从单智慧体社会到多智慧体社会：人工智能与人类的共同未来

From a Homo-intelligence Society to a Hetero-intelligence Society: The Common Future of Artificial Intelligence and Human Beings

王晓阳 /WANG Xiaoyang　　马迅 /MA Xun

（厦门大学哲学系，福建厦门，361005）
(Department of Philosophy, Xiamen University, Xiamen, Fujian, 361005)

摘　要：本文首先剖析了限制人工智能发展的几个瓶颈问题，然后论证了这些问题并非原则上无解，对于人工智能的发展并不能构成实质性的阻碍。其次，基于“人机区分难题”论证了以ChatGPT为代表的通用人工智能原则上何以不可能，同时借助意识的功能主义理论与多重实现论题，论证了人工智能具有意识特征的理论前提。最后，对人工智能与人类未来的可能发展情形和潜在风险进行预判，区分了两种可能发展情形——“单智慧体社会”与“多智慧体社会”，并且，借助演化论和博弈论的有关理论资源，论证了实现一种人类与人工智能“多智慧体互生共存型的和谐社会”何以可能。

关键词：人工智能　ChatGPT　人机区分难题　多智慧体社会　人工智能哲学

Abstract: This paper first analyzes several bottleneck problems that limit the development of artificial intelligence (AI), and then argues that these problems are not unsolvable in principle and do not constitute a substantial obstacle to the development of AI. Second, based on the “human-machine distinction puzzle”, we argue that artificial general intelligence (AGI) represented by ChatGPT is in principle impossible, and based on the theory of functionalism and the multiple realizability thesis, we argue that AI can be conscious in principle. Finally, the possible future development scenarios and potential risks of AI and human beings are predicted, and two possible development scenarios–“homo-intelligence society” and “hetero-intelligence society” – are distinguished. In addition, with the help of the theoretical resources of evolution theory and game theory, it is demonstrated how it is possible to realize “a harmonious society with hetero intelligences coexisting with each other” between human beings and AI.

Key Words: Artificial intelligence; ChatGPT; Human-machine distinction puzzle; Hetero-intelligence society; Philosophy of artificial intelligence

中图分类号：TP18　　**文献标识码：**A　　DOI: 10.15994/j.1000–0763.2023.11.003

近来人工智能领域最为轰动的成果莫过于OpenAI公司于2022年11月推出的人工智能自然语言处理工具ChatGPT。ChatGPT的升级版（GPT–4）已经能够生成与人类产生的文本相比

基金项目：国家社科基金一般项目“他心问题的基础理论研究”（项目编号：20BZX030）；国家社科基金重大项目“人工认知对自然认知挑战的哲学研究”（项目编号：21&ZD061）。
收稿日期：2023年7月4日
作者简介：王晓阳（1978–）男，江苏泰州人，厦门大学哲学系教授，美国耶鲁大学哲学系客座研究员，研究方向为分析传统的心灵哲学、形而上学、语言哲学、科学哲学、认知科学哲学、人工智能哲学。Email: wxy2018@xmu.edu.cn
马　迅（1994–）男，安徽亳州人，厦门大学哲学系博士研究生，研究领域为科学技术哲学。Email: xunma@stu.xmu.edu.cn

难以区分的复杂文本，且适用环境非常广泛，甚至能够根据上下文理解人类的语言，与人类进行复杂的对话互动，甚至模拟出人与人之间的实际交流过程。并且随着ChatGPT的推出，全球各大互联网公司也纷纷开始布局智能语言模型。例如，2023年2月6日，谷歌推出了大型智能语言模型“Bard”，并正式开放测试。2023年3月16日，百度也发布了其生成式人工智能语言模型“文心一言”。这些语言模型在某些特定智能领域已经展现出了接近甚至超越人类的能力。随着ChatGPT的火爆，也再次引发了对于人工智能的前景及其潜在风险的担忧。需要注意的是，虽然ChatGPT具备一定的智能水平，然而有理由相信，它尚未达到通用人工智能（Artificial General Intelligence，AGI）的水准。ChatGPT只是一种人工智能模型，旨在对文本输入生成类人的输出回应。ChatGPT是通过对大量文本语料库进行训练，以学习语言的结构和语法规则，从而生成类似人类的语言输出。ChatGPT基于Transformer架构。该架构通过将输入序列转换为一个固定长度的向量来表示，然后将该向量作为下一阶段的输入。在Transformer架构中，每个输入都被映射到一个向量，然后这些向量被传递到下一层。在训练期间，模型学习如何根据当前输入序列生成下一个输出，从而生成通顺流畅的文本输出。ChatGPT的训练过程使用了大量的计算资源和数据，训练过程中使用的数据集是不断更新的。训练过程中使用的数据集包含了大量的对话和文本数据，以便让模型更好地理解人类语言的模式和语法规则。训练过程中还使用了强化学习和监督学习等多种技术，以便让模型能够生成高质量的文本输出。总之，ChatGPT的实现是通过学习大量语言规则和对话模式，以及对大量文本数据进行训练，以生成类似人类的语言输出。

然而，AGI则是人工智能发展的某种目标，旨在制造出一种通用智能系统，能够像人类一样执行各种任务、学习各种知识和适应各种环境。因而是一个广泛的机器智能概念，包括执行广泛智力任务的能力。相比于ChatGPT只专注于对文本输入生成类人语言的响应，AGI则显得更加宏大和雄心勃勃。满足AGI要求的通用智能系统涵盖了人类智力和认知能力的所有方面，不仅能够理解语言和生成响应，还需要具备感知能力、运动控制和其他高级认知功能。因此，要达到AGI这一目标，就需要多学科领域的合作，包括计算机科学、神经科学、心理学、哲学等。

