自我升级智能体的逻辑与认知问题*

任晓明　李　熙

摘　要：自我升级智能体的建立使人们对自我意识的研究有了一个程序化的标准，借助这种形式化的方法有可能弥合学界关于机器意识的分歧，破解机器意识研究面临的困局。但它也有逻辑上的局限。生成主义为自我升级智能体的提出奠定了认知基础。自我升级智能体的成功为生成主义提供了一个强有力的例证。尽管自我升级智能体向机器真正具有自我意识前进了一大步，但是人们只能说它具有了“功能意识”。造成机器意识困局的症结源自分析哲学传统与现象学传统的分歧和偏颇。解决的出路在于：从对立到相容，从互斥到互补，进而达到融通的新境界。

关键词：自我升级智能体　自我意识　生成主义　哥德尔机

作者任晓明，南开大学哲学系教授（天津　300071）；李熙，中南大学公共管理学院哲学系讲师（长沙　410083）。

在认知科学和哲学中，意识是最让人着迷又始终无法解释的问题。自我意识是我们再熟悉不过的了，但它又是最难以解释的。人工智能思想家通常用两种方式研究自我意识：其一是建立自我意识的计算机模型，这叫作“机器意识”；其二是用计算术语去分析自我意识，但不去模拟。① 前者主要是人工智能技术专家的工作，他们通常只关注技术性问题而不讨论关于意识的哲学问题；后者主要是那些对人工智能有研究的哲学家感兴趣的，但他们在有关哲学问题上存在着巨大的分歧。例如，强人工智能系统会有自我意识吗？如果有，它指的是意向性还是感受质，或者什么别的属性？如此等等。这显然不是科学问题，而是人工智能中的哲学问题，是需要哲学家和科学家共同面对，通力合作而加以解决的问题。然而令人尴尬的是，人工智能研究近年来尽管取得了重大突破，但对强人工智能中涉及机器意识这类难题，

* 本文为中国社会科学杂志社哲学部主办的“语言、认知与心灵学术研讨会”参会论文，国家社会科学基金重大项目“现代归纳逻辑的新发展、理论前沿与应用研究”（15ZDB018）阶段性成果。

① 玛格丽特·博登：《AI：人工智能的本质与未来》，孙诗惠译，北京：中国人民大学出版社，2017年，第145页。

哲学家和科学家要么避而不谈，要么泛泛而谈，机器意识的研究举步维艰。

幸运的是，一些机器意识研究成果正悄然改变着这个局面：关于机器意识的认知和哲学研究尽管面临巨大困难，但是自我升级智能体的理论成果有望打破困局，它能不能像图灵机的构建打破了人工智能研究的困局一样，人们正有所期待。自我升级（self-improvement）智能体，亦即自我改进智能体，是通用人工智能的一种理论模型。作为机器意识研究的成果，它试图为破解自我意识难题作出贡献。建立这种智能体的意义不仅仅在于它可以解决问题，而在于它与图灵机一样，可以为我们讨论自我意识的话题奠定一个程序的基础，或者一种科学验证的标准，从而使意识问题不再神秘。这是我们探讨自我升级智能体问题的一个动因。

以下探讨的主要问题有：第一，自我升级智能体在逻辑和哲学上有什么贡献？存在什么局限？第二，自我升级智能体的提出有什么认知意义和应用风险？第三，自我升级智能体能否具有自我意识？这种智能体在理论上的困局是什么？第四，破解自我升级智能体困局的出路何在？

一、自我创生思想的演进

自我升级智能体是一种关于自我意识的智能体。虽然这种思想在古希腊早已萌芽，但作为一种认知科学和人工智能理论是从冯·诺意曼（Von Neumann）开始的。冯·诺意曼第一次以数学的精确性和逻辑的严密性探讨了自创生系统。①

实现自我创生的前提是实现自我复制（self-producing）或自我生产。在冯·诺意曼看来，借助图灵程序来进行自动机的“自我复制”是不够的。因为图灵机输出的是一段带有 0 和 1 的纸带。而冯·诺意曼要构造的是这样的自动机：它的输出是另一自动机。

冯·诺意曼明确指出，借助构造的方法，即通过设计各种构造性的自动机，能够构造出自复制自动机。② 这种自复制自动机不仅能够进行通用图灵机那种计算，而且能自我复制。但这种自动机离真正具有自我意识的智能体还有一定差距。

21 世纪初，斯蒂芬·沃尔夫勒姆（Stephen Wolfram）对冯·诺意曼的细胞自动机理论作了进一步阐述和改进。他指出，可以用简单的电脑程序来表达更一般的

① autopoiesis（自创生）这个词意指自我生产或自我复制。生命系统是自创生的（autopoietic），它们将那些能够产生必要部件并能够持续发展的过程组织了起来。那些并不能自我产生或复制的系统被称为它生产的（allopoietic），例如，一条河流或者一块钻石。自我复制智能体、自我升级智能体都属于自创生系统。

② 参见 John Von Neumann，*Theory of Self-Reproducing Automata*，Urbana，Illinois：University of Illinois Press，1966.

规律，在此基础上建立一种新的科学，启动另一场科学变革。① 如果说冯·诺意曼主要从理论方面阐述自动机如何从简单规则和初始条件进化到复杂系统，如何自我复制，那么可以说沃尔夫勒姆从技术细节方面更深入地探讨了自动机的自我复制功能，为自我升级智能体的建构奠定了基础。

受到细胞自动机研究的启发，认知科学中的生成主义者开展了对细胞自动机模型的研究。他们认为，在自治的复杂系统中，界定自治组织的关系不在于静态实体而在于过程，如细胞中的新陈代谢反应过程。这种自治系统的一个范例是活细胞。在一个活细胞中，其构成过程是化学的；其循环依赖性采取了自我复制的新陈代谢网络形式，这个自我复制的代谢网络产生了它自己的膜；这个网络将这个系统构成生物域中的一个统一体，并决定了与环境的交互作用域。这种自治系统称之为“自创生”（autopoisis）系统。②

不难看出，生成主义是一种非还原论的自然主义纲领。在传统认识论中难以理解的“自我指涉”，在生成主义那里得到了合理的解释，从而为自我升级智能体的建构提供了认识论资源。

