人工智能体有自由意志吗

南　星

摘　要　随着人工智能技术的高速发展，人工智能体变得越来越自主，以至于人们会很自然地对它们是否拥有自由意志产生疑问。然而，传统哲学中的主流观点以及大多数人的直觉都会否认机器拥有自由意志的可能性。在经典的“反应态度”理论的基础上，可发展出一种“实践自由”的概念，人工智能体可以满足这一概念所要求的两个重要条件，即行为的不可预测性和潜在冲突的目标，但在对期待和要求的共同认可这一关键条件上会面临十分严重的困难。通过对人工智能体拥有自由意志的可能性条件进行分析，我们能够更深入地理解（人类）自由意志的本性和基础。

关键词　自由意志　人工智能　司倬森　反应态度

作者南星，北京大学哲学系/外国哲学研究所助理教授（北京 100871）。

中图分类号 B0　　**文献标识码** A　　**文章编号** 0439-8041(2021)01-0035-13

在传统的哲学图景中，“自由意志”是唯独人类才拥有的一种特殊能力，凭借这一能力，人得以超出于地球上其他一切事物之上而获得某种独一无二的尊严。但随着人工智能技术的高速发展，人们发现，人工智能体[①]似乎也具有考量、判断和选择的能力，这些能力在传统的哲学图景下是与自由意志密不可分的。因此，追问人工智能体是否可能具有自由意志的问题就并非漫无边际的胡思乱想，而是对于理解（人类）自由意志本身的本性和基础，以及思考人工智能体可能具有的道德地位具有重要意义。在本文中，我首先会对相关概念做进一步的界定，说明我们需要何种自由意志理论来处理与人工智能体相关的问题。在此之后，我将引入并发展英国哲学家司倬森（P. F. Strawson，1919—2006，过去通常译为“斯特劳森”，但该译名与实际发音相去甚远）影响深远的“反应态度”理论，从而为探讨人工智能体的自由意志问题奠定理论基础。最后，我将尝试在这一基础上考察人工智能体是否可能拥有自由意志的问题，并对人工智能与自由意志之间的关系给出更具一般性的评论。

一

“自由意志”的概念肇端于古代晚期，在基督教思想传统中被发扬光大，成为西方哲学中人的形象的核心要素之一。在基督教哲学的图景下，自由意志乃是人类灵魂的一个基本属性，凭借这一属性，人得以在善恶之间自由地做出选择，并根据自己的选择而在死后享受永恒的福祉或遭受永恒的惩罚。在现代世界中，人们一般不再接受传统哲学中关于人类灵魂的学说。在失去了非物质的灵魂这一天然的“居所”之

① “人工智能体”系我国人工智能学界对“artificial agent”一词约定俗成的翻译，但“agent”一词实际来源于拉丁文“agere”，意思是“行动、活动”。在哲学文献中，该词通常被译作“行动者”“能动者”“自为者”“主体”等。本文为兼顾该词的本义和人们的阅读习惯，在不同语境中会采用“人工智能体”和（人类）“能动者”这两种译名，请读者注意到它们之间的内在关联。

后，自由意志便无可避免地成为了现代科学所描绘的世界中无家可归的“游魂”。尽管如此，人们却并不愿意轻易抛弃自由意志的概念，而是依然倾向于相信自由意志的存在及其重要性。这一方面是因为每个人几乎无一例外地都有自由意志的体验，都在不受外在因素强制的情况下，在不同的可能性之间进行过选择；另一方面是因为人们普遍相信，一个人只有拥有自由意志才能为自己的选择和行动担负道德责任。离开了自由意志的概念，我们的道德判断和司法实践就都无法拥有一个稳固的基础。因此，如何应对自由意志的概念与现代科学所蕴涵的决定论世界观[①]之间的张力，就成了近代以来关于自由意志的哲学讨论的核心关切。

传统哲学中关于自由意志问题的讨论早已汗牛充栋，但人工智能体的出现为我们思考这一问题开启了新的维度：无论我们是否认为人工智能体拥有自由意志，通过对人工智能体和人之间的比较研究，我们都将以前所未有的方式揭示出自由意志的本性和基础。不过，为了严谨起见，有必要在此预先规定一下“人工智能体”的含义。首先，按照通行的定义，我们将“智能体”（或“能动者”）理解为能够与环境发生互动的人或物。[②]为此他们或它们就需要传感器来接收环境中的信息，在自身内对这些信息进行加工后，再通过执行器来对环境发生作用。其次，我们将人工智能体限定在以数字计算机为核心的装置上。因此，它的“智能”或其他“心理”属性就都建立在数字计算能力的基础之上。为了处理不同的任务，计算需要按照不同的规则进行，每一组计算规则的总和就是一种特定的“算法”，而这些算法到目前为止以及在可以预见的将来，都是由人类来开发和编写的。

就人工智能体能够通过接收和处理信息来与环境发生互动而言，它们展现出和人们通常印象中的机器很不相同的性质。而当它们处理信息的内在机制趋于复杂，以至于人们会很自然地用诸如“想要”“认为”“决定”等心理概念来对其进行刻画时，将自由意志赋予人工智能体就不是一件不可想象的事情了。但另一方面，按照我们关于自由意志的直觉，仅当一个人不受外在决定时他才是自由的。由于人工智能体的一切活动都是被外在环境以及人类为它编写好的算法决定的，因此人们并不倾向于将自由意志赋予人工智能体。在目前的技术条件下，后一种直觉似乎占据了上风，这一方面是因为人们大都认为，我们只是在一种隐喻的意义上把那些心理概念应用于人工智能体之上，而后者实际上并不具有和人类相同甚至仅仅相似的欲望、信念或意愿等心灵状态；另一方面则是因为，支持后一种直觉的论证清晰明了，很容易让人确信人工智能体不可能拥有自由意志。

然而事情并没有这么简单。首先，我们在这里关心的是“可能性”的问题，即在我们定义下的人工智能体是否有可能拥有自由意志的问题。即使当下的人工智能体离拥有自由意志还相距甚远，也不能由此直接得出对可能性问题的否定答案。其次，在上文中已经提到，在近代科学的背景下，人类自身的行动似乎也难逃自然法则的决定，但许多人认为这并不意味着我们应当放弃自由意志的观念。在当代关于自由意志的讨论中，最主流的观点就是所谓“相容论”，即认为自由意志与决定论是相容的。如果相容论为真，那么将自由意志赋予人工智能体就远远没有看上去那么荒谬。当然，即使相容论为真，也并不意味着人工智能体就能够拥有自由意志。要想回答这个问题，我们必须先弄清楚当我们将自由意志归属到人身上时究竟是出于什么理由或依据什么标准。只有当人工智能体也有可能满足这些标准的时候，它们才可以被视为拥有自由意志。因此，我们接下来的任务就是找出并且证成这样的理由和标准。

