# “机器他者”与“符号自我”：论智媒的交互主体性*

胡易容

**摘　要：**随着以 ChatGPT－4、DeepSeek 为代表的自然语言大模型快速迭代，智能人机交互中包含的主体性问题成为值得讨论的重要话题。本文以 60 年前麦克卢汉的论断“有意识的计算机仍然只是人的意识的延伸”为起点，结合人工智能史上符号主义、联结主义、行为主义等重要流派，探讨人工智能在人机交互方面的演化发展，从符号交互主体建构的角度揭示了人工智能发展背后的理论依据与哲学基础，进而提出人工智能基于语言符号交互表现出的“拟主体性”之“人类意识镜像”本质。

**关键词：**智能媒体，交互主体性，AI 符号学，符号主义，拟主体性

## On the Intersubjectivity of Intelligent Media: A Semiotic Analysis of AI

Hu Yirong

**Abstract:** With the rapid advancement of conversational intelligence, exemplified by systems like ChatGPT－4, the issue of subjectivity in human-machine interaction has emerged as a critical topic. Drawing on McLuhan’s assertion sixty years ago—“Conscious computers are merely extensions of human consciousness”—this paper traces the evolution of AI through key theoretical frameworks such as Symbolism, Connectionism, and Behaviorism. By examining these

* 本文为国家社科基金重大项目“当代艺术的重要美学问题研究”（20&ZD094）阶段性成果。

schools of thought, this paper uncovers the philosophical and theoretical foundations of AI development from the perspective of semiotic interaction. It argues that AI demonstrates "pseudo-subjectivity", functioning as a "mirror of human consciousness" that reflects, rather than truly embodies, subjectivity.

**Keywords:** intelligent media, intersubjectivity, AI semiotics, symbolism, pseudo-subjectivity

**DOI:** 10. 13760/ b. cnki. sam. 202501003

一个有意识的计算机仍然只是我们意识的延伸，就像望远镜是我们眼睛的延伸，或是腹语术师的假人是腹语术师的延伸。

——马歇尔·麦克卢汉

## 引言：《她》（*Her*）已来？人工智能自然语言交互能力跃升引发的思考

2024 年 7 月 31 日，OpenAI 推出 ChatGPT－4 的高级语音功能（Advanced Voice Mode），提供更自然的实时对话。试用的网友发现，在读美国诗人狄金森（Emiliy Dickinson）的作品时，它（她）竟然带着哭腔（金磊，2024－8－1）！OpenAI 公司说明，ChatGPT－4 能够识别并响应用户语音中的情感变化，例如悲伤、兴奋等情绪。

网友们惊呼，科幻电影中的《她》已来！

人工智能对当今社会的影响方方面面，但人工智能是否具有或将会具有"意识"，始终是其中最热闹的话题之一。这个话题集中在人工智能作为媒介的交互能力方面。同样具有智能的工业机器人可能并不会直观地造成这样的困扰，而人工智能聊天程序中一个看起来不经意的笑话，却能让人浮想联翩。这表明，人们对于"意识"或"主体性"的评价，集中在其交互能力上。这意味着，当人们认为人工智能理解了诗的内涵及情绪，以及当人们跟人工智能对话的时候，人感受到了被理解，于是一个基于符号感知的"主体"似乎就悄然生成了。更重要的是，人工智能一旦作为独立一方参与人类社会生活，在人与智能媒体之间的"交互主体性"就发生了，它意味着在符号传播意义下的人工智能"意识觉醒"。

半个多世纪前，媒介理论家麦克卢汉（Marshall McLuhan）在他的广义延伸论中提及了计算机这样看上去会思考的智能媒介。对这种看上去会思考的机器，他有两个方面的论断：一方面是认为，“事实上，计算机高度专门化，完全缺乏形成意识所需的全面关联过程”；另一方面，他认为，即便计算机有某些意识，也仍然只是人类意识的延伸，就像望远镜是我们眼睛的延伸，或是腹语术表演师的假人是表演者的延伸。换言之，他认为计算机没有自己独立的“心智”。（麦克卢汉，2000，p. 431）

麦克卢汉在 60 年前写下这样的内容时，人工智能还只处于最初的发展阶段。尤其是，受限于算力和数据库，当时最好的智能计算也仅仅表现为“专家系统”。但迄今为止，人工智能发展已经历了三波高潮，并发展出符号主义、联结主义、行为主义等流派。当前基于算力与大数据驱动的深度学习及其广泛的场景化应用是第三波人工智能高潮的新动向。随着高速迭代的深度学习进化，尤其是通用大模型对自然语言处理能力的大幅提升，人工智能不仅成为全球科技竞争的焦点领域，也成为人文学科重新反思的重点领域。凯文·凯利（Kevin Kelly）等学者基于深度学习模型的复杂性和不透明性导致的“黑箱”问题，提出警惕人工智能等技术的失控风险。更有不少学者认为人工智能超越人类已经指日可待且不可逆转，如詹姆斯·巴拉特（James Barrat）认为人工智将在未来超越人类智能。在巴拉特看来，“人类面对的是一支 ASI 集体，集体里的每个成员都比最聪明的人类聪明 1000 倍，它们打败人类捍卫者会不费吹灰之力。那就等于是智力的汪洋大海对抗一滴水”（2016，p. 8）。