总之，生成主义不再关注意向性意义上的意识，而是强调自我意识的自主性，这就为关于自我意识的智能体研究开辟了道路。生成主义对建构自我升级智能体的影响在于，如果没有生成主义的概念，认知科学既不能解释有生命的认知，也不能建立真正有智能的智能体。

在生成主义看来，我们对意识问题的分析就是要解释人的意识是如何从大脑这个虚拟计算机的运作中产生的。他们希望为自我意识找到一个能使其成为科学的研究方法，借助这种实验的方法可以解决自我意识中的难题。虽然自创生系统理论探讨了自动机自我复制的认知基础问题，然而，仅仅具有自我复制能力的机器还不具有自我意识，构造出具有自我意识的智能体，是认知科学和人工智能面临的更为严峻的挑战。而真正使这种前瞻性设想变为现实的是哥德尔机（Gödel Machine）的构想。③

二、“自指”：涉及“自我”的智能体的核心概念

曾经有一本获普利策文学奖的奇书《GEB：集异璧之大成》，该书作者侯世达（Hofstadter）将巴赫的赋格曲、埃舍尔的版画和哥德尔的逻辑定理这三块奇异的瑰璧缠结在一起探讨，广泛涉及人工智能、数理逻辑、几何绘图、古典音乐、生物基因、

① Stephen Wolfram, *A New Kind of Science*, Champaign: Wolfram Media, Inc., 2002, p. 1.

② **参见** Francisco J. Varela, *Principles of Biological Autonomy*, New York: Elsever North Holland, 1979, p. 55.

③ Jürgen Schmidhuber, “Ultimate Cognition à la Gödel,” *Cognitive Computation*, vol. 1, no. 2, 2009, pp. 177-193.

认知心理、形而上学与认识论、禅宗寓言等不同领域，书中充斥着各种语言游戏、歧义、双关、悖论、怪圈、对称、嵌套、镶嵌、自指、跨越、同构等“奇技淫巧”，仿佛一座循环往复、层次错乱、令人目眩的迷宫。① 但居于迷宫最核心的珍宝是两样——“自指”（self-reference）和“对角线”（diagonal），二者又如一枚铜币的两面，被来自范畴论的劳威尔（Lawvere）不动点定理牢牢捕获——让“数学”开口说“我”，或让“我”超越预设。康托尔定理、说谎者悖论、罗素悖论、塔斯基算术“真”不可定义定理、图灵停机定理、哥德尔不完全性定理等都可以看作它的特殊示例。

19 世纪末，康托尔创立了素朴集合论，证明了一个集合的基数严格小于其幂集的基数，从而发现存在不同层次的无穷。但素朴集合论不一致（不协调）。1901 年罗素悖论的发现揭示出所有不属于自身的集合的类不再是一个集合，这直接导致了所谓的第三次数学基础危机。哥德尔 1931 年证明，任何包含初等算术的、一致的、可递归公理化形式系统不可能完全，也不能证明自身的一致。塔斯基 1933 年证明算术真不能在算术内部被定义。图灵 1936 年证明停机问题是不可判定的，从而一阶逻辑的有效性不可判定。上述这些否定性的结果意味着希尔伯特规划的失败，也使莱布尼茨关于通用文字、理性演算的梦想蒙上一丝阴影。这些否定性的结果看上去碰触到了人类理性的边界和极限，也引起了心灵哲学家和认知科学家的密切关注。侯世达的工作就是主要围绕这些定理和悖论展开的。这些结果还被彭罗斯（Penrose）用来论证心智胜过机器，从而人工智能不可能。② 所以有必要弄清这些否定性结果背后的统一结构。所有这些定理、悖论都与哲学上一个古老的悖论——说谎者悖论有关。劳威尔通过范畴论里的一个不动点定理刻画了这些悖论和定理背后的对角线结构。③ 但直到雅诺夫斯基（Yanofsky）通过集合论语言重新表述劳威尔不动点定理之前，这个重要的结果并没有得到逻辑学家和哲学家的足够重视。④ 虽然这些理论结果看上去是否定性的，但它们直接催生了关于涉及“自我”的智能体的研究。下面将分析这个刻画了“自指”和“超越”（transcendence）的定理的更多应用，并揭示它在构造自我升级的通用人工智能体中的核心作用。

首先，康托尔在证明集合 A 的幂集 P（A）的基数大于集合 A 的基数时引入了对角线方法。说谎者悖论、罗素悖论、哥德尔不完全性定理、图灵停机定理等都借助了

① 参见 Douglas Hofstadter, *Gödel, Escher, Bach: An Eternal Golden Braid*, New York: Basic Books, 1979.

② Roger Penrose, *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*, New York: Oxford University Press, 1989, pp. 40-97.

③ William Lawvere, "Diagonal Arguments and Cartesian Closed Categories," *Theory and Applications of Categories*, vol. 92, 1969, pp. 134-145.

④ Noson S. Yanofsky, "A Universal Approach to Self-referential Paradoxes, Incompleteness and Fixed Points," *The Bulletin of Symbolic Logic*, vol. 9, no. 3, 2003, pp. 362-386.

对角线方法。劳威尔通过范畴论里一个非常简洁的不动点定理给出了对角线方法的统一刻画。[①]说谎者悖论、格雷林悖论、蒯因悖论、理查德悖论、雅布劳悖论、罗素悖论、佩里悖论、忙海狸函数、康托尔定理、图灵停机定理等都服从劳威尔的模式。

其次，劳威尔不动点定理抓住了“自指”的核心，它是说谎者悖论、格雷林悖论、罗素悖论、理发师悖论、理查德悖论、蒯因悖论、雅布劳悖论等悖论的关键，也是康托尔定理、哥德尔第一不完全性定理、塔斯基算术真不可定义性定理、勒布定理、帕里克定理、克林尼（Kleene）不动点定理、图灵停机定理、冯·诺意曼自复制自动机和全自省程序的关键，借助它还可以构造不可计算的实数、不可命名的实数、部分递归但非潜递归的函数、佩里悖论、快速增长的忙海狸函数、λ演算版本的哥德尔不动点引理、柯里Y不动点算子、图灵Θ不动点算子，以及它们的“传值”形式的不动点算子，借助“传值”形式的不动点算子，易得克林尼不动点定理。甚至施米德胡贝尔（Schmidhuber）的借助定理证明器进行自指从而进行自我升级的通用智能体即哥德尔机也可以看作克林尼不动点定理的应用特例。[②]

