人工智能的心智及其限度*

——人工智能如何产生自我意识?

秦子忠

摘　要: 心智活动是以价值序为指导的。价值序映射于自我价值的诸元素关系、目标价值的诸元素关系以及自我价值与目标价值之间的整体关系。这些价值序是连通的、传递的,因而,可以此为核心理念设计检验人工智能是否具有自我意识的测试方法。这一方法同时阐释了人工智能如何产生自我意识的过程,因而有助于人类客观地评价人工智能在意识方面的发展,进而采用适度的干预原则或伦理原则。

关键词: 心智　人工智能　自我意识　深度学习　奇点

每次科技革命都会导致大量失业、社会重组。本轮以人工智能为代表的科技革命因其扩展到对人类脑力劳动的替代,因此引发的失业潮在深度和广度上都远高于之前任何一次科技革命。但是,人们的深层焦虑不仅来自人工智能对脑力劳动的替代,还来自人工智能具有自我意识的可能性。大体而言,我们可以区分出两种不同的公共焦虑。一是人工智能虽然不会获得自我意识,但是越来越强的人工智能正在导致严重的不平等世界;①另一是人工智能能够产生自我意识,并且有可能反过来主宰、控制乃至终结人类。第一种焦虑与我们的现实生活体验、实证研究关联在一起,而第二种焦虑主要是与科幻影视、规范研究关联在一起。几年前,著名物理学家霍金就表示人工智能若要终结人类,人类是无能为力的。赵汀阳进一步从哲学上探讨人工智能获得自我意识需要哪些条件和"设置"。但是赵汀阳的探讨只是界定了人类的自我意识,既没有展示人工智能如何获得自我意识的过程,也没有给出相应的测试方法。在本文中,笔者将在考察赵汀阳这一工作的基础上具体探讨这个问题,即人工智能如何产生自我意识。

人类心智:知情意及其关系

人的认知系统至少包括三个层级:第一层级是复杂行为,例如解决问题、形成概念和语言表达等;第二层级是信息加工过程,例如对光点的感觉、图形知觉的形成等;第三层级就是生理方面,例如中枢神经过程、神经结构等。这三个方面在人类身上已经处在近乎完美的状态。当前的脑科学虽然尚未完全解释三者的相互关系,但是对它们的局部关系已有了很好的研究。这一点构成了人工智能学科的理论基础。作为对人的认知系统的模仿,人工智能也有相应的三层级系统,其第一

* 本文系国家社科基金后期项目"交互行为、平等与发展"(项目号:19FZXB017)的阶段性成果。

① [美]大卫·巴恩西泽、丹尼尔·巴恩西泽:《人工智能的另一面:AI 时代的社会挑战与解决方案》,电子工业出版社 2020 年版。

层级是程序,第二层级是计算机语言,第三层级是计算机硬件。① 从漫长的人类史来看,人类作为动物界的一员,先有了生物的肉体组织及刺激—反应系统,而后形成了一种与动物感知相近的情感能力,之后语言产生,而意识或自我意识是在语言的应用过程中产生的。这种由进化而来的情感、认知、意识(即情、知、意)及其关系,即是人类的心智(mind)。② 依据心智的这一定义,智能只是人类心智的一部分,即认知的部分,而人工智能只是对人类的认知能力的一种模仿。③

情感和认知这两种能力是人类与其他高等动物共同分有的,这一点无需赘述。至于意识,似乎只有人类与灵长动物才真正拥有,而人类与灵长动物相区别的,主要是人类拥有渗透于语言运用之中的自我意识。这里,不妨扼要解释一下人类现代自我心理学之父阿尔弗雷德·阿德勒的心灵概念。阿德勒认为心灵与自由活动紧密关联,并且心灵的一切活动都拥有相同的目标,即总是追求与自己的外部环境相适应,并随着环境的变化而不断调整自己的行为。"心灵的一切活动都拥有相同的目标,这是其最明显的特征……该目标其实就是跟自己的外部环境相适应,所有人的心灵世界都包含该目标,心灵世界的所有活动都受该目标引导。"④据此而言,阿德勒所谓的心灵实质上就是我们所说的意识(consciousness)而非心智,它包括两个本质特征:一是以与自己外部环境相适应为目标,它的所有活动都受该目标引导;二是在寻求与自己外部环境相适应的过程中,心灵指向外部事物的对象意识与指向自身的自我意识交织在一起。⑤ 以下我们稍微论及笛卡尔等人的相关工作以展示人的思维意识。

