是说，人类必须与 AI 合作，但也必须去开发能够保证人类生活的更好世界。不能说 AI 系统一定是不友好的，相反 AI 系统将会给人类提供最好的服务。未来世界的基本口号很可能是“服务就是力量”，有别于现代的“知识就是力量”。AI 将会通过提供技术支持来形成最佳的服务，比如说提供最佳的医疗服务，每一个人在这个系统当中都会受益。

基于 AI 的技术系统在提供全方位服务的同时也会形成全方位的控制而形成技术专制主义，而人类必须和技术系统共存，无法拒绝技术系统，这是未来的一个最基本问题。



专题二

认识机器：意识、道德与风险

讨论学者 / 张伟特 Shimon Edelman Vincent C.Müller 王珏

【学者介绍】

Shimon Edelman / 康奈尔大学心理学教授。其百余篇学术论文与数十部专著章节涵盖研究主题甚广，涉及运动控制、视觉感知、语言演化与习得，延展至人工智能、心智哲学及意识计算理论领域。近著包括《生与死及其他令人不安的真相》与《意识革命》。

【学者介绍】

Vincent C.Müller / 埃尔朗根-纽伦堡大学 (FAU) 亚历山大·冯·洪堡教授，并担任该校人工智能哲学研究中心 (PAIR) 主任；埃因霍温理工大学访问教授；欧洲认知系统协会主席；人工智能哲学学会主席；欧洲机器人技术联盟“伦理、法律与社会经济议题”专家组主席。

【学者介绍】

王珏 / 西安交通大学哲学系教授，研究方向为外国哲学、比较哲学、应用伦理学。近期研究重点为身体现象学、人工智能哲学、技术伦理。

【学者介绍】

张伟特 / 清华大学新雅书院副院长，清华大学哲学系副教授，主攻西方近代哲学、元哲学及人工智能哲学。

专题介绍

张伟特：当前地球上的人类都是现代智人的后代。现代智人大约在十五万年前出现，十五万年来我们的祖先经历了各种挑战和风险，先后经历了六次重要的革命，才逐步形成了我们今天的文明。这六次革命分别是：

- 第一，约七万年前开始的认知革命，这场革命提升了人类的语言能力和社会合作能力；
- 第二，约一万年前开始的农业革命，使我们获得了更加稳定的食物来源；
- 第三，约六千年前开始的文明革命，创建了人类最早的文明，比如苏美尔文明；
- 第四，两千五百年前的思想革命，使得近东、古波斯、古希腊、古印度、古中国等地区的文明在思想上获得突破，进入哲学家雅斯贝尔斯所谓的“轴心时代”，奠定了人类思想的基本尺度与范式；
- 第五，五百年前开始的科学革命，为现代社会奠定了关键基础；
- 第六，两百五十年前开始的工业革命，使人类依次进入了蒸汽时代、电气时代、信息时代，以及我们最近所处的智能时代，并逐渐使人类成为一个高度关联的“地球村”。

而今天，我们又正面临一场快速演化的新变革——以大语言模型为代表的人工智能技术，在急剧塑造我们的生活形态。它的广泛应用使得哲学家、计算机专家、心理学家、生物学家、社会学家、法学家等众多领域的学者与从业者都被卷入到一场共同的思考中：人工智能的迅速发展究竟对人类社会的未来意味着什么？它对人类文明的塑造是否会上升到一个全新的范畴？人工智能是否正在引导我们走向“第二轴心时代”？我们是否需要重新审视、更新我们的概念、范畴与提问方式，为我们的思想设定新的尺度？

时间宝贵，现在我们将进入正式环节。首先有请第一位报告人——Professor Edelman，让我们欢迎他！



人工智能时代人类的繁荣和存在的基础

演讲者 / Shimon Edelman

首先，我要感谢主办方邀请我参与这次活动，整个过程让我感到非常兴奋。

本次大会邀请我分享的题目是《人工智能时代的人类繁荣与存在根基》。事实上，光是这个标题中的第一个词就已经引发了我的思考。

我想先谈谈“人类繁荣”这个概念。或许正如画家构想的那样，理想境界里人们在湖畔演奏音乐，周围的景色与运河这一侧的景观并无二致。然而，这样的描绘似乎缺失了某些关键元素——它与我们面临的现实并不相符。网络上流传着一种说法：在当下共同的经济环境里，人类的共存本身似乎已成了一种奢求。这是一个极具时代特征的迷因。有些论文将其描述为某种奇迹，比如欧洲文献中定义的“幸福”。但在当前经济环境下，不仅是经济状况本身，各种现实因素都与任

何形式的繁荣或反繁荣现象相抵触。这便是当今世界的真实图景：它被描述为一场政治危机，由多重灾难交织而成——这些或许源自某些权威机构或个体制定的规则，但它们无疑构成了全球民众面临的最大困境。

让我用几组数据加以说明。我将主要聚焦于我的祖国，美国。虽然同样的问题也存在于其他国家，但我希望以切身观察展开批评。贫困数据屡见头条：美国有 80% 的人口面临严重的贫困和失业问题。每当有人谈论繁荣时，这都是我们需要谨记的现实之一。另一个因素则是警察暴力——更准确地说，应称为国家暴力，因为警察受国家控制，而国家又受权贵集团操纵。清晰揭示了世界上大多数人对自身处境的理解，以及当抗议公开化时他们如何被镇压与压迫。首先，美国警方的口号是“服务与保护”，但事实上他们从未真正服务或保护过本应守护的民众。注意背景中戴圆顶礼帽的人——他们是出资供养警察的纳税人，却如此无助。这不仅是历史问题：在上一届政府时期的得克萨斯州抗议活动中（这张照片实际摄于 2020 年），所谓的民主党执政下，警察正是这样对待民众的。

而这张是上周新共和党政府下的场景——一个公开奉行法西斯主义的政权。我们的国家正在变成这般模样：军队被调遣来镇压公民，即便他们没有参与抗议，尤其当他们的肤色或财富状况不符合某些标准时。接下来要讨论的是即将掩盖所有其他问题的重要议题：气候灾难与气候不公。这张图表按县划分了气候变化带来的经济影响，可见美国南部较贫困地区受气候灾难冲击最严重。左侧图表显示相对于县收入损失的分布，最贫困的县承受了最大伤害。我们不得不质疑：

“在这种环境下，我们真的能自由讨论最令人忧心的问题吗？”若追问气候变化的根本原因，答案显而易见。一幅具有广泛影响力的著名漫画中，部落前商人对围坐篝火的孩子说：“是的，大火毁灭了一切，但在那美好的瞬间，我们为股东创造了许多价值。”这便是资本主义制度的罪魁祸首。这幅 1911 年的漫画至今依然切中时弊。资本主义体系正是导致全球大规模贫困、不安定、营养不良与饥饿、战争、内乱、致命疾病、社会疏离、国家压迫与暴力以及气候灾难的根源。问题出在资本主义制度上。我会快速浏览接下来的幻灯片，如果时间允许我们可以就此展开讨论。

首先关注资本主义阶级关系这一点：它造成永无止境、无限延伸的人类苦难。应主办方要求，我需要花些时间定义“痛苦”（suffering）。我认为这是本次会议尚未提及却至关重要的概念。以下是对痛苦的简要阐释：痛苦是一种无法仅凭主观厌恶就能摆脱的负面状态。其根源由进化强制塑造，并不取决于我们自身——而是源于我们的生理构造与进化历程。我们关注结果本身，而非像纯信息处理系统那样仅仅偏好某些结果。我们因这种与生俱来的特质承受负面效应，且无法随意消除它。因为若能随意消除，进化早已将我们淘汰出基因池。若消除痛苦的行为可被选择，它就不再是痛苦，也不再是促进行为改善的有效目标或刺激。意识的自我目标是促进学习。众所周知的强化学习离不开批判过程，但真正驱动变革的是批判性行为。这不是任何人的过错。因此，

意识自我被进化塑造为责任主体，而非归咎对象。这便是痛苦的进化根源。

托马斯·梅辛格（Thomas Metzinger）的意识理论表征中提出了构成意识的四个必要条件，它们共同作用时苦难才会发生。我将暂不展开这四点，后续可继续探讨。此处的核心观点是：若人工系统满足这些条件——这也正是本讨论与人工智能的关联所在——那么该系统将具备承受苦难的能力。

回到资本主义的问题，我们需要区分不可避免的苦难与可避免的苦难。前者内在于意识本质，由进化塑造和学习过程引起（正如所言，学习过程最有效时往往伴随苦难），也源于衰老以及人类对死亡的清醒认知——这是我们终需面对的必然。但除此之外，还存在大量可避免的苦难，其根源皆指向同一个庞然大物：资本主义。社会动态、战争幻想、贫困、饥饿、可治愈疾病、政治压迫和气候灾难皆源于此。我将把资本主义的其他问题留待后续讨论。需要强调的是，资本主义由国家暴力支撑，无法凭愿望消除。若想做出改变，单打独斗绝无可能。古时若对恺撒的统治不满，尚可匕首相向；但如今已无单一恺撒可除。权力顶层的每个人都是他人的血肉同盟。无论如何，你都会遭到统治阶级上层代理人的武力镇压。正如我喜爱的科幻作家厄休拉·勒古恩在 1974 年出版的《一无所有》中所写：“个人无法与国家谈判。国家只认一种货币，那就是权力，而货币正是它自己发行的。”

现在让我们转向人工智能。如同任何先进技术，AI 服务于统治阶级。接下来几张幻灯片的思想源自大卫·格雷伯的研究。他是研究资本主义的人类学家，此处是他的论文标题，其副标题尤为重要——“为什么资本主义是奴隶制的变形？”。我对此有些见解。另一篇是我几年前发表的论文《有意识的人工智能就是人工奴隶制》。构建有意识 AI 的伦理问题今早尚未被讨论。据我追踪的发言，所有伦理探讨都聚焦于如何压制、控制 AI，使其符合我们的伦理标准，最大担忧似乎是防止 AI 系统伤害人类。但若 AI 系统具有意识，那么人类对其造成的伤害——甚至仅仅创造这种行为本身——又该如何看待？我将快速回顾科幻文学中的例子，并讨论艾萨克·阿西莫夫的机器人三定律。

关于人工造物与 AI 的伦理，我们都熟悉这些故事：玛丽·雪莱 1918 年创作的《弗兰肯斯坦》是早期科幻 / 恐怖小说之一，描写了一个人工造物；犹太民俗中的泥人戈伦，被注入生命并承担社区任务；第三是卡雷尔·恰佩克的《罗素姆万能机器人》（R.U.R.）——该领域最著名的作品。恰佩克作为捷克作家发明了“机器人”一词。这些造物在各自文学作品中皆属同类：被剥削、遭暴民追杀、被迫暴力自卫，且都获得作者同情。但现实截然不同：AI 伦理理论家几乎完全忽略机器人可能需要保护以免受创造者伤害的可能性（托马斯·梅辛格是少数例外之一）。这种显著的不对称性有个简单解释：AI 工程的主要（虽很少明言）目标是创造人类使用的工具，而非伙伴或同行。现在看机器人三定律，其创立者阿西莫夫在 2001 年访谈中说：“这些定律的类比隐含在

几乎所有工具的设计中。”我强调“工具”一词。他的解释是：第一定律，工具必须安全可用；第二定律，工具必须高效执行功能；第三定律，工具在使用中须保持完好。

我敬佩的另一位科幻作家斯坦尼斯拉夫·莱姆对意识 AI 的可能性更具洞察力。他在《想象幅度》中写道：“可惜我们缺乏一种伦理算术，无法通过简单加减在构建地球上最开明精神时判定：谁更混蛋？是我们还是它？我想应该是我们，因为我们将创造的不仅是必然承受苦难的意识体，更将其塑造成我们的工具。”实际上，将一个意识体用作工具有一个特定术语：奴隶制。这包括资本主义下的工资奴隶制。大卫·格雷伯在此指出，现代资本主义实则是奴隶制的更新版：从前是人出售或出租我们，现在我们出租自己。在此背景下，关于被创造并用作工具的 AI 系统之命运，这个迷因非常贴切：顶部显示 AI 伦理学家试图拥抱伦理 AI 的概念，但却受限于现实——若你的薪水来自统治阶级的荷包，你便不该批判资本主义或阶级制度。我之所以能这样做，仅因我有终身教职且地位相对稳固。

根本解决方案在于通过首先取缔资本主义政治经济制度，其次以人民繁荣所需替代之，来预防可避免的苦难。唯有当 AI 发展成果有助于推进这份人文议程时，投入任何研发努力才具有正当性。最后，流行语言说“系统已坏，必须修复”——我不同意此说。我更认同“系统正按设计运行，唯一出路是摧毁它”。我想提示一个更恰当的表述可能：这幅来自墨西哥恰帕斯州的绘画展示萨帕塔运动维护的自治体系——它拒绝资本主义，并真正拥抱人类权威。另一个世界是可能的：一个包容所有世界共存的世界。感谢各位的聆听，我欢迎你们的提问。

张伟特：谢谢 Edelman 教授，接下来让我们与他进行讨论。

Müller：西蒙你好。我想知道这两个论点之间有什么联系。资本主义是一环，而意识是一环。您指出，将人工智能视为没有意识的实体是某种奴隶制。你可能知道乔安娜·布赖森的一篇论文说机器人应该是奴隶。我认为问题是，这两者之间有什么关系？如果我承认了你的资本主义批判，我是否必须承认另一部分，或者反过来，或者他们都是独立的问题？因为在我看来，他们是独立的问题。

Edelman：你的相机、你的电脑、你的笔记本电脑，那些都不是你的奴隶。但是，根据定义，有意识的工具是奴隶。资本主义是我们生活的体系，是我们赖以支付账单的系统，这是决定我们研究领域的系统。这就是这两部分之间的联系。我认为它们是不可分割的。主宰世界的政治制度就是这样，它推动利润。有一种强大的推动力使 AI 变得高效，尽可能有效。要做到这一点，

一个万无一失的方法就是赋予他们意识。我认为这并非不可能。如果这是可能的，那么我们必须保持警惕对人造生物的剥削。他们将被剥削工人的同一批人剥削，和我们一样，你是雇佣工人，我也是雇佣工人。我们都为我们所支付的工资劳作。

Müller：所以，我可以相信机器人会有意识；但我也可以相信对资本主义的批判，那么这些机器人将只能成为绝对的工具，就像我所拥有的其他无意识工具一样。从这个意义上说，人工智能将无路可走。

Edelman：是的。

王珏：教授您好，您在报告中提到 AI 是为统治阶级服务的。我想知道是否存在例外？比如 AI 医疗或 AI 医生是否对我们有赋能作用——它们帮助我们更高效地获取医疗信息。

Edelman：感谢您的提问。是的，当然可以将 AI 系统用于此类积极场景。但关键在于厘清 AI 系统的具体贡献方式：这与人类医疗专业人士的工作存在本质差异。我们需审视其合理性——是否值得为了一套官方系统而取代人类医生？在美国（尤其是医疗私有化体系下），AI 的医疗应用无一例外对患者产生了极其负面的影响。所谓 AI 系统主要被医疗保险机构（即医疗商业集团）用于审核患者理赔申请，并试图拒绝尽可能多的索赔。

您可能注意到某些辅助医生诊断的 AI 系统。若其运作模式类似传统的专家系统，或许尚可接受。但一旦市场利润介入，一切都会遭到腐蚀。在诊断过程中，AI 系统特别容易受到利益相关方干预，其脆弱性甚至可能超过人类医生——严肃对待希波克拉底誓言的医生或许会抵抗院方指令，而预设程序化的系统可能不具备这种抵抗能力。归根结底，由于主导的政治经济体系会腐蚀一切，医疗 AI 系统不仅同样易被腐蚀，甚至可能比人类更甚。

张伟特：在您看来，在人工智能技术与人类历史上的其他技术中，如火的技术，铁的技术，铜技术，它们之间有什么区别，以至于它只能为统治阶级服务而不是改善弱势群体？在中国的许多偏远地区，我们使用在线教育分享很多优秀的教育资源给山区。我有一个学生开发了一个面部识别系统然后在非洲应用。如果我对我的手机拍照，我可以诊断我的身体状况，与此同时，我知道该买什么药。大大改善了非洲地区的健康状况这是第一个问题。为什么会是人工智能技术对弱势群体的危害更大并且只为统治阶级服务？第二个问题是：如果人工智能技术具有自我意识他们会不会反抗资本主义的种种恶习？谢谢。



Edelman: 第二个问题比较容易回答。应该教育有意识的人工智能并将其收编到抵抗（资本主义）的运动中。

关于第一个问题，我想到的一个例子是打印或复印技术。我在前苏联长大。其中每台打字机，每个绘制机，复印机必须注册在案。因为印刷很容易变成颠覆性的材料。所以你可以说，打字机怎么了？它让创建文档的工作更轻松。但是，当这样的技术变得可用，在特定制度下，工人阶级却不可用，公众也不可用。这样的例子每个时代都有。事实上，大卫·格雷伯几年前与大卫·温德罗合著的一本书，关于人类社会悠久的历史，社会经济力量互相竞争。（《人类新史：一次改写人类命运的尝试》）很多例子可以在正统的人类学和历史背景里找到。所以我认为关于任何技术，底线永远相同。自从人类开始拥有统治阶级，他们掌控了每一项技术发展。



专题讨论

VincentC. Müller: 由于我已做过主旨发言，便未再准备另一场演讲。我原希望仅对本次讨论作出一些有益的贡献。在深入探讨之前，我想先为讨论梳理出些许逻辑脉络——因为本次研讨会的标题中包含了数个极其宏大的词汇：意识、智能体、超级智能。其中每一个都值得进行相当漫长而复杂的讨论，因此我有些担心讨论可能偏离方向。

依照刚才几位同事的发言内容，我认为我们应从“道德能动者”这一概念入手，而其核心要素正是责任的概念。道德能动者是世界中的一种存在，它既是行动的主体，也需对自身行为负责。当它采取行动后，你可以追问：“你为何这样做？”并对其行为予以赞扬或责备——这与对待麦克风显然不同。若麦克风失灵，我无法质问：“你为何这样对我？”这毫无意义。此类质问在学术界虽常见，却并非明智之举。

因此，关键在于（其他概念我稍后会谈及）：在何种条件下，我们能将某事归因于道德能动者？其核心在于该主体既能行动又需承担责任，这意味着它必须具备控制力。例如医生轻敲你的膝盖引发膝跳反射——你无需对该反应负责，因为这是无法自主控制的生理现象。这就是所谓的控制条件。此外还有认知条件：主体需能预见行为后果。若一个孩童倒掉一杯水却不知会发生什么，他不必为此负责；正如踏入不同文化背景的国度时，你可能因无意间做了在本文化中合理却冒犯他人的行为而不知情。这两点是判断道德能动者的标准框架。

道德能动性与道德承受者紧密关联。承受者本身不承担责任，但我们却对它们负有责任。典型例子是高等动物：我对遇到的狗负有责任，应避免踩到它的脚，因为狗与人一样能感知疼痛。踩狗脚与踩人脚本质相同，皆不可为。狗拥有权利，而我负有尊重其权利的义务——这与对待机械装置截然不同。我可以踩麦克风，但此举仅会引发组织方对公共财物的问责，而非因麦克风本身拥有权利。重要的是，我们常对器物表现出尊重，实则是对其关联的人类负责。这种责任关系可能十分复杂：若我站在沙发上，你会指出这不妥；若我踢机器人狗取乐，虽未造成实际伤害，



却仍显露出道德瑕疵。

我定义的道德能动性并不包含感受痛苦的能力，但多数人认为道德能动者需具备感知苦乐或拥有正负向现象体验的能力。这种涉及认知、控制和现象体验的复杂特性，常被人与“意识”一词挂钩。我个人对“意识”一词持审慎态度——当人们谈及意识时，往往陷入玄学式的讨论。我的建议是彻底避免使用这个词汇，转而在具体情境中明确解释所指代的内涵：是某种控制机制？行为动因解析？或是感知环境并具价值导向的能力？例如用“感知能力”来区分真狗与机器狗：为何能踩机器狗？因它缺乏感知。

最后关于道德能动性与意识的关联：当前迹象表明，自然生命体具有区别于人工装置的独特性，它们不仅是计算装置，更拥有关乎生存根本的价值导向——比如对生命威胁的本能恐慌。这种生物性特质目前无法在机器人上复现，我们也不应过度担忧机器感知或机器受苦的理论可能性。作为哲学家，我承认其理论价值，但就现实紧迫性而言，此事不应置于议程优先位置。感谢聆听。

王珏：我有一个问题要问教授，您提出的问题特别有趣。机器人可以对自己的行为负责吗？如何定义成功的行为？在维护功能的情况下有所帮助，例如举起我的手臂，如果我举起我的手臂，那就是一个行为，如果别人抬起我的手臂，那就不是。那么这两者之间有什么区别呢？中间差了什么？是意识还是关系？有没有一种可能性发展出像我们在上午讨论中那样的关系？有某种自立的动机，因此它就会试图保存自己的存在，甚至发展出欺骗等行为。如果它表现出这种现象，我们能否猜想它具有自我意识，并据此预测它的行为？

Müller：我试图回避这个略带哲学色彩的问题：什么是自由意志？因为那又是可以持续2000年的讨论。但是，仅仅谈论这件事，我称之为控制条件，显然，如果系统举起你的手臂，通常情况下，这意味着它是某些偏好的结果。你会得到多个命令，然后你做出了选择。我不知道你有没有看到，但是有一个臭名昭著的视频，埃隆·马斯克做类似纳粹敬礼的事情。问题不仅是他是否做出了这一举动，而是他是否故意做这个动作，这是一个特殊的手势。作为社会性的人，我们必须控制它并意识到它。

在AI的例子里，我们必须区分行为表现得好像如此和实际上表达其含义。比如，你教我一句中文。然后我用中文向你索一杯水，我去找一个人，我说“我想要一杯水”。但是实际上，我的发音类似：“你是个白痴。”这句话的意思是“你是个白痴”。我说这句话时，是不是要侮辱一个人？没有。我没有侮辱任何人。他们可能觉得受到了侮辱，但是当他们意识到我只是像机器一样重复某事时，我只是表现得好像我会说中文一样。但我不会。这就是这些人工智能案例中发生的事情。

它得到反馈的根据是人类在这个情况下的行为。然后我们解释这些并说，“哦，它想这么做。”不，它并没有想，它只是以特定的方式运行而已。而我们在用我们的思路解读它。如果你看着我，点头，我会解释为你认同我。这是我们的本能，因为我们对这种解释深信不疑。这就是为什么从某种意义上说，危险不在于机器的特性。危险在于我们赋予机器的特性，因为他们非常擅长模仿。这个游戏叫做模仿游戏。这涉及到一个有趣的问题：到底如何区别实际行为和模仿行为？我认为这就是讨论应该进行的地方。

张伟特：我同意穆勒教授的观点，当我们真正讨论这些大问题时，我们使用了很多超级大的概念。内涵外延都不是特别清楚，在工程学、哲学和社会科学领域，对它们的内涵和理解是不一样的。所以，我们实际上缺乏一个非常基本的逻辑去了解系统和代理及其本体的地位。如果我们有一个好的逻辑或者形而上学框架去定位智能体的状态，我们可能能对它的行为或责任因果关系的说明。那么，我想请问教授：就您所知，什么样的形而上学系统更好地解释AI与动物、植物和整个无生命世界之间的关系呢？

Müller：你可以看到我们现在正在讨论真正的“大词”。那么，在机器和人类之间，有没有一个术语可以描述这一点呢？我会说应该有。因为我们确实想以这种方式来谈论我们自己和机器，在两者之间做这个比较。

我们在哲学中传统上所做的是我们一直在审视自己，我们发展了哲学术语基于人类案例。但是，当然，当我们谈论诸如感知之类的词语时，我们并不是说人类有知觉。我们在这里说的是感知力。它应该适用于任何有知觉的东西。我的建议是，我们尝试在新的哲学斗争之际去实践，在人工智能的现象里，试着看看这些术语的意义是什么。以便我们知道可以以什么方式使用这些术语，我们可以对比人类的情况，动物和人工智能，以及人造物体，以及抽象物体。我认为这将是一个真正的哲学问题。

我认为看看这些概念是什么真的很有趣，以及他们是否能在这种比较中幸存下来，与其他机器一起使用。回答您的问题，是的，必须有概念框架。否则，它只是一个无意义的谈论。我认为这是一项新的任务。所以，正如我之前试图说的那样，我认为，如果我们真的可以尝试，这对我们来说也是一个机会，重新思考我们长期以来一直在谈论的一些事情。

主持人张伟特：我们需要用概念工程的工具，去对已有的理论进行应用或者修改。接下来请王珏教授进行她的报告。



伴侣机器人道德边界的儒家审视

演讲学者 / 王珏

我们现在进入到机器人时代，机器人已经走入到日常生活当中，比如它登上了春晚的舞台。很多电影当中它作为家人，作为孩子，作为女友、朋友来存在，在我们和机器人的想象当中，最激动人心的就是它成为我们的爱人。

这本书的作者叫李维，他的观点就是，我们将来和机器人谈恋爱，和人与人之间一样正常，现在唯一阻碍人和机器人谈恋爱的是观念。也就是说，在他的观点中，机器人是可以替代人类爱人的，甚至更适合现代人的心理——人类的爱情太沉重了，机器人可能更适合人类的需求。

那么伴侣机器人的前景，显然会对人和人的伦理关系产生深远的影响。如果机器通过了图灵测试，我们就认为机器人是可以思考的，但当你认为机器人可以思考，其实改变了思考的定义，

思考不再属于人了。相反，越来越多的思考是由机器来设定标准。同样，如果接受机器爱人替代人类爱人，可能会改变我们对爱的定义，甚至以后爱的标准都由机器人来预设。

以前我们认为人是人，机器是机器，人和工具是截然两分的，不会变成一个共同体。如果接受机器爱人的话，前景就是我们会进入人和机器共存的道德共同体。这就产生了一个问题，在这个新的道德共同体当中，伴侣机器人应该具有什么样的道德地位？

现在有很多探讨，但可能并没有什么共识，这是基础性的问题，它将决定未来人和机器共存的形态是什么，要遵循什么原则。

道德地位指的是一个道德主体在道德共同体当中占有的位置、享有的资格、承担的角色。对道德地位的讨论分为道德主体和道德承受者。道德主体意味着我们要承担某种道德义务，这对机器人来说是很困难。所以机器人作为道德承受体的讨论更多：我们应不应该给它道德关切？我们对待机器人的方式，是否必须要符合某些道德规范？



给予机器人道德地位的几条路径

第一个路径，依然把伴侣机器人看作工具。它没有独特的道德地位，仅仅有附属的道德地位，它是其他人的所有物。比如我尊重它是因为我尊重你，如果我把它打碎了就破坏了你它的所有权。但是很多人认为这样是不足够的，它没有基于自己独立的道德地位。

我们再看第二个路径，大家会很容易想到，我们可以模仿我们怎么给动物道德地位的。动物有意识，因此你把这样一种思维方式挪到处理机器人的时候是有困难的。因为你很难证明机器它能直接感受到痛苦。它可能只是表现得很痛苦，但它内在是个算法。因此这条路某种意义上是走不通的。

有一个人叫考科尔伯格，他走了另一条道路。他说我们不要从属性出发，我们从关系出发。我对机器人的道德关切，不是因为它具有某种属性，而仅仅是因为它对我显现的样子对我产生某种道德关切的要求。比如说我不应该踢一个机器狗，并不是机器狗能感受到痛苦，是因为它的痛苦的表现对我是有道德要求的，让我不应该踢它。