若有幸见到人工智能的最新进展，麦克卢汉是否仍会下同样的判断，我们不得而知。但从“交互主体性”的角度，则能形成一个对人工智能是否已经或将要产生交互主体的判断。交互主体性最早由现象学家埃德蒙·胡塞尔（Edmund Husserl）提出，并在后来被哲学家如马丁·布伯（Martin Buber）、莫里斯·梅洛－庞蒂（Maurice Merleau-Ponty）和尤尔根·哈贝马斯（Jürgen Habermas）等进一步发展，指的是多个人（主体）之间通过语言、行为或其他符号系统达成相互理解的过程。它强调的是不同主体之间通过沟通和互动达到共同理解的能力（Habermas，1984）。交互主体性还包括不同个体之间共享的认知、情感和社会体验，即主体之间能够相互理解和共感，从而建立起共同的意义世界（哈马贝斯，1989，p. 3）。

由此，交互主体性有两个关键要点：一是其实现交互的方式是“符号系统”（system of signs），二是其最终目标指向建立“共享的意义世界”。由此，

该话题实际上就成了一个传播符号学最适合处理的论域。符号学作为“意义之学”，指向的即是“意义世界”的建立机制。恩斯特·卡西尔（Ernst Cassirer）将人界定为“创造和使用符号的动物”，并以此区分“人”与其他物种或机器的差别（卡西尔，1985，pp. 40－41）。由此，基于传播符号学的角度探讨人工智能是否具有或可能具有某种“拟主体性”，再切实不过。这里的“拟”主要是避免进入本体论的立场之争。具体来说，沿着卡西尔在《人论》中将人定义为“符号的动物”这一线索，从主体的交互涉及的“符号观念”与具体的“语言能力”两个维度切入人工智能的三波发展及其主要理论依据，分析人工智能的符号形式及语言表现以及人工智能的“拟主体性”。

## 一、“符号主义”的理论逻辑与“语言自我”假设

当麦克卢汉写下关于“计算机”作为人的意识的延伸时，正值人工智能第一波高潮兴起。1956 年，达特茅斯会议首先提出了人工智能的定义，即“使一部机器的反应方式就像是一个人在行动时所依据的智能”（McCarthy et al.，1955）。与会者中的赫伯特·西蒙（Herbert Simon）与艾伦·纽厄尔（Allen Newell）逐渐发展的人工智能符号主义成为该流派的核心理论。这一派的先导理论是阿兰·图灵（Alan Turing）在 1950 年提出的“图灵测试”（The Turing Test），及其所依托的行为主义语言观。

### （一）图灵测试的智能判定与艾耶尔的行为主义语言观

图灵测试只要求符号的结果，即当被测试者表现得像“人”，就可被认定为具有了“人”的智能。基于此，纽厄尔和西蒙提出，“物理符号系统有必要也有足够条件来实现一般智能行为”（Newell & Herbert，1976，pp. 113－126），即如果能够构建足够复杂的符号系统，就能够模拟人类智能。

由于智能本身难以定义，图灵 1950 年发表的论文《计算机器与智能》没有纠缠于智能的本体论讨论，而是提出了图灵测试作为智能的判定标准，具体为：如果一台机器能够与人类展开对话（通过电传设备）而不能被辨别出其机器身份，那么称这台机器具有智能（Turing，1950，pp. 433－460）。图灵此文为第一波人工智能的发展奠定了理论基础。不过很少有人关注这个测试的基本方式“语言交流”所包含的深刻意义。

图灵思考人工智能的策略深受语言哲学家阿尔弗雷德·艾耶尔爵士

（Sir. Alfred Jules Ayer）影响。首先，它通过语言交流而非身体形态或生物学特征来确认“人的智能”，预设了“人”及其智能的最核心表现是语言符号能力。艾耶尔在其论著《语言、真理与逻辑》中指出“自我意识绝不包含任何实体自我的存在”（2015，p. 109），而强调语言在哲学中的核心位置。因此，“语言”实际上就构成了可被实证的“自我”之所在——这也是图灵测试的哲学基础。其次，图灵测试通过可操作的实验和经验观察来验证机器智能的表现，而悬置内部过程，排除了对机器内在意识的讨论，并不追究机器是否“真正思考”这一形而上学问题，这就契合艾耶尔所秉持的实证主义立场。总之，图灵测试的设计理念与艾耶尔行为观察方法有着共通之处，展示了对语言和行为的重视，以及对传统形而上学问题的规避，其实用性使得图灵测试成为评估人工智能的一种实用而持久的方法。图灵测试直到今天仍在持续被尝试，较著名的如，在“尤金·古斯特曼”（Eugene Goostman）程序于 2014 年举行的图灵测试中，据媒体报道，有 33% 的评估者未能识别出机器，从而认为对方是人类。