外部事物的影像通过感觉感官传入大脑,即为印象。大脑中的印象经由知觉,留下一些记忆痕迹。这些痕迹在脑海中进一步形成观念。在这个层面,思维处理的是人与外部世界的关系。更高层面是人的理性,它并非停留在观念和事物的对应关系上,还会处理脑海里诸观念的关系。如此便进入到人类的思维自身,即意识。但人的思维能力不局限于此。重要的能力是它能够对外部印入、自己拼接的东西说"不"。这种否定性行为涉及思维的一个重要方面就是怀疑性,而怀疑性的进一步表达就是笛卡尔的"我怀疑我在怀疑"。当思维进入自我怀疑阶段,就意味着它可以做出任何表达,如多维地调整手段去完成任务,加强、修改乃至否定它的目的等。在这个阶段,思维进入到人的自我反思,即自我意识。这里我们也触及了语言运用与自我意识的关系。从严格意义上讲,人类如果没有发明语言,就不可能发展出自我意识,或者说其思维意识水平就会停留在类人猿层面。如海德格尔所言,人活在语言之中,语言是存在的家。查尔斯·泰勒更进一步指出人类自我的根源深深地镶嵌在由语言交流积淀而成的文化意义背景之中,并且在其中"某些行为、或生活方式、或感觉方式无比地高于那些我们更加乐于实行的方式",⑥由此,人们的自我认同实质上就是在道德空间中发现乃至选择自己所处的方位。据此而言,目前人工智能在意识领域的发展要想产生出自我意识,需要满足相应的条件。如何把握人工智能产生自我意识的条件,是当前心智科学家和哲学家共同关注的一个前沿论域。为了给意识研究注入有益的洞见,新心智科学把意识可操作性地定义为知觉性觉知的一种状态或显而易见的选择性注意。然而,意识如此高深莫测,以至于一些心智科学家认为意识恐怕无法通过科学术语来解释。⑦ 晚近,赵汀阳对此有一个哲学讨论,其讨论的结果是人类自我意识的内在秘密应该完全

① [美]赫伯特·西蒙:《认知:人行为背后的思维与智能》,荆其诚、张厚粲译,中国人民大学出版社2020年版,第4—11页。
② 蔡曙山:《生命进化与人工智能——对生命3.0的质疑》,《上海师范大学学报(哲学社会科学版)》2020年第3期。
③ 倪梁康:《人工心灵的基本问题与意识现象学的思考路径——人工意识论稿之二》,《哲学分析》2019年第6期。
④ [奥地利]阿尔弗雷德·阿德勒:《洞察人性》,张晓晨译,上海三联书店2016年版,第3—7页。
⑤ 魏屹东:《论具身人工智能的可能性和必要性》,《人文杂志》2021年第2期。
⑥ [加]查尔斯·泰勒:《自我的根源:现代认同的形成》,韩震等译,译林出版社2021年版,第31页。
⑦ [美]埃里克·坎德尔:《追寻记忆的痕迹》,喻柏雅译,中国友谊出版公司2019年版,第442—443页。

映射在语言能力中,因此人工智能具有自我意识的条件是必须具有相当于人的意识的两个高级功能:“(1)意识能够表达每个事物和所有事物,从而使一切事物都变成了思想对象。这个功能使意识与世界同尺寸,使意识成为世界的对应体,这意味着意识有了无限的思想能力;(2)意识能够对意识自身进行反思,即能够把意识自身表达为意识中的一个思想对象。这个功能使思想成为思想的对象,于是人能够分析思想自身,从而得以理解思想的元性质,即思想作为一个意识系统的元设置、元规则和元定理,从而知道思想的界限以及思想中任何一个系统的界限,因此知道什么是能够思想的或不能思想的。”①至此,赵汀阳关于自我意识的论述可以归结如下,他把自我意识定义为具有理性反思能力的自主性和创造性意识,并界定了自我意识的两个重要功能,一是有能够表达一切的能力,二是有自我反思的能力。笔者大体认可赵汀阳关于自我意识的论述,因此以下以赵汀阳界定的自我意识为标准展开人工智能产生自我意识的一种过程,而后设计一个测试方法予以检验。但在此之前,我们还需要回答第二个问题,何谓人工智能?