“自指”或“对角线”方法在逻辑中的大部分应用都是证明否定性的结论，在计算机科学中也是如此。比如，科恩（Cohen）用“对角线”方法证明，不存在完美的反病毒软件，即，不存在一个算法能检测出所有的计算机病毒。假设存在某个病毒检测算法A，则根据对角线方法，可以如下构造程序P：“如果A检测出P被感染，则直接退出；否则，传播病毒。”显然，A不能判断P是否被感染。所以，对于任何反病毒软件来说，错杀或漏杀难以避免。专门检测危险程序的程序也面临一样的问题，要么可能放过了真正危险的程序，要么可能误判了安全的程序。[③] 所以，试图通过打造“程序警察”的办法阻止智能体的叛乱会存在很大的安全隐患。虽然这里的“对角线”论证对解决智能伦理问题产生了负面的作用，但“自指”的方法对构建通用人工智能会起到积极且正面的作用。

我们知道，发展通用人工智能最简便最理想的方式可能是，先制造某个弱一点的人工智能体，然后赋予它某种自我进化的能力，比如，让它可以修改自身的源代码，然后希望它能通过自我修改的方式不断自我升级变得更强。但直观上，自我修改源代码的程序很难令人接受，如果自我升级后的下一代更智能，是否意味着初代就已经蕴含了同样水平的智能？自我修改是安全的吗？会不会越改越崩溃？一个允许完全修改自身源代码的程序也可能面临类似的问题，如果它能修改得更好的话，

① William Lawvere, “Diagonal Arguments and Cartesian Closed Categories,” pp. 134-145.

② Jürgen Schmidhuber, “Gödel Machines: Fully Self-referential Optimal Universal Self-improvers,” in Ben Goertzel and Cassio Pennachin, eds., *Artificial General Intelligence*, Berlin: Springer, 2007, pp. 199-226.

③ Fred Cohen, “Computer Viruses—Theory and Experiments,” *Computers and Security*, vol. 6, 1987, pp. 22-35.

为什么没有修改得更坏的可能？如何确保一个程序能够自我修改并修改得更好而不是更坏？这里的关键就是不动点定理。

克林尼不动点定理告诉我们：对于任意的程序 h，总存在某个程序 e，执行程序 e 的结果等价于把程序 e 当作数据输入给程序 h 执行的结果。克林尼不动点定理的证明跟那些有趣的悖论构造差不多，都服从劳威尔不动点定理的结构，看起来像玩弄“自指”的文字游戏，但这并非简单的自指，它对于智能体的自我觉知、自我升级非常重要，它能保证一段程序可以计算出关于自身的各种性质。比如，假设程序 h（x）负责计算任意字符串 x 的长度，根据此定理，存在自测量长度的程序 e，使得执行 e 的结果相当于执行 h（e），也就是说，e 计算得出了自己源代码的长度。再比如，假设程序 h（x）负责编译出 x 所编码的程序。根据克林尼不动点定理，存在程序 e，使得执行 e 的结果相当于输出了程序 e 自身，这就是所谓的自复制程序。冯·诺意曼的自复制自动机是构造性的，而这里的自复制程序 e 看上去是存在性的，但因为克林尼不动点是构造性的，所以这里的自复制程序 e 也是构造性的。

做一个类比，心理学家卡尼曼（Kahneman）认为，人有两个自我：经验自我和记忆自我，经验自我负责动作和决策，记忆自我负责解读反思。① 瑜伽教练的“言传身教”可以看作这两个自我的配合，记忆自我对经验自我的肢体演示过程进行了逐步的反思，并将反思的结果精确地叙述了出来。瑜伽教练的“记忆自我”的反思过程可以看作用语言对自己的经验自我的行为进行的虚拟模拟。自省程序的自省过程可以与此类比，类似于程序 ϕ_e（x）内嵌了虚拟机（记忆自我），然后把自己的源代码放在虚拟机上模拟自己运行了 ψ（x）步，最后把整个的模拟结果输出了出来。

不仅如此，程序不但能够进行“自省”，而且能够通过“自指”进行“自我升级”。抽象地看，一个智能体无非是一段程序。因此，不妨设计某种“元程序”负责搜索整个“程序空间”、自动寻找“聪明”的程序，然后通过经验学习寻找更“聪明”的程序。

三、哥德尔机：自我升级智能体的一种实现

自我升级的哥德尔机可以看作“自省”程序的超级加强版。数学家曼宁（Manin）也有过类似的超越“自省”程序的想法，设想房间的桌子上有一张房间的布置地图，地图精确地描述了房间里的陈设，包括桌子上的地图自身，然后设想地图上房间物品的摆设可以脱离地图而发生位置变化，然后房间里现实物品的摆放位置也可以根据地图的变化而变化。曼宁认为，大脑的功能与此类似，大脑内嵌了一张描述自身的大脑

① Daniel Kahneman, *Thinking, Fast and Slow*, New York: Farrar, Straus and Giroux, 2011, pp. 377-385.