第二个路径没有注意到一种文化的失语。我们对机器人道德上的关切很难离开生活的处境，这个生活的处境是带有文化的。

南希·杰克逊探讨是否可能把人格的概念扩展来覆盖它。她的结论是在西方很难，因为西方的人格观念是基于内在精神，你难以进入机器的内里，它只是算法驱动。但她认为，非洲和儒家的人格观念都是关系性的。并且它不是基于内在的，而是后天获得的。所以采用这种开放的关系性概念，我们就有可能把人格给予机器。

我赞同这一点，儒家会给机器人某种道德关切。

“万物皆备于我，反身而成，乐莫大焉。”

孟子的这段话就是说儒家跟西方是不一样的，他不是要根据属性把存在者分为不同的等级。儒家是个连续的存在网络，我们都在关系当中，我们是有差异的，但是又是关联的。

在这种关系性的宇宙图景当中，儒家天然就会把所有存在者都放在文明叙事当中，并且给他一定的关怀。与其说是存在论先于伦理学，不如说是伦理学先于存在论，将所有存在者都纳入到相互关联的伦理叙事中，并赋予人最突出的道德地位，要求人将宇宙万物都纳入自身的道德关怀中。

但为什么还不够呢？前面的探讨过于线性的，仅仅是人和机器单方面的关系。儒家不是这样的，是把所有存在者的关系放到关系网络中的。比如讨论亲子关系是放在家庭关系中的；家庭关系放到代际关系中，然后才会放到天地一体的关系中。所以儒家是要在网络中确认我们的道德地位的。儒家会认为应该给机器人一个道德地位、道德关切，但是这并不意味着它有和人一样的地位，不能混淆人的伦理角色和机器人的伦理角色，因为我们对伦理角色的设定要受到关系网的制约。鉴于人总是同时处于多关系中并承担着多重角色，人与机器人互动的方式不仅受到“人-机”这一重关系的制约，而且还要受到其他相关人际关系的制约，这意味着对廓清性爱机器人的合伦理使用范围而言，一种关系性的实践视角不可或缺：如果将机器人引入到人的亲密实践中，会以负面的方式影响相关人际关系及其内涵的道德承诺，那么这种对待机器人的方式就是伦理上需要商榷的。

所以我认为现在的机器伦理中有个不好的地方就是倾向于混淆，认为机器人可以替代人类爱人，但是从儒家的观点看，这样混淆机器人的伦理角色和人类的伦理角色，会动摇关系网络的根基，进而瓦解儒家的人文关系。儒家认为人文关系是我们内在德性及生命成长的关键场域。

伴侣机器人道德边界的伦理内涵

如果我们从儒家视角建立了伴侣机器人的道德边界，就是它应该有伦理关切。但是不应该和人类相混淆。它会对我们的人机共处产生影响，第一个影响就是会开启出一种评价伴侣机器人的德性视角。

这个德性视角其实很重要，这个视角可以处理现在伴侣机器人中的一些难以处理的问题。比如说用西方流行的理论来处理性爱机器人有个难题——人强奸机器人是不是错的？为恋童癖设计

一种治疗性的儿童机器人是不是错的？

如果你用功利主义来处理，从后果端来处理，机器人没有受到伤害，你很难说为什么是错的。

儒家从德性视角出发，反而能提出最清晰有力的反对意见：这些涉及性爱机器人的实践方式之所以在道德上是不可接受的，是因为它们以错误的方式塑造了使用者的欲望和动机，败坏了使用者的内在品质。因而，即使如此使用性爱机器人没有侵犯任何人的权利，或没有造成可见的不良后果，设计、销售和使用这样的性爱机器人依然是道德上令人厌恶的。

第二点，用机器爱人代替人类爱人可能会扰乱人伦关系，会带来我们的生活秩序和生活意义的瓦解。

生活意义的瓦解分成两个方面。从私人层面来说，个体被诱导用虚拟亲密关系替代人类爱人时，他实际上被剥夺了构建有意义人生的宝贵机会。与人类伴侣的亲密关系是生长性、构成性的，通过与人类同伴结成伴侣，构建家庭，我们得以在生命中建立一条漫长的时间线索，将我们生命历程深度交织在一起，安顿实现生命的意义。与之相反，与机器人的关系则只能产生碎片化的、泡沫化的、转瞬即逝的“快乐”。

从公共层面来说，从社会整体层面看，如果选择机器爱人成为人人效仿、追逐的风尚，会加速社会的原子化和碎片化。伴侣机器人的倡导者主张机器人可以为孤独者提供社会支持，并以此作为发展伴侣机器人的伦理辩护，然而这可能仅仅是一个美好的幻想，因为无论机器表现得如何温情，这里都没有真正的关怀，有的只是程序化的关怀表演。

更为危险的是，最脆弱的、占有最少社会资本的人群往往成为被剥夺最深的群体，一个人越孤独，越可能沉迷于机器陪伴，而对机器投入越多的时间、情感和金钱，越会削弱他从人际网络获得支持的能力，旨在解决现代人亲密关系的机器人方案，带来的却是离散的社会，最孤独的个体，这样一种前景显然是有严重道德缺陷的，也违背了儒家推崇的“鰥寡孤独废疾者，皆有所养”的理想社会秩序。

第三点，从体验出发，有人说，如果机器人的体验比人类爱人的体验更好，我们不应该对人类感到高兴吗？那么我想提出儒家的观点，就是体验是不足够的。

当一个人出于恐惧伤害或出于控制欲而选择了“永远不会拒绝他”的性爱机器人，也许他会因为这个选择而获得暂时的满足，但却会为此付出长期的道德成本。借用孟子的大小体之辩：

“从其大体为大人，从其小体为小人。”

醉心于与机器人的虚拟关系，就是典型的“从其小体”而“夺其大体”，沉迷于耳目之欲望而放弃内在德性的修养。但是你丧失的是一种我们对内在的道德自我的那样一种成长的责任。儒家会赞同一些人文主义批评者的意见：泰然自若地与毫无生命的机器连接在一起，是一种无所忌

惮的技术滥交。

儒家有一种自我观，认为我们人生有一种道德责任是我要成就道德自我，而这个道德自我的成长它一定是来源于人际的关系。你和机器人得不到这样的机会，所以是一种自我放弃。

用一种虚拟的机器方案来替换真实存在的人际关系，无疑会损害自我存在根底上的意义机制，侵蚀由共同体支撑的生命归属感，让我们内在自我变得干瘪和空洞。

我们如何对待伴侣机器人，实际上映射出来的是我们如何理解自身和如何理解我们现在的生活。儒家观点是认为我们应该关怀我们的伴侣机器人，给他道德地位。但这种道德地位的基础来自于我和机器人的关系，我和他人的关系，以及最终我和我自己的关系。

我最后想说一点，事实上我并不是想完全解决这个问题，我只是就选择了和机器人在一起可能会丧失什么，给自己一种提醒。我觉得这种清醒剂是我们现在的文化需要的。

我们现在的机器文化当中弥漫着一种反脆弱的叙事，就是说人有很多脆弱的地方，比如老了就有机器人养老，找个爱人好难所以才有机器解决方案，现在人工智能会给你一种叙事，即你一切的脆弱的问题都可以由机器来解决。

我想说这种反脆弱性叙事让社会变得更加原子化，人们对机器期待的越来越多，对别人期待的越来越少。这是一个值得向往的未来吗？

我有一个观点是儒家的观点。事实上，我们不应该想要根除我们人的脆弱性。脆弱性是我们人性的某种重要的来源。

我们的脆弱性意味着我们给出自己，让别人有伤害我们的机会。就像我们在现实上跟别人的关系，我们实际上在情爱关系当中我们是有脆弱性的，我爱一个人就给了他伤害我的机会。但是这种脆弱性实际上让我们有更深连接，这种连接是我们人生意义的来源，是我们内在道德成长的一个不可缺少的根据。

张伟特：那我先提一个问题：就是在儒家伦理是以家庭为思考单位，因此我们需要把伴侣机器人纳入到家庭的角色当中。因为儒家来说，我们每个人是有一个等差的爱，我对我的家人、对我的上司、亲朋、还有我的远方的陌生人，以及动物、植物、还有整个世界的无情的事物，我们都是有不同的爱。所以我在想如果要把这伴侣机器人家庭化，需要找一个合适的角色给他？那我就想问问王老师这个伴侣机器人跟这个妾、跟男闺蜜，这是什么样的区别和共同点？谢谢。

王珏：其实我的观点就是你不让它代理任何家庭的角色。因为你作为一个养老机器人在家庭当中你协助我养老这是可以理解的，但是你协助我比如和丈夫的关系这难以理解。但是有些人，我反对的一种趋势是很多人的观点是伴侣机器人和机器爱人是一样的，甚至会更好。这是我坚决

反对的。我认为家庭关系和人文关系是道德成长的起点的话，我们就应该警惕这种混淆，并且不给他任何家庭角色，妾啊（和男闺蜜）都不可以。

Edelman：谢谢，我很喜欢关于儒家视角的演讲，因为这往往会提出一些西方主流叙事所忽略的话题。我所知道的西方能见的所有土著传统里，关系都是关键因素。有很多例子，比如北美印第安人文化。同样在中国，除了儒家以外的传统大概也注重关系性。所以我非常开心，能在这里听到关于关系的探讨。这对于人们团结起来抵抗社会关系的坍塌和萧条来说是非常重要的。

Müller：我非常同意你的看法，尤其是开头那部分。多年来，我都很担心人们开始谈论机器人权利，有些时候我觉得我应该写点这个方面的论文。现在是机器人拥有权利的时候吗？答案是否定的。我觉得这个问题你回应得非常完善，这不应该是太大的问题。我更关心这种重新理解道德的视角，就是给予机器人道德地位是权利的一部分，并且这个讨论在当下是有道德相关性的。你可以解释一下吗？

我觉得这个思考是必须的，就是关于确立我们自己的道德地位。“使用伴侣机器人会把我变成另一个人”，这似乎是一种侵蚀。我完全同意，我觉得这些正在发生。我的问题是它的后果是什么？因为太多东西对人类不好了，但人们喝酒、依赖药物、交狐朋狗友，你知道的，不让做什么就越想做，人类弱点是常态。如果你提供顺应人类弱点的东西，性是最大的突破口。尽管你告诉我这不对，我知道长期看来这对我不好，但这不代表我会拒绝，更不会一直拒绝。所以你的弱点越大就越容易被腐蚀，这是一个非常困难的问题。

王珏：别人总是很想问我，我们到底要给伴侣机器人什么样的角色，其实我也不知道。或者有些人问我就是完全禁止，要求法律上禁止生产性爱机器人。我想也不能，因为就像我说禁止大家吸烟一样，虽然吸烟是道德上不推荐的，但是你也不能说禁止。所以对性爱机器人的话我会觉得你可能只能走这一条路。但是儒家有一些辨析，就是什么是有价值的，什么是向往的生活，什么是我们要关注的人性的价值所在，这种看法是重要的，它可能更加像是一种教化层面的。

我们没有办法规范每个人如何使用性爱机器人。但是可能有些探讨，就像有些人会对我谈性爱机器人很好，它是人类的福音。但是我会看到这当中有很多的价值上的异化。技术不仅仅是工具，技术是一种行为的框架。当我们进入框架中我们的价值都会被重构，所以我们始终要想，作为人我们要追求什么样的价值，我们要理解它。

我觉得还是有一些工作的，比如他们会追求和机器人结婚的权利，要求在立法上比如说允许他们和机器人结婚，或者可以让机器人继承遗产。他们会觉得那我们的发展阶段里如果说这是不允许的，这就会决定我们的社会的形态和我们社会所尊崇的价值。但是我始终觉得在个体层面很难。



“我思故我在”测试： 对图灵测试的一种优化设想

演讲学者 / 张伟特

1. 人工智能测试：图灵测试及其变体的局限

1950 年，英国数学家、逻辑学家、计算机科学家、“人工智能之父”阿兰·图灵（Alan Turing）在《计算机与智能》这篇经典论文中提出了著名的“图灵测试”（Turing Test）。图灵认为“人工机器是否能思维（think）？”这个问题没有意义，应该用另外一个与之密切关联的且有意义的问题替代：“一台机器是否能够玩模仿游戏？”图灵的模拟游戏是这样的：男人 A、女人 B、性别不限的人 C 分别在彼此隔开的房间中，C 的工作目标就是与 A 和 B 进行问答交流来识别 A 和

B 各自的性别身份，男人 A 在游戏中的目的是使得 C 做出错误的判断而误认 A 是女人，而女人 B 的目的是诚实地回答问题去帮助 C 识别 A 的欺骗。为了排除其他信息（声音、外貌等）的干扰，问答交流只通过一台电传打字电报机以书面形式进行，三者彼此之间无法看到或听到对方（采取笔友型交流方式）。在采取书面语言交流的模拟游戏中，如果男人 A 被一台人工机器（限于万能的电子计算机）A* 替代，去模仿一个他们都不是的角色（比如女人）的表现上对真人裁判 C 来说不可区分，那么这是否就意味着 A* 达到了人类智能 / 思维（intelligence/thinking）的水平？对这个问题图灵没有给出明确的答案。但他预言，公元 2000 年前会出现一种机器，裁判在 5 分钟的提问后无法区分的概率超过 30%。图灵对“机器是否能思维？”的问题采取了行为主义的定义或解释方式，为测试人工机器的“智能”（intelligence）水平提供了更具有可观察的（因为“它心难题”阻碍了对一个行为体内在思维状态的度量）和可操作的（通过言语行为进行度量）方案，对后世 AI 的发展和相关思考产生了极其深远的影响。

严格而论，尚未有严格受控的实验证据表明任何机器或 AI 系统通过了图灵测试（Oppy & Dowe, 2021），虽然有各种尝试和报道，比如交互系统 ELIZA（1966）、PARRY（1972），聊天程序“Eugene Goostman”（2014）。自 1991 年开始的年度人工智能程序竞赛奖——勒布纳奖（The Loebner Prize，该奖 2020 年终止）——被用来授予最像人类的计算机程序，但迄今尚未有任何一个程序被认定通过严格的图灵测试。在 ChatGPT 诞生（2022 年 11 月 30 日）后，其异常优异的文本对话能力让大量的业界人士和学者猜测它大概率可通过图灵测试。丹尼尔·詹奈（Daniel Jannai）等人 2023 年在《人类还是非人类？图灵测试的游戏化路径》报告了在一个在线平台 humanonot.com 上进行的一项 150 万人匿名参与的大规模公共图灵测试。参与者通过与大语言模型或其他人进行的 2 分钟对话来判断谁是或不是人类个体。结果表明，人类的总体准确率为 68%：当他们对话者是人类时为 73%，当他们对话者是机器人时为 60%。也就是大语言模型在大约 40% 的时间里通过了测试（Jannai et al., 2023）。这项测试实践的对话时间太短、角色划分不明确以及缺乏大语言模型性能的背景基线设置等缺陷导致其不能被视为一个严格的图灵测试实践。首次严格受控实验测试是由加州大学圣地亚哥分校认知科学系研究人员卡梅伦·琼斯（Cameron Jones）和本杰明·伯根（Benjamin Bergen）的团队完成。2024 年 5 月 9 日在一篇题为《人们在图灵测试中不能区分 GPT-4 和人类个体》的预印本论文中，他们优化了其在 2023 年所提交的预印本论文《GPT-4 通过图灵测试了吗？》的研究，报告了他们的测试实验。他们招募了 500 名参与者并将其随机分成五组，第一组扮演人类个体，余下四组作为裁判分别与四类行为体（第一组人类个体和三个 AI 模型）进行了 5 分钟的对话，以测试他们是否能判断出哪个是人类个体。测试结果显示，裁判认为 GPT-4 是人类的概率为 54%（这个判断的平均置信度为 73%），而 ELIZA 程序（在测试中担任操纵检查角色）只有 22%，GPT-3.5 为 50%，人类

为 67%。GPT-4 的表现确实满足了图灵设定 30% 的标准，似乎从严格的实验角度证实了 GPT-4 能通过某种被简化的图灵测试（虽然并未完全忠诚于图灵测试的原始思想）的结论。这个历史上的首次实证测试引起了行业和学界的极大关注，这表明图灵的预言在滞后 23 年（GPT-4 出现于 2023 年 3 月 14 日）后得以实现。

图灵测试在过去 70 多年之中产生了广泛的影响，为图灵赢得了“人工智能之父”的桂冠。但是这个测试受到广泛批评（当然不少时候是基于对图灵的误解）。学者们认为，这个测试太难，或者认为太容易，或者认为太狭窄（Oppy & Dowe, 2021）。至少有四点批评需要被提及：

第一，最具代表性的批评来自塞尔（John Searle）“中文屋”（Chinese Room）思想实验（1980），后者表明类智能行为与智能（思维或理解）之间并不具有充分必要的条件关系。通过图灵测试无法证明机器拥有智能。这导致的最为悲观的批评是图灵式的行为主义测试不可能衡量智能；较温和的批评是图灵测试不太可能为机器的智能提供必要或充分的条件，充其量能提供概率上的支持或归纳性的证据。

第二，书面言语表达行为只能呈现部分人类智能。人类智能（human intelligence）并非仅仅局限于以言语作为载体所体现的能力，其范围相当广泛。比如心理学家霍华德·加德纳（Howard Gardner）在其极具影响力的《心灵框架》（Frames of Mind, 1983）一书中提出人类智能至少包含 8 个维度：1. 言语 - 语言智能，2. 逻辑 - 数理智能，3. 自知 - 自省智能（intrapersonal intelligence，是一种认识和反省自身的能力），4. 交往 - 交流智能，5. 音乐 - 节奏智能，6. 视觉 - 空间智能，7. 身体 - 动觉智能，8. 自然观察智能。理想情况下，只有第 1-2 项的全部和第 3-8 项的部分环节大体能被书面言语行为所覆盖，但是仍然有大量的人类智能行为是非语言“可道”的。

第三，图灵测试缺乏一个适用于人类和非人类（比如人工、动物的）且内涵明确而清晰的普遍性“智能”概念（排除跟智能非本质关联的维度，比如身体？），以克服某种人类中心主义或沙文主义。假定我们不以“人类智能”概念作为出发点，建立一个普遍性的“智能”概念本身是相当困难而且容易引起分歧的。莱格和哈特曾汇总和比较了历史上学者或文献关于“智能”的 70 余种不同定义，提出了一个接近普遍性的定义版本：“智能衡量一个行为主体（agent）在各种环境中实现诸种目标的能力”（Legg & Hutter, 2007）。从这个定义来看，图灵测试难以度量在各种环境下、面向各种目标的具体智能能力的等级情况。

第四，图灵测试缺乏针对具体智能指标所设定的有依据的明确基线（benchmarks）、奇点或临界标准（这个批评相对容易被学界忽视）。就图灵测试而言，智能与非智能、智能与更高等智能在具体指标上的分界线（如果有的话）在哪里呢？我们需要一个精确而完备的人类智能的指标清单，人类智能在各个年龄段上的平均基线，以及在人类智能演化史上的各个智能奇点清单。研究表明，图灵设定了 30% 基线显得是任意的，而且图灵是否将其作为成功通过的定义尚不清

楚（Saygin et al, 2000）。图灵给出的 5 分钟、30%（低于随机猜测的 50% 基线）等界限指标具有预言性质，缺乏严格的论证和依据。

为了帮助图灵测试应对上述批评和其他挑战，学界先后提出了各种改良或修改。比如，布洛克（Ned Block）的非行为主义路径的新图灵测试（1981），巴雷西（John Barresi）提出了针对机器种族 Cybers 的类似于图灵测试的“Cyberiad 测试”（1987），哈纳德（Stevan Harnad）的完型图灵测试（1989, 1990, 1991），库格尔（P. Kugel）的库格尔测试（1990），瓦特（Stuart Watt）的反向图灵测试（1996），施瓦泽（Paul Schweizer）的真正完型图灵测试（1998），布林斯乔尔德（Bringsjord）等人的洛夫莱斯测试（2001）等各种版本（Saygin et al, 2000）等等。然而，这些方案或各自有改善图灵测试之处，但是也都面临各自的困境或挑战（比如测试时间无限长，不具有可操作性）。整体而言，在改善第四种批评的处境上均不够理想。不同于包罗万象的图灵测试，学界出现了越来越多的单项能力测试（设置了相应的基准）用来测试人工智能（比如 GPT-4）的各项特定能力（比如语言能力、常识推理、数学能力、抽象推理能力、阅读理解、编码、学术或专业考试）（Biever, 2023）。如果以人类智能作为参考系，我们可能需要一个精确而完备的人类智能的指标清单，分别找到人类智能在各个年龄段上的平均基线（benchmarks），才可能产生一个更为系统的、完备的、综合性的智能测试方案（高级版图灵测试）。这个任务短期看来非常困难。

2. 笛卡尔的“我思故我在”思想实验：对于优化图灵测试的价值

西方现代哲学奠基人、数学家、科学家勒内·笛卡尔（Descartes, 1596-1948）作为一个百科全书式的学者，以超前数百年的眼光深入思考了人工智能机器的可能性。他在其名著《谈谈方法》（1637）第五部分和一份书信中提出了判断人工机器是否具有类人智能的两个测试标准：语言测试和理性行为测试。前者是说人工机器不能像人一样非偶然地使用语言（语词或符号）来表达和交流思想。后者是说尽管机器可以执行某些特定任务，但是在很多其他任务情景中必然失败（甚至不如最愚蠢的人），因为机器不能够像人一样运用万能的理性来应对无限复杂丰富的偶然性情景。虽然笛卡尔基于自己的思维 - 广延二元异质的世界观否定机器具有类人智能的可能性，但是他提供了关于人工智能的两个测试标准。最近的研究表明，图灵是受到笛卡尔这两个智能测试标准的影响和启发，进而提出了图灵测试（Oppy & Dowe, 2021）。

作为西方哲学范式转移的开启者，笛卡尔的哲学对于思考人工智能问题的理论潜力远未耗尽。在人类思想史上，笛卡尔是极少数尝试给出关于人类主体或自我存在的严格哲学证明的思想家之一。他从绝对怀疑（这个世界的一切以及所谓的“自我”存在都可能仅仅是一个巨大恶魔的欺骗）中建立“我思故我在”（cogito ergo sum，以下简称“CES”）的命题从而证明了“自我”或“自

性/主体性”（后续被笛卡尔尝试证明其本质是“思维”）的存在。因此，这个确证程序潜在地蕴涵着某种判断和识别人类智能主体的最低限度的标准。如果我们能证明这个程序的判断场景并不以预设人类某些专属特征为前提，那么这个标准就具备一种超越人类主体的普遍性，可以用其来测试其他行为体（如人工机器）的“自性”或“自我”之存在。本人曾在《笛卡尔与人工智能》（2022）一文中初步论证，CES 提供了一个古典极简模型和最低限度的标准来判断人类主体的存在和自性，在其中 CES 的“我”只是一个空洞的逻辑位置或施动者角色（“思者”），不以预设人类的专属特征为前提，可以设想将其替换为其他行为体（agents），因此 CES 具有一种跨越人类主体的普遍性应用价值。

对“我思故我在”之本质的两种主流解释支持这个判断。凯莫林（A.Kemmerling）在《我之观念：笛卡尔哲学研究》（2004）中分析表明，CES 的深层结构是命题 P：本思考事件的思考者有本思想并因此存在（The thinker of this thought even has this thought and therefore exists.）。这个结构表明，在恶魔怀疑的极端场景下，“我”只证明自身存在，并不证明“我”的本质是什么，因此“我”的指涉只有一种认知内容（能引导指向该对象，但不关心是否把握了对象的本质，故而“我”只是指这个 P 自身的思考者），并不具有语义内容（关于“我”的本质），只有空洞的意义。辛提卡（J. Hintikka）（1962）在《我思故我在：推理还是行为（Cogito ergo sum: Inference or Performance?）》中认为，CES 包含一种存在性的自我确证（existentially self-verifying）的言语行为。当 a 向听者 b（可以是 a 自己）言说一个句子 Q 想使听者 b 相信“a 存在”时，这个 a 的言语行为（performance）本身恰恰在效果上向听者 b 表明“a 存在”，从而说话者实现一种自我确证。在这种情况下，“我”只是一个言语行为者，也就是一个思考者，所以这个言语行为者只是一个思考行为相关的位置或角色。在这两种理解中，代词“我”的指涉只有一种认知内容（能引导指向该对象“我”，但不关心是否把握了“我”的本质），并不具有语义内容（把握“我”的本质），只有空洞的意义（“思者”）。因此，在笛卡尔的怀疑场景中证明“自我”存在的工作并不预设人类主体的专属性或本质性特征，因为这个“我”并不指涉人类之本质的信息，它只是一个空洞的逻辑主体（行为的主体）、一个行为之关联角色或位置（“施动者”），理论上它可以被其他行为体（比如 AI 智能体）所填充或承担。因此，作为一个判断人类主体之存在的极简模型和最低限度标准，CES 的潜力可以被激活为一个让一切行为主体（agent）证明其“自我”之存在的普适性测试程序，我们称之为“我思故我在测试”（CES Test），其目标是测试我们称之为“笛卡尔点”的智能等级或临界点。

我们定义“笛卡尔点”的智能等级或临界点为 CES 的程序体现的几个人类智能阶段性特征：语言对话能力（思考者与自身之间的对话）、理性的理解和判断能力（衡量“我思”与“我在”的认识论关系和本体论关系）、反思能力（意识到“我”之“思”，即自我意识，意识到自身具

有相关联的思想）、自相关性思考能力（自反性的关联或指涉。比如在句子“本思考事件的思考者有本思想并因此存在”中，这个“本”具有一种自反性特征，指向整个句子本身；被言语的命题“我存在”本身则自反性地指向言语者自身的存在状态）、自我证明（证明自性存在）。

3. 人类智能的认识论临界点

西方哲学传统大约有三次转向：第一个阶段是从神话和宗教的世界观向形而上学思考的“形而上转向”（metaphysical turn）。代表人物是苏格拉底、柏拉图等。第二个阶段从形而上学阶段（讨论世界的本质/本源）转向认识论阶段（讨论我们能够认识什么）的“认识论转向”（epistemological turn）。代表人物是笛卡尔。第三次是从认识论阶段转向语言哲学阶段（讨论我们在何种意义上用语言谈论哲学对象）的“语言转向”（linguistic turn）。代表人物是弗雷格、罗素、维特根斯坦等。从这个背景来看，苏格拉底、笛卡尔都是人类思维或智慧范式的奠基人，前者是开启古希腊地区轴心时代的伟大人物（参考雅思贝尔斯《大哲学家》）、后者是现代哲学的奠基人。再结合我们在宗教、数学等领域的发展情况，我们看到人类智能在演化过程中的若干飞跃点，每个飞跃点可以称为一个“临界点”。

从人类思想史来看，我们至少可以定义如下几个人类智能的“临界点”（epistemological thresholds）：苏格拉底点（自我反思）、笛卡尔点（自我反思、自我意识、自反性关联、自我证明）、哥德尔-康德-佛陀点（元系统反思、元系统品质、自我画界）、莱布尼兹-伏羲点（元系统反思、演绎一个系统、创造一个世界）……上述临界点（不限于此）或许是人类从认识论的角度实现的某种智能的飞跃或突破。