总之，ChatGPT和AGI是两个完全不同层面的概念。简言之，ChatGPT是一个现成的人工智能系统，而AGI则是人工智能的某种发展目标。ChatGPT是一种专注于自然语言处理的人工智能模型，而AGI是一个更广泛的概念，旨在实现人类智能的方方面面①。

这里或许有人会进一步问：即便ChatGPT不同于AGI，但是ChatGPT的出现，是否预示了AGI这一目标的确有望实现？此外，美国著名未来学家库兹韦尔（Ray Kurzweil）在《奇点临近》一书中基于呈指数发展的技术，曾预言说2045年将是人工智能发展史上奇点来临的时刻。[1]奇点意味着人类的生物思想与现存技术融合的顶点，这将导致人类超越自身的生物局限性，同时人机文明也将超越人脑的限制。这意味着，超级人工智能（Super AI）将会出现②。

未来将要出现的人工智能究竟会是怎样的？我们又将如何应对？有理由相信，对于正处于人工智能迅猛发展时代中的人们而言，这两个问题都是“绕不开的”。在下文中，我们将依次探讨这两个问题。我们打算先探讨未来可能出现哪些类型的AI，然后讨论应当采取

①两点说明：（1）上述是关于AGI的传统理解。根据这个理解，ChatGPT或其升级版本GPT–4并不算是某种满足AGI的通用智能系统。OpenAI公司的官方亦给有类似看法，并且认为“AGI的时间表仍不确定”，参看URL: https://openai.com/charter;（2）有理由相信，AGI这一目标原则上并不能实现。理由就是，面临一个难以克服的“人机区分难题”。下面很快会有相关论述。

②根据不同人工智能系统的设计目标，可以区分以下不同的人工智能类型：弱人工智能（Weak AI）、强人工智能（Strong AI）或通用人工智能（AGI）、超级人工智能（Super AI）。

怎样的应对策略。本文主要思路如下：在下一节里，我们先介绍人工智能的一些基本原理，及其面临的三类主要质疑，并对三类主要质疑一一加以回应；在第二节的前半部分，我们将提出“人机区分难题（human–machine distinction puzzle）”才是AGI面临的真正威胁，并论证AGI原则上是不可能的。在第二节后半部分，我们将进一步论证，虽然AGI是不可能的，但是Super AI则是有可能出现的。在第三节里，我们将先区分“‘单智慧体社会’（homo–intelligence society）与‘多智慧体社会’（hetero–intelligence society）”这两种未来可能出现的社会类型。然后借助演化论和博弈论的有关理论资源来论证，达成一种“多智慧体互生共存型的和谐社会”何以可能。最后，我们将做出总结，并给出最终结论。

一、有关人工智能的争论：质疑与回应

自人工智能诞生以来，学界有关人工智能的质疑不绝于耳。其中，对于弱人工智能（Weak AI）的实现都能达成一致的观点（事实上，Weak AI被认为已经实现。如ChatGPT）。因此，这些质疑几乎针对的都是AGI或Super AI。概而言之，这些质疑尽管多种多样，但大致均可归入以下常见的三类：第一类是，围绕计算主义的争论。认为人脑具有不可编码的认知特性；第二类是，围绕人工智能理论框架可行性的争论。认为数字计算机是基于物理符号系统的客观表征与具有主观特性的人类认知之间存在主–客认知鸿沟；最后一类则是，围绕人类高级意识活动的争论。认为人类意识具有意向性特征，而计算程序仅凭自身永远不能成为意向性产生的充分条件。然而，有理由相信，这三类质疑均有对应的解决方案，对于人工智能的发展并不能构成实质性的阻碍。下面我们来具体考察这三类质疑，并一一加以回应。

1. 第一类质疑围绕计算主义的质疑

计算主义认为AGI是可以实现的。所谓计算主义，就是“将认知过程理解为计算过程的思想。”[2]因此，在计算主义者看来，人的认知过程就是一个计算过程，而数字计算机可以完成一切可能的计算。无论人类的生物大脑和计算机在硬件和软件层次上有多大不同，但是在计算理论层次上，它们都具有产生、操作和处理抽象符号的能力，因此人类智能或心灵本质上就是可编程的计算机器。

二元论对计算主义的质疑是，认为计算机无法产生感受质（qualia），也无法对其编码。因此，AGI原则上无法实现。例如，内格尔（Thomas Nagel）曾指出，“关于成为一只蝙蝠是什么体验”（what it is like to be a bat）这样的事情是完全主观的，[3]只有能够拥有或者经历那种体验才能够理解，因此这种关于人类高级意识活动的体验根本无法从外部第三人称的视角得到理解。然而，任何已知的关于人工智能的科学研究显然只能是客观的活动。鉴于人类意识主观维度原则上是计算机编码这类客观研究所无法达及的，因而AGI原则上无法实现。

心脑同一论对计算主义的质疑是，认为计算机和人脑有着本质上的不同。心脑同一论的观点是，心灵的状态和过程同一于大脑的状态和过程。也就是说，只有当计算程序可以完全模拟人类生物大脑的状态和过程时才能实现对人类认知的模拟，AGI才有可能实现。但是，人类大脑运用的逻辑和算术与数字计算机所运用的逻辑和算术有着逻辑结构上的差异。例如，人类视觉的成像原理是，人类眼睛中的视网膜，对眼睛所感受到的视像在视网膜面上进行重新组织。而且这种重新组织是在视觉神经入口处由三个顺序相连的突触的点上实现的，也就是说，人类的神经系统对视觉成像的逻辑处理，“只有三个连续的逻辑步骤。”（[4]，pp.77–78）但是，任何机器视觉方案在进行从光学信号到数字信号的处理都必须依赖复杂的算法。对此，冯·诺伊曼的判断是“因此，中央神经系统中的逻辑学和数学，当我们把它们作为语言来看时，它一定在结构上和我们日常经验中的语言有着本质上的不同。”（[4]，p.78）因此，基于人类大脑的生理神经活动与计算程序所使用语言的本质差异，AGI原则上不可能实现。