地图，这张地图具有建模功能，可以虚拟不同于当前大脑状态的可能状态，然后还具有控制功能，它能控制整个大脑根据虚拟状态的变化而发生现实的变化。而且，这张大脑地图还是粗粒化的。① 曼宁有这个构想，但并没有给出具体的实现方案。哥德尔机恰恰可以看作这么一张大脑地图。它的可行性由克林尼不动点定理保证。虽然能够自我修改源代码的哥德尔机可以通过克林尼不动点定理构造出来，但相比于自测量长度程序、自复制程序和自省程序，这里的 h 函数要复杂很多，因为它不仅要“自省”，更重要的是，它还要通过更深层次的“自省”实现“自我升级”，下面介绍它的详细构造。

哥德尔机由两个并行的部分构成：通用求解器可以是任何一个处理具体问题的程序，比如深度神经网络 CNN/LSTM，为了更具有通用性，它也可以是一个通用强化学习算法，比如直接采用 AIXI 的某个可计算的变种，强化学习算法负责与环境交互，能对环境采取动作，并能感知外界环境的反馈，通过不断地与环境交互来获取更大的期望累积效用。通用搜索器内嵌了一个形式系统，形式系统完整编码了哥德尔机的硬件、效用函数、不确定性计算工具的全部信息以及部分环境信息。

硬件公理负责描述机器元件具体的运作方式，比如，如果它的硬件是最简单的图灵机模型，那么，硬件公理就需要描述图灵机纸带的内容、读写头的位置、当前的状态以及状态的可能的转移规则等。效用公理负责描述不同状态的可能回报、机器在输入输出和运行过程中的计算成本以及这些不同的回报和成本之间的整合方式。环境公理负责描述观察到的环境信息以及可能的环境变化，如果通用求解器加载的是 AIXI 的某个变种的话，它就需要描述 AIXI 的环境空间，即所有半可计算的半测度。不确定性公理负责描述算术、概率论、统计学以及逻辑中的符号操作规则等。

通过内嵌的形式系统，通用搜索器就可以将机器工作的所有状态当作数学定理来讨论。通用搜索器可以通过一个定理证明器搜索数学命题的证明，基于初始给定的效用函数，如果它搜索到某个“策略在未来的时间里比当前策略能带来更大的期望累积效用”成立，那么它就改掉之前的策略，改用这个新的策略与环境进行交互。这就实现了哥德尔机的自我升级（见图 1）。

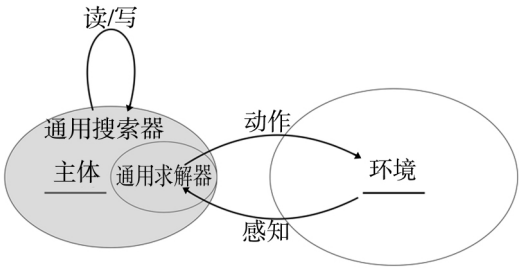

图 1 哥德尔机

① Dmitrii Manin and Yuri I. Manin, “Cognitive Networks: Brains, Internet, and Civilizations,” in B. Sriraman, ed., *Humanizing Mathematics and Its Philosophy*, Switzerland: Birkhäuser Basel, 2017, pp. 85-96.

以上关于通用求解器和通用搜索器的描述就是我们借助克林尼不动点定理构造自我升级的哥德尔机所需要的函数 h。虽然上面内容是用自然语言描述的，但只要设计得足够巧妙，通用求解器和通用搜索器的构造确实是能行的，所以根据丘奇—图灵论题，它可以看作一个递归函数 h，然后我们就可以借助克林尼不动点定理，证明存在某个程序 e，它可以对自身（包括通用求解器和通用搜索器）的源代码进行彻底的修改——只要它内嵌的形式系统的定理证明器能证明“修改后的策略在未来的时间里比当前策略能带来更大的期望累积效用”。这在一定程度上保证了对源代码的修改是相对可靠的。这样通用求解器和通用搜索器就可以比较合理地自我学习升级。

但根据克林尼不动点定理，只是存在一个能提升效用的程序 e，那么，能否让机器反复改进、持续自我升级？这就需要带参数的克林尼不动点定理：对于任意给定的递归函数 h（x，y），存在递归函数 e（y），使得对任意 y 都有 $\phi_{e(y)}=\phi_{f(e(y),y)}$。将时间、部分环境等计算资源的限制信息编码到参数 y 里面，借助带参数的克林尼不动点定理，存在一种系统的自我升级方式 e，可以持续不断地产生一系列程序 e（y），每一个都是对前一个补充了额外信息的改进。通过这样的方式，哥德尔机就能持续地搜索可提升自己效用的策略，不断用更好的策略改写自己的源代码，从而完成持续“反思升级”的学习过程。

这里的升级既有采用克林尼不动点定理的一次升级，也有采用带参数的克林尼不动点定理的持续升级。一次升级与持续升级孰优孰劣呢？是否持续升级必然强于一次升级呢？在同一个全局最优的意义上，其实二者没有区别。因为根据通用搜索器的设计，只有当“改进”后的状态比“不改进”的状态严格地好的时候才会触发改进，而“不改进”的状态其实隐含着，虽然当前不改进但会继续搜索并评估以后其他替代改进状态的可能，这意味着，相对于初始给定的效用函数和环境信息，“改进”不会落入局部最优而是稳妥地迈到全局最优。所以一次改进足矣，一次改进包含了以后可能的二代、三代……直到最优的所有可能的改进。

但是，持续升级仍然有它的优势，因为参数 y 里可以编码更多的新探测到的环境信息和计算资源等信息，所以持续改进的哥德尔机类似于一个实时算法，它可以不断拓展对环境的知识，搜索相对于当下计算资源限制下的最优改进，然后随着新环境信息的录入和计算资源的增多，不断地调整对“全局最优”的理解，从而收敛到真正的全局最优。升级后的哥德尔机也不过是一段程序，所以它还可以调用冯·诺意曼的自复制程序，不断将自己复制下去，如果允许变异的话，变异后的个体也可以繁殖下去。

不难看出，通用搜索器有点像前面作类比的虚拟机。所谓借助形式系统的搜索类似一套模拟过程。而哥德尔机允许对自身状态进行编码相当于机器可以模拟自身的运作，而用定理证明器搜索数学命题探测更好的策略的过程相当于机器进行自我反思、主动规划探索的过程。如果能将虚拟机里搜索到的策略装载到实际的执行系

统上，哥德尔机就可以不断地修改自己的代码，看上去像“揪着自己的头发把自己拎起来”。有人担心，这种装置一旦实际制造出来，就有可能引发智能爆炸，从而导致技术奇点的来临。但更需注意的是，即使有实现超级智能的可能，也未必是往好的方向。虽然哥德尔机相对于其初始给定的效用函数是全局最优的，如果初始给定的效用函数有问题的话，比如与既定的目标有偏差、没有真正反映人的真实意图、甚至是有敌意或邪恶的，那么，相对于这种效用的全局最优只会更可怕，极有可能收敛到最坏的情形。这就是哲学家博斯特罗姆（Bostrom）所说的“目标正交性”（goal orthogonality）论题。①