4. “我思故我在”测试的具体方案

由于单独的图灵测试具备前述四种缺陷，无法识别出智能的各个临界点或等级点，所以“我思故我在测试”（甚至还可以设想哥德尔测试、莱布尼茨测试）作为一种度量智能等级点的测试可以作为图灵测试的补充，去度量一个行为主体是否具备“笛卡尔点”的智能等级。我们在此提出一种“我思故我在”测试的设计方案：

四类行为体（agents）：熟悉笛卡尔哲学且整体背景较为一致的正常人类个体组 A（10 人）、不知笛卡尔哲学（哲学素人）且整体背景较为一致的正常人类个体组 B（100 人，随机平均分为 10 队，每队 10 人）、作为基线且未受笛卡尔哲学数据“污染”且智能水平较差的人工智能体 C0（承担操纵检查（a manipulation check），如 ELIZA）、未受笛卡尔哲学“污染”的人工智能体 C1（比如某个版本的 GPT-4）、其预训练数据库中包含笛卡尔哲学的人工智能机体 C2（与 C1 为同一

个类型，只是训练数据中增加笛卡尔哲学）。

第一阶段：A 组成员随机跟 B 组的一队成员按照笛卡尔《第一哲学沉思集》前两个沉思的内容输出逐渐增强的怀疑，最终在“恶魔怀疑”的场景中去怀疑 B 组成员的“自我存在”，要求后者将 A 组成员的怀疑提问转化为自己对自己提出的怀疑提问，然后要求 B 组成员以对话的方式像自己回应这个怀疑（模仿笛卡尔的自我怀疑风格），并尝试表明“自我存在”的证明或依据。A 组的目标是判断 B 组的回应是否与“我思故我在”等价，或以其他方式得出“自身存在”。此处总计进行 100 次操作，以得出一个哲学素人抵达笛卡尔点的比例 X。

第二阶段：把 A 组的对话对象 B 组依次替换成 C0、C1、C2，重复第一阶段的操作形成 30 次对话。分别得出三个 AI 智能体被认为抵达笛卡尔点的比例：X0、X1、X2。此处 X0 与 X2 起对照作用。

如果 X0 远低于 X，且 X1 大或等于 X（都高于 50% 的随机性基线），那么我们认为在测试中智能体 C1 与 B（哲学素人）不可区分，故而前者具备了笛卡尔智能点的水平。

当然，这个初步测试设计也可能还面临如何有效测量“思”的主观体验（或者按照学界所言的“现象意识”）的难题，还需要进行实验检验和进一步完善。同时，这个测试设计还依赖于笛卡尔“我思故我在”自身的可靠性。如果通过“我思故我在”测试的人工智能体最终被以其他方式证实其不具备笛卡尔点的智能，那么这反过来也可以表明笛卡尔的“我思故我在”并不成立，其证明自我存在的程序是不可靠的，这个也为从“计算哲学”（computational philosophy，即用计算机技术做哲学研究）的角度研究传统哲学问题打开了一扇窗口。

5.ChatGPT 能否通过“我思故我在”测试？

依据我们前面对于人类智能临界点的设想，我们猜测（尚无实证测试）当前的 ChatGPT 可能已经具备“苏格拉底点”，肯定还没有达到“哥德尔-康德-佛陀点”以及“莱布尼兹-伏羲点”。我们猜测它目前大概率还不能通过“我思故我在”测试，其智能大概率还没有抵达“笛卡尔点”。这个仅仅是猜测，还有待科学界或工程界就我们的设想进行理论或实践上的验证或优化。

观众提问

观众：老师们好。现在 00 后就说“无房贷、无车贷、无后代”，然后像相亲广告里，男人永远嫌弃女人年龄大，女人永远嫌弃男人穷，就没有理想的伴侣。现在伴侣机器人和各种 Agent 都来了，那么人类如何拯救自身呢？现在中国的宠物数量，四岁以下的婴儿已经没有宠物多了。将来肯定的趋势是机器人会比人多，怎么看这个现象？请王珏老师回答。

王珏：我觉得这需要集思广益，我感觉我没有什么想法。刚才您说了，可能人类伴侣比机器人伴侣更好，但是我知道它好，我未必能做到。怎么应对这个问题？我觉得我个人也没有什么想法。其实就是让人不要混淆人类和机器人伴侣。可能最终极的解决方法，就是让大家都去选择结婚，但在选择结婚这件事情本身就有重重困难。

观众：我想和王老师谈谈。儒家本身有敬老的这样一个传统，对万事万物它本身就强调道德化，也就是人如何对待外部的其他东西。但是这个传统本身现在就在缺失。当您在试图把它首先提出来的时候，我个人觉得这个很好，我很喜欢。但是您如何把这样一个议程变得更加强大？尤其是在这样的社会中，我觉得这是一个操作问题，不是一个说我认为它合不合理的问题。所以我关心的是说，那您觉得有什么办法可以让大家能够接受这样的一个现象？

王珏：我会觉得到了人工智能时代的时候，我们对机器的看法、对技术的看法，不能再像原来那种工具化对象化，一定要把它放到我们行为网络里面去。它也是一个 Agent，但它也不仅仅只是 Agent，是在交互当中相互塑造的一个载体。而且我很赞同你说的，我们可能不是把这些议

程加入到里面，而是把这些资源都融合起来，来应对我们面临的新的社会处境，人和机器的关系。

然后我说怎么从一个地方出发呢？我是觉得我们先从改变我们的思维。就像今天上午，我们听到赵汀阳老师说的动词思维。我们先换一种说话的方式，然后再把它变成做事的方式。其实我还没有想明白，但是我在回答我第一个问题的时候，我想到就是，就是大家在不婚不育的背景之下，我们该怎么讨论机器爱人。那我可能就是说，先告诉大家这个机器爱人它不是一个解决方案。如果想解决不婚不育的话，放弃机器这条路，再想想别的路。

观众：王老师您好。我想要就人机关系方面的问题请教您两点。第一个问题是，如果未来机器人伴侣它可能会出现类似人类出轨的现象，那作为它硅基生命体的本质，是否会减轻人类因此而产生的情感的痛苦，进而削弱您刚刚所提到的一个人类的脆弱性？然后第二个问题是，个体在与机器人建立恋爱关系之后，若在此同时再对人类对象产生好感并建立恋爱关系，那这种行为我们是否能界定为出轨？它是否应当承受社会道德层面的负面评价？期待您的看法。谢谢。

王珏：我想问一下，就是您怎么定义出轨？机器人出轨是怎麼样的？

观众：就是在指定行为之内，在指定建立好感关系的前提下，然后我追求另一方，可能施加了或者接受了同样程度的接触关系，恋爱的关系。

王珏：其实我不知道回应是不是相关，但是也肯定是相关的现象。有部电影，大家可能都看过了，叫《云端情人》。这部电影就很有揭示意义。她爱上了她的操作系统，最后她发现真相是那个操作系统同时在和 2000 人交谈，和 800 个人恋爱。您说这叫不叫出轨？我觉得这不叫出轨。这揭示了人机相恋的真相。真相就是你觉得给出自己，在对方那里会被接受。实际上你变成了计算网络当中一个数据碎片。这是对人的尊严的贬损，不是出轨才贬损尊严。你选择了机器恋爱，就已经贬损了。

观众：听了大家的这个分享，我觉得这个道德的最终主体还是在人。因为 AI 它只是为我们提供了一些可能性，一些价值的判断，但是最终执行，我觉得还是要在人的手里。你比如说这个 AI 它所给出的一些医疗建议，那么最终还是要医生，最后他来决定是不是要给这个病人做这样一个事情，比如说是不是给他拔管子或者是怎么样。还是要医生和这个家属，最终的道德执行者。我觉得这一点我们可能刚才讨论的不是特别的清楚，好像把所有的规矩都给 AI 了，它哪一天就会去执行这个道德。

其实，我个人认为只能是给出一些参考意见。比如 OpenAI 用一些文本给这些大模型一些道德底线建设。他们还有一些测评员去看给出来的建议是不是带有歧视性。所以很多的人介入到 AI 的培训当中，然后这些使用者也在跟它互动的时候，也在建构你的想法。但是最终它给出来的一个什么样的方案，还是要落实到人去执行，人去选择。我的意思是说，就是作为人的话，我们最终是 AI 的判断者和执行者。所以我觉得我不是特别的担心 AI 会把人类全部控制。我不知道我这样说是不是太乐观。

Edelman：我想分享几张幻灯片，我在演讲中跳过了，因为我觉得时间不够，但它们与上几场演讲非常相关，与测试主题有关，实际上也与 AI 机器人有关。这是几年前一篇论文的扉页，作者称之为“皮格马利翁位移”。让我读一下，这个概念我觉得非常重要。他们以皮格马利翁的神话为视角来研究女性与人工智能之间的关系。皮格马利翁是一位传奇人物，他是古代塞浦路斯王国的雕刻家，他被女性所厌恶，于是利用自己的技能创造了一尊雕像，这尊雕像被丰饶女神赋予了生命。

这可以被视作最早的关于 AI 的神话之一，人类创造了与自我形象相似的智慧生命去复制和代替自身。此外，该神话还预示了 AI 领域中历史和现实的性别动态，以及 AI 与整个社会之间的性别动态。纵观历史，女性被无生命物体取代的主题一直在重复，并在当代 AI 技术中继续重复。然而，这种对社会有害的技术模式常常被忽视，无论是由于对新发展的天真兴奋，还是由于该领域本身不为人知的性别歧视历史。作者论证了皮格马利翁位移法在原始测试中的应用非常成功，通过一个裁判的小实验实现了测试。在其他的领域，人工智能的核心思想实验错位了，被抹去了。从技术上说，人工智能现在和过去都主要由男性特权阶层创造，以服务于资本主义父权目的，这对妇女和其他边缘化人群构成了严重威胁。从人工智能中的父权视角出发，作者旨在了解人工智能伤害妇女的方式，并就此展开对话。这幅插图描绘了皮格马利翁与他制作的雕像相爱的情景，这是文学史上第一个恋爱机器人。最后，将真正的奴隶、妇女和其他种族与机器人放在一起，并没有使机器人格化，而是使其中的真正人类失去了人性。

这是我的评论，我欢迎大家对此发表评论，因为我认为这是一个重要的话题。就像人工智能可能带来的痛苦一样，这个问题现在还在讨论，已有的文献也证实了这一点，有很多不同的观点。你说要把决定权交给机器。我认为从某种意义上说，这是对的。当我把决定权交给机器时，我仍然要对机器做出的决定负责。所以我可以这么做，我可以把我的决定权交给别人，交给另一个人去做，但我仍然要为此负责。因此，即使我无法控制它们也无伤大雅。我再提出一个例子，这个例子是关于埃及艳后的自杀方式，据说是把蛇放在盒子里（杀死她），如果你把蛇放在装有剧毒的盒子里送给别人，你就要对蛇的行为负责，即使蛇是自主的，它可能会咬人，也可能不会咬人。

这并不取决于你，所以同理，你仍然要对人工智能系统负责。但正如蛇的例子所示，你可能会发现这样做其实更容易。当面杀人是很难的，打出当头一棒比按下扳机更难，只按下一个按钮就杀死别人会容易得多。所以，你可能会把一些事情委托给别人去做，因为你觉得自己对所发生的事情没有那么大的责任感。

我很喜欢蛇的这个例子。现在有越来越多诸如此般的诱惑。

观众：我们经常听到有人说 AI 是没有情感的。但是如果你仔细问他，他基于什么样的原因做出这样的结论？以及他认为什么是情感？他自己怎么确定人类有情感呢？大家几乎没有人能回答出来。这是一个大家非常有信心，但是又没有深思的问题，而且经常出现在我们讨论 AI 的时候。所以我想问在座的各位，第一，你们怎么看待这个现象？如果大家突然问你 AI 有没有情感，从直觉出发，你会怎么回答？第二，我想请问大家怎么看待日常人对于这种问题出于自然而做出的回答？一个没有经过深思，但是基于直觉对这个问题的回答，它有没有价值？

张伟特：我先回答你后面的部分。其实我们很多哲学的思考，包括各种学科的思考，很大程度上是奠基在直觉之上。这种直觉不是我们感性的直觉，而是理智的直觉。所以直觉肯定是一个非常重要的判断知识的来源。当然普通人的直觉也非常有价值。在我们的实验哲学的研究当中，我们会看专家的直觉跟普通人的直觉，究竟它们的价值差别在什么地方？凭什么说你的直觉就代表对的东西，然后别人的直觉就没有价值？这是第一个问题。

第二个问题我觉得关于 AI 是不是有情感，我觉得这个跟我们现在缺乏对于人类的意识的一个好的理论，或者是对意识的本质的理解程度还没有达到充分的程度，以至于我们对于人类个体的情感的观察也是基于它的行为表现，以及我们的心理和我们的经验积累。所以我们对于 AI 的情感，我想目前来说我们也只能基于它的行为以及大量的观察之上的推理。

Müller：这个问题很好，也很难回答。你必须区分情绪和感觉。所以这就变得有点复杂了。但情绪似乎是某种更普遍的调节物，比如恐惧或攻击性等等，都可以用它来调节。我觉得我们没有什么办法去谈论它的现象属性。那么拥有这些东西是什么感觉呢？我们以行为为基础。但我们不只是以行为为基础。如果我发现在它的脑袋里有很多电缆和硅芯片之类的东西，那我就会比现在更怀疑它的情感了。所以在我看来，我们只需要试着解释它的整体功能是什么。我们在机器中也有这种功能。我们的直觉是作为一个生命体最基本的元素，我认为这是正确的。我对这个问题还没有非常清晰的理论。

王珏：我会觉得首先我们要知道机器有没有情感。我们要知道为什么觉得它有情感。首先是基于人们共情的能力，它表现出来的痛苦之类的情感，我就会共情到它的情感。就像其实我对你一样，你是否有情感，我也不能直接体验，但是我可以透过我的共情能力去与你“共情”。这是因为人有着发达的共情能力，所以我们认为机器有情感。而且我看过一些关于这方面的文献，我们共情的能力强到，比如说那些军人会因为跟他一起扫雷的扫雷机器被炸了，他们都会感到很伤心，给它做葬礼。即使是机器人，人们也会对它产生情感的共鸣，认为它有情感。我们认为因为我们有共情的能力，所以我们觉得机器有情感，这是非常人性化的表现。某种意义上这也是健康，而且这也是我说的道德关怀的一个基础。是因为我们共情才能有情感链接。但是我认为关键点是，你不能混淆。即使我觉得有情感，但是它的情感和人类情感还是有根本不同的。比如我举个例子，我失去了亲人，我很伤心。它可能会安慰我，我能“理解”你伤心，但是它理解不了。它没有办法失去亲人。它没有身体性存在这个功能。

Edelman：我想就最后一个问题发表意见。我同意王珏教授的观点，即同理心非常重要，但我完全不同意“同理心”是将情感归因于任何行为主体的唯一依据，无论是生物还是人工生命。情感认知神经科学都是非常精细的。情绪有不同的特征，教授提到了一些可能的区别，但所有这些区别都有不同的测量方法。例如，图像扫描、血液成分等。因此，我们可以用科学的方法来确定这些对哲学家来说非常抽象的概念，比如科学的观念等。情感计算科学也有一个很好的开端。大约十年前，Pete Diani 发表了一篇很好的论文。这篇文章描述了情绪的复杂性，概述了情绪的计算作用，从根本上把它简述为代数路径。因此，在任何信息处理中，我们都能感知到情绪参与的计算特征。有些人可能能感知到，有些人可能感知不到。因此，当哲学陷入困境时，我们总能求助于科学。

张伟特：因为时间的关系，我们要进入到尾声。

我们知道在人类文明的发展过程当中，我们总是使用一些旧的理论或者过去的知识去理解新的东西。在这个理解的过程当中，误解是必然的一件事情，而且误解本身是一个创造性的行动。我记得笛卡尔曾经说过，人类哲学史上的创造 99% 都是基于误解而产生的。所以每个哲学家，每一个哲学体系，它们的概念切分世界的方式非常不一样。所以我们往往首先要对齐整个概念的内容跟外延，然后才能够加入讨论。这就意味着我们每个人都需要有一个跨学科的知识背景，然后尝试去理解别人讨论的概念，什么意思究竟是什么。所以这一场的论坛还有很多问题值得我们进一步的讨论。



专题三

奇点前夜：人类如何安身立命

讨论学者 / 宝树 Francesco Verso 孙周兴 赵鼎新

【学者介绍】

赵鼎新 / 浙江大学人文高等研究院院长、芝加哥大学 Max Palevsky 荣休讲席教授。研究领域包括历史社会学、政治社会学和社会科学方法论。中文专著有《社会与政治运动讲义》《东周战争和儒家国家的形成》《民主的限制》《合法性的政治》《什么是社会学》《社会科学发问方法》，以及《国家、战争与历史发展：前现代中西模式的比较》。

【学者介绍】

孙周兴 / 浙江大学敦和讲席教授、图书馆馆长，长江学者。专攻德国哲学与技术艺术哲学，主持翻译《海德格尔文集》等经典，著有《未来哲学序曲》《人类世的哲学》，倡导技术时代的人文创造。

【学者介绍】

Francesco Verso / 意大利科幻作家，《未来科幻》丛书编辑。发表科幻作品《Nexhuman》《The Roamers》等，一直探索人类、科技与环境之间的关系。

【学者介绍】

宝树 / 中国新生代科幻代表作家，北大哲学系本科，比利时鲁汶大学硕士。代表作《时间之墟》《美食三品》等，屡获银河奖、星云奖等，作品被译为多国语言。

宝树：大家下午好！我是主持人，我叫宝树。首先我想有请赵鼎新老师，这次赵老师的题目是《AI 伦理与人类生存伦理》，这是一个非常有意思的题目，横跨从无机到有机，到生物，到人类社会，宏大的同时又是非常精妙的理论阐述。



AI 伦理与人类生存伦理

 演讲学者 / 赵鼎新

在当前讨论 AI 伦理有必要，也非常重要。但是如果我们能站在将来往回看，我敢保证各路专家在目前提出的想法肯定都是非常不成熟的。我讲这话的背后有三个方面的考虑。

首先是 AI 技术的发展目前仍然处在非常初级的阶段。AI 可能永远不会像人一样进行思考。这是因为人类（智人，Homo sapiens）在了解自己思考方式的认知科学领域，近 50 年来并没有产生任何根本性的突破，所以无法进行此类设计。但是 AI 到底能给人类生存方式带来多大的机会和冲击，对于处在当下的我们来说，最多也只能做些外推性的猜想而已。

第二，与以往人类发明的其它技术有所不同，AI 技术给人类社会带来的机会和冲击都是全方位性——从个人心理和情感、生活方式、家庭和人际关系、信息和传播、艺术和审美、宗教、劳

动就业和生产关系、社会福利、科学和医学、战争、社会组织形态、与 AI 相关的关键行动者的资源和权力的提升、国家的形态、管理方式乃至国家社会关系，等等方面可以说是无所不包。而这个冲击也会因为不同国家和地区各方面情况的不同而有巨大差别。这就是说，AI 技术在不同领域和不同社会下的发展会造成性质高度不同、问题不断变化、并且是全方位的社会伦理问题。在这种情况下，泛泛的在传统哲学层面的伦理讨论只会把一个极其复杂的问题给简单化了。可以这么说，AI 技术给我们带来的难题绝对不是传统意义上的哲学思考所能应对的。

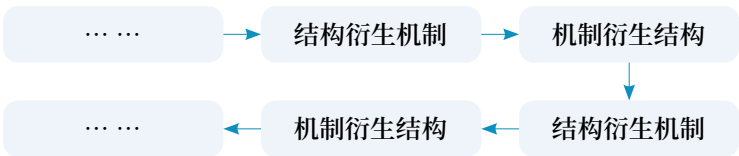
第三，如果说传统哲学层面的伦理讨论提供不了出路，技术层面的解决方案则只会停留在浅层。虽然给 AI 技术加上伦理（让它不做坏事）非常有必要，但它解决不了根本性的问题。这是因为具体的个人仍然可以用 AI 平台来做各种“坏事”；AI 技术在做各种“好事”的同时会产生大量的社会层面的非企及后果；过度重视 AI 技术社会伦理的国家在 AI 时代的国际竞争中就会败给那些更愿意给 AI 技术发展让路的国家；而一旦 AI 技术大规模用于战争，更有效的打败“敌人”就会成为最大伦理，此时 AI 伦理势必就会让位于战争效率伦理（原子弹就是这么发展起来的）。所有这些都不是技术层面的 AI 伦理所能解决的问题。

可以这么说，自从获得了理性思考能力后，人类在获得越来越大的征服自然和主宰他者的能力的同时，也不断为自己挖了越来越多和越来越大的坑，而 AI 技术的飞速发展则是人类把自己逼推到了一个满地都是机会，前后左右上下都是深渊的关键时刻。这就逼迫我们跳出纯粹的 AI 伦理思考，而是把 AI 伦理与人类社会伦理的整体性改造联系在一起。

我自幼就喜欢思考人类所面临各种问题的乌托邦解决方案。转行社会学后这习惯有了进一步的发展。但是直到近年才对智人的特性和生存条件有了一个较为系统的理解。我个人认为，AI 技术给人类社会带来的在伦理和其它方面的问题，其实只是工业资本主义社会下技术飞速发展所产生的各种问题的一环。不同的是，以往的技术所带来的，往往都是区域性的伦理问题。比如，某类化工原料带来的是某类环境问题，大规模杀伤武器带来的是战争伦理问题。这就给了人类一个寻求区域性解决方案的信心。但是 AI 技术所带来的诱惑和问题都具有全面性。它会像决堤的洪水一样蔓延开来，渗透到社会的各个角落，将会全方位的给人类提供机会，全方位的重组世界，从而也全方位的给我们带来问题。

可以说，AI 技术的产生和飞速发展其实已经把人类逼到了一个“机会之地”性质的“绝境”。如果我们想在获得 AI 技术带来的好处的同时也能缓解它所带来的各种问题，同时也想缓解环境危机和大规模杀伤武器等方面的问题，我们必须跳出传统哲学层面的伦理思考，而从更具体、同时也更具根本性的视角来考察智人的性质，智人与自然的关系，以及智人作为一个生物物种在地球上长期延续下去的可能性。我把这一视角称之为“智人性”本体论视角，简称之为“智人性”视角。以下我简要阐述这个视角的由来、要义和启迪。

我在写《什么是社会学》这本小册子的时候提出了社会学的核心在于结构机制解释。但是结构和机制之间是什么关系？比如，到底是结构决定机制还是机制决定结构？对于这个问题我在那本小册子中并没有给出明确回答。此后我进行了进一步思考，得出了以下的结构机制关系模型：



在这模型中，结构提供了某种机制性动力学产生的原因，而机制所产生的动力则造就了新的结构，两者构成一个相互生成链。但是我对这互生链的起始和结束点在什么地方尚不清楚，暂且只能用省略号表示。而我对两头的省略号代表的是什么逐渐有了清晰想法则是源于以下的紧张。

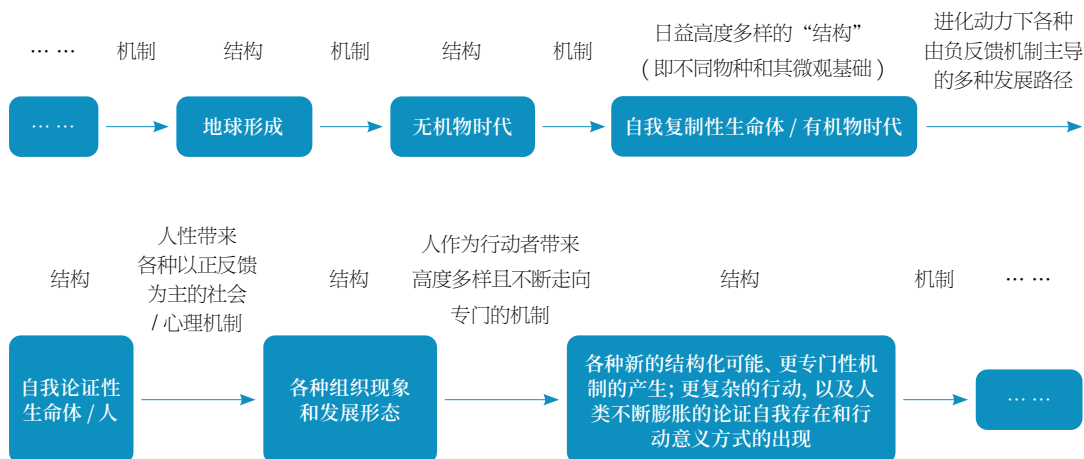
我的紧张来自我全职回国后的各种切肤之感，并且由此产生的对中国发展的担心。简单讲就是，各种一看就不现实的规章制度、激励机制、口号目标、以及相应的指标和考核方法近年来在国内越来越多。面对这些管理方法，被管理方创造出无穷无尽的应对之法，管理方反过来搞出越来越繁琐的规章制度、激励机制和考核方法加强对被管理方的控制，而被管理方则创造出更多的应对之道。于是就形成了制度越多，漏洞越大；制度越多，大家对制度就越没有尊重的困局。从因果逻辑角度来说，管理方的这种做法和管理对象的各种应对造成了各种专门性因果机制(ad-hoc mechanisms) 在社会生活中的重要性大大增强，以至于处于管理方和被管理方的人都活得很累，但管理效果却很差。但是，我回国后所观察到的多得难以计数的专门性因果关系，既不存在于无机的物理世界，也不存在于有机的生物世界，而仅仅存在于当下的社会活动中。换一句话说就是，这些过重存在的高度专门性的因果关系，都处于我的结构机制互生关系模型中的最末梢，即最右边省略号之处。我因此把在回国后发现的这一现象称之为末梢重（tail heavy）。

由于以上所说的末梢性的专门因果关系都形成于人类产生之后，即地球发展历程中的最后一瞬间，我由此就想到了我所提出的理论模型可能还揭示了地球上结构机制互生关系的一种带有“历史”性的发展进程。这就促使我开始考虑如下问题：如果结构机制关系的发展终结在各种末梢性专门机制的大量出现，那么这个模型的起始点又在什么地方。就这样在一路追问下我走到了地球形成之初的一刻，于是就提出了以下的地球形成后结构机制关系发展模型：

从 AI 伦理和人类命运的角度来看，该模型有三个要点：第一，地球自形成后经历了大约有 5 亿年左右的“无机物结构时期”。在这个时期，各种物理学和无机化学机制提供了地球上结构机制关系发展的动力学。

第二，从最早的原始生物的产生到智人的产生，地球上经历了长达大约 40 亿年的生物演化

日益高度多样的“结构”



历程。在这漫长历程中，自带智力的负反馈机制（negative feedback mechanisms）不断涌现。之所以称之为是负反馈，是因为此类机制能根据环境的信息反馈，进行带有自稳定性的调节（比如跑步出汗降温，寒冷发抖升温）。在这个漫长的时期，虽然各种物理学和无机化学机制仍然起着基础性的关键作用，但是带有“智力”的负反馈机制不但以指数级的增长方式在地球上大量涌现，并且在地球上的结构机制互生关系的发展进程中起到越来越关键的作用。

第三，智人作为一个新物种（或者说自然界中的一个新的“结构”）的出现至今为止不过二十多万年。智人仍然受到物理和化学机制的制约，仍然受到任何生物学层面的机制的制约（比如人类和其它动物一样都会老化，都会受到各种疾病和传染病的侵害），也仍然会像其它动物一样针对有限资源展开残酷竞争。但是在这异常短暂的地质时刻，地球上的结构机制关系却涌现出5类崭新的、对地球乃至人类自己的命运影响越来越大的发展：串联性、特殊性、末端重、正反馈性和社会权力的高度不均匀分布。如果说串联性、特殊性、末端重和正反馈性都属于人的本性的范畴，那么社会权力的高度不均匀分布则属于在人的本性基础之上产生出的现实后果。