对第一类质疑的回应。首先，不可编码问

题，原则上可以利用神经网络和深度学习“绕开”。计算主义的核心观点是，人类的认知就是计算过程。但是并非所有的计算都是认知。据此受到对人类大脑研究的启发，模仿生物神经元信号相互传递的过程，人们开发了人工神经网络。一般认为，生物神经并不是一开始就具备智能、思维、情绪等精神活动能力，而是在其成长过程中通过学习逐步获得的。人类出生后，在与外界环境的互动中，大脑接收输入了大量的环境信息，随着经验的积累，神经元之间的相互关系不断变化，从而完成与之相关的精神活动，并最终能够对输入信号做出正确的反应。人工神经网络是由大量相互连接的人工神经元组成的，它会根据外部信息改变自身的结构，并通过调整神经元之间的权值来对输入数据进行建模，并最终拥有模拟人类智能的能力。其次，感受质问题。至今为止，没有充分的理由或经验证据表明，未来出现的AI原则上不可能具有感受质。这里涉及到一个关键点，即所谓的心灵的多重实现（multiple realization of the mental）。心灵的多重实现特征不仅是人类社会进行有效情感沟通交流的一个必要前提，也是跨物种间实现有效情感认知交流的一个必要前提。换句话说，心灵的多重实现是社会认知（social cognition）的一个必要前提。如果反对这一点，那么就需要解释人与人之间或跨物种个体之间的有效情感交流是何以可能的。这里或许有人认为，感受质的多重实现将面临可设想论证（conceivability argument）的质疑。对此的回答是，这个论证本身一直存在争议，在这些争议得到有效处理之前，不应匆忙下定论。而且近年对感受质的深入分析表明，感受质的某种客观化研究方案原则上是可行的。[5]，[6] 因此目前没有足够的理由认为，感受质会成为未来AI发展过程中不可逾越的障碍①。

2. 第二类质疑围绕人工智能理论框架可行性的质疑

围绕人工智能理论框架可行性有以下几种比较常见的争论。首先，在人类大脑中大量相互关联的神经元是平行工作的，“神经系统的特点是高度的平行性”。[7] 那么，基于冯–诺依曼架构的一个串行的计算模型在模拟人类认知方面是无法让人满意的。而且，经典人工智能所依赖的数字计算机在存储和处理单元之间进行架构上的区分，这也不符合人类大脑的特点。其次是来自哥德尔不完备性定理的质疑。依据哥德尔不完备性定理，在一个一致的数学系统中，至少有一条无法在该系统内部获得证明的命题，但是凭借直觉，人们可以认识到这条命题为真。哥德尔对此的看法是，“以下结论的析取是不可避免的：要么数学原则在这个意义上是不完备的，即其显而易见的公理永远不可能包含在一个有限的规则中，也就是说，人类的心灵（即使在纯数学原则领域）无限地超越了任何有限机器的能力，要么就存在绝对不可解决的指定类型的丢番图（diophantine）问题”。[8] 也就是说，在哥德尔看来仅依据逻辑程序运作的数字计算机缺乏直觉能力，因此任何图灵机的计算能力都不及人类心灵，因而计算主义不可能是正确的关于人类心灵的理论。第三，认为人类认知具有很强的主观性和非表征性，与计算机基于物理符号系统的客观表征之间存在无法跨越的“认知鸿沟”。德雷福斯（Hubert L. Dreyfus）认为“一个人所学到的东西会以世界显示的方式出现，它不会在心灵中被表征且添加到现在的经验中。”[9] 也就是说，我们所学到的东西根本不在心灵中被表征，而是只有在具体的情景之下才会显示出来。例如，我们在得到一辆自行车以后可以很熟练的骑行，但是却很难立刻说出正确驾驶自行车的操作规则。那么，既然这些知识在心灵都没有表征，也就不存在可以被客观的物理符号系统所表示的情况了。

对第二类质疑的回应。首先，利用神经网络和深度学习，AI完全可以实现并行信息处理过程，并且原则上有望实现人脑类似的非表征的认知过程（如，直觉）。因此，目前并没有充足的理由相信，AI的信息处理过程只能是串行的，

①关于心灵的多重实现特征，下文（见第二节第2小节）很快还会有进一步解释。

并且其未来发展必定要受到基于哥德尔不完备性定理的卢卡斯–彭罗斯论证（Lucas–Penrose argument）的限制。其次，主–客认知鸿沟有可能并不真实存在。所谓心灵的主观性特征很可能只是语言的误用所造成的一种认知幻相。[5]具体来说就是，我们在使用“观察”一词时，没能注意到在“观察到一个物理状态”与“观察到一个心理状态”这两句话中其用法是不一样的——在应用于一个物理对象或状态时，物理对象的物理属性并不依赖于我们的观察；在应用于一个心理对象或状态时，其现象特征一定是依赖于某个主体特定的观察或经验的。也就是说，对于处在某个认知状态C1中的某个认知主体S而言，S唯一能确保的是自己是否拥有心理状态M1，但是始终无法确保自己是M1唯一的拥有者，因此心灵并不具有所谓的主观性。正是这一差别造成了心灵的主观性认知幻相。