需要预先说明的是，古往今来，人们关于自由意志的本性以及赋予自由意志的标准有许多各不相同、

① 有学者认为量子力学中的“不确定性原理”打破了近代科学蕴涵的决定论世界观，并将这一点视为自由意志在科学上的基础，但大多数学者均对量子力学与自由意志的相关性表示怀疑。关于量子力学与自由意志之间关系的一篇优秀的综述是：David Hodgson, “Quantum Physics, Consciousness, and Free Will,” in Robert Kane (ed.), *The Oxford Handbook of Free Will* (second edition), Oxford: Oxford University Press, 2011, pp. 57-83. 下文第四节中会稍微触及这个问题。

② 参见 Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach* (third edition)（影印版），北京：清华大学出版社，2011 年，第 34 页。

但都不无理据的看法，即使在相容论阵营内部也有相当大的分歧。我无意主张本文中对赋予自由意志的标准的刻画是唯一正确的，但我的确认为任何一种对自由意志的恰当刻画都需要满足一些基本要求，而本文中给出的刻画是能够满足这些要求的。具体而言，这一刻画首先应当基本符合人们关于自由意志的直觉。一个受过一般教育、但没有经过专业哲学训练的人尽管不能对自由意志给出一个严密的定义，但仍然会有一些关于这个概念的直觉式理解，如自由意志典型地体现在选择活动当中，与强制或必然性相互冲突，是道德责任的基础等。哲学家在构造理论时不需要完全符合每一个人的直觉：在哲学家看来，有一些直觉过于模糊，只有经过充分澄清之后才能获得明确的意义，还有一些直觉则因为和其他的、得到更好的理由支持的信念相互冲突，从而必须加以放弃。然而，这绝不意味着哲学家可以随心所欲地构造理论。人们关于一个概念的直觉即使不是理论构造的终点，至少也是它的起点。对这些直觉的偏移只有在有充分理由的情况下才能被允许。

其次，我们对于自由意志的刻画应当尊重现代科学的基本结论。如前所述，人们之所以对基督教传统中的自由意志观感到不满，最主要的原因之一就在于，它是与现代科学所蕴涵的决定论世界观格格不入的。直到今天，仍然有很多人相信现代科学（包括物理学和心理学等）直接证明了自由意志并不存在。其他人即使不愿意彻底抛弃自由意志的概念，在现代科学提出的挑战面前，他们也需要重新解释这一概念，乃至对其作出必要的修正。关于这一点，美国哲学家丹尼特（Daniel Dennett）打过一个很好的比方：如果把自由意志比作爱情的话，那么传统的自由意志观就好比丘比特射箭的传说；现在人们不再相信这一传说了，但这并不代表爱情本身是不存在的。我们所需要的毋宁说是对爱情（以及自由意志）的一个更加合理的解释。[①] 按照这一思路，我们与其将现代科学的基本结论看作否认自由意志存在的直接证据，倒不如将它们视为帮助我们找到对自由意志概念真正恰当的刻画的重要线索。[②]

再次，我们对自由意志的刻画应当说明它的价值。其实这一点可以说已经隐含在上述第一项要求之内，因为我们对自由意志的直觉中无疑已经包含了相当多的价值性要素。然而我们必须格外强调这一要素，因为人们之所以在发现自由意志的观念与现代科学之间存在严重冲突之后，仍然不愿意完全放弃这一观念，最根本的理由就在于这一观念承载了太多的价值。由于任何人都无法忽略这些价值，因此人们不得不严肃对待自由意志的观念。[③] 如果一种关于自由意志的构想中没有包含价值的要素，那么我们大概就不必严肃对待这一构想。举例来说，很多人会将自由意志与非决定性或随机性联系在一起。这种想法也许符合人们关于自由意志的部分直觉，但其最大的问题在于无法说明自由意志具有什么样的价值，从而切断了自由意志的概念与人的尊严和责任等重要议题之间固有的密切关联。同样，我们之所以关心人工智能体的自由意志问题，一个十分重要的理由就在于，我们想知道人工智能体可能承载什么样的价值，并进而思考我们在与之打交道时可能需要遵守的规范。

一般而言，一种对自由意志的刻画如果能满足以上三个条件，那么就可以说是相当成功的了。不过，为了有意义地探讨人工智能体的自由意志问题，我们对这一概念的刻画最好还要满足第四个条件，即它应当尽可能少地依赖于自我意识或主观体验。理由很简单：我们在人工智能体是否有可能拥有自我意识或主观体验这个问题上可以说还一无所知，更无法设想它们的自我意识或主观体验可能会是什么样子的。因此，如果自由意志需要预设特定类型的自我意识或主观体验，那么为了回答人工智能体是否可能拥有自由意志的问题，我们就必须预先回答一个至少和它同样困难的问题，这对于我们的讨论来说显然是不利的。当然，如果任何一种能够满足以上三个条件的自由意志概念都必须预设某种自我意识或主观体验，那么我

① 参见丹尼尔·丹尼特：《自由的进化》，辉格译，太原：山西人民出版社，2014 年，第 275—276 页。

② 如英国学者西蒙斯就认为，传统的自由意志观是前科学文化造成的幻象，而按照一种科学的、有意义的自由意志观，人工智能体或计算机有机体也完全能够拥有自由意志。参见 Geoff Simons, *The Biology of Computer Life*: *Survival*, *Emotion and Free Will*, Boston: Birkhauser, 1985, p. 109.

③ 这一点在人工智能领域的重要开拓者明斯基那里表现得非常明显：他一方面宣称现代科学没有为人类的自由意志留下任何空间，但随即又承认离开自由意志的概念，人的生活将变得毫无价值。参见 Marvin Minsky, *The Society of Mind*, New York: Simon and Schuster, 1986, pp. 306f.

们就必须接受这一事实，并在此基础上对人工智能体是否拥有自由意志做出合理的猜测。

在下一节中，我将引入司倬森著名的“反应态度”理论，并指出该理论可以满足以上全部四项要求。

二

司倬森于 1962 年发表的《自由与怨恨》（*Freedom and Resentment*）一文从根本上改变了人们思考自由意志问题的方式，可以说是整个 20 世纪影响最大的哲学论文之一。在这篇文章中，司倬森并没有直接讨论与自由意志的可能性相关的形而上学问题。事实上，他根本没有使用“自由意志”这个概念！与传统的论述不同，司倬森把他的出发点放在了对我们日常生活实践的刻画上。他首先区分了人们日常生活中的两种态度，即“参与者反应态度”（participant reactive attitude）和“客观态度”（objective attitude）。司倬森并没有尝试对参与者反应态度给出严格的定义，而是通过几个典型的例子告诉我们它们指的是什么：

> 一个人在试图帮助我时不小心踩到我的手，与他以蔑视我存在的方式，或抱着想要伤害我的恶意时踩我的手相比，疼痛可能同样剧烈。但在第二种情况下，我一般而言会感到在第一种情况下不会感到的某种类型和程度的**怨恨**。如果有人的行为帮助我得到我想要的好处，那么不管怎样，我都得到了好处；但如果他出自对我普遍的善意而有意让我得到好处，我会合理地抱有一种**感激**之情，而如果这种好处是他的一个有不同目标的行动计划偶然造成的结果，他对此结果并非有意、甚至感到后悔，那么我就不会抱有这样的感情。[①]