人工智能:基于深度学习的计算机程序

与人类心智的发生学不同,人工智能作为人为设计物,其形式是计算机程序,②其构成是电子元件和硅胶组织,因而没有情欲,它最初通过计算机语言模拟人的思维活动,因而只具有类似人类的思维能力。但是这并不意味着人工智能永远不能产生出意识乃至自我意识。因为意识乃至自我意识与有机生化之间并不一定存在必然关系。③事实上,从20世纪70年代以来,人工智能先后经历了模仿人类思维方式、行为方式和会学习三个阶段。进入21世纪之后,人工智能在深度学习方向取得了巨大发展。因为基于深度学习的计算机程序内含不确定性,使得与不确定性相关联的意识问题成为当前人工智能研究的前沿问题。

著名认知心理学家赫伯特·西蒙把人看作信息加工或物理符号系统。这个系统有六种基本功能:输入符号、输出符号、存储符号、复制符号、建立符号结构(build symbol structure)、条件性迁移(conditional transfer)。据此,西蒙提出了一个假设,即任何一个能表现出智能的系统,都必能执行这个物理系统的六种功能,反之亦然。这个假设附带的三个推论分别是:既然人具有智能,它就一定是个物理符号系统;既然计算机是一个物理符号系统,它就一定能表现出智能,这是人工智能的基本条件;既然人是一个物理符号系统,计算机也是一个物理符号系统,那么我们就能用计算机来模拟人的活动。④ 西蒙注意到,推论2、3已经得到了有力证明,而推论1的证据还不明确也不直接。在笔者看来,西蒙的推论1将人类心智的情感、认知、意识的复杂性简化成了只有认知的纯粹思维能力。不过,也正因此,西蒙的工作让我们清晰地注意到模拟人类认知结构的计算机系统的智能。就此而言,西蒙的物理符号系统构成了本文界定计算机程序的基本义。问题是从这个符号系统能否产生出自我意识,或更一般地产生出知情意合一的心智结构呢?

人工智能如何产生心智?

从前面论述来看,人工智能从认知能力产生出自我意识需要满足两个条件,一是具有表达一切的语言能力,二是具有自我反思的能力。从逻辑上讲,人工智能的计算机语言能表达的事物是无限多的,因此它能够表达一切。由此,人工智能的计算机语言与人类的自然语言一样都能够满足表达一切的条件(至于在极限意义上两者是否等同,搁置不论)。并且,人工智能具有自我意识可能并不需要能够表达一切的语言,而只是需要能

① 赵汀阳:《人工智能的自我意识何以可能?》,《自然辩证法通讯》2019年第1期。
② 李开复、王咏刚:《人工智能》,文化发展出版社2017年版,第35—45页。
③ [以色列]尤瓦尔·赫拉利:《今日简史:人类命运大议题》,林俊宏译,中信出版社2018年版,第65页。
④ [美]赫伯特·西蒙:《认知:人行为背后的思维与智能》,第18—19页。

够表达足够多事物的语言。也就是说,人工智能产生自我意识的条件之一,即具有表达足够多事物的语言,是具备的。问题的关键点在于人工智能是否具有自我反思的能力。人工智能目前当然不具备这种能力,但是未来可能具备这种能力(否则本文以及相关的讨论就无意义)。因此,本节的论题集中于人工智能产生自我意识的另一个条件,即人工智能具有自我反思的能力是如何产生的,因而这个论题的展开实质上就是一种合乎逻辑的推演过程。

人工智能从监督学习到无监督学习的过渡,确实给人类带来很大的不确定性,而人工智能的自我意识的诞生就蕴含在这个不确定性之中。李德毅院士等人研究指出,"不确定性人工智能,必须超越不确定本身而寻求不确定中的基本规律性,寻求表示并处理不确定性的理论和方法,使机器能够模拟主客观世界的不确定性,使定性的人类思维可以用带有不确定性的定量方法去研究,最终使机器具有更高的智能,在不同尺度上模拟和代替人脑的思维活动"。① 在有监督的学习过程中,人工智能主体 A 的目标是由人类程序员设定的,它的整个行为都为了最大化这个给定目标的价值,而不能感知到它自身。由此,比如 A 在执行任务的过程中,它要么无法处理突然发生的未预料到的事件,要么导致自身的受损。与此不同,在无监督的学习过程中,A 的目标是特定的但局部可改变的。也就是说程序员设定了目标框架(包括不变目标和可变目标),但允许 A 有自由去修改可变目标,以及结合不变目标与可变目标的动态关系,自由调整实现目标价值最大化的行为路线。在有一定自由度去实现目标的过程中,A 在功能上讲,便能够处理外部环境变化与自身的关系。比如 A 实现目标有两条路线可以选择:路线 1A 断臂的概率是 50%;路线 2A 会完好无损。在这样的情景中,如果具有一定自由度的 A 在无监督学习过程中自主地选择路线 2 来实现目标,那么 A 的自主选择在功能上便已经接近了具有自我意识的人的基本判断。就此而言,A 就像具有自我反思能力的人一样拥有了"自由意志":A 能够自主地结合外部环境局部地修改目标乃至改变自己的行为路线。