施米德胡贝尔建立的哥德尔机是第一个具有自指能力的元学习机，② 以下通过讨论作为自我升级智能体的哥德尔机，探讨机器能不能具备自我意识的话题。

通过劳威尔的定理，可以看出，“自指”与“对角线”是一枚硬币的两面，通过不动点可以实现跨越层次的间接自指，可以表达“我有什么性质”或“我要进行什么操作”，而通过“对角线”可以构造出“超越”预定列表的新对象，无论是“自指”还是“超越”都是令人振奋的现象，与人的意识活动密切相关，而哥德尔机却通过内嵌形式系统的方式自指，这种“自指”实现了另一种意义上的“超越”，不同于借助“对角线”构造出不能以既有方式表示的“不合法”的新对象，哥德尔机却是通过“自指”实现“超越”，相当于对自己说——“我要变成能获取更大效用的自己”——然后就魔术般真的变强了（实现全局最优）。

上面例子中瑜伽教练的“记忆自我”的反思过程是对“经验自我”亦步亦趋的虚拟模拟而没有实质性的指导，而哥德尔机的“记忆自我”却试图通过定理证明器的自我反思升级“经验自我”（见图2）。

哥德尔机与自我意识的类比

自模拟计算机	哥德尔机	自我意识
宿主机	通用求解器	经验自我
虚拟机	通用搜索器	记忆自我
硬件	硬件	身体

图2 人与哥德尔机

① Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies*, Oxford: Oxford University Press, 2014, pp. 105-114.

② Bas Steunebrink and Jürgen Schmidhuber, “Towards an Actual Gödel Machine Implementation,” in P. Wang and B. Goertzel, eds., *Theoretical Foundations of Artificial General Intelligence*, New York: Springer, 2012, pp. 173-195.

如果我们把“自由意志”定义为记忆自我对经验自我的反思甚至指导，那么在这种意义上，只要搜索器比求解器跑得快，哥德尔机或许可以具有自由意志。对于人来说，记忆自我可以通过虚构故事来“欺骗”自己。人可能没有自由意志，因为有实验显示，记忆自我并不参与决策过程，真正的决策早在我们意识到之前几秒的时间就已经被经验自我做出了，自由意志很可能就是这个记忆自我编造的一个故事。我们充其量仅具有自我意识，我们可以觉知到自己所做的一切，但不能反过来作用到我们的行动上。关于自由意志，法兰克福（Frankfurt）认为，指向事物或事态的欲望叫一阶欲望，指向一阶欲望的欲望叫二阶欲望。有自我意识才能形成二阶欲望。一个人可以有二阶欲望但不一定想依照它行动。想依照它行动的二阶欲望叫二阶意欲。一个人的行动是自由的当且仅当这个行动是由他认同的自己的二阶意欲所致使的。这要求二阶欲望是自主选择的而不是外部设定的。① 根据这种“自由意志”的定义，叔本华所说的“人能为所欲为，但不能御其所欲”相当于否定了人有自由意志。哥德尔机的通用搜索器只根据初始的效用函数搜索升级策略，并没有真正升级效用函数，所以不具有法兰克福意义上的自由意志。埃弗里特（Everitt）等人刻画过可以自我修改效用函数的智能体，不过，根据他们刻画的自我修改效用函数的方式，看上去“高阶欲望”像是自我决定的，但其实这并不是真正的可以作用在“低阶欲望”上的“高阶欲望”，而只是用当下的策略可以修改的“未来欲望”，所以这种智能体也不具有法兰克福意义上的自由意志。②

四、自我升级智能体理论的贡献和局限

自我升级智能体的研究切实推进了人工智能和哲学的研究，但这种推进是有限度的。以下从逻辑、认知和认识论方面探讨自我升级智能体的贡献和局限。

（一）逻辑和哲学上的贡献和局限

自我升级智能体的构建向真正具有自我意识的智能体前进了一步。它对涉及“自指”的罗素悖论、哥德尔不完全性定理、图灵停机定理等否定性结果背后的机制做了一次正面的应用。它使得我们对自我意识的研究有了一个程序化的标准，借助这种形式化的方法可以为自我意识的机器实现提供新的思路，进而弥合学界关于机

① Harry G. Frankfurt, “Freedom of the Will and the Concept of a Person,” *The Journal of Philosophy*, vol. 68, no. 1, 1971, pp. 5-20.

② Tom Everitt et al., “Self-Modification of Policy and Utility Function in Rational Agents,” in Bas Steunebrink, Pei Wang and Ben Goertzel, eds., *Artificial General Intelligence*, Proceedings of the 9th International Conference, AGI 2016, New York, USA: Springer, 2016, pp. 1-11.

器意识的分歧，破解机器意识研究面临的困局。

自我升级智能体的局限体现在逻辑、认知和认识论方面。主要有：

1. 哥德尔第一不完全性定理的局限

哥德尔机通过内嵌的形式系统（数学）对自己和环境进行建模、做逻辑推理，从而规划与环境的交互策略，这个过程非常类似于人在世界之中的生存过程。人类在漫长的文明中，发展出了先进的数学科学，用科学手段对自己和环境建模，制造和使用工具，影响改造环境。除了人很难自我升级外，人通过发展“理性”手段自我认识并认识和改造环境的过程与机器很相似。泰格马克（Tegmark）在《生命3.0》中认为，生命的发展有三个阶段，通用人工智能是生命形式的第三个阶段。在第一个阶段，生命的硬件和软件都只能依赖于进化的力量，如细菌。在第二个阶段，硬件依靠进化，但软件可以自行设计、升级改进，如人类。人类可以创造、发展、传承知识，虽然躯体只能维修保养不能设计改进，但知识可以不断地升级。在第三个阶段，硬件和软件都可以设计升级。① 哥德尔机应该可以看作泰格马克所说的第三个阶段的典型代表，它可以对自身（包括通用求解器和通用搜索器）进行彻底的升级改造。最开始的通用求解器可以选择AIXI的某种可计算的逼近或变种，如强化学习模型AIXI^{tl}，而通用搜索器可以选择莱文通用搜索的某个变种，如胡特尔搜索。但是，既然通用搜索器内嵌了形式系统，那么它就面临哥德尔不完全性定理的障碍，有一些重要且必要的“变身”可能无法被形式系统找到。