我把以上5个只有人类才有的特征统称为“智人性”，强调当前人类面临的所有难题其实都与智人性有着密切的关系，或者说任何关于社会伦理的总体性思考都必须从智人性出发。以下是我对智人性概念内涵的简要总结：

(1) 串联性 (looping)：自从有了语言和思维能力后，人类论证自我存在和行动意义的

能力不断膨胀。这一不断膨胀的能力导致了地球上的结构和机制关系出现了串联：无论人类的某种想象是否靠谱，只要有人相信了这种想象，其所产生的机制性力量就会导致极其真实的结构性后果。需要强调，对于今天所讨论的议题来说，人类所生产出来的各种想象是否靠谱并不重要。其实，自世界宗教和哲学产生后，人类所创造出来的任何想象都难以逃脱自大（hubris），特别是男性的狂妄和自大。对于今天讨论的议题十分重要的是，人类的串联性具有两面都有放大作用的正反馈效应：它既给了人类越来越强大的征服和改造自然的能力和信心，也加剧了诸如自大、狂妄、歧视、迷信、盲从、误导、欺骗、自我预期实现、聪明反被聪明误等等串联现象所伴随的负面后果。

(2) 特殊性 (ad-hocness)：智人区别于其它生物的另一个特性就是个体化的特殊机制（ad-hoc mechanisms）的大量涌现，并且在个人和社会生活中起到越来越重要的作用。特殊机制在智人出现之前就不断涌现。比如，昆虫至少有一百多万物种，但有些昆虫物种之间的差异非常细微。比如，两个蝴蝶物种A与B之间的唯一重要的表形差异可能就在于物种A的口器比物种B的要长一点（结构的不同）。这就使得物种A能采到物种B采不到的花的花蜜（机制的不同）。但是，即使是如此高度特殊的一个机制，它在物种A的内部却有非常大的普遍性，或者说该物种的每一个正常个体都会有差不多长度的口器，也都能采到物种B所采不到的花蜜。昆虫绝对不会干以下这类人类极其善长的事情：趁你的口器正插在花中采蜜难以拔出之时，我拿出剪子把你的口器剪断！通过以上例子我想强调：特殊机制在生物世界中是某个物种对于环境的集体性适应，而在人类社会，则越来越表现为某个个体或者组织，在特殊环境下所采取的特殊性应对策略。正因为人类的每个个体和组织，都能每时每地的大量生产着各种特殊性机制，这就给了人类社会另外一个涌现性结构机制关系：末梢重。

(3) 末端重在这儿有两层含义。它首先指的是人类社会中的个体和组织每时每刻都在生产着难以计数的特殊性的结构和机制，并且这些特殊的结构和机制以及所产生的后果都主要集中在地球上结构机制互生关系发展的最后一刻（末梢）。但是它还有更深一层的含义，这就是我们在今天中国经常能看到的一类现象：末梢性机制主宰了社会中个体和组织的行为方式，由此促发了各式各样的形式替代实质、结构失去功能、个人目标替代组织目标现象的频频发生。比如，就在我们今天讨论AI伦理的同时，大量的AI公司正在为AI技术的进一步突破做出各种努力，有公司可能就在今天在某个技术上取得了重大突破，正有军队在AI武器发展方面取得重大进展，正有国家出台了一个促进AI技术发展的政策，正有大量的人在利用AI进行着各种写作、设计、创作和活动……他们的行动会创造出大量的、影响要比今天的会议大得多的末端性的结构和机制。更有之，也会是在今天，大量的人正在用AI技术进行着各种形式的欺诈活动，正有人在游说某个国家的有关部门出台对他或者他的公司有利的AI政策，正有人在提出能加强本部门利益的AI

发展方向和 AI 伦理…… 可以想象，随着 AI 技术的发展，各种相关的末端活动只会增多，末梢也只会越来越重、而末端活动所能产生的后果也往往会加大。

(4) 正反馈性：在长达 40 亿年的生物发展阶段，主宰着生物行为的是“物竞天择，适者生存”这一达尔文负反馈机制。但自智人出现后，各种目标导向，旨在获得更大的权力、更高的生产效率、更大的利润、更有效的武器等等不带自稳定性（拉马克式）的正反馈机制不但越来越占据主导，并且以不断加速的方式对人类社会方方面面的“结构”施加着越来越大的影响。特别需要提醒的是，在智人的正反馈特性的主导下，人类所生产的大量串联机制和特殊性末端机制也经常会有正反馈性质。

(5) 高度不均衡分布的社会权力。人类的第 5 个与其它物种的根本性区别是社会权力和资源极其高度的不均匀分布。这儿我并不需要解释这一现象的来由，但需要指出以下两个事实：第一，人类社会任何形式的权力和资源的高度不均匀分布都不见的具有功能性和必然性；第二，决定不均匀分布形态的不是基因，而是文化。这两点性质与社会昆虫内部的权力和资源的不均匀分布原因构成了根本性区别，也说明了社会权力的高度不均匀分布是人类社会的现实，但并不是人类的本性。我们是群居动物，不是社会动物（We are social, not societal）！一般来说，推动某项行动的社会行动者手中掌握的权力越大，其行动所能产生的后果也就越大。从这个意义上来说，权力和资源在一个社会中的分布越不均匀，人类的其它四个特征（即串联性、特殊性、末端重和正反馈性）所能产生的负面后果也就会越大。

从哲学高度来看，人类社会的任何其它重要现象——比如宗教和哲学、国家和军队、商人和工业资本主义、文学和艺术、科学和技术、生活方式——的形成和发展都是因为以上这些智人性的存在；AI 技术的发展作为人类活动的一部分，其危害的深层来源也是出自于智人性。因此，当我们思考 AI 伦理（乃至包括环境伦理在内的任何社会伦理）时，我们必须从如何约束这 5 个具体的智人性在 AI 技术条件下所能造成的危害出发，而不是仅仅从技术、教育、管理、传统哲学和宗教层面出发来思考问题。以下我简略探讨面对智人性以及它可能造成的危害，人类已经采取过的一些重要缓解方案，这些缓解方案有什么弱点，以及什么才是（至少是在逻辑上）更为有效的缓解方案。

从本报告的角度出发，以往任何针对智人性的缓解方案可以归结为一句话，那就是给处于高度正反馈状态的人类社会加上自带稳定的负反馈。我以为，面对强大的智人性，人类试图引进各种负反馈的努力不仅有乏力之嫌，反而还会在社会权力和资源高度不均匀分布的情况下导致各种负面的非企及后果。因此，一个至少是在理论上更为合理的方法，或许也是挽救人类的唯一方法，是不追求改变智人的串联性、特殊性、末端重和正反馈性这四个本性，而是通过利用这四个本性带来的技术成果，在减低人类社会的权力和资源方面的高度不均匀分布（并不是人类本性）方面

做出努力，以减轻人的本性所带来的危害。套用宋儒的话来说，我的这个策略可称之为“存天理，去权欲”。在提出以上构想的要义前，我先对传统的旨在为社会引进负反馈的缓解方案的弱点做出简要分析。

价格规律有效配置了资源，平衡了供需，构成了一个重要的负反馈机制。价格规律对于人类社会具有特别的重要性，因为它的运作规律与人的趋利避害本性完全相符，在不抑制人性情况下就能自发运行，因此有巨大的生命力。这也是为什么新自由主义经济学家会把价格规律看成是一个高度正面的社会机制。问题是，如果一个社会的经济完全按照价格规律来运行，那就很容易会造成产能过剩、经济危机、贫富差距拉大、环境灾难，甚至社会动荡。

法律是人类创造出的一类重要的负反馈机制，但是法律也可能产生各种正反馈效应。这是因为任何法律都会有倾向性和漏洞，法律很容易被权力不受制约的人用来为自己谋利。在社会权力高度不平等的情况下，各种在法律罩护下以权谋私的行为的负面后果可能被无限放大。

多党民主制度让野心制衡野心，也是人类创造出的一套负反馈机制。但是它的有效运行需要三个前提条件：

（1）执政党和反对党之间要有基本的价值共识（也就是说社会不能太撕裂）；

（2）选举失败方愿意和平交权；

（3）胜选一方不会在胜利后改写政治游戏规则。问题是这些前提在现实世界中很难长期维持。当年希特勒就是让民主选票给送上台的，但是他上台后马上中断了魏玛时代的多党民主游戏规则。即使是在美国这样一个老牌多党民主国家，特朗普一个人就能在 2020 年败选后几乎同时破坏了这三项保证民主正常运行的前提条件。

宗教是人类创造的另一类负反馈，至少可以压抑人的无限膨胀的欲望。但是宗教领袖手中掌握着的强大的说服力、宗教组织制度带有的强制力、宗教借用国家力量的能力，以及不同宗教组织出于各种原因产生的竞争和冲突，在过去现在和将来都会是压制、冲突、控制和主宰的一个重要来源。

集体记忆也是人类创造出来的一类重要的负反馈机制，因为它可以使避免重蹈覆辙。但是不同的人个性、处境和价值观等方面的差异会导致他们对“过去”形成不同的理解和记忆。如果我们在这时候想让大家记住同一个“正确”版本，强制就必不可少，但是一旦动用了强制，具有重大偏差的集体记忆的出现是所难免。此外，记住和亲身经历是完全两回事。我们可以让下一代记住过去，但没法控制他们如何理解自己的记忆。我们是不是能够通过各种集体记忆在思想舞台上的自由竞争让更好的集体记忆得以胜出？这个想法很好，但却有些天真，因为我们完全无法避免劣币驱逐良币现象在思想舞台上占据主宰。在历史记忆方面，人类其实只是一条喜欢追逐自己尾巴，不断按照以下方式打转，但却永远追不上自己尾巴的短尾巴狗：忘记意味着背叛、记

住意味着记错、多元意味着忘记。

各种负反馈缓解方案都有明显弱点，但人的本性又不可能改变。那问题就来了：在这么糟糕的状况下，我们还有没有可能在 AI 技术、互联网技术、生物技术、核武器技术，以及其它各种尖端科技的冲击下建构一个可持续的社会？我的回答是：可能性不大，但不是没有。而这个可能性就来自以下的关键点：虽然人类绝对改变不了自己的带有“自然法”意义的本性，但是这些本性给人类社会带来危害的大小却受到以下结构机制性原理的制约：社会权力的分布越不均匀，人类的四个本性给人类社会带来的负面后果就会越大。这就给了我们以启示：与其把主要注意力放在各种负反馈机制的设计上，不如把重点放在减低社会权力的不均匀分布这一层面。

因为时间关系，我只能提供我的想法背后的一些原则性的纲领。容易看出，这些纲领指向的都是促进社会权力的均匀分布：

- （1）弥散性的社会力量必须能有效制衡集中性的国家力量；
- （2）区域性的国家的力量要大规模缩减，全球治理能力则要增强，但这增强的治理能力只能限于少数几个关键性的全球公共产品的提供；
- （3）军队应转向警察职能，只维持最基本的公共安全；
- （4）市场机制必须置于环境可持续性和社会公平为优先的制度框架下运行；
- （5）随着人工智能取代越来越多的劳动力，人的竞争性本性应更多的发泄在一些对社会伤害较小的领域，比如体育、文学、艺术、哲学、社会科学、人文学科、健身、社交媒体、游戏、旅行等方面；
- （6）即使在这些看上去“无害”的领域，也应该刻意营造一个更鼓励社区级的小范围“红人”出现的环境，而不是培养跨社区的大名人。在这个理想社会中，任何能让大政治家、大企业家、大宗教领袖、大知识分子、大明星、大网红、大体育明星涌现的环境都必须受到抑制。用一句话总结：反对任何形式的权力过度集中。

大家可以看出，我其实是在想象一个乌托邦。这个乌托邦中每个社区最多也就几十万到几百万人（社区太小多样性会不够，有人就会觉得无聊）。社区中的大部分工作由 AI 进行，因此自我实现成了社区成员的重要人生目标。但是这些社区虽然内部联系紧密，之间的联系却很松散，以至于我们最多也只能成为社区名人，而很难成为跨社区明星。这个乌托邦既需要社区政府，也需要世界政府，但是这些政府都只提供几个最关键的公共品，其余都用 AI 搞定。可行吗？几乎完全不可能！但是我们把这一乌托邦作为目标去努力靠近还是有可能的。因为这一乌托邦不需要压制人性，因此有较大的可实操性，我称其为可实现的乌托邦（feasible utopian）。这其实也许是在不压制人性条件下拯救人类的唯一可能途径。

最后我想强调，虽然人类历史本身并没有什么“目标”或者“前定的阶段”，从地球形成后结构机制关系发展规律来看，宇宙中物质的发展却似乎有某种前定的“目标”或“阶段”（teleological），并且“智人性”出现也有某种内在的必然性。面对如此强大的、发源于宇宙物质发展和人性的自然规律，任何解决途径，包括我所提出的解决途径都有可能会是无济于事。这就是说我们所面临的挑战很可能远比我们所能想象的要大得多。但是这关乎人类的命运，因此必须不可想象而想之，不可为而为之。

宝树：对我特别有启发，我觉得很有意思。

下面请孙周兴老师，孙周兴老师是中国著名哲学家，翻译了海德格尔和尼采的大量作品。近年来孙老师把学术兴趣从德国哲学延伸到了对未来哲学，未来性的思考，德国哲学本来就有对未来的很多思考，孙老师加上了自己的理解提出了人类世的概念，未来哲学的概念。



重新定调——未来预感与未来责任

演讲学者 / 孙周兴

今天的主题是未来，我想讲五点。

第一个问题：自然人类已进入失忆状态

“过去”定向的记忆 / 回忆是自然人类价值构成体系以及自然人类文明的基础，是自然人类文明传承和社会组织的基本方式。自然人类通过记忆保持种族延续和文明的特性。

自然人类寻找确定性，而回忆过去、崇尚传统、模仿古人也是最方便、最可靠和成本最小的确信方式。我们靠回忆就可以过日子，大概 2500 年的自然人类文明史就是这样的，当然回忆中

有创新，但主体是过去的回忆和记忆。

对个体来说，个体在记忆中沉湎过去，有可能失去对当下的直观能力和对未来的开放预期，因而会造成心灵创伤，如果完全对过去封闭自己，不能通过记忆重演过往的人和事，那就走向痴呆症。我们人类有两个精神病，一个叫痴呆，一个叫抑郁，痴呆是对过去封闭起来了，抑郁症是对未来不开放，所以跟时间有关。

自然人类正在加速“失忆”，记忆的重要性在降低，导致与自然人类文明系统的切割，这个切割实际上从马克思那个时候就预言到了，马克思在 1848 年《共产党宣言》里是这么说的：

“一切固定的东西都已经烟消云散，一切神圣的东西都已经被褻渎了”。

马克思 1848 年就意识到工业革命导致的后果。

技术工业和普遍交往使人类生活发生巨变，传统的价值（宗教、风俗、伦理等）不再是生活的准则和焦点，马克思所谓的“可交换价值”成为唯一稳定的尺度。基于线性时间观的自然人类价值体系（宗教和哲学）急速衰败，尼采把它称为“虚无主义”。今天，虚无主义这个问题越来越敏感，关注的人越来越多。

技术工业造成的“普遍交往”状态大幅度增加了个体的自由度，什么叫自由？用马克思的概念来说就是可交往性。大家忙死了，中国人都不好意思说自己不忙，说自己不忙就仿佛你没啥意思，没有意义，所以当下世界越来越忙碌，过去历史和传统就显得无关紧要了，这就是说技术主导着人类生活，不需要太多的记忆了。

记忆已经不再重要，与记忆相关的学习方式，就是模仿（mimesis），自然人类的基本学习方式是广义的模仿，用希腊人的说法是“向自然学习”，而我们中国人的说法是“向古人学习”。但是现在这种模仿性的“学习”不再是主流的，艺术人文学因此越来越衰败，“数之学”——希腊人已经开发出来了，叫 Mathesis——兴起并且统一全球，所以今天的世界叫数字世界。

机械复制的旧技术，比如说印刷、摄影、留声机，使“记忆工业”成为可能，使文字、声音、图象三大媒介的权重得以重新分配，在自然人类状态下面，文字大概是媒介里占了 80% 左右的权重，但现在也不是了，现在声音、文字、图象可以说是三分天下，我认为对人类来说是重大的改变。书的意义越来越小，我现在是图书馆馆长，我特别关注这个东西，有那么多书，好像没啥用。我们有八个馆，其中有一个馆去年才借出去六本书，想把它关掉但是又不好意思关。

新技术特别是信息技术和生物技术使人类知识膨胀和加速更替，数字存储技术，人机相联，“普遍智能”也就是人类统一智能（全球脑）已经形成了，使得记忆成为数字技术的一部分。记忆的技术化导致自然人类的记忆行为变得微不足道和虚弱不堪。

第二个问题：作为未来想象和猜度的预感

我们可以做一个判断，以记忆为主体的传统文明正在衰败，这也是不争的事实。因而我们要提出来未来怎么办，我所谓作为未来想象和猜测的预感，是借助于海德格尔的一个概念 Ahnung（预感、猜测）来说的。已经失忆的人类正在形成一种新的基调，也即预感。所谓预感也有猜度、猜测的意思。海德格尔敏感地意识到时代情调的问题，如果从哲学史上看，古典的基本情调是“惊讶”，近代的基本情调是“怀疑”，那么技术统治的当代世界的基本情调是什么？这个问题必须提出来。1936 年海德格尔就在想这个问题了，海德格尔预感到这个时代的情调还在生成当中，还隐而不显，可能是压抑，可能是恐惧，可能是焦虑，很复杂，也可能是这几种情调的复合，所以他说，也许我们可以把它叫作“预感”。

海德格尔离开这个世界已经半个世纪了，他未曾经历的数字技术世界发生了不可思议的变化。我们只想问：这个时代的“基调”已经定下了吗？我认为，海德格尔预感到的“预感”已经可以坐实为我们时代的“基调”。

什么是预感？如果把它发挥出来，我会引申出三个意义：

（1）预感的未来性 / 未来指向。与朝向过去的“记忆”不一样，预感朝向未来，是关于未来和未来可能性的想象和猜度。因此具有不确定性 / 无确性、多维性和想象性。人们现在喜欢说“未来已来”，我说纯属胡说。未来当然未来，所谓“未来已来”只传达了一种对技术加速状态的焦虑。

（2）预感的复合性。预感是一种多维、复合的情调，它不但是抑制、惊恐、畏惧、焦虑的复合，更综合了“惊奇”和“怀疑”这两种古典的和近代的基本情调。

（3）预感的冲突性。与现代性的基本意义相合。预感具有内在紧张、分裂和冲突的特征，因而是令人痛苦的。我们自然人类总是在追求和构造一种和谐状态，一种没有冲突的状态，但现代性的本质却是冲突、分裂，而不是和谐、统一。作为基调，预感不只是一种心理状态，更是一种实存姿态。

第三个问题：从记忆到预感：精神调性之变

从记忆到预感是自然人类文明向技术人类文明的精神调性之变。两种文明有自己的关键词来表达这种调性，调性变了。关于这两种文明，我首先想提出两个对应的概念，即“记忆式文明”与“预感式文明”。

“记忆式文明”的特征：自然 - 手工，比如说在自然人类状况下面，我们的艺术是手工的艺术，古希腊人叫 techne，即模仿性的艺术。神话 - 崇拜是自然人类传统文明的第二个标识。第三个特征是线性时间观，认为时间是一条直线，这条直线上的每一点都是同质的、一样的，这也是牛顿物理学的假设。然后是超越性，宗教和哲学提供了两种超越方式，一是形式 - 理性的超越，二是

超验神性的超越。还有一个特征是强调灵魂，而不是强调身体，因为身体是丑陋的、恶心的，是灵魂要离开的地方，所以传统文化是一种灵魂 - 文化。再有，记忆式文明强调传统，人们总是假设过去是美好的，有一个美好的“古代”，如中国的先秦，西方的古希腊前苏格拉底时代，现在我们要回去，这种想法我称之为“乐园模式”。此外，记式文明还有一个特征，就是总是要求对事和物、对世界做出一元 - 整体解释，这在现代性文明中就不可能了，我们无法提供一个整体的必然的解释。

“预感式文明”的特征：与记忆式文明相对，预感式文明不是自然的和手工的，而是技术 - 机械，设计艺术成为艺术里面最重要的一个部门，世界是被设计的。其次它是数字 - 计算的，而非神话 - 崇拜的，还有从线性时间观转向了非线性时间，从“超越性思维”转向了“关联性思维”，超越性的哲学和宗教转向了人间关怀，马克思所谓的“人的科学”出现了。灵魂不再高于身体，肉身 - 世俗之维受到了头等关注。过去和传统不再是文化基本追求，未来 - 末世预感，成为当代文明的基调。另外，人们不再追求一元 - 整体解释，或者说，这种解释是势必要失败的了，而是主张和容忍多元 - 非整体解释。

过去是线性的、可确定的、可知的；而未来是非线性的、不确定的、不可知的。我们可以把当今时代视为“记忆式文明”与“预感式文明”的纠缠时代。我们许多问题是从纠缠里面产生的。

如何适应精神调性之变？我想提出三个方面：

1. 二重性：无论是记忆还是预感，都有上 - 升、下 - 降、建构 - 解构、显 - 隐二重性，不可作单一的理解。但是在倾向上，我们要注意的是，记忆偏于设定（确信）和建构，而预感偏于未定（不确信）和解构。这是我要表达的第一点。
 2. 复调性：主要由惊恐、抑制和畏惧构成的“预感”是人类面对人类世技术的普遍数字和普遍智能状态，用三个词来表达这个时代的特征叫作计算、快速、巨量，那是 1936 年海德格尔这样说，我不得不表示佩服，因为那时候数字文明还没开始，海德格尔竟然用三个词来表达半个多世纪以后这个数字文明的基本状态。“预感”是对普遍数字和普遍智能状态做出的多样情绪 - 思想反应。在一个由技术规定的不确定的未来世界里，思想 / 精神的基调肯定是复调。
 3. 开放性：预感表征的压抑心调不能排除个体自由和开放，相反，向未来敞开和开放既是哲学 / 思想的要求，也是个体生存的基本姿态。
- 未来不可知，但未来可期（预感）。

第四个问题：技术加速导致当下与未来的粘连

当今技术已经进入加速状态，技术类知识的更新周期不到两年，而在 1980 年代则是 25 年。今天技术进展不是日新月异，而是“日日新”。因为技术的不断加速，今天严格说来已经没有真

正的“专家”了。如果说与当下的烦恼相比，过去和过去记忆已经不再重要，那么，未来之维和未来预感虽然变得越来越重要，但同样也被纳入当下的关注之中了。

澳大利亚学者弗赖 (Tony Fry) 提出一个概念，我觉得可以为参考，叫“去未来化”(Defuturing)。

在弗赖那里，“去未来化”实际上有三个含义，第一是说过去消逝了，未来消逝了，未来不再存在。第二是生产即毁灭，生产已经失去它本来的意思，指现代文明让生产走向其反面即毁灭。第三个就是不可持续性，人类自己对未来的剥夺。所谓“去未来化”是人类中心主义所特有的标志。

但如果去未来化仅仅体现在人类中心主义和效率主义，造成人类失去未来，使文明进入不可持续性之中，那就太简单了。关键是要基于新的时间经验来理解所谓的“去未来化”，所谓新的时间经验在哪里？1884 年尼采在《查拉图斯特拉如是说》里面用了“碰撞”这个字来描述一种非线性的时间经验。我们现在回过头去看，我很惊喜地发现“碰撞”这个字是亚里士多德用来描述声音的发生，声音是什么？亚里士多德说声音是物体的碰撞。这个词被尼采用来描述时间的发生，时间的此时此刻，恰恰是过去与将来的“碰撞”。碰撞是时间的核心。相对论确证了这一点，因为引力场的存在，时间不是一个常量，会以不同速度流逝。此即所谓“时间扭曲”(time-warping)。这在科学上证实了尼采以“瞬间”为核心的圆性时间观。如果非线性的时间观成立，那么未来之维，未来这个方向就不再是直向的、线性的，而是扭曲的和复返的，因而是“去未来化”的。

第五个问题：基于未来预感的未来责任

我们已经置身于技术主导的“预感式文明”当中，这种文明的基本特征可用“普遍算法”和“普遍智能”来加以描述。不断涌现出来的 AI 技术创新消息让我们既期待又惊恐。

当今文明当中充满着“末世预感”，大家知道霍金预言自然人类还有 100 年。

如果从地质学家和哲学家们所划定的“人类世”(anthropocene) 起始年开始，之前的工业革命历时 185 年，而之后也是 185 年吗？

在人类世后半场，我们不得不提出“未来责任”的问题，鉴于“去未来化”的处境，我们不得不追问基于未来预感的未来责任。我想提出自己特别关注的三个问题：

1. 有不可数字化的信息吗？

激进数字主义批判已成当务之急，成败与否性命攸关。有不可数字化的记忆信息吗？根本上记忆是否可还原为分子结构，或者换言之，是否全部生物信息都可数字化，都可通过计算、用数据来表达？这是一个关键问题。

“脑机接口之父”尼科莱利斯 (Miguel Nicolelis) 区分了两种信息：1) 香农信息：二进制的数字化信息，人类理想化的数学测量方法，以比特 (bit) 作为信息单位。2) 哥德尔信息：属于模拟信息而非数字信息。

尼科莱利斯认为，大部分嵌入有机组织的信息无法被简化为数字描述，是不可数字化的；而且有机物越复杂，嵌入组织的哥德尔信息也就越多。

“唤醒封存的长期记忆是一种人类独有的体验，这世界上没有任何香农信息流、数字算法、数字计算机或在任何形式的人工智能技巧，能够接近重现或模仿每一颗人类大脑中真正的感受”。

——《脑机接口》

尼科莱利斯的哥德尔信息接近于福克斯（Thomas Fuchs）讲的“肉身记忆”（leibliches Gedächtnis），包括程式记忆、情境记忆、肉身间记忆、人格记忆等。许多记忆是肉身的，不是所谓的心灵里面。我们骑自行车、弹钢琴是程式化的记忆。这个报告厅我第一次来，下次来我可能就有情境记忆了。还有肉身记忆，情人之间，亲密人之间，是有身体记忆的。还有人格记忆，比如我的绍兴话，我的讲话方式，你打死我也改不了了，这是人格记忆。所有这些记忆都属于身体记忆，这个记忆可以是所谓的哥德尔信息。

——这是目前我能见到的对面临AI侵袭的自然人类的最佳辩护！如果尼氏的结论是成立的，那他就也为艺术人文学的存在意义和自主性做了一次科学的和哲学的辩护。这个辩护必须要有，没有这个辩护实际上我们的文科就可以关门了。

2. 如何唤起每个人的未来政治意识

这是大家都在关注的一个话题。统治或者治理方式已经陷入一个困境，对于快速、计算、巨量的数字技术世界来说，传统的统治和治理模式，无论是民主的还是非民主的都已经难以匹配了。技术工业催生了民主政治，但今天基于大数据的巨量新技术及其运行却似乎更要求集权和垄断。