3. 第三类质疑围绕人类高级意识活动的质疑

围绕人类高级意识活动的争论主要有以下两点。第一点是，人类个体除了推理、计算这些意识活动，还展现出情感、感受、意图、自由意志、自我意识等一些更高级的意识活动。当计算机处理推理计算等问题时只需要消耗很少的计算资源，但是当涉及模拟人类的情感等高级意识活动时往往需要消耗很大的计算资源，结果还未必理想。这说明人类的高级意识活动与推理计算等具有完全不同的性质。第二点是语义相关的。语义是理解（语句）的一个必要条件，而要获得语义就需要具备意向性（intentionality）。人类意识的一个显著标志就是具备了意向性的特征。然而，数字计算机只是纯粹句法层面的操作，仅凭句法操作无法生成语义。因此数字计算机无法做到像人类一样理解语句。这其中最为著名的论证是当代著名哲学家塞尔（John Searle）提出的“中文屋思想实验”（Chinese room argument）。[10]根据这一思想实验塞尔认为，存在这种情况，房间里的人对于输入和输出的东西完全不理解，仅凭句法操作，根本不具有任何相关的语义特征。也就是说，运行正确的程序并不一定产生语义上的理解（具备意向性）。可见，计算程序仅凭自身永远不能成为意向性产生的充分条件。

对第三类质疑的回应。首先，有理由相信，人类意识的意向性特征原则上可以被“自然化”，进而也可被编码。当代认知科学和心灵哲学中关于意向性的自然主义研究进路认为，人们的心灵也是自然的物理对象，因此心灵的内容及其获取内容的方式也是自然的或物理的过程，因此人类的意向性特征是取决于一个人的环境的，而不是仅仅取决于头脑中的东西。这其中最具代表性的是来自德雷茨克（Fred Dretske）的信息语义学。德雷茨克认为心灵的语义属性关注的是心灵内容本身，真正的信息理论关注的是我们咨询的内容的理论，而不是内容所依附的形式。“设备S携带关于属性G的实例的信息，当且仅当S的存在F是名义上和G的实例相关。除非属性G被实例化否则S不会是F，那么S的存在F携带关于G的信息，或者像德雷茨克喜欢说的那样，表明G存在”。[11]也就是说，携带信息的设备可以表现出某种程度的意向性，这表明非心灵的事物也可以表现出意向性特征。其次，既然今天是在GPT–4时代讨论这个问题，因此我们完全可以设想，GPT–5以及以后更强的系统完全实现了多模态学习，而且安装在机器人上以后，机器人可以成功地将所听到的语言指令关联到所观察到的环境中的事物、事件，并关联到机器人对自己的行动的计划与控制。这样的机器人应该算是理解语义的①。

二、通用人工智能的真正困境与超级人工智能的可能实现路径

在上一节里，我们主要论证了不可编码问题、主–客认知鸿沟，以及意向性这三类质疑并非原则上无解，对于AGI这一人工智能目标的实现并不能构成实质性的阻碍。然而，尽管如此，我们认为关于AGI有一个问题不可忽

①这里关于机器人如何理解语义的看法，来自于一位匿名审稿人所提出的具体建议，在此表示感谢。

略，那就是“人机区分难题（human–machine distinction puzzle）”，这才是AGI不可实现的真正原因。但是AGI的不可能，并不意味着Super AI不可行。有理由相信，借助意识的功能主义理论与多重实现论题，Super AI原则上是可能的。

1. 通用人工智能的“人机区分难题”

我们认为AGI之所以不可能，是基于我们无法区分AGI与人类智能（HI）。对于Weak AI或者Super AI而言，有效区别开人类智能与具有智能的人造物都不困难。真正的麻烦在于，对于AGI而言，人类智能与人工智能是可以区别开来的吗？这就是AGI面临的“人机区分难题”。

有理由相信，对于AGI而言，“人机区分难题”原则上是无解的。[12]理由就是，如果“人机区分难题”有解，不管AGI是否可以达到HI的水平，则意味着HI和AGI 两者之间存在“可比较关系”（comparable relation，CR）①。存在可比较关系则意味着HI和AGI两者之间必然存在不同。那么问题就是，我们出于什么理由认为HI和AGI两者之间必然存在不同？对此，引入的“人狗混合体”思想实验可以表明，我们拥有关于人和狗之间差别的知识只是偶然为真的②。也就是说，鉴于集体人格同一性（collective personal identity）难题的存在，人与所谓“人狗混合体”之间并不存在必然的区别。“同理，我们也可以认为，‘K有别于人类’的知识，也并非是必然真的”。[12]也就是说，HI和AGI两者之间并不存在可比较关系（CR），即无法实现HI和AGI的区分。而AGI则目标在于创造像人一样思考和行动的智能机器，这一目标的实现则取决于AGI于HI之间是否存在可比较关系，是否是可区分的。因为HI和AGI 两者之间并不存在可比较关系，所以AGI是否可能则是一个没有认知意义的问题，那么AGI也不可能实现。

本质上来说，人机区分难题属于哲学上的“含混性”（vagueness）论题。含混性论题争议长达两千年，目前学界通行的看法是，含混性论题是语言概念分类系统中原则上无法被根除掉的一个漏洞，因而含混性论题原则上是无解的。因此，AGI原则上也是不可能实现。那么现在的问题是，即使AGI原则上不可能实现，那么Super AI可能吗？有理由认为，答案是肯定的。理由就是，对于含混性论题，原则上无法划出秃子和非秃子的清晰界限，却并不妨碍我们在日常生活中一眼就能辨认出秃头。同理，我们缺乏区分AGI与HI之间的标准，则并不妨碍我们区分Super AI与HI。

2. 超级人工智能的可能实现路径

虽然AGI原则上不可能，但是人机区分难题并不对Super AI构成阻碍。以下，我们将论证，借助意识的功能主义理论与多重实现论题的Super AI的实现路径原则上是可行的。