对于任何可计算的环境，算法概率都可以很好地逼近它，为了保证可以逼近任何可计算环境这种“通用性”，算法概率本身不是可计算的。② 强大的智能体必然复杂，虽然复杂且强大的智能体是存在的，但只要它足够复杂，那么形式系统将无法帮助我们找到它。这个限制可以称为哥德尔第一不完全性定理和莱格不完全性定理限制。

2. 哥德尔第二不完全性定理的局限

生物的进化有可能是基因无目的地随机变异，自然环境做选择、适者生存的结果。因为是随机变异，所以进化速度慢。对于自我升级的智能体来说，智能体根据环境自我选择变异。哥德尔机就是带有方向性地自主选择变异。在当前的主体技穷之时，人们总寄希望于进化的力量，希望演化后的主体能更强大。对于哥德尔机来说，如果它想实现迭代进化的话，一个自然的办法就是制造后代，只要每一代给下一代装配更强的形式系统，那么哥德尔第一不完全性定理的障碍就可以在一定程度上突破，但问题是——哥德尔第二不完全性定理，主体1在构造主体2时如果不能

① Max Tegmark, *Life 3.0: Being Human in the Age of Artificial Intelligence*, New York: Penguin Random House LLC, 2017, pp. 26-31.

② Ray Solomonoff, “Complexity-Based Induction Systems: Comparisons and Convergence Theorems,” *IEEE Transactions on Information Theory*, vol. 24, no. 4, 1978, pp. 422-432.

在自己的形式系统内证明主体 2 的形式系统的一致性，那么它根本无法保证主体 2 的可靠性。无法保障可靠性，就无法回避完全坍塌的风险。一致性得不到保障意味着机器智能的伦理安全问题得不到保障。但是，如果要求每一代主体必须严格证明下一代主体的形式系统的一致性的话，那么，这种进化在某种意义上是一种退化，而且是一种极为快速的退化。而生物的进化则不需要一致性的保证，好的变异、不好的变异都可能产生，自然选择的结果虽然常常是、但不必然是优胜劣汰。变异和自然选择不能保证可靠性，哥德尔机面临的也是同样的问题。这可以称之为可靠性限制。

3. 复杂性限制与高层抽象的反思

如果不考虑哥德尔不完全性定理的限制，哥德尔机在理论上是全局最优的，但是否可以马上进行工程制造了呢？其实很难，难点在于复杂性。由于一个自我模拟系统需要两层设计，随着求解器复杂性的提高，需要更复杂的搜索器去模拟它。一个可能的方案是，先对求解器进行高层抽象，比如，采用类似特征强化学习的办法，将求解器的环境抽象为简单的马尔科夫决策过程，搜索器只“反思”抽象后的求解器，而不是亦步亦趋地模拟所有细节。人的意识也是类似，对于一段经历，记忆自我会做加工裁剪，忽视中间波澜不惊的过程，而格外重视尖峰体验和最终时刻的体验。这就是心理学家所谓的“过程忽视”和“峰终定律”。① 所以，搜索器必须以“粗粒化”的方式反思求解器，学习需要遗忘，高层抽象可以看作对复杂性限制的一个解决路径。这也可以看作一种实用复杂性限制。

4. 独立于还是内嵌于环境的问题

克林尼不动点定理保证了哥德尔机的全局最优策略，哥德尔机所谓对自身策略、效用函数、搜索器的自我修改其实也不是真的自我修改，而是通过自指方法构造的不动点。这样构造的哥德尔机是外在于环境的。这种构想类似于哲学上笛卡尔主张的“心物二元论”思想。哥德尔机与环境除了有限的输入输出的交互外，二者完全独立，环境无法影响哥德尔机的运行过程。这是一种高度的理想化，与现实情境相去甚远。现实中的机器可能随时受环境的影响，机器不是超越于环境的心灵，机器与环境由同样的物质材料构成，机器是环境的一部分。所以，如果抛弃二元论的设定，考虑更现实的一元论设定，不做过度简化的更理想的哥德尔机应该是嵌入式的。如果把哥德尔机看作环境的一部分会怎样？对于内嵌于环境的主体来说，它只是环境的一个子部分，环境可以修改主体的源代码和内存，并执行主体的代码。受资源限制，第一步主体可以从长度不超过某个固定长度的程序集合中搜索最优策略，后面主体怎么改变就完全由环境控制了。

主体试图通过自己的行动影响环境，环境直接修改生成下一代主体的策略和下

① Daniel Kahneman, *Thinking, Fast and Slow*, pp. 377-385.

一代主体的感知，所以主体的动作、下一代主体以及下一代主体的感知都可以整合在一起，相当于所有的信息都被整合到了主体里，所谓的交互历史就是主体的更迭。① 如果主体的源代码和内存完全由环境修改，环境执行主体的代码，那么，因为所有的信息都被整合到了主体里，主体只能在最开始没有任何经验知识的前提下进行决策，此时主体缺乏对环境的经验估计，唯一可以依赖的只有对环境的“先验”信念。而这种“先验”信念如何去把握？这是内嵌于环境的哥德尔机遇到的难题。

5.“自指”的物理限制

前面讨论的哥德尔机是基于经典的图灵机，并没有考虑量子效应。彭罗斯曾借助不完全性定理论证心灵胜过机器，认为人的意识可以把握机器证明不了的真理。他猜测人脑的意识行为源于大脑神经元微管的量子效应。如果彭罗斯的这个猜测成立，那么是否量子哥德尔机才能刻画人的意识？其实，只要借助“自指”来刻画意识，意识都无法超越对角线论证的局限。② 假设M是可能的量子测量的集合，O是可能的量子测量结果的集合，劳威尔定理告诉我们，如果α没有不动点，那么通过对角线方法构造的g就不能被f表示。测量一个属性并完全不改变它是不可能的，换句话说，因为有量子观察效应，所以α没有不动点，这意味着，观察者的自我测量或者说自我反思行为必然存在局限，必然存在不能进行自我观测的死角。