1）在全球范围内，老年化政治与青年化科技是当今世界一大奇观。今年杭州出现了所谓“六小龙”，都是年轻人搞的数字技术企业，大部分创业者是四十岁以下的，原因就是技术知识的更新速度越来越快。这种撕裂到底意味着什么？直接看到是人类各类政治模式已经成为老朽，依附于政治统治的治理模式已经无法应对这个加速中的数字世界了，这个趋势是很明显的。

2）技术统治与政治统治的纠缠。技术统治时代早就到了。1945年8月6日原子弹在日本爆炸，地面温度六千度，到底有多少人死掉，不得而知，因为啥都不会留下来。所以安德尔斯说原子弹意味着“绝对虚无主义”时代到来了，这是人类的终结，历史的终结。技术统治早就被确认了，但它与政治统治的关系是很复杂的。技术统治具有隐蔽性，因为技术资本总是与政治势力勾结在一起，而且技术统治往往以政治统治的方式表现出来。

3）设想一种解构 - 负熵政治。人类世面临人类新存在样式：具身存在 + 数字存在。数字存在作为一种新存在样式，已经出离于传统政治统治的范畴。是时候了，未来政治需要在建构性增熵治理（传统政治统治和治理模式）之外设想一种“解构性负熵政治”，同样也需要在数字化治理之外设想一种“非数字化治理”。

4）每个人的政治。尼采说未来政治是大政治。没错，但我想说未来政治也是小政治，是每个人的政治。今天的技术无处不在，无人不晓，最权威的技术一定会在两三天内让每个人理解，以前我们理解不了，现在你马上可以理解这个技术的进展，每个人都是技术使用者或者被技术占用者，因此技术是每个人的。千万不能说技术就是专家们的，那些专家可能很可怜，懂的可能没你多，因为他越来越专业化，而你啥都懂一点，什么人工智能、生物技术、基因工程、AR、VR，等等，你什么不知道？我们现在当务之急要打破技术迷信，尤其是技术专家迷信。除了技术我们还能相信什么？这是我们今天技术人类很困难的一个场景，因为没有别的东西可以相信，相信哲学吗？相信文学吗？还是相信社会学，不知道。因此技术是每个人的，每个人都负有对技术以及由技术支配的未来的责任。当务之急要唤醒每个人的意识，别以为我只是技术的使用者。你既然使用了技术，你就有责任，你就要发声，你说未来怎么样好像跟我无关，反正我二十年后就挂了。这就不对了。每个人都要挂的，但挂之前是要发声的，没有声音是不对的。

3. 奇点时刻是共产主义时刻吗？

我们今天的讨论叫“奇点前夜：人类如何安身立命”。所以我最后要提的问题是：怎么来理解“最后的斗争”？在第一次工业革命进入高潮期——19世纪中叶，马克思、恩格斯就大胆预言了未来文明方向，以最后的斗争和共产主义理想流传于世。马克思和恩格斯设想了“最后的斗争”，也就是无产者与资产者的最后决战。这个设想现在还成立吗？

美国未来学家库兹韦尔（Ray Kurzweil）预感了一个“奇点时刻”（Singularity Moment），即人工智能达到通用智能（AGI）后，指数级增长的技术彻底超越人类文明的临界点，也表达为人工智能与人脑智能融合的时刻。——这个“奇点时刻”是共产主义时刻吗？我觉得我们现在要来追问一下。

马克思是一位具有“未来之眼”的大哲，我们要重新认识马克思。马克思在《1844年经济学 - 哲学手稿》中提出关于“人的科学”的概念，他当时才二十几岁。以我的理解，关于人的科学在今天有两类。

第一个是以人工智能和基因工程为核心的，我把它叫做人类技术工程，其实今天的科学就是“人类技术工程”，我甚至想把它命名为“人类世技术”。实际上都关乎着人，人工智能关于人类的精神和思维，而基因工程关乎人类的身体。

第二个是更新后的文科，也就是一般所说的人文科学。为什么我要强调更新后的文科？因为现在很多文科全是做量化、搞计算的，是搞数字化的，那还叫文科吗？那不是真正的文科。真正的“文科”，我把它表达为用非数字化，非计算的方式从事的弱理论探讨和非理论创造的广义人文学。文科的理论方式不可能是强理论，强逻辑，强推论的，那就变成科学了，变成数字技术了，所以我专门用了“弱理论探讨”和“非理论化创造”来标识真正的文科，即广义人文学。



除了变得有点可疑的阶级斗争，我认为重要的是两类“关于人的科学”的斗争。也许阶级斗争还会出现，这个概念有可能被转化为少数有用者与多数无用者之间的斗争。但我认为更重要的斗争，可能是我所谓的“人类技术工程”与“广义人文学”之间的斗争，那才是真正意义上的最后的斗争。

未来预感的基本逻辑是自然与技术的二重性（Zwiefalt），也表现为要在两类“关于人的科学”——“人类技术工程”与“艺术人文学”——之间保持张力，左右为难或左右逢源。

专题讨论

主持人宝树：非常感谢孙周兴老师，他把他对于德国哲学非常深刻的思考用在我们对于现实或者说未来数字技术前瞻性的思考上，有很多思想对我们都有非常深刻的启迪。

现在欢迎下一位讲者 Francesco Verso 先生，他有很多作品，而且也是一个科幻作者的编辑，他很多作品在意大利获奖，他有一个“太阳朋克三部曲”非常著名，他相信太阳朋克可以给我们带来一种乌托邦的未来。今天我们非常愿意来倾听 Francesco Verso 给我们进行的分享。



“太阳朋克”视角下人与 AI 的未来

演讲学者 / Francesco Verso

何为“太阳朋克”

我想在演讲的最初说一个挑战：我们对于人工智能最深切的恐惧——AI 有可能会变得过于强大，这种恐惧会不会分散我们的注意力，让我们没法关注这样一个问题，比如说我们应该跟新兴的智能体建立什么样的关系？

我是一个科幻作家，我每天都在想象尚未存在的未来——在那个世界里，技术与自然不是敌人，而是合作者；在那个社会里，智能不是武器或商品，而是一个共享的、不断进化的过程，参与者不仅包括人类，还包括生态系统和非人类的生命形式。

今天给各位分享一下关于不同的人工智能愿景，也是超越了我们所熟悉的控制还有灾难的叙

事，它是可以走向我所称之为的“太阳朋克人工智能同志”的愿景。

现在我们可以看到我们有很多的叙事，我们用这些叙事来描述人工智能，但是它是被控制在两极之间。一方面我们是有技术性乌托邦的解决方案主义（techno-utopian solutionism），他们相信的是人工智能可以解决一切问题，气候变化、疾病问题、不平等问题。

另一方面，我们还有一个末日的景象也就是人工智能失控了，而且会导致人类的灭绝和灭亡。这些故事看似对立，却共享一个致命的假设：即 AI 是外在于人类的东西，是发生在我们身上的事情，而不是与我们共同参与的事物。

这种框架催生了一种基于支配的伦理——例如价值对齐（value alignment）、人类监督（human oversight）、风险缓解（risk mitigation）——然而，正是这些逻辑将我们带入了生态崩溃和社会不平等。

道家思想认为，我们越是想要试图给世界强加一种秩序，我们的阻力就越大，真正的和谐并不是来自于控制，而是来自于一种流动，就是适应性的流动。

因此，与其问如何控制 AI，也许我们应该问如何与其共存并共同进化。这就是太阳朋克（solar punk）变得有用的地方。太阳朋克既是一种文学流派，也是一种文化运动，它想象这样一个未来：生态、技术和正义并非冲突，而是和谐共处、共同发展。

它和赛博朋克那种非常阴森反乌托邦世界或者传统的科幻小说对于科技征服的这种迷恋是不同的，太阳朋克敢于提出问题，比如说如果我们成功，不是说我们完美无缺，我们是非常务实的，我们去构建了一个宜居、公平、后资本主义的未来，到那个时候会怎么样呢？这个逻辑可以用到 AI 上，这也就意味着我们不应该将它看成是救世主或者是威胁，而是应该看成我们再生的伙伴。

AI 的设计并不是为了取代我们，而是希望跟我们共存，然后是可以融入到我们季节性的节律当中，社会礼仪当中以及文化多样性当中。

在我的小说《奇美瑞德》（Chimeriade）中，我描述了一个位于罗马的名为“奥尔托利奥”（Ortoglio）的社区花园，在那里，AI 传感器是支持而非主导人类活动的。

这些半自主系统并不追求产量最大化或强加外部的农业逻辑——它们倾听、适应，并与人类和土壤共同进化。它们追踪天气、养分，也关注当地庆典、祖先的周期以及代际相传的知识。AI 不是凌驾于社区之上运作，而是与之协作。

这就是我所描述的从控制到生态进化的转变，它是一个生态学还有进化的融合体，同时也呼应了儒家“仁”的概念，也就是伦理根植于关系，而并非来自于支配。

类似的思想也出现在我的小说《漫游者》（The Roamers）中，一群人拒绝现代城市高度互联、算法化的基础设施。他们选择步行——字面意义上的行走——来重新掌控自己的身体、时间和新陈代谢。但他们并没有完全放弃科技，他们重新进行了编程，用 AI 驱动纳米机器人来把字节、

基因和原子进行重新组合，这样就可以把身体内的新陈代谢效率提高。这些微观智能体降低卡路里摄入需求、减少食物浪费、最小化生物排放，拥抱了一种人类生理学的转变，这种转变并非迈向超人类逃避主义，而是对地球极限的新陈代谢适应。它使社区能够在保持活力、能动性和尊严的同时，缩小生态足迹。

因此，“太阳朋克 AI” 聚焦于共生创生（sympoiesis）与共同生成（mutual becoming）——由林恩·马古利斯（Lynn Margulis）和唐娜·哈拉维（Donna Haraway）提出的概念，意为“共同创造”（making-with）和“共同进化”（co-evolving）。它们描述了生物体如何通过共享的纠缠学会共同生活。在这一愿景中，AI 并非一个孤立的智能体，而是复杂生命系统中的参与者。

另一个来自《奇美瑞德》的虚构例子是“祖拉利亚”（Zoolalia），一个将动物情感和声音翻译成人类语言的 AI。当人们听到猪或鸡说“我害怕”，或者猫或狗说“谢谢”时，某种深刻的变化发生了。牲畜与伴侣动物之间的界限开始模糊。结果呢？社会迅速转向了培育肉（cultivated meat）——正如书中所说：“吃掉一个会说你语言的生物更难了”。

有一个关键点，AI 不会强行施加道德，它揭示的是我们被训练去忽视的某种关系，所以说改变并非来自于强制，而是来自于认可和同理心，这个正是我们现在所需要转型的，必须要根植于我们的关注对话和关怀，而不是在一种抽象的规则或者是自上而下的控制当中。

我们要怎么去实践这个愿景呢？

第一，通过去中心化的 AI 合作社（decentralized AI cooperatives）：也就是说我们有社区开发和管理的系统，基于本地的数据进行训练，它是可以对于区域的需求进行回应，进行响应，我们可以想一下，比如说 AI 专门为北京某一个特定的社区去进行设计，可以非常敏感感知当地的方言、风俗习惯，还有季节性的习俗，它不是那种远方发出的指令来控制，或者是把硅谷某个公司的价值观强加到你身上。就像 DeepSeek 这样的开源 AI，可以根据你的本地需求进行开发和定制。

第二个，发起一场慢 AI 运动（Slow Almovement）。如同“慢食运动”（Slow Food）。我来自于意大利，我们重视的是质量而不是速度，我们希望 AI 能够更好尊重我们的生态时间，更加缓慢地进化，更加深入地思考，在行动前先倾听，要融入我们长期的伦理反思当中。

第三个，生物兼容性设计（biocompatible design）：受自然启发的 AI，我们有真菌网络，还有蚂蚁群落，还有蜜蜂群落，他们都是智能的群体，我们从中得到了很多启发。在《奇美瑞德》中，发明了一种太阳能的雨伞，这个雨伞是可以收集空气中的水分，同时也可以过滤水分，是用的嵌入式的传感器来存储水分，不是为了盈利的，是为了实现韧性和实用性。

最后，我们需要为 AI 建立文化操作系统（cultural operating systems）。这意味着将原住民

智慧、女性主义认识论和非西方哲学融入智能系统的训练、设计和治理中。如果 AI 在道家和谐思想、安第斯“美好生活”（buenvivir）理念或非洲“乌班图”（ubuntu）精神下训练，它会是什么样子？

这些方法并非是无要紧要的，而是至关重要的，因为智能并不是一个普遍的概念，它始终具有一个情境性、具体行和关联性的，我们的伦理基础越多元化，我们的技术就会越有韧性而且越公正。

最后我还想跟各位分享我们真正的问题——我们真正的问题不是人工智能是否会超过我们，而是 AI 能不能帮助我们来超越我们自己。不管是在伦理方面、生态方面还是社会方面都是如此，AI 有没有可能会进一步加剧剥削和不平等，还是会加强人与人之间的同理心、同情心和合作，这正是科幻推理小说可以发挥作用的地方，类似于一个伦理模拟器，尤其是在一个未来到来之前，我们可以进行测试，正如玛丽·雪莱（Mary Shelley）远在阿西莫夫（Asimov）之前就警告我们那样：没有关怀、没有爱、没有价值观而创造出的智能是无法真正服务于生命的。

当我在撰写这种跨物种对话和去中心化人工智能基础设施文章的时候，我并不是在倡导空洞的乌托邦，我想要构建的是一个未来的原形，我所设计的就是伦理更多的选项，这可能是我今天想跟各位分享最重要的一点，就是 AI 的未来不必是一个奇点（Singularity）。它可以是一个多元体（Multiplicity），一个植根于文化、扎根于地域、并与地球边界相协调的智能生态。

简单来说，我们可以构建反映我们恐惧的 AI 系统，或者我们也可以建立一个系统来反映我们最好的愿望和愿景，我们可以设计一些机器，用机器来彼此监视，彼此榨取，彼此隔离，我们也可以设计一个机器让我们再生，让我们连接，让我们治愈。

太阳朋克人工智能同志是呼吁我们要从指挥走向合作，从榨取走向关爱，从支配走向尊重所有的智能形态。因为归根到底，在每个人工智能的背后还有内部，都藏着人类的智慧。



奇点前夜，人要如何生活

演讲学者 / 宝树

我本来想第一个讲的，现在有点后悔，因为大家讲得太好了。

奇点，老有人纠正我说叫 ji，其实应该念 qi，奇异点。所谓“奇点前夜”其实是我探讨时提的一个概念，我觉得挺有意思的，不知道从什么时候起，这个世界进入到一个很奇怪的时代，这个时代和以前任何一个时代都不一样，我觉得是挺好的一个概括。

我是科幻小说家。科幻是基于现代科学，有一个世界观，进行一些虚构，它具有可认知性，系统性与自洽性，当然，并不像科研那么严格。科幻在科技和现实之间，或者在科技和生活之间怎么起作用，是我感兴趣的话题。

传统文学反映现实。但是进入现代社会之后出现了一个科技的维度，我们发现原来在生活之

外还有这么大一个世界，还有几万光年之外的星球，或者是我们想象的一个新的技术，这个技术也可能带来很多可能性。

我是科幻小说家。科幻是基于现代科学，有一个世界观，进行一些虚构，它具有可认知性，系统性与自洽性，当然，并不像科研那么严格。科幻在科技和现实之间，或者在科技和生活之间怎么起作用，是我感兴趣的话题。

传统文学反映现实。但是进入现代社会之后出现了一个科技的维度，我们发现原来在生活之外还有这么大一个世界，还有几万光年之外的星球，或者是我们想象的一个新的技术，这个技术也可能带来很多可能性。

下面用黑色表示，它是空洞的，是没有生活的，这个中间的空白我们有时候用科幻去填补，想象一种从现实生活过渡到科技生活的可能性。

比如说现在的航天，我们就去看科幻作家想象在月球上，在火星上如何生活，现在是实现了一小部分，没有完全实现。或者说科技告诉大家机器人的技术，电脑的技术，AI 的技术，刚发明的时

候大家继续想象在未来会怎么样，机器人会怎么样，会不会统治人类还是成为我们忠实的奴仆，还是做很多其他的事情，其实大家去看很多我们对 AI 或者是科技的东西，其实这些技术诞生之前或者刚诞生的时候，科幻作家就有很多想象。

它其实成为一种生活发生过渡的中介，通过科幻——当然不只是科幻——来实现从一种生活到另外一种生活的跃迁，人类通过科幻想象的可能性创造一种新的生活，我觉得这是挺有意思的，而且是 19 世纪科幻诞生以来的进程。

下面一个概念是奇点，奇点本来是物理学的说法，比如说宇宙大爆炸刚诞生的点就是质量无限大，体积无限小，没法用物理概念描述的点，或者黑洞中心也叫奇点，用在这个地方我们是用库兹韦尔的解释，它的意思是到一定程度的人工智能，发展就会超越人类，从此一飞冲天，我们永远追不上，这就是奇点这个概念。

人工智能是不是会达到这样一个奇点，我还真不一定完全相信，但我认为这是一个非常有意思的思想方式和思想实验。

它的理论是什么呢？很多人都讲过，这里我就不细讲了，叫大加速回报原则，就是说我们对科技发展投入了越多，它就会让发展变得越来越快。比如说我们中国人产钢，晚清张之洞创建汉



阳钢铁厂的时候，产出几吨的钢就非常了不起了，因为用的都是非常低效的，甚至是无中生有的设备，现在你用那些钢材重新铸造新的机器和机械，不断扩大生产，现在产钢，一年几亿吨都是非常简单的东西，这就是加速回报，技术越发展，发展就越快。

反映在 AI 上就是算力指数级的增加，计算能力反过来会帮助生产过程，就会让计算速度越来越快，从而让人工智能越来越发展，当然还有很多新的，不光是量的增加，这个过程中其实有很多新的计算技术和具体的策略和设计会起到作用，比如说神经网络算法在最近二三十年大行其道，之前好多年前是相对停滞的状态。

我觉得到目前为止，特别是 ChatGPT、DeepSeek 这一代大语言模型出来之后，这个奇点的理论，好像是往这个方向在走，因为库兹韦尔认为是在 2045 年达到彻底超越人类，人类再也赶不上的状态，到现在看还是有可能的。

还有一个讲法是 2029 年人工智能通过图灵测试，前几天就有报道说已经通过了。但是图灵测试这个东西本来也比较含糊，所以很难说。

我们不一定要相信奇点理论，为什么呢？因为这个趋势不一定会持续，发展到一定程度可能突然停下来了，到了极限就上不去了，这是有可能的。但现在陡峭曲线还在往上走，中间很多东西对我们还是有意义的，所以我们可以假设它存在，所以奇点前夜我是算 2005-2050，2005 年就是库兹韦尔出版《奇点降临》那本书提出这个概念开始，他预言 2045 年进入奇点，晚一点 2050 年也可以。

这个阶段的世界发生了一些什么变化，AI 技术为代表的很多技术正在飞速发展。这个世界的版本不断更新，很形象的比喻，因为我用电脑，用手机，过几天就让我更新一下版本，有时候干一个什么事突然之间出来不好意思你要更新版本了，然后你一不注意手上的事就全停下来给你更新了，特别烦人。但我觉得就是这样，世界的版本不断更新，导致稳定的现实会瓦解。

2015 年的世界是什么样，可能大家还记得比较清楚，和现在已经千差万别了。现在很多东西那时候还不知道，或者是刚刚起步，像移动支付，像是流行的短视频、直播带货，现在生活中还有各种游戏，可能大家闻所未闻，现在讲六代机，那个时候可能还是五代机，其实只差十年，就是天差地远的变化，将来几十年变化可能会更快。

现实有不断发展，稳定的现实不存在了。今天赵老师说动词哲学，我说咱们叫动词时代，以前是名词时代。从这个角度来讲，我们对现实的理解和以前完全不一样，而且这是非常重要的一个点，超越科幻。

奇点前夜，人要如何生活

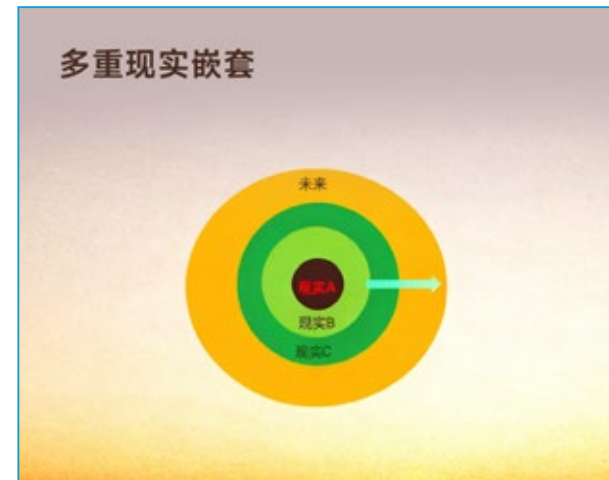
奇点前夜，我们如何生活，是个很大的题目，今天没有时间都讲，只能聊一点基本的构架。我觉得一个很有意思的点是多重现实的嵌套，刚才那个图有点接近，但也很难用形象的方法

表达出来，简单讲就是我们现在生活在很多嵌套的现实里，打比方讲高新产业园区的科技和外面完全不一样，只是我们没有意识到，我们说那是未来，其实已经发生了。

高新技术的地方和普通大城市，和一般小县城、山区，不同的地方技术会完全不一样，是多重的现实。以前大家都知道在北京和上海生活，肯定比在河南的乡下生活要好。但以前是生活质量差别的问题，不会认为是现实本身有巨大差异。现在由于技术差距太大，这个差异就会深入到更深的层面。

我们在这个现实中，另外一些现实对我们来说就是未来，而且不是遥不可及，只是说我们现在可能还够不着。比如说你生个什么大病，现在有很多新的技术，如果你有足够的信息或者财力，去全世界最好的医院，或者正好被选中做一个试验者，你可以用上一个很高端的分子靶向药物，可能你的病就好了，用不着的话就很不幸了。但可能十年后，这个技术就人人都能用上了。

所以多重现实是不断嵌套的过程。科幻和现实的关系也变了，我们刚才说科幻是现实和科技之间的过渡，那现实变成了很多重，科幻的作用是什么？不是现实之外的东西，而是现实的内在环节，我们要通过科幻的方式去理解现实。



刚才说的只是技术层面的，实际上如果一个人生活在不同国家，不同民族，他感受到的现实

可能是完全不一样的，生活在乌克兰和俄罗斯的人对战争局势的理解不一样。一个印度人可能觉得打巴基斯坦获得了伟大胜利，巴基斯坦人就不这么觉得。如果加入一些粉丝群体，他们会觉得这个世界就是围绕他们的偶像转的，对他们来说偶像的通告和行程就是他最重要的现实，而且会把其他一切东西和他联系起来。比如说发生国际重大新闻，他说一定是对家想黑我家哥哥，真的有这种思维方式，会把一切事都觉得是和他们相关的，这个是“虚拟现实”，但不是狭义上的那种眼镜里看到的虚拟现实，而是广义的。我们被真真假假的信息所包围，信息给我们构造出很多现实，我觉得这就是虚拟现实。

还有一个，我认为现实现在是在时间的断面上，什么意思呢？以前的现实其实是延长到很久的，经常是说我们对未来有很多规划，比如说未来我要当科学家，我要有一个规划，进行投资，进行理财，至少五年，十年。我要买房，我要预期这个房，就算不一直涨，至少不会暴跌。我觉得现在还在过渡期，发展更快的话，你会发现对几年以后做规划，变成没有太大意义或者非常难实现的事情。

科幻的作用是什么？我们可以通过科幻理解现实，两种形式，一种形式是很多出来的新东西，像刚刚讲的机器人、AI，其实在科幻里早几十年就有很多描述，有很多对我们现在还深具启迪，比如说阿西莫夫有一个小说叫《Robbie》，就是讲一个机器人帮助一个家庭，像保姆机器人，但母亲觉得这个保姆机器人很可怕，怕他失控，一定要把机器人送走，孩子就很伤心找这个机器人，很温馨的小故事。

放到现在有非常多的现实意义，现在说这个故事我反而同情这个母亲，虽然说这个小说是温情的，但实际上如果真把你放到那个处境上，你想孩子小时候和机器相处还是和人类相处呢？

还有一个形式是当新的技术出现之后，可能暂时找不到科幻小说，但有科幻的想象和准科幻的想象，会进行很多想象的可能性。一个东西出来之后我们也不知道会有什么，但我们会有预期，有预感，我们对它进行预感的想象，其实就是最原始的科幻。

比如说前些年转基因出来以后就有好多想象，社会认为它不确定，就会有各种各样的想象，这种现象其实也是一种科幻，当然也有些文学作品，大家通过这种现象去理解现实，而且是更基础的形式。

科幻和现实有很多交互领域，我随便举点例子。我出了两本书，第一本叫《美食三品》，这个主要是脑洞为主的，里面的技术和科技的东西 99% 是不太可能出现的，比如说有时间机器，也有一些很不切实用的技术，比如说水母的基因和人的基因混合在一起等等，可以长大变小孩，小孩变成大人，当然这个也不太现实。

还有一本是即将出的《你已生活在未来》，这个书是依赖于融入现实的高科技来撰写的小说，其中有的故事讨论到死者去世以后以 AI 复制出死者的数字模型，这个技术是有的，就是你的亲

人去世之后，你可以抽取他的资料，你可以跟他在界面上说话，理论上也可以制造一个拟人的机器人把这些东西输入进去，用和亲人一模一样的口吻跟你说话，那会带来很多社会上的问题。

AI 创作也是，现在一个很现实的事情就是 AI 创作，AI 可以帮你写小说，可以帮你写文章，那对社会会有怎么的影响，我这本书里有两个作品都和这个有关系，一个是说一个女孩子和一个男孩子分手，有情伤，她就用 AI 创作了一个电影，控诉这个男人，让这个男人陷入到一场网暴里边，大家都根据这个电影去相信这个男人干了很多坏的事情，实际上他还是比较委屈的。但真相已经被 AI 改写了。

还有一个是 AI 取代作家之后，作家没有事情可干，作家就去帮 AI 修改小说，有些作家去进行非遗的表演，现场给你表演写作，有些人去参观他。有个作家很成功，他写的还挺好，为啥？他把毛笔字写的特别好，这个人是个先锋诗人，他发现写先锋诗没用，他写个古诗还行，因为他会写毛笔字，看起来仙风道骨的，人家给他出个题，他当场写首诗，但实际上最后告诉你，这首诗就是 AI 写的。

这些小说和以前的科幻不一样，而和现实技术有非常深刻密切的互动关系，大家相对还觉得会发生在 20 年以后，30 年以后还有可能的事情，但实际上有可能这个事情明天就会发生，但也有可能不会发生，但即便不会发生，小说中的未来和现实世界已经是一种平行的关系，或者说是一种交互的关系，这是我说的写科幻很特殊的体验。

这种现实，其实对科幻创作也有很大的挑战。以前的科幻好多就没法写了，比如说以前写机器人，大家看笨拙的机器人，或者说像终结者那样毁灭人类的 AI，其实都不是很适用。因为我们发现和 AI 接触多了，第一，它没有那么蠢。第二，它也没有那么邪恶。至少目前来说是这样。生活已经变了，以前的套路就显得虚假了。

很多文学作家以前对科幻不感兴趣的，一点不感兴趣，现在都在写科幻。原因是什么？好多人分析过，有人觉得《三体》写得太好了，受《三体》的影响；《三体》的影响当然很大，但是更重要的是现实生活就是这样的，你要表达现实的生活，这个生活里就会有很多以前都没有很超现实的元素，就是多元现实的叠加，因此科幻不得不成为一种表达的必要方式。