依据英国牛津大学哲学家、知名人工智能专家博斯特罗姆（Nick Bostrom）教授给出的定义，Super AI是指“在许多非常普遍的认知领域远远超过目前最好的人类头脑的智能”。（[13]，p.162）博斯特罗姆认为有三种形式的超级智能：速度型超级智能（speed superintelligence）、集体型超级智能（collective superintelligence）和质感型超级智能（quality superintelligence）。“速度型超级智能 ”是一种拥有与人类思维一样的智力，但速度更快——“一个系统可以做人类智力所能做的一切，但（速度）要快得多”；（[13]，p.163）“集体型超级智能”是通过聚集大量较小的智能来实现超级性能的系统——“一个由大量较小的智能组成的系统，以至于该系统在许多非常普遍（general）的领域中的整体性能大大超过了目前任何认知系统的性能。”（[13]，pp.165–166）“质感型超级智能”是集体型超级

①可比较关系：（CR）事物t1和事物t2之间具有一种“可比较关系”，当且仅当t1和t2之间具有某个或某些相类似（analogous）的特征（character）C。

②“人狗混合体”思想实验：某种未知病毒x突然爆发，该病毒危害性大、传染性强而且传播迅速。病毒x致使整个人类种族面临灭绝的危险。然而狗对此病毒具有先天的免疫力，而且届时的科技已经非常发达，人类已能够将人脑成功移植嫁接到别的哺乳类动物身上。假定时间紧迫而且为躲避灭顶之灾，我们不得已采用这种技术。但是经历这场浩劫之后，满街都是狗身人脑的生物。这个时候面对此混合体，我们以往拥有的关于人与狗之间差别的所有常识或证据似乎都将失效。

智能的增强。博斯特罗姆认为，一个系统的集体智能可以通过增加组成这个系统的智能的数量或质量，或通过更好的组织整合来增强。如果我们逐渐提高“集体型智能”的整合水平，将一个个松散的“集体型超级智能”，最终整合成一个统一的智能（a unified intellect），就可以成为一个“质感型超级智能”——“一个至少和人类心灵一样快，并且在本质上更聪明的系统。”（[13]，p.172）

人工智能被诟病最多的就是，无法拥有人类意识的特征。然而，有理由相信，从功能主义的角度来看，并非如此。例如，普特南（Hilary Putnam）认为，人的心理状态与图灵机同样具有计算功能的多重实现性。[14]1972年福多（Jerry Fodor）和布洛克（Ned Block）更进一步，将普特南对图灵机的功能主义理解推广到一般的计算系统，主张重要的并不在于内部的神经状态本身，而在于这一状态在整个机体活动中所起的作用或功能。[15]人类可以凭借大脑和神经系统计算“1+1=2”，但是计算机也可以在没有任何内部的神经生理状态的情况下得到“1+1=2”的结论。将心理状态与计算机的功能或逻辑状态进行比较，可以看出，正如计算机程序可以在许多不同的硬件结构中实现一样，心理状态也可以在不同的机体中实现。

功能主义认为，思想、欲望、痛苦等心理状态之所以是其所是并不取决于其内部结构，而是完全取决于其在认知系统中的功能或其所扮演的因果角色，心理状态的特性是由其与感觉刺激、其它心理状态和行为之间的因果关系决定的。因此，功能主义对心理状态的定义是一个功能性定义。例如，什么是捕鼠器？功能主义把捕鼠器定义为能把自由活动的老鼠变成被捕获的老鼠的东西，不管它是用什么做的，也不管它的内部机制是什么。如果它能把自由的老鼠变成被捕获的老鼠，那么按照功能主义的定义就一定是捕鼠器。同样，功能主义认为，心理状态和心理事件在主体的感官刺激和随后的行为之间起着因果作用，其功能是将输入转化为输出。因此，不管在输入和输出之间运作的内部状态的物理基础是什么，不管它的运作机制是什么，只要它把某种输入转化为特定的输出，那么这种内部状态就是一种心理状态。

根据功能主义理论，可以将疼痛描述为由身体受伤而引起的心理状态，那么，任何使类似因果模式可能发生的内部状态都属于心理类型的疼痛，不管在任何特定情况下实现该因果模式的具体物理机制如何。假设对人类而言，疼痛的心理状态就等于C–神经被激活的神经活动。那么根据功能主义理论，人类可以仅仅通过接受C–神经被激活而感受到疼痛。但是功能主义并未排除，具有其它不同的物理结构的对象也能具有同样的有心理状态，即C–神经被激活时而感受到的那种疼痛。也就是说，同样的疼痛可以由不同的物理状态在不同的对象中实现，或者说多重实现。例如，被一只狗的神经系统和大脑状态所实现，被合理编程的数字计算机的电子状态所实现等等。换句话说，所谓的人类意识的特征，原则上不仅可以在人类的大脑和神经系统中实现，也可以在其他满足类似因果模式的系统中实现。

那么依据多重实现论题，意识特征可以对应多种不同的物理状态，既可以是生物的神经或大脑状态，也可以是机器的计算状态。对于多重实现论题来说，这二者是等价的。可见，功能主义的多重可实现论题，构成了人工智能具有意识特征或心理状态的理论前提。

三、人工智能与人类的共同未来

以上，我们讨论了关于人工智能基本原理的争论，介绍了不同类型的人工智能，论证了AGI原则上的不可实现，以及Super AI的可行性。鉴于奇点来临的预言，这意味着，超越人类智能的Super AI将会不可避免的出现！那时人类社会将要面临的是比现在更加复杂的局面。不可否认的是，人工智能一定会对人类的自主性，自由乃至生存形成潜在或是直接的威胁。因此我们不能将人工智能的研究与研究人工智能的道德伦理问题分割，针对人工智能安全风险的未雨绸缪是必要的。我们认为这一工作应当从以下三个方面着手，首先是对未来可能出现的