图灵测试的另一个重要的基础原则就是后面被称为“符号主义”（Symbolicism）的逻辑计算模拟。图灵用一个理想化的数学模型模拟计算过程。它是一个能够读写符号的机器，通过一组简单的规则进行计算。图灵机理论为第一波人工智能浪潮奠定了理论基础，但符号主义的哲学基础则可追溯到 19 世纪末和 20 世纪初的数学逻辑和哲学发展，尤其受到逻辑学家和哲学家如戈特洛布·弗雷格（Gottlob Frege）、伯特兰·罗素（Bertrand Russell）、阿尔弗雷德·诺斯·怀特海（Alfred North Whitehead）等学者的影响。这些学者致力将逻辑符号化，用符号系统表达数学和逻辑命题。在他们看来，似乎所有的问题都可被归纳为符号逻辑。因而，符号主义形成了一个可表示为三段论的工作假设：首先，符号是一切语言、逻辑命题或其他内容的形式化表示；其次，人类思维过程可以用符号表示；因此，可以通过符号的组合和操作来模拟人类的认知过程。换言之，人的智能理论上可以被具有计算能力的机器还原。

### （二）符号主义基于逻辑符号的拟主体交互

符号主义也称“逻辑主义”（Logicism），带有强烈的还原主义色彩，它将人类视为“逻辑动物”。其所称“逻辑符号”与卡西尔说“人是符号的动物”中所使用的“符号”一词内涵有差异。符号主义所说的“符号”基于固定明确的指示关系用于逻辑运算和推理，而卡西尔“符号使用”强调的是人

的社会文化心理可能产生的多样意义，是从符号角度丰富了对人的本质属性的认知，是对哲学终极问题“认识你自己”的里程碑式贡献。他在专著《人论》中专门以“巴甫洛夫的狗”为例说明了人与动物对信号形成的反馈机制的根本差异，提出动物只能进行信号的反馈，而人则能提供自由的阐释（Cassirer，2006，p. 32）。与卡西尔类似的人文学观点，有如：符号学家艾柯（Umberto Eco）将“信号”视为符号的下阈限，赵毅衡则将信号视为“不完整符号”。综合来看，信号与符号在研究中的差异体现为三个不同倾向：

首先，信号无需解释而倾向过程性要素并引向精确量化，符号因注重解释而倾向受者感知。学界广为接受的定义“信息即负熵（逆熵）”源自 1948 年美国数学家、信息论的创始人香农在《通讯的数学理论》的界定——“信息是用来消除随机不定性的东西”。符号则是“携带意义的感知”（赵毅衡，2011）——诉诸意义维度。其次，信号指向反应，指向固定的反馈机制，而符号因其本质的任意性而指向不确定的阐释，并最终引向无限衍义。因此，信号在逻辑上符合这一阶段的物理系统假设，但符号的意义却最终受制于背后的社会规约，是人类文明传承的本质性力量。在卡西尔看来，人类文化演化是（文化的）符号生产与自然交往两个过程及其交织的结果。最后，信号作为信息的基本单元，所属的学科背景是信息论，其本质是物理主义的，而符号是意义之学，属人文主义的维度。

符号主义路径与 20 世纪 60 年代符号学在人文学科中的主流发展分道扬镳。以符号主义为基础的人工智能，发展到专家系统阶段后，在特定领域产生了一批具有影响力的产品，如战胜象棋大师卡斯帕罗夫的电脑“深蓝”。符号主义的许多应用和原理迄今仍然影响着人工智能的发展，但符号主义加行为主义的局限也逐渐暴露出来。在应用层面，一方面，是受到当时计算机算力的限制，不足以支持复杂符号系统的运作，一旦面临组合爆炸，系统就无法应对；另一方面，符号主义的逻辑符号抽离了意义的复杂性，使得人工智能并不能很好地反映现实世界中意义的模糊性，这使得它尚不能很好地处理自然语言。

由上，第一波人工智能尽管名为“符号主义”，但若说这种理论假设与符号学所说的“符号”有某种交集的话，除了名称上的关联，则是用了符号的“指示性”部分，而排除了其中的社会规约成分，是纯粹的逻辑符号或信号——至多是符号的下限。这个阶段的人工智能所呈现的交互主体性主要是由“去实体化”的行为主义结果来呈现的。在理论层面，这个阶段的智能计算内部过程是没有黑箱的，任何一个结果都基于特定程序计算。但以图灵测

试为代表的交互智能观承认基于逻辑符号交互的“拟主体”，恰恰是对过程不加追问的，唯其如此，机器才被赋予了仿佛与人共通的交互主体能力。但在后续的发展中，机器是否真正“理解”人的语言这一问题将被重新提出。