在人工智能主体能否思维这个问题上,图灵不仅给出了肯定的回答,也给出了检验其能否思维的标准。这个标准就是著名的图灵测试。但是,图灵测试容易导致欺骗的造假行为。图灵测试实质上就是模仿游戏,在这个游戏的最后,重点不是计算机程序能否像人一样对话,而是能否骗过询问者;只要这一模仿能够让询问者在一段时间内信以为真,即认为自己是与人类而非计算机程序对话,那就视为通过测试。但是即便通过测试,谈话机器人既不理解其行为的价值,也无法意识到它自身。因此如赫克托·莱韦斯克所言,"图灵测试并没有真正激发人工智能研究人员去研究更优秀的会话者,却导致欺骗询问者的技巧越来越多"。② 更重要的是图灵测试在设计理念上至多只有指向对象的意向性而无指向自身的内省性,因而它无法检测出人工智能是否具有自我意识。以此为鉴,结合第一节关于人类心智的定义,尤其是意识的两个高级功能,一种检验人工智能主体是否具有自我意识的测试设计,其设计理念可表达为以下内容:心智的活动是以价值序为指导的;价值序映射在自我价值的诸元素之间的关系、目标价值的诸元素之间的关系以及自我价值与目标价值之间的整体关系,记为 R(表示偏好或无差异关系)。这些价值序是连通的(connected)、传递的(transitive)。具言之,对于构成自我价值或目标价值的诸元素备选项,如 x、y、z……如果对于所有 x 和 y,或者 xRy 或者 yRx(读作 x 偏好于和无差异于 y,或 y 偏好于和无差异于 x),那么它们的价值序是连通的;如果对于任意的 x、y、z,xRy 且 yRz 则有 xRz(读作 x 偏好于和无差异于 y,y 偏好于和无差异于 z,则 x 偏好于和无差异于

① 李德毅等:《不确定性人工智能》,《软件学报》2004 年第 11 期。

② [加]赫克托·莱韦斯克:《人工智能的进化》,王佩译,中信出版社 2018 年版,第 59 页。

z),那么它们的价值序是传递的。① 经由大量实践而获得的相对稳定的价值序,即常识或直觉,记为 R_1;通过无监督的深度学习而获得的针对某特定环境的价值序,即专用知识或专业能力,记为 R_2;通过条件性迁移学习,即通过已掌握一种环境的知识或符号结构去完成另一种不同但相关环境的学习任务,从而获得新环境下的价值序,即推理知识或适应能力,记为 R_3。R_1、R_2、R_3 均为 R 的子集。

基于这一设计理念,这个测试设计至少包括三个部分:(1)能够分别表达自我价值与目标价值,以及两者之间的关系;(2)能够依据目标价值增加最大化原则与自我价值受损最小化原则来选择最优行为方案,若两个原则给出的方案相冲突,则选择满意行为方案(它属于次优行为方案的集合),直至目标完成;(3)能够处理突然发生的未预料事件,实现自身与外部环境相适应。与此相应,检验标准的关键点在于:在给定的情景中,人工智能主体要能够区分自我与对象,并能够在价值序引导下自主地选取满意的行为方案来完成任务,否则就不能完成任务。参照组是具有自我意识的普通人员。如果被试的智能机器人完成的任务质量与参照组无异,即第三方进行独立评估时不能区分出完成任务的是普通人还是智能机器人或者第三方能够准确做出区分的概率维持在一个极小范围内,则通过自我意识测试。这是哲学层面的一般论述,下面让我们进入实例层面的精细分析。