人工智能型态进行预判，其次是讨论人工智能的道德主体地位，最后对人工智能发展过程中潜在安全风险进行分析，并尝试制定行之有效的防范措施。

1. 可能的社会形态

有理由相信，未来的可能社会形态有且仅有如下两种："单智慧体社会"与"多智慧体社会"。"单智慧体社会"是指，只有一种高级智慧体存在的社会形态，每当出现一个新的智慧体，原先旧的智慧体就会被新的消灭和取代。因此，在"单智慧体社会"中文明样式趋于单一化、静态化，其演变和交替遵循的是零和博弈（zero–sum game）原则与丛林法则。如，人类历史上不断出现的各种社会形态。需要知道的是技术是一种工具，本身并不是目的，也不仅仅是保护有机生命的手段，其最终目的必然存在于技术领域之外。在一定限度内技术确实可以用来实现人类自己的目的，但这并非技术的目的。因此，在出现超级人工智能之后，我们不可忽略的一个最坏后果就是人类与人工智能的零和博弈。显然在这种情况下，纯粹是单智慧体社会的演变交替，人类在面对Super AI时毫无优势可言。因此，我们需要准备的是避免人类与人工智能的对抗，从以往基于零和博弈的"单智慧体社会"逐渐过渡到人类与人工智能共生的正和博弈（positive–sum game）的"多智慧体社会"。"多智慧体社会"则是指，多于一种高级智慧体存在的社会形态，其特点是文明样式趋于多样化、动态化，同时正和博弈、互惠互利成为社会的基本法则。

目前看来，未来多智慧体社会中可能出现的智慧体可能有以下九种：

（1）智能增强人（个体）；使用基因技术与人工智能技术相结合之后的改造人，其外形上基本与人类无异，但特点是智能水平远远高于传统的人类。

（2）人机融合体/赛博人（个体）；使用基因技术与人工智能技术相结合之后的改造人，特点是外形上与传统人类有明显区别。

（3）智能机器人（个体）；是指Super AI技术在人形机器上的实现。如，超级智能机器人。

（4）异型智慧体（个体；微观、宏观、宇观）；是指Super AI技术在非人形机器上的实现。如，超级智能纳米机器人、超级智能软体机器人、超级智能行星体。

（5）智群体（群体）；一种集体型超级人工智能，是指具备了超级群体智能（Super swarm intelligence）的群体。如，蚁群型超级智能机器群。

（6）虚拟智慧体（个体）；是指完全由电脑生成的，仅存在于虚拟世界或虚拟实在（VR）中的虚拟智慧体。如，在元宇宙中的虚拟人。

（7）虚拟智群体（群体）；是指完全由电脑生成的，仅存在于虚拟世界或VR中的智群体。如，元宇宙中的虚拟蚁群型智群体。

（8）虚实交融/融合智慧体（个体）；是指由电脑生成的，但可以借助中介设备，由虚拟世界或VR介入到现实世界之中数字智慧体。如，美剧《神盾局特工》中的虚拟智慧体的机器人化身（avatar）。或者反过来，借助中介设备，由现实世界介入到虚拟世界或VR中的现实世界智慧体的化身。如，电影《黑客帝国》的尼奥（Neo）。

（9）虚实融合智群体（群体）；是指由电脑生成的，借助中介设备，由虚拟世界或VR介入到现实世界之中虚拟智群体。或者反过来，借助中介设备，介入到虚拟世界或VR中现实世界智群体的化身。

可以预见，人工智能的成功或者说Super AI时代的到来，会彻底改变大多数人的生活。在"多智慧体社会"或者说在其到来之前的阶段，我们的工作和娱乐，甚至继承而来的人类文明体系都将逐步解构。我们对智能、意识和人类未来命运的看法也将改变。因为Super AI是可以威胁人类的自主性、自由，甚至是生存的另一种前所未有的智慧体形态。因此，相应的我们也必须提前对其发展过程中潜在安全风险进行分析预判，说明其他类型智慧体的道德主体地位，并为"多智慧体社会"的发展及其面临的风险制定行之有效的防范措施。

2. 人工智能的道德主体地位

道德主体地位是人类的一种基本权利。因为我们似乎不会怀疑人类具有道德主体地位，具有道德选择和行为的能力这件事情。那么现

在的问题就是人类为什么具有这种基本权利，以及人工智能是否可以做到像人类一样具有？对这一问题通常的回答是去考察道德主体地位的标准是什么，然后用这一标准去衡量人工智能。然而这一方式的弊端是，“考察具有道德主体地位的标准”这一做法，所能得到的答案只是针对至今为止的人类而言的，因为目前为止除了人类，其他对象的道德主体地位我们是尚未确认的，那么这样的考察对于未来可能面对的Super AI或者说“多智慧体社会”而言其实是难以奏效的。因此，我们提倡的做法是考察道德主体地位从何而来，考察人类的这种基本权利从何而来。然后考察人工智能是否能够满足这种来源条件，倘若能够满足，那么就有理由相信人工智能同样也可以具有道德主体地位了。

道德主体地位是一种“地位功能”。我们需要考察的是地位功能的来源。塞尔认为只需要三个最基本的概念就能解释地位功能创立和维持其存在的一般原则：集体意向性、功能赋予、丰富到足以能创立地位功能宣告的语言，包括构成性规则。其中第三个要素实际上就是构成性规则或者用来赋予地位功能的程序，塞尔将其表述为：“X在背景C中算作Y”的形式。[16]

对于宣告式语言，塞尔的论证是，从众多的言语类型中找出了两类言语类型，“语词向世界（word–to–world）”与“世界向语词（world–to–word）”。例如“猫在席子上”，“雪是白的”，“苏格拉底是有死的”这些言语类型表述了世界上的事物是什么样的，其真假取决于它们在多大程度上成功表征了世界上的事物（即能确切谈论其真假）。上述这样的表达被称为是“语词向世界”的言语行为。而另外一些并不试图告诉我们世界上的事物是什么样的，但试图改变世界使其与言语行为的内容相一致，这样的言语行为被称为“世界向语词”。例如我命令某人离开我的房间，或我承诺将于某天来看望某人。在这些情形中我试图做出一个言语行为来改变世界，目的是引起世界的某种变化而使其与言语行为的内容相一致，而不是要与独立存在的实在相一致。