## 二、联结主义：试图通向“生命体验”的神经网络

### （一）来自“中文房间”思想实验的反驳：“意义理解”的生命体验前提

约翰·塞尔提出的“中文房间”思想实验（Searle，1980，pp. 417 - 457），就是针对图灵测试及其行为主义理论基础的反驳，具体设想如下：一个不懂中文的人被关在一个房间里，房间里有一本规则手册，这个人可以根据手册将输入的中文符号转换为输出的中文符号。从外部观察来看，这个人的输出似乎是理解中文的，但实际上他并不懂中文，只是在机械地进行符号操作。塞尔旨在批判分析人工智能特别是强人工智能（Strong AI），其攻击的要点即图灵测试所依据的“行为主义”所观察到的“智能现象”并非真正的智能。塞尔指出，虽然计算机程序可以在形式上处理符号，但这并不意味着它们真正理解这些符号的意义。

艾耶尔从语义（Semantics）与语法（Syntax）角度阐述了关于“理解”的区别。其中，在“中文房间”实验中，房间内的人只是根据语法规则进行符号形式操作。语法仅仅是符号之间的形式规则和逻辑结构，而语义指的是符号内蕴的意义。塞尔认为，计算机程序只能处理语法，但不具备语义理解能力。语法本身不足以产生语义，计算机即便能够处理复杂的语法操作，也无法从根本上理解语义。由此，他进一步提出，心灵状态具有内在的主观性和意识体验，这种体验是符号形式操作无法捕捉的，他因此得出“机器不能具备心灵状态”的结论。

塞尔论点的核心主张是，意识是基于生物过程的，因此，心灵现象必须通过生物学机制来解释，而人工智能系统不能仅通过符号操作和计算模拟来达到与人类相同的心灵状态，心智能力是生物大脑的产物，而不是任何形式程序的产物。他的观点在语言哲学、认知科学和人工智能伦理等领域产生了深远影响。这一阶段的语言符号观念也在各种因素驱动之下进入了带有“生命科学”色彩的联结主义时代。

### （二）生物神经模拟与生物的符号“环境界”

这一时期，以 DNA 双螺旋结构发现为标志的生命科学取得了长足进步，其发展影响了几乎所有学科，也成为人工智能与符号学的共通背景。20 世纪 80 年代，人工智能进入了第二波发展高潮。这一阶段，人工智能研究重点转向了联结主义（Connectionism），强调通过神经网络模型来模拟大脑的学习过程和信息处理方式。

实际上，联结主义最早就是源于 19 世纪末 20 世纪初美国心理学家爱德华·李·桑代克（Edward L. Thorndike）的动物实验研究。通过动物实验，桑代克观察到了学习过程中的联结现象，并试图用这一理论来解释学习过程（Thorndike，1898，pp. 1 - 109）。这一阶段，人工智能向前兼容了物理符号系统，但更强调符号间的关系，并突出了符号链与行为结果的反馈机制。居于主导地位的理论假设是大卫·拉姆哈特（D. E. Rumelhart）、杰弗里·辛顿（G. E. Hinton）等提出的，用于训练多层感知机的“反向传播算法”（back-propagating errors）（Rumelhart，Hinton & Williams，1986，pp. 533 - 536）。反向传播算法通过有效的梯度下降优化策略，实现了多层神经网络的训练和应用，成为联结主义浪潮中至关重要的技术突破。它不仅为第二波人工智能浪潮提供了强有力的技术支撑，还为后续深度学习的兴起奠定了坚实的基础。

生命科学也同样推动符号学内部分化出了“生物符号学”，其秉持的连续论等观念，并不是要处理人类之外的动物的偶发符号现象，而是雄心勃勃地将整个符号学纳入其中，正如柯布利（Paul Cobley）所说“生物符号学就是符号学”（2024，p. 45）。从符号本身的意指关系来看，这个阶段“符号”不再是单个“信号”的逻辑推理，而被视为一个多层级、多关联方式构成的“神经网络”。换言之，联结主义的侧重点不在于单个符号的纯粹指示性，而更注重符号链的彼此联结关系，这是对符号主义的重要补充。这一阶段人工智能进入医疗领域以及语音识别，为很多后续人工智能发展奠定了基础。不过，塞尔的论点实际上将“生物性”作为“智能”或“理解”的前置条件，成了绝对化前提，使得后续讨论失去了必要性。要客观地理解“交流”本身，需要从生物性作为必要前提的必然性，以及语义、语法与语用交流之间的总体关系来继续展开讨论。

联结主义提出的神经网络等概念，形式上比较接近于符号学常说的“意义之网”。不过，联结主义的生物神经之网并非包含人类社会规约的意义之

网。在生物符号学的视域下，每种生物因自身所能感知和理解的全部内容而形成意义之网——这构成了此种生物的“环境界”（Umwelt）。“环境界”包含了动物所感知的“周围世界”。西比奥克指出，“对于任何有机体来说，在其气泡般的私属环境界之外，什么也不存在”（Sebeok，2001，p. 19）。这意味着，这个私属的周围世界与人类意义之网存在一些区别。人类意义世界不仅需要一个个符号联结成符号链，更需要置入人类特定社会文化语境之中。正如布里尔（S. Brier）指出，“‘意义’是基于共通体验的耦合结果，是所有语言符号过程的重要基础。而词（word）并不携带意义；相反，意义是基于感知着的背景体验而被感知到的”（Brier，2008，p. 87）。