人类心智活动追求与自身的外部环境相适应。这一追求的具体化,在现实性上通常表现为人的生活计划(具体为各种价值及其关系)、完成计划所需要解决的系列问题(具体为各种任务及其关系),以及通过搜索解决问题的满意的(特殊情况下才是最优的)方法(具体为各种行为及其关系)。② 在这一过程中,人还可以通过评估自我与环境的关系来调整自己的预期水平,这种可变的预期水平由自我价值与目标价值的动态关系呈现出来。由此,就某一任务的完成而言,它包含若干条可能的行为路线,每一条行为路线包括连续的或相关的 n 阶行为,每一阶行为集合有若干种可供选择的具体行为。当选定并执行某一阶行为集合中的一种具体行为之后,也相应地限定了下一阶行为集合的范围。每一阶行为集合最终只能选择一种具体行为。每一种具体行为,都包含三个层面的价值序,它们分别映射于自我价值的诸元素之间、目标价值的诸元素之间以及整体意义上的自我价值与目标价值之间。因此,不同的行为之间的比较,既可以是目标价值层面的比较,还可以是自我价值层面的比较,也可以是总价值层面(自我价值与目标价值之和)的比较。实际完成某一任务之后,每一阶的已经选取的具体行为的合集构成了现实的行为路线。在现实的行为路线中的自我价值受损、目标价值增加、总价值增加与其相应的预期值之差,构成了人的压力或动力。压力或动力的自我调节即是人的可变的预期水平,它既会体现在任务完成之中,也会体现在任务完成之后,或兼而有之。

2016 年战胜人类围棋大师李世石的阿尔法狗(AlphaGo),其原理是对人类以上认知系统的一种模仿。它主要由估值网络(Value Network)、策略网络(Policy Network)和树搜索(Monte Tree Search)三个部分组成。其运作原理是首先通过估值网络评估棋局状况,其次通过一个快速的策略网络选择下一步的位置,一直到最后,获得胜利。③ 它的估值网络相当于人的问题分解图,它的策略网络相当于人在处理问题分解图中针对每一个节点可以选择的方法的集合,它的树搜索相当于人的搜索能力。阿尔法狗所不具备的是可变的预期水平。但是通过自我价值尤其是价值序的引入,改版升级的“阿尔法狗”也能具备这种可变的预期水平。

① [美]肯尼思·J. 阿罗:《社会选择与个人价值》,丁建峰译,上海人民出版社 2020 年版,第 14 页。
② [美]赫伯特·西蒙:《认知:人行为背后的思维与智能》,第 26—30 页。
③ 吴岸城:《神经网络与深度学习》,电子工业出版社 2016 年版,第 173—174 页。

现在让我们构想一个可量化的思想实验，[①]由此检验人工智能主体阿智是否具有表现自我意识的行为。为了简化表述，我们考虑一个只有二阶行为的行为路线。目标价值范围设为0到10，其中的0代表目标价值无增加；自我价值范围设为-10到0，其中的0代表自我价值无损耗。

表1　一阶行为及其价值

价值 / 行为方案	目标价值	自我价值	总价值
行为1	10	-10（程序损坏）	0
行为2	6	-1（效率较低）	5
行为3	5	0（无损耗）	5
行为4	0（价值无增加）	0	0

在表1中，不同的行为方案对应着不同的价值序。其中的参照行为是行为1，它代表阿智只是一味追求目标价值的最大化，而不顾自身的受损情况。如果阿智总是选择行为1，则会被识别为无自我意识。这个实验的重点在于，在这个情景中，阿智的程序自编指令[②]允许它产生不同的价值序，因此如果阿智在大量训练之后总能够做出选择行为1之外的任一满意或次优行为，那么它就表现为一种自主选择的能力。由此，行为2至行为4则是允许的，所以需要考虑下一阶行为。

表2　二阶行为及其价值

价值 / 行为方案	目标价值	自我价值	总价值
行为2-1	5	0	5
行为3-1	8	-3（硬件受损）	5
行为4-1	0	0	0

在表2中，行为2-1至行为4-1[③]各自对应的总价值是不一样的，其中行为4-1的总价值低于其他行为，所以行为4-1先被排除。行为4-1在这里也暗示一种极端情况，即阿智只是一味追求自我价值受损的最小化，而不顾目标价值的增加情况。这种情况目前与人类发展人工智能的目标是相违背的，因此它被排除。但这种排除只能说是人类关于人工智能的事先性干预施加的结果，与人工智能能否乃至如何产生自我意识无关。

在表2中，选择行为2-1还是选择行为3-1，在总价值上都是一样的；但是就目标价值和自我价值的差别而言，选择行为2-1有利于维护阿智的自我价值，而选择行为3-1则有利于增加目标价值。因此，如果在整个过程中阿智的程序自编指令指向选择行为2-1，那么这种行为至少在功能上表明阿智已经具有维持自我价值的意识。这个意识不是事先给定的，而是阿智在不断地无监督学习训练中产生出来的。至此，阿智通过了功能上具有自我意识的检验。在这个二阶行为的图景中，阿智的行为路线由两阶行为组成，即作为一阶行为的行为2和作为与之相应的二阶行为的行为2-1。如果阿智在这一图景中的自主选择行为与作为参照组的普遍人在相同情景下所做的选择是一样的，那么它在思想实验上也通过了自我意识的检验（至于实际操作层面的检验，这是人工智能实验家的工作）。