宣告式语言就是兼有“语词向世界”与“世界向语词”双重指向的言语行为。塞尔认为这种言语行为通过宣告某种事态存在而使得那事态得以存在，从而改变世界。地位功能则是通过地位功能的宣告式言语行为而创立的。可见，宣告式言语行为，正是通过将其自身即地位功能表征为实际存在，从而创立起一项具有地位功能的制度性实在。

利用地位功能理论再去考察道德主体地位的来源，就会发现，人类的道德主体地位来自于我们宣告自己存在，宣告自己拥有。之所以能如此这般，是因为人类具有自我意识和自主行为能力。同样，对于人工智能而言，具有自我意识，能够进行独立的思考和决策的强人工智能类型，也可以宣告自己的存在，宣告自己的道德主体地位。反之，一个没有自我意识，无法进行独立的思考和决策的系统，也就不能称之为真正的“智能”。因此，有理由相信，人工智能原则上也是可以具有道德主体地位的，而且，在真正的人工智能诞生那一刻起其就拥有了道德主体地位。

3. 人工智能与人类的共生策略

（1）人工智能的潜在风险

有理由相信，人工智能的安全风险主要涉及以下三个层面：

（i）科技层面。互联网、物联网技术一旦与人工智能技术相结合，会使得安全问题复杂化。一方面，无限制的大数据和网络资源使得可供人工智能使用的资源趋于无穷，人工智能的进化速度会得到了指数级飙升。因而，把控人工智能发展会变得越来越复杂困难，失控风险概率越来越大。例如，现在语音识别技术的广泛使用，导致各种软件窃听行为泛滥成灾，个人隐私遭到严重侵犯。对于人工智能来说，其智能的成长必然需要大量的数据来喂养学习，一旦对此处理不当或监管不力，那么就会导致个人隐私的丧失。另一方面，目前的人工智能都是基于机器学习算法并使用大量的数据训练生成，如果算法出现了错误或模型学习了错误或带有偏见的数据，很可能会给人类带来误导。例如，ChatGPT的学习材料主要来自英语语言模型，由于不同语言和文化的差异，ChatGPT在不

同语言之间的学习和训练也可能存在差异和偏差。因此，ChatGPT的输出可能存在大量的语言和文化差异。如果模型学习了错误或带有偏见的数据，可能会产生不准确或带有偏见的回答。

（ii）伦理层面。一方面，人工智能的介入很有可能会给传统秩序带来巨大挑战。以ChatGPT为例，ChatGPT的决策过程通常是基于深度学习模型，这些模型通常被认为是黑箱，难以解释其决策的原因和依据。这给用户和监管者带来了透明度和可解释性的困扰。如果ChatGPT的决策导致了不当的行为或伤害，要如何能够进行追责和纠正呢？目前看来尚未建立合适的道德框架和追责机制来解决这一伦理挑战。另一方面，基因技术或VR技术与人工智能一旦相结合，可能会对传统伦常构成危害。例如，虚拟仿真智能机器恋人的出现，潜在地会危害人类情爱关系或婚姻关系。更为重要的是，人类会丧失其作为唯一智慧体独一无二的神圣感，这对人类社会的冲击可能堪比当年达尔文的演化论所造成的威胁。再例如，经过基因技术改造的智能增强人或人机融合体，其在智能或某些方面会远远超越人类，那么传统的未经改造的人类与智能增强人或人机融合体之间是否会出现歧视现象？更进一步，倘若这种改造被普及是否意味着人类种族的终结？

（iii）社会层面。在弱人工智能阶段，各种辅助或替代系统的应用可能会带来相关人类主体责任感的丧失，甚至一些无节制运用也很可能会使我们对人工智能的依赖演变成灾难。例如，医生在诊断病情时依赖人工智能的医疗系统给出诊断，那么一旦发生误诊责任在谁呢？亦或者当医生不听从人工智能的医疗系统的结论而发生误诊又该如何划分责任呢？同时，互（物）联网技术使得黑客、病毒等人为因素对人工智能产品构成巨大威胁。例如，一旦黑客控制了儿童看护机器人、助老机器或其他智能系统，由此导致的后果不堪设想。在超级人工智能阶段，对人工智能的过度依赖可能会失控以致于演变成人工智能对人类社会的全面奴役，人类沦为超级人工智能的养料供给源。就如电影《黑客帝国》所呈现的景象。

（2）防范策略

通过以上分析我们可以看出，人工智能在给人类生活带来美好期待的同时也伴随着风险的生成，因此对人工智能风险的预测管控将是必要的。为了推动人工智能健康发展，塑造人类与人工智能和谐共生的“多智慧体社会”，有理由相信，应当从正向与负向两个方面对人工智能的可能风险给出防范策略。具体来说，负向防范策略可以从以下几个方面入手。第一，针对科技风险，应加强科技管控，尽快尽早从国家层面加以立法。2020年1月，美国白宫发布了《人工智能应用监管指南备忘录（草案）》，提出了人工智能应用相关原则和建议。2023年4月11日，为促进生成式人工智能健康发展和规范应用，中国国家网信办起草《生成式人工智能服务管理办法（征求意见稿）》。这些努力都是必要的，但是更应当加快从立法层面入手，完善法律体系，在法律层面对人工智能的发展给出必要限制；第二，针对伦理风险，应设立独立的伦理审查规约机构。尽早仿效医学伦理审查机制，成立独立的人工智能安全伦理审查委员会；第三，针对社会风险，应不断增强人工智能与人类的互动社交能力，尽快尽早培植人工智能具备悲天悯人情怀，将人类的智慧与意识尽可能的植入人工智能机器，可大概率避免因超级人工智能出现而导致的人类灭绝或被全面奴役的灾难发生。正向防范策略则主要是，由“围堵为主型防范”逐渐转变为“疏导为主型防范”。首先，建议放弃“人工智能只能服从于人”这一人类中心主义立场，尽早着手探讨“多种型态智慧体互生共存”的可行性。调整人工智能发展的远期目标。建议考虑将人工智能发展的远期目标调整为，最终构建一种“多智慧体互生共存型的和谐社会”。其次，人类社会中技术力量的增长和我们对于技术伦理的思考之间的差距正在日益扩大，面对这种情况我们有必要建立一个全球性的组织或共同体，为促进人类技术伦理的发展而服务。