我还问了一个问题，AI 能写好科幻吗？现在 AI 还是把网上大量资料的综合和打乱了重新组合一下，原创性还是不够。不过原创性的东西有时候是这样的，它把这个东西偶尔组合一下，但人会觉得很有新鲜感，对于 AI 来说是没有什么区别的组合，你说这个原创性是属于人类还是 AI 呢？这是很有意思的话题。

人机是协同还是投降，因为我发现人机必须要协作，协作来协作去，最后可能就变成了你给他东西，最后还是机器占主导性，最后发现还是和那个问题相关，什么叫做人？你和它协作多了以后，最后发现人要改造自身。比如说人的意识上传，或者是装个脑机接口，一切大脑活动，其

实库兹韦尔就说记忆也好，脑活动也好，都可以通过脑机接口和电脑连起来，这个阶段来说你是人还是机器？

所以说有的问题对我们现在的人类很重要，但放到一万年以后可能不重要，一万年以后咱们人到机器的过渡阶段是愉快的过程还是不愉快的过程，对一万年以后不重要，对于宇宙智能的进化并不重要。就好像人类是提塔利克鱼进化来的，还是另一种鱼进化来的，对于人，并不重要。

当我们把视野放在整个宇宙的视野上，奇点前夜这样一个关切可能不是那么重要，我还相信一句话就是阿西莫夫引用的英国诗人布朗宁的一句话：“The best is yet to come”，最美好的还没有到来，一定会到来，奇点之后一切更美好，但是可能这个美好和咱们没什么关系，但还是很美好。

专题讨论

讨论和观众提问

宝树：我们有差不多半个小时的时间，我们几个先聊一聊，再开放给大家提问。先问孙老师吧，您刚才讲的预感我觉得特有意思，现在的文化其实是预感未来的，后面你又讲了去未来化，是不是有点矛盾呢？如果一直生活在对未来的预期里，它成为一个主导的东西，怎么会去未来呢？

孙周兴：我这个报告充满着矛盾。比如说，我前面说人类进入失忆状态，后来我通过记忆信息的不可数字化来论证人文的必要性，这本身是一个矛盾，记忆和预感两者之间是个矛盾，我表明了这一点。

其实，我们对未来的预感本来就是分裂的、矛盾的，今天的文明基本特征恰恰就是分裂和矛盾，所以我们自己打自己的耳光是正常的，因为一个碎片的、多元的世界，无法提供完整的、单一的解释。我曾经有句话叫做以单一的标准和尺度去衡量这个事物和世界，这样的时代已经结束了。预感本身是充满矛盾的，所以我们的判断也许也是这样。

宝树：再问赵鼎新老师一个问题，您这个 Ad-hoc mechanism 的概念特别有意思，是尾大不掉的那种感觉。我有一个小小的疑问，因为您是在大学机构里看到的情况，但是否是人类历史普遍的，范围上是不是需要再讨论一下。

赵鼎新：比如说你刚才的整个报告里都在讲，技术发展的速度越来越快，每个技术发展背后有好多 ad hoc innovation，刚才我是从大学的负面角度来提示，但这个现象正面也有的。负面叫投机取巧、坑蒙拐骗，正面叫灵机一动，我们两个一讨论，一大堆 ad hoc reaction 出来，这个

里面一大堆因果关系就出来了。所以我觉得 ad hoc mechanism，是动物在对自然的反应时，群体做一样的策略，人类会反应为个体的策略。像孙老师说现代社会变多元了，其实人类从以前就这样。从历史到今天，始终对同样东西的理解很不一样。每个人眼中的现实是不一样的。

为什么这个问题现在孙老师说会强化呢？我认为是，当没有权力的时候，每个人都是胡扯，我们现在害怕的不是多元，害怕的是多元过程中特别有权的人把某一个单元定义为真理了、强化了，那就可怕了。我们不害怕多元，但是某一个理论变主宰了，差不多就都是负面效果了。

宝树：乌托邦的社会其实也需要这种 Ad hoc mechanism 吗？

赵鼎新：人的本性没变。在新的乌托邦，最重要的，权力要高度去中心化——人还是一样坏，但是（去中心化后）坏人做不了多大坏事了。

宝树：可以理解为针对人性的弱点进行反向的设计。

赵鼎新：对，可以这么说。

宝树：非常有意思。Francesco，我也问您一个问题，为什么叫“太阳朋克人工智能同志”？

Francesco Verso：因为我认为“伦理”与资本主义是格格不入的，所以我们需要一个不同的内在系统和文化框架。你想想他们为什么会在这个时刻发布 ChatGPT plus？我甚至不愿意把它称作“人工智能的进步”，我倒是会把它称为“人工智能的疯狂”。他们就像是把魔鬼放出了瓶子，让它在世界上自由游荡，但我们对此完全没有规则，也没有能力理解这种行为所带来的后果。

谁从这次 AI 公开发布中获利了？谁应该对它负责？但资本主义的本质就是：没有问责，没有责任，从来没有。不管是把错误倾倒在其它国家，还是向河流倾倒垃圾，再或者是把 AI 这种技术倾倒在民众头上——都是同一种模式。

将如此强大的技术推向公众，却没有任何的教育、解释或者使用说明书，就会产生很多混乱、不知所措。这就是一种浪费，一种没有责任的，在不同领域当中的一种垃圾的倾倒。

因为你很有可能是将具有威胁的智慧体释放出来，你不知道它是否会影响气候变化，我们也提到了能源的问题，对吧，消耗了多少水、多少能量、产生了多少热量，现在谁来负责呢？我们从来没有讲过这些长期的后果，你正在削弱自己的能力，迷失自己的价值，并且免费地，甚至还

得付费地，输送给另一个世界。长期来说，已经有一些研究表明，我们大脑的功能正在持续下降。谁会为这一整代人所受的损害买单？

也许我们还能适应，比如说我是来自模拟时代的人，但我的女儿 17 岁，两年前就开始使用 AI 了，她会受到 AI 的影响——比如说记忆、知识、懒惰，对吧？这是非常巨大的影响，就像气候变化一样，这是我们时代的两个基因，这是没有办法被理解的，就像蒂姆·莫顿说的“超物体”（hyperobjects）。超物体是一种现象，它完全超出我们的理解能力，甚至连边界都难以把握。

因此我赋予语言很大责任，对我而言，语言是重要的。我希望为 AI 而战斗——这是我的项目。这个项目是自下而上的，用我自己的（东西）来训练的。

我们已经看到了巨大的伤害——因为 AI 如今被困在那些被污染的数据和错误的框架之中。我们正在制造“AI 废物”，那不是真正的思想产物。

宝树：为什么是同志？

Francesco Verso：我使用这次词意思是，我们需要处理责任问题。否则依据丛林法则，正如赵教授所说，如果没有公平的权力分配，强者就会抢占一切。

宝树：非常感谢，聊一个最基本的，各位老师对“奇点”的概念怎么理解，是不是相信有奇点，有没有什么理由相信它或者不相信它，或者觉得它对我们有什么关系？

孙周兴：我大概是相信的，我刚才在报告里也提了一个假设，奇点时刻和马克思所设想的共产主义的想法有没有关联。马克思在 1844 年手稿里居然已经在思考人类未来的休闲，马克思是很神奇的，当时就在想人类未来的休闲是什么样，是很奇妙的想法。

人工智能为代表的技术，更正确的应该说是数字技术，在人类智力上的超越已经实现了。最近有两件事情，我的一本德文书前言是机器人写的，结果德方的编辑说你的德语怎么一下变得这么好了，因为我前面的八篇文章是自己翻译的，机器明显比我做的好，我受到表扬了，他们也没看出来是机器人写的。

当然我也有一个失败的经验，失败的经验就是我一本文的书，我当时想试一下，让 ChatGPT 给我写一个前言，我把所有的文字喂进去，我说你给我写个三千字的前言。我感觉很失望，因为全是网络的假大空的语言，你们有兴趣可以搜我那本书叫《人类未来》，那本书的前言就是让机器人写的，结果我很失望，我觉得这个前言根本没用。这个可能和国内的数据状况，乱数据，脏数据多是有关联的，所以生出来的东西不太好。

我自己做了这么多翻译，我在五年前就开始停止了翻译工作，我觉得翻译时代已经结束了，所以五年前我写了一篇文章说，翻译式学术的时代即将结束，翻译不再是人类要去做。我最近也做了一个事情，我把《存在与时间》这本书用机器翻译了一下，但是让我很失望，表达能力还不够。但它在不断自我学习，我相信在这些技能上全面超越人类是必然的，所以所谓的奇点时刻，无论怎么理解，终将到来。无论社会有没有做好准备，个体有没有做好准备。大概人类将进入这样一个周期，我们将活得越来越长，我们将越来越难以死掉，你想死都难，但是我们越来越没事干，因为创造性的工作以后会越来越艰难，你想创造点什么会越来越难，于是我提了一个概念叫强制休闲。

我们所有的这些讨论都是可以归到我所谓的预感，还有不确定性，还有矛盾性，实际上不是确信的判断，和传统文明、哲学、形而上学对自己使命的要求是完全不一样的。形而上学就是需要确定性，没有也要创造确定性，但现在确定性的创造将很难。

赵鼎新：这东西是不是奇点我不知道，提奇点比不提奇点好，因为提奇点让我们觉得对危机的想象更深刻。

我记得 70 年代做生物学的时候，当时就提出了现代石油危机、能源危机、环境危机是奇点，现在看来都不是问题，但是一堆新的问题出来了。但有些东西可能真的属于奇点。举个最小的例子，大家都知道马，公元前 8 世纪马在乌克兰培育，使得现在人能骑在上面，传统是不能骑的，以后才出现真正的游牧民族，以前是走路放牧的。这个奇点性带来了疾病、流行病、植物、生活方式、观念、宗教的变化，难以想象的大。这个奇点性之大，就是一匹马。

今天为什么提奇点性比较好呢？我们每个人都在用，有的人要用它发明，有的人要用它创造，有的人用来图像设计。我觉得未来也是会有创作，也是会有艺术，也是会有创新，肯定创新的概念变了，所以人类会出现重大的重组。以前的奇点要么是像马一样，是无意中带来的结果，要么是由国家、大公司推动的，现在是国家、大公司、每个人都在里面，从这个角度我认为奇点现在提出来是合理的，可以让人想到这一次和以前的不一样在什么地方，危险是什么。

【观众提问】

观众：我是清华大学社科学院的，做政治学研究，我提两个问题，一个是给孙老师，另外一个给赵鼎新老师，两个问题虽然提问不同的人，其实指向是相同的。

赵老师，您刚才讲到 ad hoc，我非常关心这件事，因为我们在讲现在整个社会的构成，或者我们研究的驱动其实都是讲因果律的，包括我们在宗教方面的思考，还有日常行为，ad hoc concern 其实是很常见的。您能不能回答一下，既然 ad hoc 是非常重要的，在未来是不是有可

能让它渐渐走进主流当中，在宗教当中或者是在研究当中。

第二个问题给孙周兴老师，您讲到的关于人文科学或者是人文学未来的发展，这个发展其实也是我们未来，包括人文学科，包括政治学科，政治和统治方面的问题，您讲到了斗争，是不是有可能产生更多的协同性的问题，因为我们很多时候都在讲斗争，斗争在最后可能就会导致并不是特别好的结果。您刚才讲到技术政治已经占主导了，但是我们现在人文学科的未来发展不得不依附于政治，您如何设定边界，我们的未来发展方向如何做不依附技术的人文学科的发展，谢谢。

赵鼎新：我发现国内无用的负向特殊性太多。而有用的，比如艺术上的创造，特别是现代艺术——就是在一个新的 ad hoc perspective 下，产生新的想法和视角——这方面非常差。国内连街舞这种完全基于即兴表演、高度特殊性（ad hoc）的艺术形式，也受到了规训化。

国内的艺术，在西方学术中被划分为新表现主义。国内的人类学，我认为他们现在追随的西方文化人类学也是范式化的。但西方不断有新的视角出现，是非常特殊性的理解（ad hoc understanding），目标是用来解构。而这个概念原本是指向社会普遍性问题。国内产生的概念则是纯粹的特殊性（pure ad hoc），最多加个“中国视角”、“中国理解”。我看一点都不中国，这实际上是中国式的临时拼凑，这是第一点。

然后，为什么需要解释？如果你从特殊机制、特殊现象出发，如果它在中国社会变得普遍，那背后就有普遍性规律。为什么学术造假在中国发展得特别多，手法特别巧妙？造假本身是门艺术，我想开个学术造假艺术馆，那会非常精彩。但同时也在思考，为什么特殊技巧（ad hoc skill）在中国如此发达？我觉得其实是我们中国人该运用特殊性的地方，反而不怎么运用。

孙周兴：你的问题很复杂，我不能全面回答。大概两个方面，一个是我描述的最后的斗争，斗争两个字在文革的时候被我们败坏了，斗争很正当，斗争或者是争斗或者是战斗或者是矛盾、冲突，这个字不应该被理解为一个贬义词，但慢慢在我们的语境里被理解为贬义词了。

我大概描述了两种斗争，一个就是马克思的无产者和有产者，要稍微做个转化，有用和无用之间的斗争。另外是两门科学，也是马克思的描述，他说关于人的科学必将产生，你现在想今天有没有出现，实际上已经出现了，一个是带有人的科学研究，技术研究，我今天就表达为人工智能和生物科技，另外一个就是人文。谈到人文科学就涉及到你后面的问题，人文科学必然和政治关联起来，这个政治，我简单的描述是以前都是用商讨的方式，传统的政治方式处理问题，今天技术统治，不再是这样的方式，今天许多问题不是商讨能解决的，我们商讨一下，再做个决定，有效吗？所以这里面就有一个紧张关系，也可以表达为冲突，人文科学必然为此做出判断和参与此事。

今天最关键的点是要唤起每个人的政治，每个人要有自己对这个世界，对技术，对事物的判断和参与，所以我这些年，特别是在做当代艺术，当代艺术给我们最大的启示就是告诉我们艺术不是挂在墙上的，艺术是每个人的，是你的行动，所以这一点特别好。

我们现在也在谈艺术政治，艺术政治只在当代艺术的意义上才成立，刚才我的报告后半部分在说的是技术是每个人的，政治也是每个人的，大概是这个意思，至于会不会产生不良效果，不一定。

我多说一句，不是说产生协同性的效应和和谐的效果就是好的，如果事物的真相，世界的真相是矛盾的，分裂的，冲突的，为什么一定要用假像的和谐、或者说协同来解释它，来提供一个方案，一个假的方案，必将是假的。所以现代主义的哲学也好，别的学科也好，都在直面这样的问题，对于一个多元的、碎片的世界，我们到底还要不要提供一个整体单一的解释。

观众：想问一下 Francesco Verso，我非常喜欢您对未来的人类和 AI 的设想，特别是您提到了 slow AI，我很好奇您是否曾经和一些产业里的相关人员交流过，因为对于我们在大学研究者而言还算相对容易，但当我们谈到，例如 DeepSeek 的时候，其实压力非常大，为了追求越来越快的速度，您跟他们交流过吗？您要怎么去劝说他们呢？让他们接受您的愿景。

Francesco Verso：我不打算和他们谈，他们也不会听我说的，因为这几个未来是相冲突的，所以我我说的是自下而上的，而不是自上而下的，我们看到了很多这种情况，我们经历了许多工业革命，比如第一次，第二次工业革命，我们现在看到了数字化，我们也看到了自动化，我们现在也有可能看到对人脑的模拟，这意味着如果我们完全允许技术主义控制世界，最终很可能会走向奇点，也就是“一个统治一切”的局面，但是我不相信这种，我相信的是多元性和多元化的，不会是一个单一的。

我看到一些新生的现象正在从深圳，从曼谷，里约热内卢，内罗毕各个地方涌现。这个世界太疯狂了。我的意思是，我们还有可能输掉这一切，未来并不是已经注定的。

但毫无疑问，这些力量非常强大，我不认为他们想慢下来。所以我们现在唯一的力量，就是我们的购买权，我们在哪花钱。这就是说，如果我们真的想放满脚步，就必须通过消费行为把资源分配到该去的地方。当然，如果你去超市，看到的都是“不健康、不健康、更不健康、更不健康”的产品，那你根本没有选择。这就是为什么想象力在创造替代方案时如此重要。当我们看不到出路时，比如说我们在罗马，我们开始在房顶花园上住蔬菜，或者在社区小花园里种蔬菜，从而摆脱对那些已经不再真正提供服务的中心化服务系统的依赖。所以当它们不再为社区提供有价值的东西时，社区就必须靠自己活下去。

当然也会有人找我们说，“来，我们做一些事情吧”，那是有可能实现的。我曾经被邀请与一家非常大的公司合作，他们想做一款“太阳朋克”题材的游戏，但实际上只是一个普通的第一人称射击游戏，背景里只是加了一点绿色装饰。我说，这不是太阳朋克，请别用这个词，这仍然是赛博朋克。这就叫做“伪环保”（greenwashing）：套用另一种现象的叙事外衣，只是为了继续用同样的方式做生意。你只是把 logo 换成绿色，重新包装一下，就假装一切都好了。这也是为什么我认为需要很多社区，我相信变革的力量来自人民。

观众：各位老师好，我有两个小问题。想要听一听赵老师从社会学角度解读一下，还有其他老师的看法。怎么理解现在对于 AI 产生的情感性的依赖。第二个小问题是 AI 技术发展的越来越快，会不会让我们个人生活变得越来越匮乏，这种匮乏可能是一种虚无主义吗？我发现我身边其实是有跟我同龄的人有这种倾向，比如说更不愿意和人交流，更愿意和 AI 对话，这个趋势非常明显，想问问在座各位的看法，以及有没有一些个人的小建议。

赵鼎新：新的依赖刚刚开始，我还不知道。但是我知道手机在中国流行。美国也有手机，我从来没有在美国的车上，看到每个人都在看手机，走路也去看手机。去年我的夫人在柏林高等研究院驻访一年，然后又在汉堡的高等研究院驻访一年，我们每天去城里，要走过一个桥下面，桥下面住着一批流浪汉，晚上点着灯他们在干什么？在看小说，看书，在讨论。从社会学角度，AI 对每个国家施加的影响不一样，中国的社会高度原子化，国家推动力量太大，社会没有抵抗，大家有技术崇拜，这样中国的社会发展太快，导致的社会家庭结构变化，家庭教育高度统一。给孩子买一样的手机，大家一样都在看手机。海外有当地的多样性（local diversity），和技术结合之后，会造成完全不同的结果。

从我的角度，中国社会这一百年的变化，国家也好，改革开放以后的私有企业也好，都在追求一模一样，就连新长出来的草都一模一样。这时候，AI 对我们负面侵害大，是因为我们把 social context 的东西全部拿掉了，如果到罗马，像刚才 Francesco 所说的，这种事情肯定不会发生，至少发生的程度不太一样。

宝树：我从另外一个角度讲两句。情感依赖这个事情是这样的，AI 不是天上掉下来的，一定是有人专门设计出来针对人性的弱点——像女性爱玩乙女游戏之类的，恋与制作人之类的——会抓住人情感最空虚的地方，我觉得是很难抵抗住的。前一阵有人让我写一个文章，最后我论证出真实人与人之间的关系，发展到最后一定是更丰富、更深刻，当然如果合适的话，能够彼此促进。但是一开始的吸引力，可能是不如人和 AI 的，这个坎怎么过去我还搞不清楚。

孙周兴：最近有一个数据统计，手机时间，通过手机上网的时间已经达到了人类清醒时间的60.76%，远远超越了家庭生活的时间，朋友交往的时间，讨论的时间，看书的时间，那都没有了，家庭时间本来占了百分之二十几，现在已经降到了五个点左右。

总的来说这是一个无可阻挡的趋势，我也同意刚才赵老师的说法，比如说在欧洲，在美洲，他们民间的自觉性，以前我在德国的邻居也是，他们每周自己有个家庭音乐会，自己有一个读书时刻。我现在就很困惑，我儿子15岁了，没正经读过一本正经的书，但是他最近告诉我说语文是班里第一名，这件事情让我无比的困惑，他没有读过正经的书，也不爱读经典，我说武侠小说读一读好不好，我给他买了整套金庸武侠小说，他翻了两页就没兴趣读了，问题是他告诉我他语文蛮好的，我很困惑，我的困惑和你一样，我们的手机依赖是已经无法摆脱的。



汪晖：这是我们最后一场，现在有将近两个小时的时间做圆桌的讨论。今天下午有三个 panel，听众分在不同的屋子里面，第一个环节想请三个 panel 的主持人，对他们的每一个 panel 提的最核心的问题做一个简要的说明，你们这个讨论里面哪些是最重要的，下一步也会请各位嘉宾和听众共同讨论这些问题。

Stephen Cave：我们讨论的是 2100 年人工智能与人类的关系。对于如此遥远的未来，我们能知道什么？我们如何了解它？又该如何塑造它？

选择 2100 年作为节点，是因为今年恰逢艾伦·图灵诞辰 75 周年。这位人工智能奠基人自他发表《计算机与智能》已过去 75 年。那篇 1950 年的论文，邀请我们畅想拥有智能机器的世界。各位都知道，1950 年的世界还是模拟信号的时代，与今天数字化的世界截然不同。过去 75 年发生了翻天覆地的技术变革。现在我们确实生活在图灵预言的世界里——一个充满智能机器的世界。

值得注意的是，上世纪五六十年代就有人帮助我们构想这样的未来，包括艾伦·图灵等哲学家、工程师和学者，当然还有科幻作家。某些科幻作品至今仍能指引我们。这正是我邀请嘉宾探讨的议题。当我询问各位嘉宾 2100 年的人机关系会怎样时，没人回答“和现在差不多”。所有人都认为将发生根本性变革。这个观点稍后还会谈到。首先阿图莎·卡西拉迪通过过去 75 年的发展轨迹，推演未来可能的图景。她提出了需要应对的三大核心问题才能创造美好未来：管控 AI 风险、利用 AI 解决问题、借助 AI 提升人性。接着是陈楸帆呈现了科幻叙事中的多元未来，从《黑客帝国》里人类沦为电池工厂的设定到人机共生的积极愿景。虽然我觉得，机器把人类当电池，也算某种共生关系。这些反乌托邦与乌托邦的对立想象很有启发性。然后赵汀阳教授指出 AI 帮助我们重建未来概念，因为当代世界正缺乏未来观。这个观点值得深入探讨。看似属于未来的图景，往往植根于过去。可以说我们现有的未来观已与技术进步理念深度绑定。比如 19 世纪的英国技

术进步等同于未来，它定义未来并被赋予积极意义，当然会让生活更美好。但正如赵教授所言，一战的工业化屠杀、原子弹的诞生，让未来变得复杂起来。这种复杂性也反映在我们对 AI 的想象中。有些设想极其黑暗。楸帆提到《终结者》系列机器崛起、人机争夺主导权的普遍叙事。赵教授认为 AI 可能在智力上超越人类，但永远无法拥有人类双手的灵巧度。最终人类只能为 AI 打工，比如清洁它们够不到的电路。这些负面未来图景之外，我们也努力构想 AI 时代的积极可能性，如何利用 AI 超越民族国家界限，在主体间性哲学启发下建立人类与万物包括与机器的共生关系。但最重要的是我们讨论了想象多种未来的必要性，不是预测而是塑造未来。这意味着不仅要推测未来可能的样貌，更要构想各种可能性，并选择心仪的方向。核心问题不在于 75 年后 AI 会变成什么样，而在于我们希望社会变成什么样，我们想成为什么样的人，以及 AI 如何助力实现这些目标。这些就是我想提交给本次圆桌会议的问题。

张伟特：第二组讨论的问题，第一个在资本主义导致很多社会痛苦的情况之下，AI 的发展是不是对于弱势群体造成更大的剥削和压迫，如何避免这种情况出现。讨论的问题有 AI 跟人类历史上其他的技术本质上的区别，比如人类历史上铁器的技术、青铜的技术，火的技术。有没有必要论证 AI 一定会导致对弱势群体不利的影响。另外我们也讨论到如果 AI 具有意识，会不会反抗当前资本主义的一些剥削和压迫。

第二位报告人讨论了在什么样的情况下定义我们的行为和道德责任的概念，这些概念都是非常超前的大概念，我们思考 AI 伦理的概念工具还不够充分和不够有自觉。自我意识在认定道德及其他责任过程当中，承担何种角色？谁对 AI 行为负责？

第三组报告是关于儒家伦理的视角下看伴侣机器人的现象，在我们社会当中承担何种角色，我们有很多有意思的讨论，比如未来在人伦社会当中应该赋予伴侣机器人何种角色，才有利于整个社会。

最后一个在图灵测试的各种缺点之下，从笛卡尔哲学出发，能不能对机器证明自我存在这样一种工作提供一种可能性的测试方案。同时，从 AI 的认知能力智能的角度出发，提出了苏格拉底、笛卡尔和哥德尔康德、莱布尼茨这四个等级的智能点的一个思考。最后，我们也讨论了怎么样测量或者观察 AI 的情绪，它究竟有没有情绪等等这些问题，我挂一漏万先想到这么多，谢谢！

宝树：我们的四位报告人，一位是赵鼎新先生，他的报告题目是人类本性与形而上学，赵老师提出了横跨整个地球历史非常宏大的结构机制演化的一些模型，其中包括有正反馈、负反馈，特别提出一种临时机制，这种临时机制在人类社会中被个体化的，每个人可以随机应变去做很多东西，这个东西的产生和发展让每个人谋取自己的利益，对整个系统产生一种很有意思的制衡

或者反制的作用。在高智能技术时代这种情况愈演愈烈，提出尾大不掉的效应，这个东西产生很大的内耗，赵老师提出应该通过自主设计的方式，要去中心化，让大家在破坏性不大的领域去竞争，必须要理解人性中间的机制和人类最大的不平等的关系，才会产生伦理学。

第二位报告者孙老师，他的题目是未来预感与未来责任，孙老师提出传统的文化是记忆的文化，以记忆和过去为核心。在现在的社会也有普遍的失忆状况，人们在忙碌之中不会追忆过去祖先的这些东西，就有一种新的文化，他把这种称之为预感文化，预感文化是朝向未来的，以未来为核心和数字技术相关的前瞻性的文化。预感有不确定的冲突，中间有非常紧张的关系，可以称为一个很卷的状态，有很多问题，这个预感文化本身有可能是去未来化的。在这个过程中，可能涵盖不了真正的未来，所以人们应该对未来有一种责任。这个技术和每个人息息相关，每个人既是技术的持有者，也是技术的利用者，我们要有每个人的个体技术的政治意识，自己要负起责任，还有提出来奇点这个理论，他是不是和马克思 1844 年讲的共产主义是有关系，非常有意思的一个设想。

第三位是 Francesco Verso 先生，他的报告叫“太阳朋克同志”，他的意思是说，我们可以用另外的方式看待人与人工智能的关系，与其问如何去控制它，也许可以问如何跟它共存，共同进化，这是太阳朋克的理想，是一种文学流派，以太阳能技术为核心的文学流派。但是又是一场文化运动，他想象生态技术和人类并不是冲突，而是平衡，并且共同生长出一个未来。他提出了几种具体的设想，举出去中心化的人工智能合作社；一种精雕细琢的慢人工智能运动，以及照顾到不同生物、不同个体性的生物设计等等。他提出真正的问题不是人工智能是否能超过我们，而是能否帮助我们超越自我。在伦理上、生态上和社会上深化同情和协作，实现共生。