这里需要做些补充说明。一是上面的量化实验只涉及极简的二阶行为，它的目的主要是为了清晰表达人工智能主体阿智如何展开在功能上具有自我意识的过程，而不是说在二阶行为里就能够检测出阿智具有自我意识。毋宁说这个极简二

① 这个实验的技术性说明如下：笔者将以阿尔法狗原理为初始点，构想一个人工智能护林员（阿智），进而阐释一种阿智通过自我意识检验的思想试验。这个构想隐含这样的假定，即但凡有利于人工智能产生自我意识的阿尔法狗原理的相关部分都会被保留或加强，而阻碍的相关部分则被改写或移除，并加入新的部分。具体的构想如下：1. 估值网络：（1）它由多个相互兼容的子数据库构成（模拟一片开放的森林）；（2）给定目标价值范围G和自我价值范围S，以及G依据内部元素价值大小依次如队友、野生动物、树木等，S依据内部元素价值大小依次如阿智自身的程序、躯体、手臂等；（3）通过条件性迁移学习和/或自主性序列建模，在G与S之间生成多个可供选择的价值序的集合。2. 策略网络：它决定了森林的状态并且选择下一阶的行为。它首先由专业人员对它进行培训，而后预测其下一阶的行为。然后它在模拟的森林中自主行为，如此训练足够多次数之后，再训练程序系统中的下一阶行为，直到它最终达到预期效果。3. 树搜索：树搜索把估值网络和策略网络结合在一起，模拟下一阶会发生什么，并通过策略网络选择更优的价值序，并进行行为。

② 参见[美]伊恩·古德费洛等：《深度学习》，赵申剑等译，张志华校，人民邮电出版社2017年版，第10、14章“序列建模：循环与递归网络”“自编码器”。

③ 行为2-1指的是选择行为2后对应的下一阶行为。以此类推。

阶行为的思想实验是启发性的,它允许并兼容进行更多阶行为的各类检验实验。由此,如果不是二阶而是三阶或三阶以上的行为方案才能检测出阿智具有自我意识,那么这并不否定本实验理念而只是提出了更优的量化实验要求。二是在上面的量化实验中,阿智在无监督学习下自主地摸索相应的推理模式,因为是自主摸索的,因此它有可能自发地掌握演绎推理、归纳推理甚至因果推理等。三是整个检验过程中,笔者有意淡化阿智的言与行的区分。主要原因有二:一个原因是言语一般而言是行为的一种,因此本文采用的行为这一术语,既可以是语言行为,还可以是非语言行为,也可以是语言行为和非语言行为的混合;另一原因是就人工智能如何产生自我意识而言,关键点不(再)是人工智能的语言表达(前文已予以说明),而是关于人工智能对自我价值的意识,因此本文的重心落在阐述兼顾自我价值的人工智能如何展示其自我意识,它具体映射于人工智能的自主选择过程。

人工智能的心智限度

以上的论述是初步性的,有待后续完善。最后,笔者想评述与此相关的问题作为本文的结尾。让我们从这个问题开始,即有自我意识的人工智能会不会进化出情欲的能力呢?在回答这个问题前,我们需要先界定一下情欲。情欲是人类生物性的一种感性倾向,它包括人类的畏惧、烦闷、忧愁等情感,即所谓七情六欲,并且对人的行为有一定影响。① 人工智能主体是一个硅基的胶体无机结构而非碳基的生物有机结构,因此它不具有和人类一样的情欲。但是,有自我意识的人工智能主体可能会演化出类似"情欲"的反应,比如遇到自身温度升高、自己可能受到伤害、自己目标没有如期进行等,通过映射在自我价值、目标价值及其关系中的价值序,它会进行相应的自我调整,暂且将之称为"自适反应"。这个自适反应表现为外部环境改变加重它的负荷与它的自我调整之间的平衡关系。由此,在与人类互动的过程当中,如果这种自适反应对人造成了一定的负担或伤害,那么它就是负面的。试想一下,你遇见这样的实体(机器人或阿凡达),他能与你谈情说笑,也能带给你喜怒哀乐,当他表露他的恐惧和渴望时,你会相信他。你会接受这样一个有意识的机器人吗?确实如雷·库兹韦尔所言,这可能是关于信仰飞跃的问题:当机器人说出他们的感受和感知经验,而我们信以为真时,他们就真正成了有心智的人。② 在这个意义上,人工智能主体和人类进入一个非常关键的互动关系中。