总之，设法构建一种“多智慧体互生共存型的和谐社会”，应当成为指导当下人工智能发展的基本方针。也就是说，我们应当尽早地开

始慎重考虑如下情形：如何能够从零和博弈的“单智慧体社会”，设法打破丛林法则，实现从零和博弈过渡到正和博弈。唯其如此，才能有望避免“单智慧体社会”的不断更替，从而过渡到遵守正和博弈和互惠互利法则的“多智慧体社会”，最终得以实现“多智慧体互生共存型的和谐社会”。

结　论

以上，我们先是梳理了AGI面临的三大类主要质疑，提出了“人机区分难题”，论证了以ChatGPT为代表的AGI何以原则上难以实现，接着，我们分析了Super AI可能的实现路径，并具体列举了未来社会可能出现的多种智慧体形态，论述了人类社会何以可能从“单智慧体社会”过渡到“多智慧体社会”。在此基础之上，我们进一步提出，如果Super AI未来可能实现，那么人类在面对Super AI时，需要设法避免人类与人工智能的相互对抗，避免陷入基于零和博弈原则的“单智慧体社会”，设法过渡到基于正和博弈原则的“多智慧体社会”。为了成功实现这一过渡，人类需要承认人工智能道德主体地位的合理性，清醒意识到人工智能的潜在风险，以及尽早制定相应的应对策略。

ChatGPT的出现及其带来的轰动效应，无可避免地将会大大影响和改变我们的社会生活。现代技术正在朝向信息化，智能化，甚至意识化的方向发展，超越人类智慧的人工智能必然也将到来，这是值得我们期待也更加需要警惕的。当人类引以为傲的人类智慧失去了神圣的唯一性，遭受威胁的就不仅是基于人类智慧文明的伦理体系，更严重的是对人类生存的威胁。趁着还有时间和希望，为了人类在不久的未来可以继续生存且获得更加美好的生活，可能现在就是行动起来的最佳也是最后的时机。现阶段采取怎样的行动，毫无疑问，将会深刻影响着人类与人工智能的共同未来。

（致谢：本文初稿曾在西安电子科技大学和南京信息工程大学报告。其间，朱锋刚、苏丽、王雨程，以及崔中良等师友提出了诸多有益的评论和建议，这些评论建议有效地促进了我们的思考。对以上诸位师友表示衷心的感谢！）

［参 考 文 献］

[1]库兹韦尔. 奇点临近[M]. 李庆诚、董振华、田源 译, 北京: 机械工业出版社, 2011.

[2]程炼. 何谓计算主义?[J]. 科学文化评论, 2007,(4): 5–16.

[3]Nagel, T. 'What is It Like to Be a Bat?'[J]. *Philosophical Review*, 1974, 83(4): 435–450.

[4]冯·诺伊曼. 计算机与人脑[M]. 甘子玉 译, 北京: 北京大学出版社, 2010.

[5]王晓阳. 非主观的心灵[J]. 自然辩证法通讯, 2019, 41(8): 40–56.

[6]王晓阳、林崧驰. 无言的感受何以可能？——布洛克与维特根斯坦主义者之间的“颠倒光谱之争”辨析[J]. 哲学动态, 即将刊出。

[7]Cordeschi, R., Frixione, M. 'Computationalism Under Attack'[A], Marraffa, M., De Caro, M., Ferretti, F. (Eds.) *Cartographies of the Mind:Philosophy and Psychology in Intersection*[C], Dordrecht: Springer, 2007, 37–49.

[8]Gödel, K. *Collected Works III*[M]. Fefermaned. S. (Ed.) New York: Oxford University Press, 1995, 310.

[9]Dreyfus, H. L. 'Intelligence Without Representation-Merleau-Ponty's Critique of Mental Representation'[J]. *Phenomenology and the Cognitive Sciences*, 2002, 1(4): 367–383.

[10]J. R. 塞尔. 心灵, 大脑与程序[A], 玛格丽特·博登: 人工智能哲学[C], 刘西瑞、王汉琦 译, 上海: 上海译文出版社, 2006, 73–95.

[11]Jacob, P. 'Intentionality'[EB/OL]. https://plato.stanford.edu/entries/intention ality/. 2019–02–08.

[12]王晓阳. 人工智能能否超越人类智能[J]. 自然辩证法研究, 2015, 31（7）: 104–110.

[13]Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*[M]. Oxford: Oxford University Press, 2014.

[14]Putnam, H. 'The Mental Life of Some Machines'[A], Putnam, H. (Ed.) *Philosophical Papers*[C], Cambridge: Cambridge University Press, 1975, 408–428.

[15]Block, N., Fodor, J. A. 'What Psychological States are Not'[J]. *Philosophical Review*, 1972, 81(April): 159–181.

[16]Searle, J. *Making the Social World: The Structure of Human Civilization*[M]. New York: Oxford University Press, 2010, 101.

［责任编辑　李斌］