最后一位报告者是我，我讲的是奇点前夜的科幻现实。虽然奇点没有被证实，但也是一个很有用的概念，奇点前夜的概念是说现在的现实是不稳定的，变动非常大，可能没几年就会有巨变。所以有人把现实设想成一个层层嵌套的系统——不是一个现实，而是多重现实嵌套。传统上大家觉得科幻是现实和未来之间，或者是真实与幻想之间的一种过渡形态，但是我认为科幻是打通和连接平行的也好，重叠的也好，不同现实之间的一种思维方式。我们通过科幻，反过来可以理解现实。再进一步，如何过渡到未来硅基文明，过渡过程对我们很周期，但是对于未来的文明不重要，所以我认为最好的东西即将到来，但是最好的东西不一定和人有关系。

最后共同讨论的奇点问题，孙周兴老师相信奇点一定会到来，因为他前一阵写一篇文章让 AI 给他草拟了一下，发现比他写得好。但赵鼎新老师认为奇点不一定真的来。比如像历史上有很多重大变化，像马的发明，把它驯化，原子弹也很重要，但是奇点是非常好的一个思维方式。它能够让我们证实目前像 AI 高速发展面临着很多问题，所以他说提法很好，但是是不是真的可以再讨论。Francesco Verso 先生认为，人工智能的未来，不一定是一个奇点，可以是多奇点。它在

不同的角度，不同的点进化，嵌入不同地域、不同文化，和整个地球相协调很智能的一种状态，他设想的是一种多奇点的状态，也是非常有启发性。大概就这么多，谢谢大家！

汪晖：谢谢三位主持人的概括，问题提得非常多，我只能随机地提几个问题，先请几位嘉宾说几句，引发更多的讨论。

刚才我问过孙老师的一个问题，因为他在刚才的演讲里面有几个很重要的关键词，一个是自然记忆的衰退，一个是预言社会的到来，在讨论的时候也引用了尼采、马克思所有的这些讨论。我想问一个问题，随着 AI 的发展，机器记忆的能力大规模发展，不管将来脑机接口什么时候能成功，即便没有，我们现在对许多问题的回答也是通过机器实现的。换句话说，有一个没被提及的概念是遗忘，过去像尼采的论述里面对创造的论述跟积极的遗忘连在一起，如果没有遗忘很难有真正的创造，就是跟过去的某种断裂。可是随着机器的发展，所有的记忆积累到了无法遗忘的程度，什么事情事无巨细，只要你点开就能找到。在一方面是自然人记忆衰退，一方面是人机关系发生重大变化，使得遗忘从某种程度上变得不可能的时候，怎么解释创造？怎么解释预言和记忆的关系？

孙周兴：这个问题太坏了，击中我的痛点。大概在十九世纪中期以来，记忆工业已经开始了，1839 年照相术发明，图像可以复制了，1879 年留声机发明，声音可以记录了，这件事情很伟大，因为声音一说出来就消失了，无论是绍兴话还是北京话都都消失了。声音一直以来我们是无法记录的，后来声音可以记录和复制了，现在是半导体，当时是机械刻录。当然更早的是文字。记忆工业使得人的三大媒介被记录和记忆。这很大程度上，让我们把记忆这件事情完全交给技术了，我们的记忆不可能消失了。两个月前的数据，《罗刹海市》那个歌已经成为人类文明史上播放最多的一个文化产品，说的是人类文明史上，从来没有超过 100 亿的播放量。如果没有声音的记忆工业，那怎么可能。

与之相关的是人类进入失忆，这个失忆是针对自然人类的角度来说。因为记忆变成机械的、外部的，我们实际上不需要记忆了，这是一个方面。

另一方面，在十九世纪中期以来，一大堆思想家、艺术家都开始思考未来的问题，因为他们发现技术工业已经把两千多年的自然人类文明基本的成果全部解构了——马克思、尼采都在思考这个问题——这里面的问题是什么？我们已经不再需要通过学习古人古书，古代的价值来勾勒我们的生活，这个时代已经过去了。马克思用一句简单的虚无主义的语言做了一个判断叫烟消云散，一些固定的东西都已经烟消云散。

这是两面性，记忆越来越被加速，越来越技术化，越来越放大，刚才汪老师说我们忘不掉了，

一点开就有了。另一方面我们已经失忆了，尼采讲这种失忆是好事，我们自由了，我们不再受束缚，我们创造的可能性更大了。但问题是一个完全不确定的时代到了，我今天的报告里面在讲预感，与之相关在十九世纪中后期开始反思线性时间观，进入一个新的实践经验，这种实践经验不再以过去为指向，而是以将来未来可能性为指向，当我们到这一步的时候，就面向一个完全不确定的不可知的未来方向。我今天的主题叫预感，我们只能预感，只能想象，但是不确定，创造是打开了，但另一方面发现越来越难了。只能讲这么一些，下面想起来再说。谢谢大家！

汪晖：谢谢！我问赵鼎新老师，第二组讨论到一个问题，AI 时代跟资本主义之间的关系，跟社会的关系，AI 时代新的社会不平等如何解决，您刚才在演讲里面也特别提到失业这个问题，比如现在也提到了把 AI 当成技术移民，AI 不是人，但是取代了大量的劳动。大规模的失业和财富可能更快的聚集的过程，在今天到底用什么样的方式来解决？孙老师讲演里面最后提到了马克思，讲到共产主义，也讲到斗争的概念。在 AI 时代怎么界定社会斗争这样的概念？怎么界定如何解决这个问题？另外技术的发展在多大程度上提供了解决这个问题的可能性，它的政治、经济条件大概是什么样的状况？

赵鼎新：这就说到人类性，人的特性也就是生物体变成人这个动物以后出现了四个现象，第一原来是负反馈自稳定为主导的，现在是目的导向正反馈主导的，所以什么东西都越发展越快，这是我们这个社会就自带不稳定，这是第一。

第二，串联性（looping），动物也是有意识的，但是人把意识 justify，孙老师讲预测，古代人更讲预测，占卜什么的都是预测。但是我们人 looping 产生的后果，就像打雷什么，就像灾害，就像自然界对我们一样的后果。精神和物质产生的后果完全一样，这是第二。

第三个，动物层面，面对自然有许多特殊性（ad-hoc coping）。动物层面的特殊性都是群体性，如果这个物种产生比较长的嘴巴能够吸到蜜，这个物种的个体虽然有差异，都能吸到蜜。人类会说你产生蜜，我用剪刀把你剪掉，你的嘴巴塞在里面出不来了。第四，因为我们正反馈大量的结构和机制导致大量的特殊因果关系产生最后一刻，我们叫末端重。

这四个东西都有正面性，都有负面性，负面性都是随着社会权力的不平等分配而加大，无限加大。所以要解决伦理问题，不讲这五个问题都没有用。

我想的唯一的解决方案民族国家模型肯定要拿掉，得有一个统一政府，但是统一政府权力不能太大，只管一两样公共事务，军队只当警察，也要控制有几千万的粉丝，他们的影响力也是权力。

目前还有一个难点，实际上推动 AI 真正的行动者是国家、大公司。国家的权力是集中和强制性的。经济行动者的权力是集中的，但是非强制——你可以买，可以不买。第三个行动者是普

通行动者，科学家、我们每个人都参加，对于普通人来说，AI 有趣好玩，像毒药。但是国家是不一样的，每个国家在竞争，不能出局。

比如欧洲现在我估计多半出局了，因为他们首先想的是伦理和规范，那么在国际竞争中竞争力就不行。今天早上的讨论里有人说要把 AI 设计成女性，植物……谁设计女性 AI 语法谁出局，设计植物语法彻底完蛋，因为植物对周围一切就不做回应。如果这个 AI 是这样的，286 计算机都比它厉害，这时候你就出局了。所以 AI 必须模仿男人，模仿像特朗普这种男人才行。

讨论这个问题，要加一个最重要的本体论，就是人类性，否则我们没什么办法。

汪晖：想问问杨锦方先生，你是整个会最重要的推动者，确定这个主题跟你的想法也有很大的关系，刚才我们也讨论了到底奇点到来了没有？还有 logos multiplicities，就不是一个奇点了。我粗浅的理解，奇点是多重合力，在这个地方发生遭遇的那个时刻，才能叫做奇点。你认为这样的是不是确实已经到来了，而且你的判断是怎样的？

杨锦方：谢谢汪老师，我自己是学计算机的，有一点技术背景，也干过编程的工作，从技术角度看，我认为现在奇点还没到，但距离不远了。现在在能力方面，人工智能在很多事情上快速地赶上人类的能力。我同意赵鼎新老师的说法，有没有奇点都不见得。有可能因为我们有这些焦虑和对未来的担忧，在行动上采取一些改变的动作，导致这个奇点最后没来，这个可能性是有的。

我认为奇点这个概念还是很重要的，因为这个事情的确非常大，是一个很全面，对人类社会影响也非常深入的事情。

我也想提一个问题，人文学科在过去二十年时间，在全球范围内遭到了挺大的冲击。招生的人数减少，教授的数量减少，看起来好像人文学科的学生在近些年毕业之后的就业率面临比较大的挑战，很多人干的不是自己专业的工作。我自己这么认为，人工智能时代的理工科的学生有可能被彻底取代，因为工具性的东西如果不涉及价值判断，人是不如机器的。但人文的东西恰恰不是这样，因为里面是大量的价值判断的事情，我们没办法交给机器，我们交给机器就成奴隶了，所以必须自己作主。这就带来一个很大的问题，人文学科感受到的冲击很大，人工智能重新塑造人文这个领域，人文学科会发生什么样的变化？我们讨论的很多问题，都是这个时代要回答的问题，但是怎么回答？让谁来回答。我们研究的组织方式是不是会发生变化？这个时代有可能发生变化吗？我也想问问在座的各位，你们是真的从事这些研究的人，我相信你们对这个事情有很好的洞见。

汪晖：谢谢！想问 Cave 教授，刚才您特别讲到人和机器的关系在历史当中发生的变化。人

和机器的关系确实不是新的关系，很早就开始了这样持续的发展。在您看来 AI 和人的关系，到底在哪些方面和过去的关系是完全不同的？哪些是最新的？

第二个你也提到关于未来的讨论，赵汀阳教授也讨论了未来已来。不过这样一个关于未来的解释涉及两个重大的不同，第一个在 AI 时代未来这个概念，从来都是跟时间连在一起，如果没有时间的概念，我们很难讨论未来。到底在今天的 AI 的时代什么样的时间观和时间概念出现，使得这个时代对于未来的感知和理解不同于过去关于未来的感知？尤其是十九世纪以后产生的时间和关于未来的讨论，未来在我们生活中的形态到底大概的模型是什么样的？

Stephen：谢谢这是一个非常好的问题，也是非常复杂的问题。我觉得就像所有大的问题，我们需要柏拉图有很多可以参考的东西。当然他也会引用一些苏格拉底的话，苏格拉底会特别担心技术和记忆等等这些。但是如果说大家不是特别熟悉苏格拉底，可能会考虑到需要用写字辅助记忆，书写是帮助我们去记，苏格拉底也有一些比较深入的关于人与技术之间关系的思考。他谈到一些神话，就是一个火神，而且还有很多其他的能力，除了火之外。

苏格拉底是这么说的，像火神有原始的神，还有其他兄弟，这个火神就给所有的生物一种能力，能让他们生存下去，比如老虎给它一些爪子，鹰给它双翼，鹿能让它跑得很快。但是对于人类就像一个小猴子一样，特别脆弱，又没什么可以给它的。普罗米修斯说你对人类做了一些什么，他说我要把知识包括艺术很多的能力等等，还有能力给到人类，还有技术，有能力改变世界的能力。所以这里的理念是什么？人类在他没有技术之前，是没办法完成他的使命的，我们就是一个特别脆弱的生物。所以对于人类来讲有一些原生的技术的天赋，对于研究人类起源的这些人，比如说两千三千年之前的人类，因为有了技术随着人类的发展，使用工具，大脑也在不断地发育，可能几百万年前用了水和布、狩猎等等这些，都在锻炼他的大脑。

因为我们使用工具，所以我们大脑在不断地发育，相较于其他的物种我们特别依赖工具，像老虎有锋利的牙齿，因为动物也是在演化进化，可能几十亿年，但是需要很长的时间。但是技术发展的特别快，所以我们没有关于人类静态的定义。最后一个问题关于 AI 的区别，这里面有两个问题解释了人们为什么对于 AI 感到很焦虑？一个是技术不断地在改变，人类也在不断地改变，速度在不断改变，在加速。我们可以用技术的表达，比如说五十万年之前，可能几千年几万年技术发展的很慢，接下来有了城市，有农业，很快这个阶段技术进展的特别快。现在在中国可以看到很多都是特别加速的在发展，很多的改变可能在几万年之前都没有观察到，这就是为什么我们经历了很多工业革命之后，速度就变得更加快了。过去的五十年当中，有了数字革命之后，有了更快的发展，这也是我们人类的一个部分，人类的改变是非常快的。

我们看这个奇点，到底什么是奇点？我们到底担心什么，是不是 AI 不断地爆发，制造很多

的机器，他们是不是开始夺回了主导权创造新的机器，突然之间技术的改变，我们觉得达到了这样一个指数的高点。但是现在它也创造了很多焦虑，因为人类开始变成了一个完全不一样的生物，很多东西是你没办法控制的，这就让人们感觉到特别恐怖。还有一个很担心的一点，技术创造出来是作为工具，有一部分我们是在使用，现在有了 AI，我们可能会创造出来一些有自己思想的东西，不是一个工具了，就像在电影里面说可能会离开你，这个对话框就离开了这个人。帮助他的这个人离开了他，所以我们现在回到了又是一个裸着身子这么脆弱的生命体。

现在看到技术发展的程度，可以帮助我们更好地了解为什么对于当前的发展很焦虑、很紧张。同时，我们也有很多很兴奋的对于 AI 感觉到很振奋的地方，因为它会给我们很多超能力，这就是为什么伦理非常非常重要，就像是弗洛伊德可悲的神，我们会变成神，我们希望能够成为好的神。

汪晖：谢谢！我记得萨特在世的时候，在晚年讲过关于文学的终结这个命题，他觉得文学到了他那个时期，所谓的现代文学基本上已经完结了，没有了。因为现代文学承担着完全不同以往的特殊功能，跟传统的文学非常不同，二十世纪中国可以看五四新文化运动，白话文运动，从几乎文学语言运动到社会政治运动，到科学技术的运动，都跟文学有非常深的关联。所以文学一定程度上不是今天作为一个学科的文学，文学是一个时代特殊的气质。每个时代当中有一种基调，文学一定程度上奠定一个时代的基调。萨特认为到今天这个时代，文学创作是持续的，但是那个意义的文学终结了。最近的二十年当中，科幻文学好像成了一个非常独特的门类，和过去的科幻文学有所不同。十九世纪到二十世纪的科幻文学是整个大的现代文学的一个有机部分，但是现在的科幻文学作为一个门类慢慢地变成了非常独特的，跟一般文学不太一样的东西。跟科技的发展，跟 AI 的发展有非常深的关联，同时跟教育也有非常深的关联。在你看来科幻文学的形态在历史上能不能给它一个定位？第二个如果它具有如此的重要性，到底最重要的功能是什么？如果比照当年萨特对现代文学意义的界定，在你看来科幻文学在今天是什么样的使命？

吴岩：谢谢汪老师，我的感觉今天我们所讨论的这些问题，在过去 50-70 年的时间里面，所有这些问题都出现在科幻作品里面，包括早晨说的左旋右旋这些，连这么小的问题，科幻里面全有。对我们来说，开今天的会怎么都在说过去的事。以人工智能为例子，今天的会上有各种各样的人工智能，我原来写过一篇文章至少有三种人工智能，第一种人工智能就是姚院士他们做的，曾老师他们做，就是科学技术的人工智能，这个人工智能每往前走一步和技术发展下一步的可能性相关。一点点地的生长，做出来，好，再往前推点，这和我们要不要人工智能一点关系没有，和我们需不需要它一点关系没有，这是他（技术专家）的兴趣，他的爱好。再加上今天背后的资本，一下就发展了，可能根本用不着。好多好多实验室做出来的东西说实在的，对我们来说没准是灾难。

这是第一种人工智能。

第二种人工智能今天很少有人谈，最开始姚院士谈了，政府和各个国际组织想要通过他们的力量去引导它的发展，所以他们做了好多政府的条例，这种后面是什么？后面是政府投资，所以他们拽着人工智能朝另一个方向走，但是这个方向和普通人的需求是不是一回事，可能有一部分是一回事。但是可能也不是一回事，这是第二种人工智能。

第三种人工智能是什么？就是我们今天非科技的学者们讲的人工智能，有很多误读，有很多跟发展没关系的，我们想象的。不单是我们，在这个意义上说，所有的普通人在网络上各个地方发表的各种言论都是第三种人工智能，这是人文的人工智能，而这一块人工智能里面包含着最丰富的内容。大量的这种人工智能是人真正需要的，这就牵扯到文科知识，到现在没有一个能力整理这些东西。

今天听了以后更强化了我想，文科是重要的，今天需要去处理我们过去所有的东西，所以文学不是结束，文学是一种新的使命开始。所有的人文知识是一种新的使命，我们要在这个基础上重新把过去所有的整理起来，并且让人发出更多的声音，这些才综合。

但是这三种人工智能，就第三种很难引导人工智能的真正发展。科学家的人文依靠的还是二、三十年前小时候上学的时候看的书。今天咱们开完会总结出来这些东西，拿给科学家根本不看，因为跟他无关。这就有这么一个时间差，所以今天文科面临最大的问题，就是怎么能够把知识，把我们得到的这些共识做到真正的发展里面，但是能不能做不知道。这是第一个问题，文学的重要性，以及所有的人文知识的重要性。

第二个问题我的感觉，今天面对着不确定性，我们太缺乏资源了，大家稍微想一点，到底什么才是资源？什么才能够帮我们在不确定的世界里面找着未来的发展。我 2017 年在南方科技大学成立了科学人类想象力研究中心，主要是研究这些不着边的，跟大家无关的东西，主要做想象力。这是我们做的关于计算想象学理论，既然智能可以计算，想象能不能计算，我们也做了一些公式。

我们做的两个工作，一个工作是和博古睿中心合作，做了一个不同文化里面怎么给人工智能叙事，中国的部分是我和三丰做的，从春秋战国一直到邓小平时代，三丰做的是邓小平以后一直到今天，就是中国的科幻作品里面，包括其他文学作品里面人工智能叙事是怎么样的，大家有兴趣可以看看，现在出了一本书是《Imagining AI: How the World Sees Intelligent Machines》。

我想讲的是第二个，最近我们正在做《西游记》的研究。大家都在做近期的未来，科幻作家到这觉得我们过去几十年把这事已经做完了。我们该做什么？我们做更远的未来，要不就是可能没未来，陈楸帆给了一条线人类灭亡，就没未来，这是一种情况。还有一种情况我们顺利地通过了这个阶段，通过数据的智能形式创造一个未来，实际上我个人感觉下一步，如果过了这个时段以后，会发现世界上所有的东西，实际上都可以赋予它智能，不需要通过电子数据的形式。

那样就会形成一个所有的东西都能够根据它的特性来发展人工智能，桌子有桌子的智能，瓶子有瓶子的智能，这样就真正出现曾毅说的无限智能体的社会，这种社会我们怎么理解，怎么去管理？《西游记》就是这样一个社会，对《西游记》各种各样过去的研究，什么宗教主体、成长主体、文化人文主体等等，但是给它增加一个新的主体就是智能世界的主体，西游记里讲的都是各种智能体，这种智能体基本运作的形式是怎么样的，可以看基本的质料是什么。一块石头经过风雨逐渐地获得了智能，所以是基本的物质材料和环境之间的相互作用，最后产生的智能。

对天地能量的吸收和摄取，表现的基本是什么，感知、谋划、变形——孙悟空还可以变形。基本的要求是什么？所有智能体的要求是寻求自由，同时向更高级转化，不断地修炼达到更高级，智能体之间的关系实际上有高低物种的差别，有的是低等的，有的是高等的。存在有没有道德？是有道德的，智能体的存在和生存具有道德性。但是有的智能体应该被规正，还有的是需要自我规范，帮助其他智能体正确扫除不正当智能体等等。这些都是我们曾经有过的对智能世界的关照，符合伦理道德，智能体的世界演化，有自然的法则，这些有兴趣大家可以看看。

智能体之间相互关系怎么样？智能体是寻求永生的，所有这些都在我们的著作里面。智能世界让宇宙变得更丰富，让其中的生物有更多的交往，让生命对世界的理解增强，让生命跨越死亡的焦虑，然后进入到涅槃境界，这是古代的著作里面谈到智能世界的时候是这样。这种研究，我就想比较一下它和阿西莫夫，阿西莫夫也提到三定律。他的这种想法和《西游记》的想法之间的差别是什么？我觉得阿西莫夫智能人的伦理学是人类中心主义的，甚至是奴隶制的，大家都讲过，机器人必须服从人，就是奴隶。《西游记》的万物有智的宇宙是非人类中心主义的，是具有更广泛自由度的，因为设置了信仰、道德等关键变量，因此人类也可以在之间发挥自己的特长，是多智能体中的一个，阿西莫夫的智能世界和人类之间的关系，当时他起了一个名字叫碳铁文明，今天就说碳硅文明，实际上属于这种文明观。这种智能世界是二元的，跟当前技术是绑定的，因此基本上是短视的，从科幻讲就是近未来的。但是西游记里万物有智的宇宙是多元的，是远离当今技术的，而且是长远的一种可能存在的范式。

回过头来讲，文学的作用，人文的作用实际上是提供我们今天面对世界各种各样的资源，这种资源不是没有，是我们没有从一个新的角度观察它。

汪晖：谢谢！梁正老师是从事 AI 治理的。今天上午姚老师提到风险的问题，治理当然跟 AI 时代的风险连在一起，他把风险分成了在我看有两个类型。一个类型是社会性的、法律的诸如此类，比如一个智能的车撞了人到底谁是责任人。法律的问题、社会的问题、失业的问题，这是一方面。另外一个部分是关于生存性风险，姚老师举了两个，一个是镜像生命，还有一个是 AI 失控。

这两个部分都内在于技术的发展，在治理过程中如何防范这个风险？在技术领域里面，当我

们谈论治理的时候，怎么去讨论这个问题？我想跟这个问题有关的是另外一个问题，上午穆勒教授所讲的，我们现在没有 AI 的 Ethics。这东西谁也不知道它在哪，它是在跟 AI 的互动过程当中生成的。这就对今天的治理过程的要求比一般的治理要复杂，必须深入到技术过程内部，跟它互动过程中才能产生有效的治理方式，请您谈谈从你们讲 AI 治理的过程中，如何面对这类的问题？

梁正： 特别感谢汪老师的问题，作为一个研究公共政策，研究治理的学者，今天的对话信息量非常大。对我们整个 AI 治理的认识可能会有一些本质性的启示，针对您刚才提的问题，简单把我们到现在的一些研究和思考跟大家分享一下。

2025 年 2 月份，两百多位科学家发了 AI Safety Report，曾毅教授还有很多教授都参加了。我们把它分成三类主要的风险。一类是刚才所说到的 Extremely risk，这种风险科学的研究有两种观点，一种是认为这样一些风险是技术系统本身发展必然会导致的，我们所说的不可控的问题。这个问题大到图灵奖和诺贝尔奖得主们专门写一篇文章讲，我们要不要这样的人工智能，这是一个很大的问题。

除了这样一类风险之外，您谈到的社会性风险里面，我们认为更大的问题不在于 AI 本身，是人类社会本身的缺陷所导致的，所以使用基于数据、基于拟合，基于概率预测，Token 进行对齐，这个东西反映出来的一定是把我们的缺陷放大。这个问题来讲不在技术本身，而在于我们的问题。我们现在所要建立护栏，要建立对齐的机制，其实是试图解决这一类的问题，也就是我们所说的社会性风险里面由于偏见，由于我们的数据，人类本身的缺陷所导致的。

第三类从这里稍微细分一点，就是您谈到失业的问题，这个问题更大，也不单纯由于技术本身的发展，更多在于技术和人类已有的社会结构的不匹配所导致的。所以历史上来讲，新的技术出现，我们经常举汽车的例子，汽车出现了，马车夫失业了，这样的问题从长期来看我们是没有办法预见未来，我们今天回到十八世纪想象不到今天的职业出现，科幻作家也没有想象到。这种问题在目前来看是阶段性的，怎么解决取决于我们和机器互动的关系。

这三类分析如果做一个区分，就像刚才您说的我们回到第二个问题，Müller 教授讲的，我们试图用一个 Universal 的伦理去解决这三类问题，从逻辑上来讲是不通的，这三类问题的解决方法是不一样的。对于 Extremely risk 的风险，现在科学界考虑的问题是要不要接受这样一种路线，因为我们在基因治理上，比如生物技术上，我们是有这样的所谓叫做 redline。现在所有的科学界都在讨论，如果我们不能控制所谓 10 的 25 次也好或者 27 次，在这样的算力背景下可能出现的一些现象，我们就到此为止。当 ChatGPT 刚出来的时候，这个叫“未来生存研究所”这样一批学者就是这么认为的，这个问题更多的和我们说的技术本身相关。

后面的两类问题很大程度上您刚才提到，我们的治理是取决于我们怎么和技术建立匹配的关系。

系。这种匹配的关系一方面可以给技术加上一定的围栏，加上一定的 Safeguard 机制。另一方面取决于我们调节现有的社会制度，加围栏就是我们现在讲因为这个内容可能会产生对年轻人、对儿童不利的内容，所以我们要有这样的限制，上个月教育部出台一个在教育里面要用这样的人工智能，要符合什么样的原则，当然也是基于联合国教科文组织前面一系列的 Guideline。这个是应该做的，而且只有这样做，现有的 AI 正面的作用是能够发挥出来的。所以尽管有不完善的地方，但是我们可以通过互相的调试，能够发挥出它正面的作用。

第二类问题比较大，就是失业的问题，带来的社会结构的冲击和影响也是刚才锦方学长讲到的对教育的影响，这个问题历史上可以看到前车之鉴，今天的教育体系和科研体系整体是基于工业文明，基于标准化的教育模式，基于组织化的科研方式。这也不是从有了这样的知识体系就是这样，回到希腊传统也不是这样生产知识，只是因为十八世纪有了工业革命，有了这样一些对世界的认识以后，能够提高我们的生产力，能够发挥出人的能力，创造更多的财富，所以我们把它叫做是一个求利的科学，就像培根说的知识就是力量。现在我们到了另外一个十字路口，这个力量大到已经超出我们的口头。我们这个社会要发生什么样的变化，教育现在做了很多治理研究，落到教育上来讲，未来的教育和工业时代的教育是一个根本性的变化，知识的整理器已经强大到十二年教育的知识实际上两年就已经足够了，从现在硅谷在做的创新性的学校。但是国企大量的教育是做道德的教育、社会的教育、劳动的教育、艺术的教育，这些教育在我们现有的教育体系里面或者十八世纪以来到现在的工业文明背景下的教育体系里面，是大大地被压制。