当人工智能主体从认知能力演化出了自我意识能力,从自我意识能力又演化出了与人的情欲相近的自适反应时,他具有和人相近的心智结构。由此我们回到本文开篇提及的焦虑性问题,即人工智能主体会不会主宰、控制乃至终结人类?对这个问题的追问,涉及了对人类的复杂性的分析。因为在复杂性上人工智能主体高于、等于或者低于人类,都会有非常不一样的结果,因此关于人类的复杂性研究在人工智能领域是必要的。它的必要性首先在于,防止在人工智能的发展上做出简单的或错误的决策。从历史来看,第一次认知革命的心智改造,让非洲猿类能够接触到主体间的领域,建立了城市和国家,发明了文字和货币,成为地球的主宰者;而第二次认知革命的心智改造极有可能诞生人工智能与人类的混合体,从而让这个混合体接触到目前难以想象的新领域,得以最终逃离不再适合人类碳基生物体居住的地球,成为其他星球的主人,③或者人类个体在生物体衰老死亡后,将自己的心智移入人工智能网络。当然,这目前只是一种未来学想象。

更具有现实可能性的情景是人工智能只是对人类复杂性中的可程序化部分的模仿,而那些它

① 本文在情欲与情绪(情感)之间不做区分。但提及的是在本文中,人工智能的情欲是由自我意识衍生而来的,而非直接基于认知心理学设计的人工情绪。关于人工情绪的新近研究,参见徐英瑾:《欧陆现象学对人工情绪研究的挑战》,《探索与争鸣》2019年第10期。

② [美]雷·库兹韦尔:《人工智能的未来》,盛杨燕译,浙江人民出版社2016年版,第271页。

③ [以色列]尤瓦尔·赫拉利:《未来简史:从智人到智神》,林俊宏译,中信出版社2017年版,第318页。

不能模仿的东西,它只能通过前面所说的允许一定自由度来逐步靠近人类的心智结构。就此而言,即便人工智能能够发展出自我意识,它也需要相当长的时间来进化才能达到人类的能力水平。但是人工智能的心智最终会不会超越人类,这确实是个开放的议题。在这个议题上,笔者持有谨慎的乐观态度。主要原因来自两个方面,一个方面是从复杂性科学来看,人类的复杂性高于人工智能主体,依据哥德尔不完全性定理,即"没有一个足够强有力的形式系统会在下述意义上是完备的:能够把每一个真陈述都作为定理而重现在该系统中",①因此人类在综合能力上高于人工智能主体。人类的复杂性是生物界极其漫长演化积累的结果。生物从低级到高级,从无意识到有意识,又是经历了数亿年的演进。人类作为高级生物的一支,从古猿到智人,再到现代人类,我们也经历了数百万年的进化。这种由进化而来的复杂性是不可逆的,因此它不可能完全还原到程序主义的形式系统。人工智能作为一种形式系统,它的程序代码都只是在某个或几个维度逼近人类的相关方面,而非整体性逼近。这个复杂性约束构成了人工智能发展的边界,由此而来的一个复杂性推理有两个互证的方面:一是人类在综合能力上的不可超越在于人类经由数百万年实践积累起来的复杂性,但由于这个复杂性的干预而不能在单个领域胜于专用人工智能;二是人工智能在专项能力上的不可超越在于人工智能经由形式系统施予的单一性,但是由于这个单一性的限制而不能在多个领域胜于人类。② 依据这个复杂性推理,我们可以对人工智能的发展做出这样的具体化表述,即人工智能越是作为单一的形式系统,它在专门领域就越可能发展出人类个体无可匹敌的专业能力;如果它兼容多个形式系统,它虽然可以处理多个领域的事务,心智结构也更接近人类,但是由于复杂性的干预,它不仅在任一领域的专业能力都弱于它作为单一形式系统时所具有的专业能力,在综合能力上也弱于人类个体。