由此可见，所谓“反应态度”，指的就是我们在与他人正常交往中，针对他人在行为中表现出的意图和态度自然产生的反应。除了怨恨和感激以外，典型的反应态度还包括原谅和愤怒等。只要我们采取这些态度，就意味着我们投身或参与到了正常的人际关系当中。客观态度则与之不同：

> 对另一个人采取客观态度，也许就是把他看作一个社会政策的对象；看作一个在宽泛意义上可以说是需要处置（treatment）的对象；看作一个需要加以考虑，或许是预防性考虑的东西；某种需要管理、处理、治疗或训练，也许要干脆避开的东西…… 如果你对一个人的态度是完全客观的，那么尽管你可以与之争斗，但你却不能与之争辩，尽管你可以与之交谈甚至谈判，但你却不能与之说理。你至多可以装作与之争辩或说理。[②]

不难看出，这两种态度的区分和自由意志问题之间的密切关联：当我们对一个人采取反应态度时，我们似乎就认为他拥有自由意志，而当我们对一个人采取客观态度时，我们似乎就认为他没有自由意志。更粗略地说，当我们以反应态度看待一个人时，我们就将他看作一个完整意义上的人，而当我们以客观态度看待一个人时，我们则将他看作一个物。不过，这一说法是十分粗略的，因为它似乎在暗示，当我们对一个人采取客观态度时，我们就是在“物化”他，在把他“仅仅当作手段”。但这种理解是不够准确的：我们完全可以在道德中立乃至出于善意的情况下对人采取客观态度，如一名教练在教学员练习特定技能时，或者病人家属在哄劝身患绝症的病人时，都需要采取客观态度。从两个例子中我们还发现，反应态度和客观态度并不是相互排斥的。我们在对待同一个人的时候，完全可以同时既采取反应态度，又采取客观态度。但这一点并不能抹煞这两种态度之间的差别。在司倬森看来，这一差别最明显地表现在我们的情感上：“客观态度可能以很多种方式得到情感上的表达，但不是以所有的方式……它不能包括那些属于介入或与他人一起参与到人际关系中的反应性感受和态度；它不能包括怨恨、感激、宽恕、愤怒，也不能包括两个成年人有时可以说是相互感受到的那种对彼此的爱。”[③]

在阐明这两种态度间的区别之后，司倬森考察了我们是如何在它们之间进行切换的：当我们受到冒犯或伤害时，我们会自然产生一种怨恨之情，但有两种考虑可以“减缓、平息或完全消除”这一感受。在第一种情况下，我们无需切换对一个人的态度，而只需要把特定的伤害和这个人本身在某种意义上切割开来。在上面的例子中，我之所以对一个在试图帮助我时不小心踩到我手的人一般不会感到怨恨，是因为我

---

① P. F. Strawson, *Freedom and Resentment and Other Essays*, London: Routledge, 2008, p. 6. 着重号为引者所加。

② Strawson, *Freedom and Resentment and Other Essays*, pp. 9f.

③ Strawson, *Freedom and Resentment and Other Essays*, p. 10.

相信他的行为不是故意的，相信他踩我手这件事不是出于他的性格、行为方式或对我的一般态度的，而只有这些才被认为是真正属于他的。这类情形“并不促使我们将能动者看作一个完全能负责任的能动者以外的东西。它们促使我们将伤害看作一个他不能完全为之负责乃至根本不能为之负责的伤害”[①]。与此相反，在第二种情况下，我们需要彻底转变对待他人的态度。当我们得知对我们造成伤害的是一名幼童或精神病人时，我们（作为文明人）就不会对他们产生怨恨之情（或者至少会减弱这种感受），但这并不是因为我们想要把特定的伤害和造成伤害的人切割开来，而是因为我们根本无法将这个人看作负责任的能动者，无法将他看作正常人际关系的参与者，从而不得不对他采取客观态度。

从这两种情况的区别出发，司倬森便提出了他支持相容论的第一个论证。他指出，“如果有关于决定论的一个融贯的论题，那么一定有一种‘被决定’的意义，使得如果这个论题为真，那么无论什么样的行为都是在这个意义上被决定的”[②]。换句话说，决定论作为一个一般性的论题，对以上两种情况是同样适用的。因此，我们之所以在前一种情况下可以保持反应态度，在后一种情况下则需要采取客观态度，这必定是出于决定论之外的理由。我在上文中已经列举了一些相关的理由，在它们之外当然还有许许多多其他的可能性，但无论如何，决定论这一一般论题是不可能包含在这些理由当中的。这也就意味着，决定论的真或假，对于我们是否对某个能动者采取反应态度、是否将其视为担负道德责任的能动者而言是毫无影响的，从而相容论就得到了证明。

不过，这一论证未必能够说服那些视决定论为对自由意志的威胁的人。这些人也许会提出反驳说，我们之所以在第二种情况下对他人采取客观态度，正是因为我们认为他们的行为是被决定的，而我们之所以在第一种情况下对他人采取反应态度，是因为我们并不认为他们的行为是被决定的。然而，如果决定论是真的，那么在这种情况下他们的行为事实上就同样也是被决定的，而我们通常的看法只不过是由于我们认识能力的不足而造成的错误而已。一旦我们学会正确看待包括人在内的整个世界，我们就应当彻底放弃反应态度，而改用客观态度去看待一切。[③]针对这一反驳，司倬森提出了他支持相容论的第二个论证。他写道：

> 我想，人类对参与日常人际关系的投入（commitment）是太彻底，太根深蒂固了，以至于我们不能严肃对待下面的想法：一般的理论上的确信可能会如此这般改变我们的世界，使得其中再也没有任何像我们通常理解的人际关系那样的东西；而进入到我们通常所理解的人际关系当中，就意味着向正在谈论的一系列反应性态度和感受敞开。[④]

在这段话中，司倬森试图表明，完全放弃反应态度是一件心理上不可能做到的事情。决定论作为一个纯粹理论命题，并没有力量从根本上改变人类的生活实践。一个人即使在哲学上毫无保留地相信决定论为真，在日常生活中，他大概也不可能对他人故意对他造成的伤害无动于衷，而将它们与由自然原因或意外导致的伤害视作同一类东西。在后来出版的《怀疑论与自然主义》一书中，司倬森再次强调了这一点，并明确表示他给出的是一种自然主义的论证：“我们自然地是社会存在者；与我们对于社会存在的自然投入相伴随的，是对于由我谈到的那些人类个人的和道德的反应性态度、感受和判断组成的整个网络或结构的自然投入。我们朝向这些态度和判断的自然倾向自然地为我们提供了保护，以防范那些暗示这些态度和判断在原则上是没有保障的或未经证成的论证。”[⑤]

然而，司倬森观点的反对者或许对上面的论证仍不满意。他们会坚持说，即使人们事实上无法离开自由意志的观念来生活，自由意志仍然有可能只是一个幻象。这正如我们无法不将放在水杯里的直棍看作弯

---

① Strawson, *Freedom and Resentment and Other Essays*, p. 8.

② Strawson, *Freedom and Resentment and Other Essays*, p.

③ 这一观点在哲学史上最有影响的代表人物是斯宾诺莎。

④ Strawson, *Freedom and Resentment and Other Essays*, p. 12.