总体上，这一阶段人工智能交互的方式是强调联结的，但它未打开社会文化这一总体意义之网，而任何人类的交流总是在具体的社会文化语境之下完成的。即便悬置“生物性”这一前提预设，文化语境也是此阶段的人工智能交互主体性所欠缺的重要部分。文化语境也可以理解为人类的整体知识或全体文化这一“总体数据库”。由于互联网尚未大规模介入，这种包含人类全部知识的大数据库在联结主义发展的 20 世纪 90 年代时代并未形成，这也成了联结主义的瓶颈，限制了人工智能在实际应用中的表现。

## 三、深度学习的“过程模拟”与交互主体性新曙光

对于人工智能的第三波浪潮的发生时间，不同学者的看法略有不同，但大致时间是以 21 世纪为起点，也有学者明确为 2006 年。这一年加拿大多伦多大学的辛顿等人提出了深度信念网络（deep belief networks，DBN），并通过无监督训练算法在图像识别等领域取得了突破（Hinton et al.，2006，pp. 1527－1554）。人工智能第三波浪潮实现突破的大环境，则被归结为互联网大数据和计算机芯片摩尔效应共同催生的软硬件发展。它们共同解决了前两波人工智能面临的算力限制和数据源限制问题。几乎所有不同立场的人，都同意人工智能正在飞速进步之中。对图灵测试的努力也从未停止。最近，美国加州大学圣地亚哥分校的喀麦隆·琼斯（Cameron R. Jones）和本杰明·伯尔根（Benjamin K. Bergen）发起了一项实验，共有 652 名志愿者参与，完成了 1810 场次游戏。在这些游戏中，志愿者被要求通过聊天界面与对方互动，然后判断对方是人类还是人工智能。测试结果显示，ChatGPT－4 在最优情况下实现了约 49. 7% 的成功率。这一结果远远超越了历代人工智能对话程序。总体结果显示，尽管 ChatGPT－4 与人类在测试中依然存在差距，但

在模仿人类对话方面已经取得长足进步（Jones & Bergen，2023－10－31）。这些进步意味着，人工智能交互能力在继续迭代向人类靠近。

### （一）基于深度学习与大模型的“过程模拟”

主导人工智能第三波浪潮的“深度学习”看上去在理论上似乎是对前两个阶段的延续和整合，并没有提出全新的理论，但这种整合性恰恰是人脑与类脑智能之间作为复杂系统的共同特征。同时，由于互联网提供的知识总库补足了机器在社会文化语境方面的缺失，这一时期通用人工智能在符号能力方面表现出了惊人的进步。以 ChatGPT 为代表的大语言模型表明，人工智能正在从专家系统向通用智能转变；同时，人工智能对自然语言、图像等多模态符号的处理能力有了长足进步，人工智能一旦获取、学习各种方言的语料，就能够获得相应的处理能力。这种从无边界的外部环境学习的方式，接近包括人类在内的生物学习过程，其结果也类似“人的智能行为”，并取得了惊艳的表现。

可以从两个具有同类性质的标志性事件——深蓝战胜卡斯帕罗夫和 AlphaGo 击败李世石——来比较“深度学习”与“符号主义”的理论基础，以理解人工智能当中符号观念的演变。IBM 的电脑深蓝主要是依托当时的超级计算机，通过穷举搜索和启发式评估函数来决定棋步，它通过存储尽可能多的人类下棋的策略来作为自己策略选择的参照依据，是一个专家系统，本质上属于符号主义的方法。它展示出人工智能在规则明确定义、信息完全的任务中取得了超越人类的能力，但缺乏通用性和自适应能力。但对于 19×19 大小的围棋盘来说，其决策的可能性是 $10^{170}$，迄今为止任何超级计算机都没有能力以穷举法来分析围棋的所有走法。DeepMind 开发的 AlphaGo 程序在围棋比赛中击败世界冠军李世石，依靠的是深度神经网络能够从大量棋局数据中自动提取特征。通过进行复杂模式识别，后期再通过自我对弈进行无监督学习，其能力进一步提升。后面战胜柯洁的人工智能程序 AlphaZero 的升级更加彻底，它没有人类棋局的输入，仅仅了解围棋的基本规则，从零开始进行学习，通过自我对弈，不断探索和积累经验，提升棋力。AlphaZero 的设计没有特定领域的棋局规则，只需基本规则便能学习任何棋类游戏，展现了人工智能技术在多领域、多任务环境中的通用性和适应能力。