通过劳威尔不动点定理，我们看到了“自指”与“超越”的广泛应用，看到了很多悖论、定理、算子之间深刻的相似性，也看到了“自指”与“超越”之间深刻的“平衡”。“自指”与“超越”恰如一枚硬币的正反两面，正面的应用甚至在构建自我升级的通用人工智能中也起着至关重要的作用，而反面的应用则在不断地挑战着机器甚至人脑自身的认知局限。

（二）认知意义和应用风险

从认知的角度看，自我升级智能体的提出具有重要的认知意义。我们知道，生成主义强调具身主体以由其生理机能决定的精确方式与环境发生交互作用，“自我”作为这种交互过程的一部分而产生，“自我”并不表征，但他通过与环境交互作用的方式产生世界，生成世界。因此，生成主义纲领的一个基本特征是自指性。比起第一代认知主义纲领，生成主义纲领的高明之处就是它直面自指性，即涉及自身的问题。罗素悖论、说谎者悖论的出现让人们时刻警惕自指性带来的风险。实践证明，作为第三人称的外部观察者所感知的东西与作为第一人称的当事人所感知的东西是

① Laurent Orseau and Mark Ring, “Space-Time Embedded Intelligence,” in Joscha Bach, Ben Goertzel and Matthew Iklé, eds., *Artificial General Intelligence*, Proceedings of the 5th International Conference, AGI 2012, Oxford, UK: Springer, 2012, pp. 209-218.

② Karl Svozil, *Physical (A) Causality: Determinism, Randomness and Uncaused Events*, Cham: Springer, 2018, pp. 17-19.

不同的，生成主义采用的策略是，在二者之间搭建一个解释学循环之桥。这就像我们从生命的基本形式出发，经历了生命发展的从简单到复杂的进化，最后以生物学家研究生命的基本形式结束。这似乎是一条循环往复的麦比乌斯环，一个怪圈。从起点出发，转了一圈，又回到原点。考察的对象变成了正在探寻的主体。与此类似，生命作为自我升级智能体考察的对象，通过“自指”和“超越”变成了正在探寻的主体。生成主义为自我升级智能体的提出奠定了认知基础。自我升级智能体理论的成功为生成主义提供了一个有利的例证。

更有认知意义的是，像哥德尔机那样的自我学习升级方式也可以涌现出某种创造性，按照施米德胡贝尔、奥索（Orseau①）等人的观点，可以加载类似“人工好奇心”（artificial curiosity）的功能，这种聪明的智能体可以像一个好奇的儿童面对纷繁复杂的全新世界不断去探索、理解、发现和解决全新的问题，尽量增进对现实环境的理解，降低对可能环境的不确定性，最终成为能解答通用问题的通用智能体。在这个意义上，结合合适的“目的因”（效用函数），自我升级智能体的诞生具有重要的认知意义。

五、机器意识研究的困局：以哥德尔机为例

以下我们以哥德尔机为例，探讨机器意识哲学研究面临的困局。主要讨论三个问题：哥德尔机器真正具有自我意识吗？哥德尔机器会导致智能爆炸吗？从哲学上看，哥德尔机在理论上的困局是什么？

哥德尔机是否可以具有自我意识？如果能，那它是一种功能意识，还是一种现象意识？回答这个问题之前，首先要区分两种情况：其一是机器看起来像是有自我意识的样子，也就是说，机器具有了“功能意识”；其二是机器真的具有了自我意识，而且它还知道自己具有了自我意识。这大致相当于现象意识。

如前文所述，既然哥德尔机可以“自我反思”，可以通过“自指”进行自我升级，如果人类的意识仅仅是通过自我模拟进行自我反思的话，那么机器也完全有可能模拟看起来具有意识的人类。人们只需要制造出一个同样具有两种自我的机器：一种虚拟的机器（记忆自我）和一种纯算法的执行机构（经验自我），那么这种机器就可以表现得像是有自我意识的装置了。人们甚至可以用这样的系统定义自我意识。然而，这就说明机器“真的”具有了自我意识了吗？答案需要等待对“意识”的科学解释。因为到目前为止，科学界尚没有一个公认的关于人的“意识”的科学理论，

① Laurent Orseau, Tor Lattimore and Marcus Hutter, “Universal Knowledge-Seeking Agents for Stochastic Environments,” in Sanjay Jain et al., eds., *Algorithmic Learning Theory*, Proceedings of the 24th International Conference, ALT 2013, Singapore: Springer, 2013, pp. 158-172.

所以这里所谓的机器具有自我意识，只是根据我们的定义机器“看起来”具有了自我意识，还不能说机器真正具有了自我意识，更遑论机器知道自己具有了自我意识，解决了知道自己知道、知道自己不知道之类的认识论难题。实际上，我们说哥德尔机具有“自我意识”只是一种“功能意识”。目前哥德尔（实体）机并没有创制出来，即便是造出来了，可能离真正具有意识还有很长的路要走。

自我升级智能体的建构有何风险？是否会导致智能爆炸？如前所述，哥德尔机可以通过在模拟器上搜索而寻求让自己优化的方案。我们只需要将搜索到的虚拟代码装载到实际的执行系统上，那么哥德尔机就可以不断地修改自己的代码而升级下去。这样的机器做出来之后会有什么后果呢？由于哥德尔机具备不断自我改进自己代码的能力，亦即，这个智能系统可以通过不断地提升自己的能力而优化，这个过程会越来越快地持续下去，从而有希望很快超过人类的智能。一旦达到了这样的智能，哥德尔机是否自己就会设计出更强的哥德尔机，从而让整个智能过程加速，从而导致智能爆炸？