⑤ P. F. Strawson, *Skepticism and Naturalism: Some Varieties*, New York: Columbia University Press, 1985, p. 39.

折的一样。也许人类无论如何也无法摆脱这些视觉幻象，但它们毕竟是幻象，这就意味着我们应当把看到的东西和事物真实的属性区分开来。司倬森预料到了这一反驳，并针对它给出了自己支持相容论的第三个论证：

> 人们也许会说，所有这些都没有回答真正的问题……因为真正的问题不是关于我们实际做什么，或者为什么做这件事。它甚至不是一个关于如果某个理论上的确信得到普遍接受，我们事实上会做什么的问题。它是一个关于如果决定论为真，那么做什么是理性的问题，一个关于对日常的人际态度进行理性证成的问题。对此我的回答是，首先，只有对于一个完全没有领会前面的答案的要旨，即没有领会我们对日常人际关系态度的自然的、人性的投入这一事实的人来说，这个问题才显得是真实的。这种投入是人类生活总体框架的一部分，它不能像特定案例在这个总体框架内被审视那样被审视。我会回答，其次，假如我们能想象一下我们不能拥有的东西，即在这件事上我们可以选择，那么我们也只有根据对人类生活的得与失，它的丰富或贫乏进行评估，才能理性地做出选择；而决定论这一一般性论题的真或假不会影响这种选择的合理性。[①]

司倬森的这一回应有两大要点：首先，他区分了人类生活的总体框架和框架内部的具体问题。对具体问题的证成只能在总体框架内进行，而框架本身却是既不可能、也不需要加以证成的。以上面提到的视觉幻象为例：无论我们怎么样观看事物，有一些基本条件是始终需要得到满足的，如任何可观察到的事物必定在空间中处于一定的位置，必定具有某种颜色（包括“无色”）和形状等。我们对于特定位置、颜色或形状的感知当然可能是真实的，也可能是错误的，但这些问题只能在我们视觉空间的总体框架内才能得到有意义的讨论；与之相反，框架本身作为讨论的前提，必须被接受下来。[②] 尽管司倬森没有明言，但这一论证策略显然与他在《个体》中对于我们认识世界的基本概念框架的论证一脉相承——用后者中的一个重要概念来说，司倬森以“先验论证”（transcendental argument）的方式，证明了反应态度乃是人类生活实践的一个“先验条件”。

其次，司倬森指出，即使我们能够（反事实地）选择我们生活的总体框架，我们的选择所依据的理由也绝非是决定论这一抽象而一般的形而上学命题，而是对于人生意义和价值的具体分析。这一论点实际上已经超出了关于自由意志问题的讨论，而是指向一般而言的哲学方法论。在司倬森看来，传统哲学中将形而上学命题当作出发点或第一原则的做法是一种对我们生活实践的事实的“过度理智化”。[③] 这种过度理智化的倾向也许对科学来说是必要的，但在哲学中，以及在富于人性的生活实践中，我们有很好的理由来抵制这一倾向。司倬森用颇为吊诡的语言表达了这一想法：“假如这种选择是可能的，那么去选择比我们［实际］是的那样更加纯粹的理性，这未必是理性的。”[④] 司倬森的这一想法会让人们很自然地联想到康德关于“实践理性的优先地位”的著名观点。[⑤] 不过在司倬森看来，我们的“实践理性”（借用康德的概念）所关注的与其说是道德法则，倒不如说是人类生活本身。

通过以上论证——概念分析、自然主义描述、先验论证以及“实践先于理论”，司倬森阐明并且辩护了一种相容论的自由观。回顾上一节中提出的几项条件，我们可以得出以下结论：首先，司倬森的自由观充分体现了自由的价值。自由作为植根在人性深处的一个基本要素，是正常人际关系的先决条件。司倬森认为，当特定的人际关系被普遍化之后，就展现出道德的要求，而当正常的人际关系被反身地应用之后，就展现出道德自我的观念。因此，这种意义上的自由同时也是道德和自我观念的基础。传统哲学中对于自由意志的价值的根本性和绝对性的强调，从而也就包容在了这种自由观当中。其次，通过对价值因素的强

---

① Strawson, *Freedom and Resentment and Other Essays*, p. 14.

② 在《自由与怨恨》接近结尾的地方，司倬森对归纳问题做了类似的处理：“人类对于归纳式的信念形成的投入是原初的、自然的、无关理性的（不是非理性的），它根本不是某种我们选择或可以放弃的东西。”（Strawson, *Freedom and Resentment and Other Essays*, p. 28n. 7）关于具体的归纳式推理的证成或修正只有在这一背景之下才是可能的。

③ Strawson, *Freedom and Resentment and Other Essays*, p. 25.

④ Strawson, *Freedom and Resentment and Other Essays*, p. 28n. 4.

⑤ 参见康德：《实践理性批判》，韩水法译，北京：商务印书馆，1999 年，第 131—133 页。

调，司倬森的自由观抓住了我们关于自由意志的直觉的一些重要方面。毋庸讳言，这一理论并没有容纳我们关于自由意志的直觉中的全部方面，如意志对行为的控制、自由选择过程中常见的纠结等。或许正是因为这个原因，司倬森并没有用“自由意志”这个词，而是一般地谈论自由和道德责任。但由于价值因素在关于自由意志的传统理论中的核心地位，我们有充分的理由将司倬森的见解看作传统自由意志理论的延续和修正。再次，司倬森的自由观不仅与现代科学的基本图景并不冲突，而且也无需陷入“自由意志论晦涩而令人惊惶的形而上学”。[①] 后者无疑指的是哲学史上以康德和瑞德（Thomas Reid，1710—1796）等人为代表的所谓“能动者因果性”（agent causation）理论，该理论主张人类能动者具有一种和一般自然事件不同的因果作用力，这种作用力构成了人类自由意志的基础。能动者因果性理论的支持者同样力图在现代科学的基本图景与自由意志之间做出调和，他们的理论构造也不可谓不精巧，但司倬森恰当地指出，这种理论带给人们的困惑与不安丝毫不亚于它所解决的问题。因此，司倬森自由理论的一个重要优势就在于，让我们能够避开这些形而上学构造。最后，司倬森的自由理论并不直接诉诸任何内在体验或自我意识，因此非常适合用来讨论人工智能体是否拥有自由意志的问题。

三

不过，在把司倬森的自由学说应用到人工智能体上之前，我们还要对它做进一步的考察和发展。尽管有上面列举的那些优点，但与历史上一切影响深远的哲学理论一样，司倬森的学说也受到了来自方方面面的批评。[②] 在这里我将只考察其中与本文主题关系最为密切的两个。首先，在上一节中我们已经考虑了这样一种反驳，即我们之所以对他人采取反应态度，仅仅是出于我们的无知而已。司倬森对这一反驳的直接回应，是他对于人类心理的自然主义描述：人类的自然天性让我们无法放弃对建立在反应态度基础之上的正常人际关系的投入。然而，即使我们承认司倬森的回应成功地辩护了一般而言的自由意志（这一点绝不是没有争议的），但他的理论却无法说明我们在特殊情形下应当如何做出判断。毕竟我们在特殊情形下采取哪种态度，在很多情况下既不是人类天性中不可改易的事实，也不是人际关系或道德生活的先验条件。但如果缺乏在特殊情形下做出判断的标准，司倬森的理论就将无助于回答人工智能体是否可能拥有自由意志的问题。[③]