人工智能在复杂博弈等问题上的突破，及在自然语言处理方面的发展显示出当前人工智能的符号表意策略在两方面悄然改变：一方面，人工智能对人脑的模拟过程正在从此前的“结果模拟”向“学习过程模拟”转变，在围

棋博弈中，人工智能并非调取所有棋局来实现对弈，而是基于“复杂模式识别”的自我学习；另一方面，恰恰是通常被视为深度学习的两个缺陷——模型“黑箱”和数据“伦理偏见”得到了重要发展。从工具应用角度来看，它们无疑是缺陷，如“黑箱”效应会导致司法、医疗、自动驾驶等具体情形下不可预知的结果，还会导致追踪模型错误成因变得比较困难。但从符号与人的关系来看，“黑箱”与“伦理偏见”更符合人类符号表意的特性。

首先，人工智能大模型内部的参数剧增，系统参数过于复杂，并涉及多线程决策系统，复杂系统的人工智能决策系统内部一个微小的参数就可能导致的不确定性增加，再加上与外部环境变量接触，“黑箱”效应凸显。其表现形式可能引发“犹豫”这种拟主体的行为特征。换言之，深度学习的“黑箱”及其本质上反映出的“复杂系统”的决策机制恰恰通常是人类决策才会呈现的。在大多人文学者看来，人工智能被视为“无心智”的机器，很大程度上是由于它被认为是“纯逻辑”，并且在“后台完全编程”的。而深度学习的过程正在越来越接近人类大脑的模拟思考，它可能有犹豫、错误和偏见，甚至也会胡诌。当人类在与人工智能交互过程中感受到机器“不可信”时，交互中的拟主体性就悄然涌现了。“黑箱”所呈现的那种“人心不可测”对人类来说是一个悖论——它既构成人类对人工智能深深的恐惧，也是人类自身主体性的本质特征之一。

人工智能决策的外部行为结果的“伦理偏见”同样非常具有类人特质。目前各种大模型已经以多种方式接入了互联网和物联网，它的知识源理论上就成了人类社会全部知识。这些知识本身不是专家系统时代那类被高度遴选过的，而是人类社会在互联网上的全部投射。模型在训练过程中尽管会存在人类后台干预和加权特征，但它不可避免地，会基于与性别、种族相关做出偏见性决策。作为缺陷，它的极端情形就是人们常说的深度学习的“垃圾进，垃圾出”。这个问题非常类似于儿童教育问题。悖论的是，人们一面批评人工智能是“无情的逻辑机器”，一面却试图消灭它身上所折射出来的“人性”，让它回归那个中性、无偏差的媒介工具。

### （二）“符号交互主体性”与跨语际交流的敞开

由于深度学习的深入开展，人工智能具有的不确定性使人们无法再用简单的“机器性”否定人工智能的“拟主体性”。那么，生物性是否必须作为交流或理解的前提？

首先，可将生物性视为一个交流他者，作为交流主体双方的基本预设。

换言之，交流主体必然存在一个逻辑上的他者，否则就无需发生交流。其次，此交流双方之他者必须存在可能的交集。由此，交流或传播的意思就是：两个逻辑上存在距离的相异主体实现其交集的过程。但这个界定并不能直接排除人与非人的“异体”交流。既然交流存在的逻辑前提是意义主体的相异，则从人的角度出发否定其他“被感知到交流”的异体，就颇有“子非鱼安知鱼之乐”的意思。每个主体的内心本身就是一个“黑箱”过程，所谓“此之甘饴，彼之砒霜”，是不同主体对同一事物的价值观念不同。

对于异体之间的跨界交流，中西方思想史多有论述。《圣经》中用“对空而语”（speak into the air）（哥林多前书 14：9）来形容交流之难，如同徒劳之举。因此当年何道宽在翻译彼得斯（John Durham Peters）的 *Speaking into the Air: A History of the Idea of the Communication* 时，直译为《交流的无奈：传播思想史》（2003）也正是凸显完全交流之不可能。在这部书当中，彼得斯分析了交流的三种意向分支，分别是：给予或联结（imparting）、（物质或思想）的传输或迁移（transfer or transmission），以及交换（exchange）。这三种意向都并未预设一个人类或其他物种边界。不仅如此，这部书还广泛地涉及跨越生死的交流（招魂术），人与动物、机器及外星人的跨界交流。

当我们以物种为壁垒，将人设定为唯一的交流对象，人类将失去与其他生命形式交流的可能，而当人类以动物或全部生命为边界，排除其他形式非生命世界的交流，我们会失去与宇宙中未知可能的对话。由此，关于生物作为“智能”的前提可调适为：生物演化产生了生命感知的独特方式。生物的生存本能让特定生物发展出了某种更适应于该物种生存的感知方式。这就意味着，这种感知或理解方式必然是“特别的”，而不是“绝对的”。因此，跨越主体的交流也是相对的、特别场域下的有限交集。比如，人们认为自己理解了他的宠物狗或小狗懂了人们的心，无非是在跨物种的交流之中，我们基于某个意义场域达成了临时性的跨语际交集。此时去追问整个“狗生”与“人生”的体悟之差别就没有必要了。即便是在人类这个物种内部，这种有限交集的获得也是我们实现交流的全部内容。