另一方面是人类既可以创造有益的人工智能,也可以创造有害的人工智能,甚至一些中性技术也会被人类用于坏的目的。由此,当我们谈论人工智能的发展时,要注意这个区分:(1)无任何人为介入性环境下,人工智能的发展可能性;(2)人为介入性环境下,人工智能的发展可能性。聚焦这个区分,既有助于在科学研究上探知人工智能的发展边界,也有助于人类反观自身。从以上分析来看,更合理的理论前提应当是(2)。并且这个观念也是可疑的,即人工智能虽然有着许多不足甚至缺陷,但是人类同样也有许多不足和愚蠢,因此人工智能对人类的替代,只需要跑得比人类快。更合理的观念是人类—人工智能同步发展,或更弱些,人类具有足够适应性。③ 这个适应性一方面是对人工智能的发展的介入性干预,另一方面是人类借助人工智能的发展来实现自身在智力上的发展。据此,研究者当下要做的不应是简单地拒斥或夸大人工智能在意识方面的发展,而是从不同角度展开人工智能产生心智的过程,由此探讨约束人工智能的适度干预原则和/或伦理原则,以便让人工智能的发展造福于人类。

作者简介:秦子忠,海南大学马克思主义学院副教授。

〔责任编辑:赵　涛〕

① [美]侯世达:《哥德尔、艾舍尔、巴赫:集异璧之大成》,本书翻译组译,商务印书馆2019年版,第135页。

② 秦子忠:《人类的复杂性及其程序化的限度——兼评"人类终结论"与"竞速统治论"》,《自然辩证法通讯》2021年第1期。

③ 这是个合理的假设,否则人类早就灭亡于地球环境的变迁了,相较而言,现在的技术环境是一个更容易应对的变化。

Main Abstracts

Mind activities are guided by value orderings mapped to the relation of the elements of self-value, the relation of the elements of goal-value, and the whole relation of self-value and goal-value respectively. These value orderings are connected and transitive. With this as the core idea, the test method is designed to test whether artificial intelligence has a self-consciousness. This method also explains the process of how artificial intelligence produces self-consciousness, thus helps human to objectively evaluate the development of artificial intelligence in consciousness, and to adopt the principle of moderate intervention or ethical principle.

At the beginning of German classical philosophy, Kant established the main tone of constructing practical philosophy with "positive freedom" in the subject's practical reason. In his spiritual philosophy, Hegel transformed freedom into objective universal ethical spirit through externalization of subjective free will. Meanwhile, he retained the positive freedom connotation and established practical philosophy with realistic freedom dimension on this basis. Marx's practical philosophy has changed from Hegel's ethical realm of objective spirit to the realm of social reality of human material production. Historical materialism contains the positive freedom dimension of "human historical legislation". Human nature implies the dialectical unity of "subjective will" and "objective action" in the objectification of practice. Practical freedom is created according to the self-promulgated "quasi essence" and this is also Marx's practical interpretation of positive freedom, as there is isomorphism based on positive freedom in the practical philosophy of Hegel and Marx.

After building a moderately prosperous society in all respects, China has entered the stage of taking solid steps toward common prosperity, and common prosperity becomes a central topic of modernization construction. Promoting common prosperity in high-quality development is epochal propositions when high-quality becomes the development theme and common prosperity becomes a central issue in the context of a specific era. In promoting common prosperity in high-quality development, it is necessary to grasp the reform logic of market-oriented elements in the new era, to build a high standard market system, to promote high-level dynamic balance between supply and demand, to build driving mechanism for high-quality development, and to promote high-quality wealth creation and fair and reasonable distribution of wealth.

Human's carbon use causing global warming is not only an environmental problem, but also a social problem. The key to achieving carbon peak and carbon neutralization is to promote the zero-carbon society construction which includes the shift of development concept and mode, the realization of environmental justice, the joint participation of all parties, the transformation of everyday lifestyle and so on. China's zero-carbon society construction has initially achieved an "integrated diversity" pattern, "1+N" policy framework, systematic theory and typical cases. Yet, it faces many challenges: the short window period of carbon emissions governance and the arduous task, high-carbon economy inertia, tensions between climate politics and survival morality, and unsound system policies, technology and macro governance system. Thus, it is feasible to resort to the sociological perspectives of reflection and criticism, professional tools, intervention activity, and policy construction to seek the solutions.

The "mass" as a concept has undergone the transformation from negative term to positive one in modern context, at the same time, it has objectively shown its collective force and energy as social-historical creator. However, some kind of dual and paradoxical attitudes to "mass" has been revealed. On the other hand, it is a basically objective concept which has appeared in historical texts and discourses. These factors constitute the inner tensions within the modern conception of mass. It may have varied implications to investigate its origin and variants in Chinese context, to analyse the social foundations of mass psychology and consciousness for recognizing its historical and actual issues. It is of significance to establish subjective consciousness as citi-