对司倬森自由理论的第二个批评，是认为它没有充分把握我们赋予自由意志的全部意义。站在司倬森的立场上，我们可以回应说这一理论抓住了自由意志概念中最重要的方面，即它是人类道德生活和相互交往的基础，而其他方面要么相对而言并没有那么重要，要么根本就是应当抛弃的错觉。但这一回应是不能令人满意的。正如司倬森之子盖伦·司倬森（以下简称“小司倬森”）指出的那样，“一个人对于人类自由的理念的投入，其真正核心在于他对自身的态度以及关于自身的经验。对于一个人对自己最深切的感受——将自己看作自我规定的规划者，行为的实施者，一个能够创造事物、作出牺牲、犯下错误的人——来说，这一理念是不可或缺的一部分”[④]。在小司倬森看来，司倬森对自由的刻画虽然抓住了我们直觉中自由意志概念所具有的某些重要价值，但却遗漏了一些更重要的价值，即自由与自我之间的内在关联。小司倬森的批评无疑是有道理的。事实上，除了与道德责任之间的关联以外，自由意志的重要性至少还表现在它与以下几方面价值之间的联系：真正的创造性，自主或自我规定，尊严，个体性或独特性，对

---

① Strawson, *Freedom and Resentment and Other Essays*, p. 27.

② 对司倬森自由理论最有影响的一些批评和回应收录于 Michael McKenna and Paul Russell (eds.), *Free Will and Reactive Attitudes*: *Perspectives on P. F. Strawson's "Freedom and Resentment*," Farnham: Ashgate, 2008.

③ 这一反驳的力度比表面上看起来要更大，因为如果缺乏在两种态度之间进行切换的具体标准，那么人们就有理由对这一实践本身展开怀疑，进而陷入对自由意志或道德责任本身的怀疑论当中。参见 Thomas Nagel, *The View from Nowhere*, Oxford: Oxford University Press, 1986, pp. 124ff.; Paul Russell, "Moral Sense and the Foundations of Responsibility," in Kane (ed.), *The Oxford Handbook of Free Will* (second edition), pp. 205ff.

④ Galen Strawson, "On 'Freedom and Resentment'," in McKenna and Russell (eds.), *Free Will and Reactive Attitudes*, p. 105.

未来的希望等。[①] 这些价值的实现看起来的确会受到决定论的威胁。因此，如果司倬森的相容论无法为它们奠定良好的基础，我们就很难将其视为对自由意志概念的充分刻画或恰当修正。

这两方面的批评看上去似乎很不相同，但它们实际上都指向人们对相容论自由观的深刻不满：如果想要为客观态度和反应态度之间的切换找到一个可靠的标准，那么我们似乎最终不得不诉诸像自我规定或自我控制这样的观念，而这些观念与我们刚才提到的传统自由意志观中的那些要素显然是密不可分的。这些要素与决定论之间有着显而易见的冲突，司倬森本想利用“反应态度”的概念来绕过这些要素，从而避开决定论这个难缠的幽灵，但我们如果进一步追问反应态度的根据，最终似乎就仍然要回到这些要素上来。换句话说，即使司倬森证明了我们是“自然的相容论者”，传统自由意志观中这些要素的不可消逝性也表明，我们同样是“自然的不相容论者”。[②] 因此，如何化解这两种自然倾向之间的紧张关系，就成了关于自由意志的论争中最根本、也是最困难的问题。

在本文中我并不打算尝试解决这一巨大的难题，不过我们可以转换一下考虑问题的角度：也许人类未必具有某种终极的、形而上学意义上的自由意志，但在日常生活和人际交往中，人们的确会预设自己以及其他正常人都具有某种自由。这两种自由之间并非毫无关系——哲学家和喜欢思考的普通人之所以会对前一种意义上的自由感到困惑，正是因为他们看到了它对于后一种意义上的自由具有理论奠基的作用。尽管如此，我们却不必因为在理论问题上的悬而未决而放弃我们的实践——事实上，作为人类，我们根本不可能彻底放弃这样的实践。我们不妨将这种实践所预设的自由称为“实践自由”。[③] 在是否将实践自由归属给某个能动者这件事情上，人们也许很难给出确定无疑的标准，在某些情况下甚至还存在不小的分歧。但由此并不能推出实践自由的归属是完全任意的，或实践自由的概念本身是毫无意义的。这一点可以类比于认识论中怀疑论的挑战：按照笛卡尔式的怀疑论（或它的现代变种，“缸中之脑”假说），我们任何一条关于外间世界的信念都有可能是错误的。人们也许无法从理论上彻底反驳这种怀疑论，但这并不妨碍人们在日常的和科学的认知实践中对不同信念抱有不同程度的确信。如果我们能够对人们在这些认知实践中证成信念的基本模式给出比较系统的描述，那么就能够获得一种认知证成的理论。这种理论即使不能从根本上回应笛卡尔式怀疑论的挑战，但它也足以说明在认知实践的层面上，人类的确可以说是拥有知识或得到证成的信念的。相应地，为了说明人类拥有实践自由，我们也不妨将那些极端的和富有争议的情形暂且放到一边，先尽可能地去描述这一概念在人们实际的道德生活和人际交往中究竟发挥着什么样的作用。在我看来，这就是司倬森的“反应态度”理论带给我们最为深刻的洞见。我们接下来的任务就是进一步发展司倬森的理论，以更好地理解人类实践自由的意义，并在此基础上考察人工智能体是否可能拥有实践自由的问题。

事实上，司倬森已然说明了我们采取反应态度的根据是什么：“个人的反应态度基于并反映一种期待和要求：我们期待和要求他人对我们自身表现出某种程度的善意或尊重；或者至少期待和要求他人不要表现出主动的恶意或漠不关心。（在特定情况下，什么东西会算作善意或恶意或漠视的表现，这将根据我们与另一个人所处的特定关系而有所不同。）”[④] 当这里所说的特定关系被理解为最普通的人与人之间关系的时候，相应的期待和要求就成了道德的期待和要求。如果我们认为这些期待和要求对于某个人来说并不适用，那么这个人就“没有被看作一个道德上负责任的能动者，看作道德关系的一端，看作道德共同体的一员”[⑤]。在我看来，司倬森的这些论断无疑是正确的，但却不够完整。因为我们对于动物乃至对于像钟表或温度计这样的人造物也会有相应的期待和要求，但这些显然不是反应态度所适用的对象。问题的关键

① 参见 Robert Kane, *The Significance of Free Will*, Oxford: Oxford University Press, 1996, p. 80.

② 参见 Galen Strawson, “On ‘Freedom and Resentment’,” p. 88, p. 104.

③ 这个概念的引入无疑会很容易让人联想到康德在“先验自由”和“实践自由”之间的著名区分。这一区分与这里讨论的问题无疑是密切相关的，但由于篇幅和主题的限制，本文将不涉及康德的具体分析。

④ Strawson, *Freedom and Resentment and Other Essays*, p. 15, cf. pp. 6f.