塞尔的“中文房间”批判图灵测试过于集中在“符形操作”或许是有力的，但他自身又陷入了狭义语义论之中。既然所谓的“生命体验”本就是千差万别的，我们何以用个体“我”之理解去否定“他者”体验，以及去断定我并不能与之共情的非生命体的“非体验”？我们只能知道我们不同，而不能否定他者的存在，即所谓“子非我安知我不知鱼之乐?”语言哲学转向以来，语言背后的那种“本体”就弥散了，所留下的唯有语言或符号现象。这

意味着，语言需要从狭义向广义拓展。人类定义的语言，不过是以特定物种和文化社群为基础形成的狭义语言，它仅仅是丰富多样的自然语言和符号中的“特定方言”。

某种意义上看，交流的本义就是将人从幽闭的意义囚笼之中解放出来。因此，广义的语言和符号就不仅仅是单个符号的语义问题或符号形式操作的问题，而是一个包括受传双方、媒介传输形式在内的整体语用问题。例如，借助超声波人类可以在蝙蝠的意义上与蝙蝠交流；使用二进制代码，人类可以跟图灵机直接对话。反过来，动物也在学习“理解”人类的语言。例如我们养的宠物基于与人类的相处，就会对人类的手势或语言做出我们所期待的回应。从人工智能的发展来看，机器也在以人类所期待并实质上作为后台力量推动的方式学习这种“理解”。从当年的图灵机到今天的 ChatGPT－4，人工智能也一定程度上学会了对机器而言的“外语”——人类自然语言。这是一个跨语际交流问题，至于询问这些语言背后是否有“人类的灵魂”或“生物的心灵”，这本身就包含了人类中心主义的预设。说人工智能没有生物学意义上的进化史，实际上陷入了一种空洞的本体论立场之争。更切实的方案或许是，从人工智能在学习过程中所呈现的全部语料背景出发，承认其表现出来的“拟主体性”，分析它所有语言符号表达现象对人类语言的模拟特征。这样，就可以清晰地理解，人工智能并没有那种基于生物进化发展而来的“心灵”，而只有通过模拟和变异表现出来的“心智”——一种人类集体意识的文化的镜像。

## 结论：作为意识镜像的人工智能“符号拟主体性”

回顾人工智能半个多世纪以来的发展，其演进历程清晰地体现在它对语言、图像及其他行为等符号系统的处置上。在符号主义阶段，符号是一个个被赋予指示性对象的逻辑符码或信号；在联结主义前半段，符号之间的邻接关系凸显，不再是单一的符号，而符号链成为接近生物神经感知的意义之网的尝试。综合前两阶段成果发展而来的深度学习阶段，侧重点已经从单个符号和符号链向模式识别这种“符号语篇”转化，人工智能的新能力凸显为模式特征识别。人工智能对人类模仿的重点从逻辑结果向学习本身的过程转化，其模拟的不是人脑的结果，而是包括社会发展在内的演化进程。

早期以符号主义为基础的人工智能没有“黑箱”，是完全可解释的。人工智能发展到深度学习阶段，“黑箱”出现，导致内部解释不完全。为了实

现人工智能自主性增强，人工智能算力必然持续增加且算法将进一步复杂化，这意味着人工智能“黑箱”有可能持续扩大。若这个发展成为现实，则行为主义学派将在未来人工智能中有较大潜力，进而也意味着人类对人工智能的解释性将持续走低——更低的解释性意味着更加自主的“拟主体性”。行为主义似乎回到了图灵时代所强调的外部观察。这种行为主义的方法，通常只对独立的生命个体或人类才适用。从行为观察理解人工智能，并通过奖惩措施来激励机器学习的方法，在原理上符合皮尔斯符号学的“试推法”（abduction）（皮尔斯，2014，p. 109），也与人类或其他一些哺乳动物培养孩子的过程相仿。当我们面临人工智能输出的符号、所说的话、绘制的图像以及其他行为结果内在决策机制上的不确定性时，这本身就意味着人们正在以有限程度的“准主体性”处理与人工智能的交流。

行文至此，回望麦克卢汉 60 年前的判断“即便计算机有某些意识，也不过仍然只是人类意识的延伸”，它是否仍然有效？新的图灵测试中，人类之所以仍然“占优”乃是人类在元认知上为自己提供了“我知道我在测试”的某种警戒[①]。若将今天的人工智能提供给阿兰·图灵在他那个时代开展测试，人工智能恐怕早已成“人”。人类只是在退守“何以为人”的过程中不断提出诸如艺术创作、灵感、元意识等新的标准去要求人工智能。或许，只要人类能够提出具体标准，从逻辑上看，人工智能并没有不能实现的理论限制。但问题的关键不在于人工智能有多么聪明，而在于人工智能存在的全部原因是人。无论人工智能表现得多像人，它都没有基于生物生存演化的那种“自足的目标”，其所实现的都是人类意识对人造智能的预设、想象和自我投射的结果，也即麦克卢汉所说的“人类意识的延伸”。