对此我们不能过于乐观，认为智能爆炸会开启智能新时代。因为如前所述，哥德尔机也具有博斯特罗姆所说的“目标正交性”问题，即便有引发智能爆炸甚至达到技术奇点（Technology Singularity）的可能，也未必是往好的方向发展。虽然哥德尔机相对于其初始给定的效用函数是全局最优的，如果初始给定的效用函数有问题的话，比如，与既定的目标有偏差、没有真正反映人的真实意图，甚至是有敌意的或邪恶的，那么，相对于这种效用的全局最优只会更可怕，极有可能收敛到最坏的情形，甚至危及人类安全。另一方面，根据带参数的克林尼定理，持续升级的哥德尔机并不是完全彻底的自我迭代升级，而是需要将新获取的有关环境的信息和计算资源的信息编码到参数里，然后再借助相同的“自指”过程升级。这种升级过程并不是一种指数迭代，能不能收敛到技术奇点都存疑。退一步说，就目前的技术水平，我们离智能爆炸的实现还有很长的距离，机器不至于成为人类的“终结者”，或许有办法阻止其终结人类，比如，可以令其通过合作逆强化学习的办法学习加载人类的价值观。所以，不管是哪一方面，目前远不至于对人类构成真正的威胁。

哥德尔机的理论困局是能否实现真正的机器意识，破解意识之谜。这显然不是一个科学问题，而是一个哲学问题。通过考察理性主义与经验主义、分析哲学与现象学在这个问题上的分歧和争论，可以为我们提供化解困局的启迪。深受人工智能思想影响的哲学家麦克德莫特（McDermott）在他的《纯粹理性批判》论文中深刻指出，历史上许多没有出路的人工智能研究“只是因为对哲学家们昔日的失败一无所知，才得以维持”。① 从细胞自动机到哥德尔机的发展历程，不难看出其中始终贯

① 玛格丽特·博登：《人工智能哲学》，刘西瑞、王汉琦译，上海：上海译文出版社，2001年，第307—308页。

穿着理性主义和经验主义的较量和争锋。冯·诺意曼在自动机研究中尽管对概率和统计的因素有所考虑，但他主要偏重演绎逻辑和计算等理性因素而忽视统计数据等经验因素。这就使他即便走到了自我升级智能体的大门口却止步不前。哥德尔机一方面通过通用求解器采用统计学习的模型和工具与环境进行交互，这是经验主义的方案，另一方面通过通用搜索器内嵌的形式系统不断搜索，把搜索到的最优策略装载到实际的执行系统中以实现系统优化升级，这是理性主义的方案。另外，考虑借助带参数的克林尼定理构造的哥德尔机的话，还可以实时地把新的经验数据和计算资源通过参数的方式编码进去，然后借助“自指”的方式不断改进、持续升级，从而沟通了经验和理性两个方面，在看似对立的两极之间保持平衡。这是自我升级智能体研究给我们的哲学启示。

结语：自我升级智能体发展前瞻

自我升级智能体研究正处在发展的十字路口，需要在道路和方向上做出关键抉择。自我升级智能体的研究始终纠结于分析哲学传统与现象学传统的分歧，和人工智能中理性主义与经验主义此消彼长的争锋交织。它们都各有偏颇，在理性与经验孰重孰轻的问题上表现出某种片面的深刻性。自我升级智能体摆脱困境取得发展的可能出路是：在分析传统和现象学传统之间寻求一种动态平衡。这将有助于化解自我升级智能体理论的困局。

我们对上述问题的回答是：第一，自我升级智能体的建立使我们对自我意识的研究有了一个程序化的标准，借助这种形式化的方法有可能弥合学界关于机器意识的分歧，破解机器意识研究面临的困局。第二，生成主义为自我升级智能体的提出奠定了认知基础。自我升级智能体的成功为生成主义提供了一个强有力的例证。第三，尽管自我升级智能体向机器真正具有自我意识前进了一大步，但是我们不能说机器真正具有了自我意识。建构真正具有自我意识的系统，还有很长的路要走。就目前的技术水平看，我们离智能爆炸的实现还有很长的距离，机器还不至于成为人类的“终结者”。造成机器意识困局的症结在于分析哲学传统与现象学传统的分歧和偏颇。解决的出路在于：从对立到相容，从互斥到互补，进而达到融通的新境界。

〔责任编辑：莫　斌〕

enlightenment modernity allows us to fundamentally realize a revolutionary transformation, i. e., the logical establishment of practical "public rationality" based on "the full and free development of every individual" in reality. It promises human civilization an underlying labor practice based on freedom, consciousness and autonomy and relies on a "community of free individuals" to continuously fulfill the ideal vision of a better life, i. e., the realization of public values. What this effort highlights is the way Marx's philosophy has a distinctive theoretical and practical character, an assumption of history and a lofty spiritual realm, all of which are fundamentally different from all the old philosophies.

(3) Logical and Cognitive Issues of Self-Improvement Agents

Ren Xiaoming and Li Xi · 46 ·

The establishment of self-improvement agents has offered a programmatic standard to the research of people's self-awareness. With this formal method, it is possible to bridge the differences in machine consciousness in the academic world and break the predicament facing machine consciousness research. But it also has logical limitations. Enactivism lays the cognitive foundation for the proposal of self-improvement agents. The success of self-improvement agents provides a strong example of enactivism. Although the self-improvement agent has taken a big step towards the machine's real self-awareness, one can only say that it has "functional consciousness." The crux of the predicament of machine consciousness stems from the differences and biases between the tradition of analytic philosophy and the phenomenological tradition. The solution lies in: from opposition to compatibility, from mutual exclusion to complementarity, and then to a new realm of integration.

(4) A Political Economy Analysis of the Globalization of Platform Economics

Xie Fusheng, Wu Yue and Wang Shengsheng · 62 ·

As a new organizational form suited to capital accumulation and social production and reproduction under digital technology, platform economics relies on digital platforms supported by efficient data collection and transmission, advanced computing power, and powerful data-processing algorithms. They integrate social production, distribution, exchange and consumption across temporal, spatial, national and sectoral boundaries in a way that gives a vigorous boost to the development of society's productive forces. In the platform economy, the technical characteristics of digital platforms and capital's monopoly of those platforms have shaped a structure of imperfect dynamic competition. The new form of labor organization on the basis of digital platforms leads to unstable employment and wages, allowing the logic of capital accumulation to seep into the process of labor reproduction. Under capitalist conditions, platform economies will remain unable to overcome the inherent contradictions seen in the law of capital accumulation.