⑤ Strawson, *Freedom and Resentment and Other Essays*, p. 18.

不仅在于能动者要服从于某些期待和要求，还在于能动者要以特定的方式来服从这些期待和要求。现在的问题是：应当如何理解这一特定的方式呢？

要想回答这个问题，最好的办法就是去观察一下人们实际的道德生活和人际交往。我们不妨再回顾一下司倬森在引入反应态度概念时所举的那几个例子：我之所以会对带着恶意踩我手的人感到怨恨，是因为在最普通的人际交往中，我有理由要求他人不要故意伤害自己，并且我确信道德共同体中的其他成员也完全认可这一要求。只有在这一背景下，我才有理由对他人伤害我的行为产生怨恨之情，而不是仅仅感到遗憾，就仿佛自己是被野狗咬伤一样。感激之所以有别于单纯的幸运，也可以以类似的方式得到说明。原谅的情况则稍有不同：我之所以会原谅在试图帮助我时不小心踩我手的人，是因为我相信他在此之前并没有预料到自己的行为会伤害到我——假如他预料到会踩我的手，他很可能就不会用原先的方式来帮助我。此外，我们也可能出于另外的理由原谅或部分原谅一个对我造成伤害的人：如果他对我造成的伤害并不严重，并且如果不伤害我，自身就会面临无法承受的损失时，我们也理当对他的行为表示谅解。总之，道德生活和其他人际关系中蕴涵的那些期待和要求，首先必须得到人们的认可，但这种认可并不一定导致人们会按照这些期待和要求来行动：由于人类行动中固有的不可预测性，以及常常出现的与这些期待和要求相互冲突（或至少不尽相同）的其他目标，人们的行动常常会超出通常的期待或低于正常的要求。人性中的这一基本事实可以说就是人们采取反应态度的真正根据。

除了作为反应态度的根据以外，这一基本事实与上文中提到的传统自由意志观中包含的其他价值之间也有密切的关联。以创造性或原创性为例，任何一件被认为是真正具有创造性或原创性的作品似乎都要满足两个条件：一方面，它的产出过程一定是不可预测的；另一方面，它的意义必须能够在人类生活的总体框架中得到解释。更加通俗地说，它必须既在意料之外，又在情理之中。如果不满足前一个条件，那么我们会认为这一作品只不过是机械复制的产物，毫无创造性可言；而如果不满足后一个条件，我们则会将它看作单纯的偶然或随机事件的结果，而不是特定的能动者创造出来的作品。再以自主或自我规定为例：一个行动如果是自主的，就意味着它是出于能动者本身的选择或默许而发生的。换句话说，即使给定在某一时刻与某个能动者相关的一切信息——包括他的信念、欲望、目标和性情等，如果不考虑这一（明确的或默认的）选择活动本身，人们就无法严格推论出该能动者究竟要采取什么样的行动。之所以做不到这一点，原因就在于，一个正常人一般而言在某一时刻总是具有不同的、潜在地相互冲突的目标或冲动，在做出选择以前，没有任何人能够确切推论出他最终会选择实现哪些目标或满足哪些冲动。当然，上面提到的那些信息或许足以让我们以很高的概率估计他将要采取什么样的行动，但这样的估计毕竟不同于严格的推论。与之相反，对于一个毒瘾发作或被催眠的人，我们则有充分的理由准确预测出他的行为。因此，潜在冲突的目标与行动的不可预测性似乎就构成了自主或自我规定的前提条件。[①]

综上所述，我们可以得出结论说，人类行动的不确定性或不可预测性、潜在冲突的目标或冲动以及对人类生活中种种期待与要求的认可，这三者共同构成了与人类自由相关的种种价值——成为反应态度的适当对象、真正的创造性、自主和尊严等——的基础。因此，人工智能体是否拥有“实践自由”的问题，就转化成了它们是否能够满足这三个条件的问题。我们接下来就对这三个条件逐一做进一步的考察。

---

① 这个结论可能会面临来自两个方面的反驳：一方面，人们或许会认为这两个条件对于自主来说是不必要的，因为一个德行完满的圣人毫无疑问是自主的，但却并没有潜在冲突的目标或冲动，且因为其行动总是与道德的要求完全一致，故而也是可以预测的。另一方面，人们或许会认为这两个条件对于自主来说还远远不够，因为单纯随机或偶然做出的选择似乎也满足这两个条件，而真正自主的行动却必须保证其根源来自能动者本身。针对前一方面的反驳，我的回应有两点：首先，德行完满的圣人只是一个虚无缥缈的理念，在实际生活中，不管多么高尚的人也很难说彻底排除了私心；其次，也是更重要的，许许多多的道德两难和价值冲突都告诉我们，即使是单纯为了追求道德价值，目标之间的潜在冲突也常常是无可避免的。因此，自主行动必定要求人们在不同的目标和价值之间做出选择和调和。针对后一方面的反驳，我的回应是，在我们描述的情况下，由于潜在冲突的目标或冲动本身就已经在某种意义上属于该能动者了，因此无论他做出何种选择，他的行动都是来源于自身的。如果对这个解释不满意，那么人们就难免要把能动者理解为某种超验的、形而上学的主体，但这样就势必会陷入“自由意志论晦涩而令人惊惶的形而上学”。

四

我们首先来考察不可预测性的概念。在传统的、机械论的哲学图景下，世界上发生的一切原则上都是可预测的；如果人的行动构成了例外，那也只能借助于非物质的灵魂这样的概念来加以解释。由于人工智能体无论如何不可能具有非物质的灵魂，因此它们的一切活动原则上都是可预测的。正如计算机科学的先驱洛芙蕾丝夫人（Lady Lovelace，1815—1852）论证的那样，“分析机并不宣称能创作（originate）任何东西。它能做任何我们知道怎样命令它去执行的事”[①]。图灵在他的名文《计算机器与智能》中严肃对待了洛芙蕾丝夫人的这一反驳并给出了两条回应：首先，当人类认为自己能够做出某种“原创性的工作”时，他所做的很可能“只不过是教育在他心中植下的种子的生长，或遵循人们熟知的一般原理而产生的结果”[②]。其次，人们或许会抱有这样的看法，即只有原创性的工作会让人感到惊喜，而机器则无法做到这一点。图灵认为，这种看法“来自一个哲学家和数学家都特别容易犯下的错误”，即“只要一个事实呈现在心灵面前，这个事实的全部后果也就同时涌现到心灵当中”。这一看法尽管并非完全没有道理，但却严重忽视了从原始数据和一般原理出发推导出后果这一工作的重要意义。[③]