图灵测试看似在要求机器成为人，而实际上是让智能机器折射出人类自身是什么，成为人类的一面镜子。人要求这面镜子照出自己的“美”，但镜子也会折射“恶”的一面。而人工智能当前及未来在全部互联网世界中汲取的知识、文化中也可能包括陋习，它只是更全方位地成为人类集体意识的镜像。“镜像”本就是一种关于人类意识的元符号形式，它具有突出的自指性，并包含着生物发展演化“自我认知”的关键阶段，涵盖人类“个体发展”与“社会身份”的关键内容（胡易容，2024，pp. 112－115）。只不过，生物越过镜像阶段，是“自我”的觉醒，而人工智能这里折射的并不是“智能机

① 霍桑效应（Hawthorne Effect）：当被测试者意识到自己正在接受测试或观察时，他们的行为可能会有所改变，通常表现出更积极或更有效的行为方式。

器”，而是人类。因此，这是一个不对称的镜像。从“交互主体”这个关键词来看，人类实际上是与一个自我的分身做交流游戏。人类通过深度学习的方式训练人工智能，当人工智能接入互联网的时候，它吸纳的就是人类的意识——习得语言、语法以及社会文化，并成为人类意识的镜像，只是这镜像有可能延展、扭曲，甚至形成一些看上去“独立”的观感，但其实背后都是人的符号自我的技术化投射。因而，当人们讶异于人工智能的惊艳表现时，他们讶异的是人类自身意识的外在延伸。懂得我们的，不是一个机器他者，而是另一个符号自我。

**引用文献：**


艾耶尔，阿尔弗雷德（2015）. 语言，真理与逻辑（尹大贻，译）. 上海：上海译文出版社.

巴拉特，詹姆斯（2016）. 我们最后的发明（闾佳，译）. 北京：电子工业出版社.

彼德斯，约翰（2003）. 交流的无奈：传播思想史（何道宽，译）. 北京：华夏出版社.

哈贝马斯，尤尔根（1989）. 交往与社会进化（张博树，译）. 重庆：重庆出版社.

胡易容（2024）. 文化传播符号学论纲. 北京：科学出版社.

金磊（2024－8－1）. ChatGPT 版《Her》被玩疯：哭着读诗，中文表现也很亮. 获取自 https://mp.weixin.qq.com/s/YAC56GANpNfrrpc8q3Jf6w.

卡西尔，恩斯特（1985）. 人论（甘阳，译）. 上海：上海译文出版社.

柯布利，保罗（2024）. 生物符号学的文化意涵（胡易容，等译）. 北京：社会科学文献出版社.

麦克卢汉，马歇尔（2000）. 理解媒介：论人的延伸（何道宽，译）. 北京：商务印书馆.

皮尔斯，C. S.（2014）. 皮尔斯：论符号（赵星植，译）. 成都：四川大学出版社.

Brier, S. (2008). *Cybersemiotics: Why Information is Not Enough!* Toronto: University of Toronto Press.

Cassirer, E. (2006). *An Essay on Man: An Introduction to a Philosophy of Human Culture*. Hamburger: Meiner Verlag.

Habermas, J. (1984). *The Theory of Communicative Action, Vol. 1: Reason and the Rationalization of Society*. Boston: Beacon Press.

Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, 18 (7), 1527－1554.

Jones, C. R., & Bergen, B. K. (2023－10－31). Does GPT－4 Pass the Turing Test? Retrieved from https://arxiv.org/abs/2310.20216.

McCarthy, J., Minsky, M., Rochester, N., & Shannon, C. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. Dartmouth College.

Newell, A., & Herbert, A. S. (1976). Computer Science as Empirical Inquiry: Symbols and Search. *Communications of the ACM*, 19 (3), 113 - 126.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning Representations by Back-propagating Errors. *Nature*, 323 (6088), 533 - 536.

Searle, J. R. (1980). Minds, Brains, and Programs. *The Behavioral and Brain Sciences*, 3, 417 - 457.

Sebeok, T. A. (2001). Nonverbal Communication. In P. Cobley (Ed.), *The Routledge Companion to Semiotics and Linguistics*, 19. London: Routledge.

Thorndike, E. L. (1898). Animal Intelligence: An Experimental Study of the Associative Processes in Animals. *Psychological Review Monograph Supplements*, 2 (4), 1 - 109.

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, 59 (236), 433 - 460.

**作者简介：**

胡易容，四川大学文学与新闻学院教授，四川大学符号学 - 传媒学研究所所长，主要研究领域为传播符号学、文化与艺术学理论等。

**Author:**

Hu Yirong, professor of College of Literature and Journalism, director of ISMS Research Team, Sichuan University. His research interests are semiotics of communication, culture and art theory.

Email: hyr@ scu. edu. cn