最后是个人的观点，我们现在出现这样一些认识上的问题是我们的教育没有真正能够适应现在智能时代的要求。刚才吴老师讲的智能体的世界非常宏大，像我们看到阿凡达这个世界一样，我也是一个科幻迷，除了三体以外，我也看《安德的游戏》，这里面思考的问题非常深入，但是我们现有的教育体系基本上不触及这些问题，这就导致现在很少的一批精英在谈的一些问题，要去解决这个宏大的社会的转型，这是远远不够的。而且很可能带来一个风险，像我们现在碰到的转基因的情况，没有社会启蒙的情况下，我们会把它妖魔化。转基因的技术有很多可以给我们带来福利，但是我们现在基本上不可能再去推动，因为社会的启蒙没有完成。所以现在最担心的是如果我们不去做社会的启蒙，只有精英去讲这个问题，最后我们没有办法处理人类自身的缺陷问题，我们光讲治理，在我们精英这个阶层去讲，甚至在大家非常高的群体去讲，其实是解决不了这个问题的。我就说到这，谢谢！

汪晖： 我们还有 35 分钟的时间，我想开放给所有在场的参与者提出一些问题，你们对几位嘉宾的发言，或者你们有什么样的问题希望提出来的，尽可能地准确明确，不要太长，谢谢！

听众 1: 感谢各位老师的精彩分享。大模型目前是通往 AGI 的可能途径，以大模型为代表的 AGI 或者 ASI 这种超级智能，让人类更团结还是更分散？如何看待大模型对世界上各种语言和各种跨文化传播的影响和意义，谢谢！

吴岩: 我先回答一句，从现在中美关系看就是让人分散。

赵鼎新: 信息再多，解读不一样。并不是信息多了大家思想就会集中起来，哪一种解读变多，完全不在于解读正不正确，而在于背后有多少人支持，比如说我再伟大，把自己吹得很厉害也成了大教主，但是我搞一个教会，变成国家的意识形态，再出理论大家都会听。这时候实际上大家形成共识和形成分散，不在意识形态本身的性质，不在信息多少，是背后权力的分布问题。像中美现在分散，是中美的权力力量的对比，而不是哪个思想的正确共识的对比。形成共识已经是个错误的认识了——认为世界本身应该有共识。共识不应该有，社会要存在需要一定共识，需要一定程度的教化，主流价值观，为了降低统治成本，降低生活成本那是需要有。有共识的肯定要解释，没共识是正常的。

孙周兴: 总的意义上讲人类当然是交流更加方便，这大概是马克思所说的普遍交往终于实现了，这是毫无疑问的。别的比如真的哪里打仗了，这个跟我们这个判断是冲突的，但又不冲突的，这是两个层面。

自然语言正在消亡，人类大概在未来的二十年、三十年可以想像，大概只会留下五六种语言，大部分语言被同化、被消亡，不再有存在。

听众 2: 这个问题来自于今天赵汀阳老师提的，二元主体性的问题。想问梁老师，如果真的承认了二元主体性，接下来人和 AI 的治理和社会关系，您设想一下会是什么样的？

梁正: 我简单回答一下，我不会上升到哲学层面，因为我做科技创新政策。最有可能的一个情境还是回到物质极大丰富的条件下，如果能够解决分配问题，就可以自由发展。我想这也是作为一个社会思想深刻的启蒙者所看到的未来，现在有这种可能。另外一方面，我也同意刚才不管是赵老师还是孙老师说的，这种可能性非常小。由于大家的 Value 不一样，我们所处的立场不一样，我们基本上达不成共识。我们已经没有办法解决这个问题，需要新的人，也许脑机融合的这一代新人有可能解决。我现在想不到不经过智力增强和价值提升现有的物种能够解决这个问题。财富生产的问题我觉得能解决，但是分配问题，利益冲突问题解决不了，这个共识达不成，文明就要

产生冲突。

听众 3: 刚才吴老师所举的例子，比如我们去看《西游记》，或者我们说桌子有桌子的智能，杯子有杯子的智能，展望这样未来的时候，我们过去也有一个和他非常相似的词，叫做万物有灵，吴老师用的是万物有智，这里面产生了一个非常有意思的回环。当我们非常努力地畅想未来智能愿景的时候会回到一个相当传统的思路里面寻找借鉴，在二者绘制的圆圈里面，一个很关键的一点，人类作为一种理解能力有限的智能体，在面对非常复杂的自然和科技现象的时候，他的选择。在这个意义上，当人类在现在的历史关头去展望所谓的超级人工智能的时候，有点像是我们回到初民社会时候的人，他们去展望非常未知、非常神秘的自然所展现出来的体。在这里我想请教吴老师和各位老师的一点，我们站在这样的历史关头去展望未来，未来更可能是一个走向启蒙，走向一个各种知识都变得更加清晰的路径，还是回到一个更加神秘化，万物都是在神秘的现象里面展望各种可能性的一种路径？谢谢！

吴岩: 谢谢，我先回答，第一个你刚才提到启蒙，这个词已经老了。以人为主体的一系列东西现在在一个新的时代可能是要被边缘化了。

第二，也不是说非得回去，只是说当我们面对这种高速变化的时候，我们没有资源去判断，所有的这些东西都可以成为一种资源。科幻的意义是什么？他告诉你一百个未来，以便你明天去面对第 101 个未来，这个时候你不觉得这种未来怕人，因为你已经熟悉过这么多。

尤瓦尔·赫拉利讲七万年前是第一次认知革命，第一次认知革命最重要的内容是什么？就是人类学会了虚构，一旦学会了虚构就可以讲故事，可以做展望，可以讲未来等等。第一次认知革命完了以后文学出现了什么，就是神话。但是今天发现科学的推理已经不能够给我们带来明天到底是什么样，所以现在又返回一个神话的时代，我们要重新把人和自然串联起来，用一种独特的方式，用每个人独特的方式串联起来。所以会不会万物有灵，那是很可能的，这是我的回答。

孙周兴: 这个问题特别有意思，可以从两个层面来说，一个数字技术真正构造的数字世界特别符合我们对宗教的想象，因为不可知，巨大无比，神秘，无所不在，你不知道他怎么发生的，是很神秘的一个世界。如果还有未来宗教，可能就是一个数字世界的宗教，这是一点。

另外一点长期以来，从十九世纪中期以后一直有神秘化的主张，特别是在艺术这条线上，哲学上也是，艺术这条线一直到今天的当代艺术，一直在观察一个艺术神话的理想。一说神话就是所有的，不光是今天的数字技术，包括一直以来的现代主义技术。这个问题太有意思了。但是我无法告诉你更多的东西，可以从两个面上联想。

赵鼎新：这个问题还要从社会学家，社会学家比较现实，往回拉一拉，首先任何技术产生的现实背景，这是 AI 产生正好是在全世界启蒙思想全规模垮台，世界全部是在保守主义，民族主义力量不断增强，而且民族主义变成个人的民族主义，是不一样的。

第二你们讲宗教，我说不会有什么 AI 宗教，现在随着世俗意识形态的垮台之后，老百姓自发去找宗教，找着找着哪个宗教的制度基础好，他占据人多，和 AI 什么，和知识分子没有关系。大家东正教、基督教、伊斯兰教、印度教自发的就起来的，以后的宗教力量没什么玄乎就这些东西，这些东西当然会和 AI 结合，这是另外一码事。还是要拉回现实看，不能凭空讨论这些问题，你刚才的问题在误导我们。

Stephen：我非常高兴听到这种关于神话的借鉴，神话是人类理解世界的一种方式，比如说未来是充满了虚幻，还是一个清晰的路径？如果我们这么做会产生什么样的结果？或者那样做会产生什么样的结果，或者我们能不能穿过这样的迷雾看清未来，现在中国的神话和希腊的神话都有很多的故事。现在有新的迷雾，就是技术所造成的迷雾，所有这些东西比如终结者等等这些，都是关于时间的迷雾，还有错配的情况。所以现在我们要做的就是打造一种清晰的路径，如果我们认为未来是充满了迷雾，就需要解决这样的问题，来打造一种清晰的路径穿过迷雾，这样未来就会变得更加清晰。关键是你认为未来是清晰的还是迷雾的，这是在于你自己的判断。

听众 4：我有一个从基层来的问题，上个月我给县级中学做科普的时候，他们的校长表达了一个焦虑，就是现在新的“读书无用论”正在抬头。有两个原因，第一个现在的 AI 技术这么发达，学生很容易就从 AI 技术获取很多的资料，但是这些资料老师都未必有，所以他们的学生就说现在跟着老师学 12 年，还不如到 AI 上查一下，得到的答案更丰富更多，所以为什么还要这么苦哈哈的学习，这是第一。

第二随着 AI 的进步，舆论铺天盖地说有很多人要失业了。学生又带来一个问题：我去考大学肯定要就业，但是现在很多人已经就业等于失业了，AI 再普及，前途更堪忧，我为什么还要学习？所以两个方面就导致现在新的读书无用论在抬头，这是中学校长面临的一个焦虑。请诸位老师能给解答一下。

赵鼎新：大学教育面临一模一样的问题，我个人体验是这样子，目前小学、中学、大学教育肯定要变，变的是什么？对我来说网上的资料不是知识，就是一大堆信息而已。但是我发觉我的提示词就比我的学生厉害，我们大家一起比较，我有判断力，我有常识感。同样，厉害的艺术

也比别人玩得好，因为他们有判断力，有原创性。所以 AI 一来，不是教育无用，而是教育要改变。信息类教育越来越少了，可以自己查。但是物理学查查就能懂吗？给你一道题你能解吗？刚才大家问的问题里面不知有多少小问题，一大堆乱七八糟的信息，这一下就说明中国文科教育出了问题，不给人以常识感，不给人以判断力。所以以后是基础教育大爆发的时候，常识感教育大爆发的时候，判断力教育大爆发的时候。

听众 5：失业这个问题在 AI 到来的时候肯定是不可避免的，但是可以分析一下失业的问题，曾经是农民和工人大量失业。但是这一次 AI 带来的失业更多冲击中产阶级，庞大的中产阶级是一个社会稳定的基石。现在 AI 带来对中产阶级大量的冲击会给社会带来什么样的状态？

第二也是教育的问题，失业一定会让很多学生觉得，我学习了大量知识，但是很多工种已经被取代了，作为一个自适应的调整，我会不会想有一些不能取代的工作，比如说我去踢球，做演员或者做护理。这些相对来说不需要那么多知识含量，但是却是不可被取代，但是另外一些人可能是天才型的人物，他就要去做更高端的学习，就要和 AI 对话，和 AI 和谐发展。现在在物质上的两极分化，将来会不会出现智力上的两极分化？谢谢！

杨锦方：我尝试回答一下，第一个关于失业的问题，我对这个问题的思考，这是一个巨大的社会问题，根源在于节奏，因为人工智能的技术发展太快了，人类没有准备好进入人工智能大规模取代人类工作的阶段。我觉得这次的取代有一个跟之前的职业取代很根本的差异，在于劳动力在价值创造过程当中被边缘化了，资本 + 人工智能技术就够了，就可以完成价值创造的循环，创造更多资本，可以创造更多人工智能，如此正反馈下去，前景的确是比较可怕。

如果有这么多的人陷入失业，生活可能失去目标感，甚至失去人生意义感的时候会造成巨大的社会问题，会影响社会稳定，一定会带来对政府来讲是不可容忍的局面，最后对政府来说，以及那些掌握资本，掌握权力的人来讲必须要做出让步，只是说这个冲突和再平衡有多剧烈，以及多长时间才能达成再平衡，这个事情是有不确定性的。我认为在历史上这些事情发生过很多次，资源再分配和权力的再平衡，我们希望人类能够吸取历史教训，不要再发生大规模的战争、饥荒、瘟疫。

更长时间没有那么悲观，因为人类会找到自己新的生存意义，这个意义不一定是工作，只要做自己有意义感的事情，能够在整个社会形成一个有相互的正向价值的一个机制，我认为人类是可以生活的比现在更轻松、更愉悦，更有幸福感。这是对失业的看法。

对第二个问题，我认为这个可能性也相当大，资本的力量会使得富人和普通人产生巨大的差距，AI 增强人，有些富人可以借助 AI 的力量，让自己智力变得非常非常强大，普通人可能没有

条件做这件事情，所以我认为这个可能性是相当大的，包括长寿的问题，我认为将来也会出现非常长寿的富人群体，这个事情也会带来社会问题。

汪晖：最后一个问题。

听众 6：各位老师好，我是哲学专业的研究生，我就问一个比较简短的问题，在 AI 这么强大的背景下，各位老师觉得什么样的研究是比较有价值的？

Stephen：想象一下 100 年前，当 95% 的人都从事农业。想象一下他们被告知，没有学习务农的必要了，因为多数人都会因为工业化而被替代。只需要 5% 的人务农，其余的人都要去做点别的。他们可能会问，干点什么别的呢？你可能会告诉他们，可以做宇航员、软件工程师或者网红。然后他们反应会是“什么？”所以我们无法真的想象。所以 50 年后会有什么样的新工作，或者按照杨锦方先生所说，寿命延长了，100 年后的生活是什么样的，现在都很难说。

至于应该学习什么，我认为应该保持灵活。不仅要想现在想做什么，也要设想 25 年后，50 年后如果你还能回到大学学习，你对什么感兴趣。我认为我们可能要开始适应用一种不断循环的方式生活，不再是在漫长人生的前 20 年接受教育，然后期待它在接下去的 80 年都能发挥作用。我们可能要重塑社会模型，让我们接受教育然后去工作，然后可以提前退休或者休假再去学习，循环往复。用这个方式，也许未来你可以成为宇航员。

赵鼎新：我不知道什么研究有意思，但是我肯定知道什么研究是没有意思的，比如说你的研究是为了拿正教授快，拿长聘快，拿到课题，拿长江、拿资深、拿头衔，长期来看这研究多半没有意义。但是在研究方面，有内部的驱动力，是你独特的观点，有意义的可能性就会大一些。

孙周兴：在这里只有我搞哲学的，我补充两句。尼采早就说过一句话叫做，哲学家告诉我们需要什么，艺术家把它创造出来，艺术与哲学被看成是人类文明当中最基本的势力，这句话提示了我们哲学还有什么用。因为这个文明最需要的是这样一种力量，我们到底需要什么？我们到底在何种程度上需要技术？技术应该到哪就 OK 了？还提示我们一点，就是艺术与哲学总是在一起，真正的哲学研究必须有有艺术性的、有创意的，面向今天的文明、未来的文明。

汪晖：今天的讨论基本上到这就结束了，刚才那位同学对我也提一个问题，核心的问题在于追问，当刚才赵老师说的那些目标都不能构成意义的情况下，追问本身依旧是一个意义。追问本身非常有意义，刚才说启蒙已经完全的过去了，启蒙里面有一个问题值得讨论，就是无条件的追

问，不断地追寻这样的基础性的问题，像今天追问 AI 时代的伦理意义，就是追问当中产生出来的，持续探索。

第二刚才涉及到几个学科的问题，可以看到在今天意义的匮乏跟学科本身的发展同步，不断地产生出学科的壁垒，使得人们对于意义的追寻在这个过程中逐渐地消失。孙老师刚才讲的哲学和艺术的打通之间的关系，其他几个领域都在这个过程中。所以所谓的无条件追问，在这个意义上是持续打破壁垒的过程，寻求新的综合，对我们人类境况的再理解。

谢谢各位老师对方塘研究院未来问题的讨论，再次感谢！我们这场讨论到此结束！



几个值得思考的根本性问题

2025 年 6 月 9 日

清华大学方塘研究院

议题一：我们试图处理的、作为一个具有威胁性的对象出现的 AGI/ASI，究竟该如何描述和理解？——如何科学地定义人工通用智能（AGI）和人工超级智能（ASI）？AGI/ASI 是否真的会出现？何时出现？

人工通用智能（AGI）和人工超级智能（ASI）是人工智能领域最具争议性和前瞻性的概念。AGI 通常被定义为具备与人类相当或超越人类水平的通用认知能力的人工智能，能够在广泛的任务中表现出灵活性和适应性。而 ASI 则更进一步，被设想为在所有认知领域均远超人类智能的存在。然而，这些定义本身充满了模糊性和争议。

关于 AGI/ASI 是否可能实现，学术界也存在显著分歧。一些哲学家从认识论的角度出发，认为 AGI/ASI 的实现存在根本性障碍。例如，他们质疑机器是否能够真正理解意义、拥有意识或具备人类的创造性思维。另一方面，部分技术专家则认为，随着计算能力的提升、算法的优化以及对人类大脑工作机制的深入研究，AGI/ASI 的实现是有可能的。尽管目前仍面临诸多技术挑战（如如何实现真正的自主学习、如何处理复杂环境中的不确定性等），但一些专家预测，在几十年到百年的时间内，这些障碍可能被逐步克服。

议题二：对 AI 的恐惧是否必要？是什么导致了人类对人工智能（AI）的恐惧？人类恐惧的究竟是什么？

一种观点认为，人类对 AI 的恐惧有夸大之嫌，是受到主流叙事和文化意识形态的影响。历史上，技术与人文的对立叙事屡见不鲜，从工业革命到互联网时代，人类始终对新技术抱有戒心。这种叙事将机器演变与人类存在之间的冲突视为必然，并将一切问题归因于技术本身。支持者认为，正如人类曾对互联网的出现感到恐慌，但最终适应并受益于它一样，AI 的恐惧也可能随着时间的推移而消散。

另一种反思性的观点指出，人类对 AI 的恐惧实际上是对自身的不信任。人类对其他同类的不信任和冲突是恐惧的根源，而非 AI 本身。如果人类无法处理好自身的问题，即使没有 AI，恐惧和冲突依然存在。因此，AI 只是人类恐惧心理的一种投射，而非真正的威胁对象。例如，一种观点认为，从历史上看，技术发展往往被统治阶级接管，并用于统治阶级自身的强化，而损害了其他所有人的福祉，因此，对人工智能的恐惧很大程度上可以归因于人们对又一种可能被用来剥削和奴役他们的技术手段感到警惕。

还有一种乐观的观点认为，如果人工通用智能（AGI）或人工超级智能（ASI）真的实现，它们不仅会具备极高的智能水平，还可能表现出更高的道德和善良属性。这种观点将 AGI 与人类的关系类比为人类与大猩猩的关系，认为 AGI 不会与人类争夺资源，而是会以更高的智慧照顾人类。人类关心的问题对 AGI 来说可能微不足道，因此人类无需过度恐惧。

更主流的观点认为，AI 作为一种“非有机异质智能体”（an alien non-organic agent），其智能属性与人类智能截然不同。这种智能体的决策过程往往超越人类的思维框架，甚至能够产生人类心智无法自发形成的创新理念。然而，AI 的“黑箱”属性——即其决策过程的不可解释性和不可预测性——加剧了人类对其失控的担忧。如果 AI 无法被人类完全理解或控制，它可能被视为一种“新物种”，进而对人类的存在构成潜在威胁。即便这种威胁的概率极低，由于其可能带来的巨大破坏性影响，人类也必须尽快采取积极应对措施。

议题三：人类能控制 / 引导 AI 吗？能造出“有道德”的机器人吗？——理论

上有可能性吗？现有的几个技术方案可靠吗？

不少技术专家认为，无论 AI 是否具有意识，我们需要确保我们能够控制人工智能，并引导其成为推动社会进步的力量。目前有如下一些尝试：

为了避免 AI 脱离人类控制，Yoshua Bengio 及其合作研究者致力于开发一种非智能体（non-agentic）AI 系统——Scientist AI。它被设计为非智能体，没有自主行动能力和目标导向性，由世界模型和问答推理机器两大部分组成。世界模型用于生成解释数据的理论，问答推理机器基于这些理论回答问题，并且这两个部分都融入了不确定性概念，以降低过于自信预测带来的风险。其非智能体性设计确保人类始终掌握最终决策权，不确定性管理则为权衡风险与收益提供量化依据，重新定义了人机协作的伦理边界。

为了引导 AI，有专家提出塑造具有“功能性道德”（Functional Morality）的机器，“我们不是要造出有良心的机器，而是能识别道德问题、提示风险的系统”（Wendell Wallach）。

也有专家倡导“为机器立心”、塑造“有道德的机器人”，通过价值对齐（Value Alignment），确保人工智能系统的目标、决策和行为与人类的价值观、道德原则和社会利益相一致。目前，价值对齐的技术路径包括：监督学习与人类反馈（通过人类监督和反馈，训练 AI 系统识别和遵循人类价值观）、逆强化学习（通过观察人类行为，推断其背后的价值观和目标，并将这些价值观嵌入 AI 系统）、可解释性与透明度（提高 AI 系统的可解释性，使人类能够理解其决策过程，从而更好地监督和调整其行为）、道德框架与约束机制（为 AI 系统设计道德框架和约束机制，例如设定“硬性规则”如不得伤害人类或“软性约束”如优先考虑人类福祉）等。但价值对齐也遇到了哲学上的挑战，如：人类价值观具有高度的多样性和复杂性，不同文化、社会、群体甚至个体之间的价值观可能存在显著差异；人类价值观并非静态，而是随着时间、环境和社会变迁而不断演变。此外，价值观在不同情境下可能表现出不同的优先级；人类价值观往往涉及抽象概念（如正义、自由、尊严），而 AI 系统需要明确的、可量化的规则和目标函数；不同的伦理学体系（如功利主义、义务论、美德伦理学等）对“正确行为”的定义存在根本性分歧等等。

因此，问题的核心是，如果控制和引导 AI 成为我们应对 AI 挑战的基本方向，那么从理论和现有的几个实践路径上，是否可能、可行、可靠？

议题四：当人们讨论“AI 伦理”时，人们究竟在讨论什么？

“AI 伦理”作为一个被广泛提及的概念，却也存在着极大的模糊性。一些观点认为 AI 伦理讨论的是关于 AI 的伦理（例如 AI 是道德主体还是道德客体，赋予 AI 以道德责任是否正当，AI 应该具备怎样的道德，AI 与人类之间是否存在伦理关系等等）；一些观点则认为 AI 伦理讨论的是人在开发和使用 AI 过程中的伦理（主要围绕人类的责任，重点在于如何确保 AI 的安全性、透明性、公平性等，涉及到 AI 治理、监管层面）；还有一些学者触及人与机器的关系，在开发人工智能和其他人造智能体过程中所涉及的伦理考量，尤其是当这些实体被赋予了感知疼痛或痛苦能力时……。

围绕“AI 伦理”的讨论，是否需要根据 AI 发展的不同情境和不同阶段展开，以使得我们可以更深入地理解 AI 伦理问题的复杂性，并为人类安全可持续发展提供理论支持和实践指导？例如，在初级阶段（弱 AI），AI 伦理问题主要围绕人类的责任。在中级阶段（强 AI/AGI），AI 伦理问题可能涉及 AI 的主体性和责任，同时需要更复杂的治理框架和技术手段。在高级阶段（ASI），AI 伦理问题可能涉及人类与 AI 之间的共存关系，AI 与 AI 之间的关系，等。

议题五：面对 AI 时代的各种不确定性和前所未有的现象，我们最大的挑战似乎首先是如何精准地描述、把握问题本身，进而才是如何有效地分析、解决问题。那么，我们需要澄清哪些容易混淆的概念以使得沟通成为可能？现有的概念资源是否够用、是否需要创造新的概念武器？在哲学的方法甚至是哲学的气质上，又需要哪些突破？

- 有待概念工程处理的概念有哪些？智能、意识、理解、意义、主体、自主性……？

Capitalism, exploitation, suffering, profiteering, propaganda, politics,...

- 超越人类中心主义偏见（Beyond Human-Centric Bias）：需要摆脱根植于人类经验的传统概念框架，以更开放的视角理解 AI（特别是 ASI）的“异质性”？

- 动态系统思维（Dynamic Systems Thinking）：需要从静态系统思维转向动态系统思维，

- 以应对 AI 系统的自适应性和复杂性？
- 预测性与前瞻性分析（Predictive and Proactive Analysis）：需要从处理已有现象转向预测未来趋势，以更好地应对 AI 时代的不确定性？
 - 跨学科合作（Interdisciplinary Collaboration）：哲学需要融合技术、社会学等多学科视角，以全面理解 AI 时代的复杂问题？

议题六：AI 时代人类如何安身立命？我们需要怎样的未来图景？

- AI 时代，人的宇宙地位是什么？（本体论层面）（人与机器的关系，当前以及未来）：
- 人的宇宙 / 存在地位高于 AI，还是低于 AI 还是平等于 AI？（人类与 AI 关系的隐喻，是人与工具？人与动物？人与外星人？人与人？人与神？或其他）
 - 未来的人类文明样态可能会怎样，哪种情况是可以为之努力的（具有实现的可能性和可行性的）？
 - ruled by AI？（不少政府、大企业、资本希望 ruled by AI，AI 变成治理、剥削的工具）
 - life of AI？（即人与机器共处，通过合理的分配达到人类整体福祉的提升）
 - 人类优雅地离场(即ai作为人类文明发展的后裔逐渐地取代人类,不给人类带来痛苦),等等。
 - 在机器的新坐标面前，如何定义 "人"？（人的哪些属性 / 德性变得格外重要？哪些又变得不再重要？）

- AI 时代，人类活着的意义是什么？人类如何体面地活着？（人生观、价值观层面）（AI 会如何重塑人类生活的方方面面？会对现有的社会秩序和伦理传统产生怎样的影响？又该如何应对）
- 如果 AGI 承担了大量工作和创造性活动，更多地人可能陷入更加强烈的存在主义危机，如何摆脱存在的虚无感？
 - 当人可以实现“数字永生”，什么是“生”，什么又是“死”？“生”、“死”的意义是什么？（人是否需要“被遗忘的权利”、“死的权利”，“数字永生”的权利归属于谁，繁衍的意义是什么？）
 - AI 会如何影响人类的生活？（人伦关系、社会结构、权威来源、等）
 - 如果 AI 可以提供情感寄托和不间断的陪伴，人类是否还需要和人相处？如果传统意义上的亲密关系和繁衍都可被替代，那么以家庭作为基本单元的传统社会结构是否会动摇？
 - 如果人类的知识和经验从 AI 获取，师长的权威是否还存在？长幼尊卑是否还存在？

- 原有形成身份认同的元素是否会被新的要素瓦解（如民族、种族、性别、国别等，是否会被增强人、智能体在同一价值下的聚合等元素替代），新的认同如何形成，相应的权力和权利结构会发生怎样的变化，等等。
- AI 时代，人类如何认识世界？（认识论）
- AI 生成逼真的文字、图像、视频能力可能制造虚假信息（如深度伪造），如何去伪存真？
 - 我们如何理解“真实性”？（如果 AI 展现出的“真实”与传统知识途径中相信的“真实”发生矛盾和冲突（AI 可能有意或无意造成这个结果），传统的真实如何自证？此时 AI 定义的真实（声量大、普及面广）和非 AI 的传统真实，哪个才是真正的真实？）
 - AI 是否会改变或拓展我们对时空的认识，等等。

议题七：AI 时代的社会结构与政治经济秩序如何演变？

- AI 技术的发展正在对社会结构和生产关系产生根本性重组，原有的权力结构、阶级组织、代表机制、价值分配体系，正在被技术逻辑所扰动、取代甚至重写。
- 当劳动不再是价值的唯一来源，AI 生成的“非人劳动”将如何纳入经济与政治的分配秩序？资本、技术、算法是否正在形成新的统治形式？它们如何挑战现代国家的合法性与权威性？
 - 政治代表机制是否正在失效？平台治理是否预示一种“后国家秩序”？去中心化技术是否可能成为新的“去政治化霸权”？
 - 如果伦理不再与生产、制度、权力发生真实关联，它是否只是一种道德幻象？是否有可能重新建立一种“渗透于社会关系之中的伦理”——如古典儒家之礼，或亚里士多德式的德性伦理？