图灵的见解无疑是引人深思的。事实上，如果我们希望避免诉诸非物质的灵魂这样的观念，那么有关人类“真正原创性”的来源问题也会让我们感到无比困惑，而图灵的评论则给我们带来了重要的启发。传统哲学中人们倾向于把“因果关系”“决定论”“可预测性”等概念混淆在一起，但我们应当将它们区分开来：一个事件是另一个事件的原因，并不意味着前者就决定了后者，因为这其中还有许多其他事件可能发挥影响；而即使一个或一组事件决定了另一个事件，也并不意味着我们从关于前者的知识中就能严格推论出后者来。换句话说，决定论与不可预测性之间未必存在冲突，而如果“真正原创性”或“自由意志”只需要以后者为条件，那么它们也就不会受到决定论的威胁。英国著名科学家罗杰·彭罗斯试图在科学层面上说明决定论与可预测性之间的区别：在他看来，在量子世界和经典世界的边缘处存在一种他称为“正确的量子引力”（CQG）的过程，这一过程构成了我们意识的基础。该过程包含一个“本质上非算法的要素”，因此“即使未来或许会为现在所决定，从现在出发也不能计算出未来”[④]。不过，由于彭罗斯把决定论与可计算性之间的区分建立在某种非算法的过程的基础上，因此他主张任何以算法为基础的数字计算机都不可能和人一样拥有自由意志。

我在此无法对彭罗斯的科学假说给出恰当的评判。不过，我认为，我们根本没有必要接受彭罗斯为自由意志设立的如此高的标准——在彭罗斯看来，一个自由的或真正原创的活动必须是原则上不可计算的——而不妨回到图灵更加贴近常识的洞见。一方面，在实际生活中，人工智能体早已呈现出许多完全不可预测并让人感到惊喜的事例，如 20 世纪末战胜国际象棋世界冠军的“深蓝”，以及不久前横扫人类围棋高手的“阿尔法围棋”等。它们的活动无疑是不可预测的，因为任何人（包括它们的设计者）事先都不知道它们会如何行棋。另一方面，我们之所以认为自己的（某些）选择是不可预测的，并不是因为我们洞见到在这些选择背后有某种非算法的机制在发挥作用，而是因为我们确信在做出选择之前，任何人（包括做选择的人自身在内）在严格意义上都无法知道我们究竟会做出怎样的选择。正是这一确信构成了我们“自由感”的重要来源。现在的问题是，人工智能体是否也有可能拥有这种自由感呢？

美国计算机科学家赛斯·劳埃德在这个问题上的评论颇有启发性。他认为，我们可以在数学上证明以

---

① 引自 Alan Turing, “Computing Machinery and Intelligence,” in B. Jack Copeland (ed.), *The Essential Turing*, Oxford: Oxford University Press, 2004, p. 455.

② Turing, “Computing Machinery and Intelligence,” p. 455.

③ Turing, “Computing Machinery and Intelligence,” p. 456. 在一篇生前未公开发表的广播稿中，图灵表达了类似的观点：“如果我们给机器一个程序，结果它做出了某种我们没有预料到的有趣的事，我会倾向于说机器创作了某种东西，而不是主张它的行为都隐含在程序当中，因此原创性完全在我们这里。”[Copeland (ed.), *The Essential Turing*, p. 485]

④ Roger Penrose, *The Emperor's New Mind*: *Concerning Computers*, *Minds and the Laws of Physics*, Oxford: Oxford University Press, 1999, p. 558, cf. p. 220.

下结论：预测一个决策过程的结果的问题，在计算上要比执行这个过程本身更加困难。因此，即使决策或选择的过程是机械的或决定论的，任何人原则上也无法绕过这一过程而直接预测其最终结果。[①] 劳埃德认为，这就是人类自由感的真正基础，而由于对这一结论的证明并不依赖于人类的任何特征，因此它对人工智能体也是成立的，换句话说，人工智能体也可以拥有某种自由感。在我看来，劳埃德的论证如果能够成立，就可以看作对图灵第二点回应的实质性的深化和发展，这对于我们理解人类和人工智能体的自由意志及创造性的基础具有相当重要的意义。不过，即使劳埃德的核心论证是正确的，要推论出人工智能体拥有自由感也显得有些操之过急，因为除了他揭示出的机制以外，自由感作为一种感受很可能还需要其他条件，而这些条件或许是人工智能体不能满足的。但无论如何，在不可预测性或原创性这一点上，人类与人工智能体之间似乎并不存在根本的区别。

我们接下来把目光聚集在人工智能体是否能够具有潜在冲突的目标这个问题上。对这个问题似乎很容易给出肯定的答案：只要我们为一个人工智能体设定多个相互独立的子程序，并通过比较这些子程序产生的结果来决定该智能体的行动，该智能体就可以说是具有潜在冲突的目标。不过人们或许会追问，为人工智能体设定潜在冲突的目标这件事有意义吗？在这里需要注意的是，有多个不同目标不等于有潜在冲突的目标——如果我们为不同目标设定明确的先后次序，那么在它们之间就不会有任何冲突。为了让人工智能体发挥多方面的作用，为它设定多个不同的目标这件事显然是有意义的。但如果不同目标之间存在潜在的冲突，那么事情就很不一样了。假设某一台机器人的唯一目标就是照顾老人，另一台机器人则有照顾老人和训练棋艺这两个目标，并且人们不知道它在某个时刻究竟会追求其中的哪一个，试问有任何人会对第二台更加偏爱吗？如果人们实际上并没有任何动机去制造具有潜在冲突的目标的人工智能体，那么我们似乎就须对这一理论上可能的问题在实践上予以否定的回答。

然而，更进一步的考虑会让我们认识到，为人工智能体设定潜在冲突的目标也许并不是一件没有意义的事情。首先，当人工智能体变得越来越“通用”，其智能水平变得越来越高的时候，我们就很难为它的不同目标设定明确的先后次序。这一点从人类身上就可以看得很明显：越是简单的任务，就越容易分解成次序分明的几个步骤来完成，而越是复杂的任务，就越没有一定的规律可循，而需要能动者自己想办法化解完成任务过程中可能遇到的种种潜在的冲突。其次，人们或许会认为，我们至少可以仿照阿西莫夫“机器人三定律”的方式，将道德的目标设定在最优先的地位。然而，在人类面临的许多道德两难的情况下，我们必须在若干本身都具有内在价值、但却处于潜在冲突中的目标之间进行选择，而无法按照一套普遍接受的规范体系来指导我们在特定情况下的行动。因此，如果我们希望人工智能体在这种情况下能够帮助我们——我们有理由期望，与在棋类游戏中一样，高度发达的人工智能体在类似情况下也会做出比人类更加合理的决策——那么我们似乎也应当把这些潜在冲突的价值都“教给”人工智能体，并由它在两难状况中自主地做出选择。[②] 再者，我们知道，多样性对于人类来说是不可或缺的宝贵财富，因为人类各种各样的才华、智慧和创造力都建立在多样性的基础上。因此，我们似乎有理由相信，为了让人工智能体更好地发展它们的能力，也有必要让它们变得更加多样化。而设定潜在冲突的目标，再以随机的方式让不同的人工智能体各自自主地寻求实现这些目标的方式，似乎是实现人工智能体多样性的一条可行的路径。即使人们没有被上面这些理由说服，但除了它们以外，我们一定还能找到其他重要的理由来支持为人工智能体设定潜在冲突的目标。因此，无论是理论上还是实践上，我们都应当对这一做法给出肯定的回答。

最后，让我们来考虑一下人工智能体与人类生活中的各种期待和要求之间的关系。在上一节中我们已经指出，成为反应态度的适当对象，不仅意味着要服从某些期待和要求，而且还意味着要和道德共同体中的其他成员一道认可这些期待和要求。这种认可的基础是什么呢？答案无非是人的共同天性和生活形式。

---

① Seth Lloyd, “A Turing Test for Free Will,” *Philosophical Transaction of the Royal Society A*, 370 (2012), pp. 3597-3610.

② 事实上，有很好的理由表明，为人工智能体制订一套特定的道德规则的做法根本是行不通的。参见温德尔·瓦拉赫、科林·艾伦：《道德机器：如何让机器人明辨是非》，王小红主译，北京：北京大学出版社，2017 年，第 78—84 页。

毋庸讳言，由于时代、文化和个体上的巨大差异，不同的人对同一件事情的看法可能千差万别。在此时此地广受称赞的行为，在彼时彼地也许会遭到普遍唾弃。尽管如此，人类共同的天性和生活形式使得不同的人在最基本的价值和行为模式（如趋乐避苦）上总是表现出广泛的一致，而人与人之间的差异则主要体现在对这些价值的结构以及它们之间先后次序的理解上。由于分享有共同的基本价值，因此人与人之间、文化与文化之间的沟通和相互理解在原则上总是可能的；但由于人们对诸价值的结构以及它们之间先后次序的理解常常是根深蒂固的，因此这种沟通和理解常常是相当困难的。如果相互之间的沟通和理解比较充分，那么人们就会共同认可一些基本的规范，从而也就会倾向于用反应态度来看待对方；反之，如果相互之间的沟通和理解十分匮乏，那么人们就很难在任何实质性的期待和要求上达成共识，因而人们就只能用客观态度来看待对方。

如果此说不谬，那么人工智能体似乎很难成为反应态度的适当对象：问题并不在于人工智能体无法理解人类的思想——从近年来人工智能技术的发展来看，人工智能体已经可以和人类在相当复杂的议题上展开实质性的对话和论辩，其展现出的沟通和说服的技巧让绝大多数人都望尘莫及。[①] 真正的问题在于，我们和人工智能体并没有任何共同的天性或生活形式，因此也谈不上分享哪怕是最基本的价值。我们无法想象人工智能体的快乐或痛苦——即使人工智能体有朝一日能够完美地模拟出人类的快乐或痛苦，这种模拟出的感受对于它们来说，似乎也完全不具有快乐和痛苦本身对于人类来说的那种意义。受到快乐和痛苦的驱使是人之为人不可改易的天性，而是否以及如何在人工智能体上模拟类似的感受则似乎是一个单纯的技术问题。正是因为我们与人工智能体之间这种巨大的不对等关系，让我们不知道对它们究竟应当有怎样的期待和要求，更不知道在什么意义上可以说它们也认可了特定的期待和要求。按照我们目前的想象，即使我们确定了某种合理的期待和要求，当人工智能体满足或者超出期待时，我们似乎也只会像看到一台机器运行良好那样感到满意或惊喜，而不会产生感激之情；而当人工智能体没有满足要求时，我们似乎也只会像看到一台机器出现故障一样感到不满或遗憾，而不会产生怨恨之情。总之，即使人工智能体表现出对人类种种基本价值及其内在结构的充分把握，乃至对不同价值之间的冲突以及可能和解的深入分析，我们仍然会自然地认为，人工智能体向我们传达的一切都缺乏某种“深度”或“厚度”，而这种“深度”或“厚度”对于反应态度来说似乎是不可或缺的。

当然，我并没有在逻辑上证明对人工智能体采取反应态度是不可能的。也许随着人工智能体愈发融入人类生活中，以及随着人的性质本身可能发生的变化[②]，人类将在与人工智能体的和谐共生中自然地发展出一套相互之间的期待和要求。假如这一天真的会到来，那么我们到那时当然可以（甚至应当）对人工智能体采取反应态度，并认为它们和人类一样拥有上文中所说的“实践自由”，进而赋予其相应的道德乃至法律地位。不过，从现在的眼光来看，这一天似乎还遥不可及。而人工智能技术在当下对自由意志造成的最严重的冲击，则体现在它对于人类实践自由的三个条件的影响上：人类的行为变得越来越容易预测；原本相互冲突的目标或冲动变得越来越容易同时被满足；人们共同认可的期待和要求也在发生着微妙的变化。在这样的条件下，我们无疑需要对自由意志以及与之密切相关的种种价值——如原创性、自主、自我实现、尊严等——重新进行审视。

行文至此，我们不妨来做一个简短的回顾。文章的前半部分致力于在司倬森“反应态度”理论的基础上发展出一种“实践自由”的观念，这一观念既能容纳传统自由意志观中所包含的绝大多数价值，又与现代科学的基本图景和谐一致。在此之后，我们分析了支撑这一观念的三大要素，即行为的不可预测性、潜

① 在这方面表现最突出的是 IBM 公司开发的 Project Debater 系统，该系统可以在各种复杂的议题上与人类展开辩论，乃至帮助人们做出决策。详情见 https://www.research.ibm.com/artificial-intelligence/project-debater/。

② 关于人工智能和其他现代科技可能造成的人的性质本身的变化，可参见韩水法：《人工智能时代的人文主义》，《中国社会科学》2019 年第 6 期。

在冲突的目标以及对一系列期待和要求的共同认可。从原则上来说，前两大要素都可以为人工智能体所满足，但它们是否能够满足第三个要素则是颇为可疑的。

自由意志的问题困扰人类已逾千年之久，而与人工智能相关的一切也都扑朔迷离，因此，本文中的许多论证和结论都注定是试探性的。或许这些讨论并不能让人对本文标题提出的问题获得确切的答案，但我希望借助于人工智能体这面镜子，我们得以廓清自己对自由意志的理解。自由意志的价值并不在于它是来自上帝的恩赐，也不完全在于它是我们自主选择、创造性和个体性的基础，而是很大程度上在于它与我们那不起眼的、但却充满人性的道德生活和人际交往之间密不可分的联系。或许这就是司倬森带给我们的最重要的启示，也是本文最终的结论。

（责任编辑：盛丹艳）

Do Artificial Agents Have Free Will?

NAN Xing

Abstract: With the new developments in artificial intelligence research, artificial agents (AAs) become more and more autonomous, thus it is natural to suspect that they might, to certain extent, have free will. However, according to the mainstream view in traditional philosophy as well as to most people's intuition, it is impossible to attribute free will to any machine. This paper engages in this debate by developing a concept of "practical freedom" on the basis of the classic "reactive attitudes" theory. It argues that, while AAs may satisfy two important conditions required by the concept of practical freedom, i.e., the unpredictability of behavior and the potentially conflicting goals, they can hardly satisfy the key condition of mutual recognition of expectations and requirements, thus are unlikely to be regarded as having freedom of will in the near future. Through the analysis of the possible conditions of the freedom of will of AAs, this paper also aims to deepen our understanding of the nature and foundation of the freedom of will (in humans).

Key words: free will, artificial agents, Strawson, reactive attitudes