神经符号人工智能——为何、是什么以及如何实现

Amit Sheth南卡罗来纳大学人工智能研究所，美国南卡罗来纳州哥伦比亚，amit@sc.edu

Kaushik Roy
南卡罗来纳大学人工智能研究所，美国南卡罗来纳州哥伦比亚，kaushikr@email.sc.edu

Manas Gaur
马里兰大学巴尔的摩县分校，美国马里兰州，manas@umbc.edu

arXiv:2305.00813v1 [cs.AI] 1 May 2023

摘要——**人类通过感知与认知相结合的方式与环境互动：感知将来自环境的感官输入转化为符号，而认知则将符号映射为关于环境的知识，以支持抽象、类比推理和长期规划。在人工智能领域，受人类感知启发的机器感知，是指利用通过自监督学习目标训练的神经网络，从原始数据中进行大规模模式识别；这些学习目标包括下一词预测或物体识别。另一方面，机器认知涵盖更复杂的计算，例如利用对环境的知识来指导推理、类比和长期规划。人类还能够控制并解释自己的认知功能。这似乎要求保留从感知输出到环境知识的符号映射。例如，在医疗保健、刑事司法和自动驾驶等安全关键型应用中，人们能够遵循并解释指导其决策的指南和安全约束。数据驱动的神经网络人工智能算法能够有效地模拟机器感知，而基于符号知识的人工智能则更适合模拟机器认知。这是因为符号知识结构支持对感知输出与知识之间映射关系的明确表示，从而能够追溯和审查人工智能系统的决策。此类审计记录有助于通过追踪人工智能系统的输入、输出和中间步骤，落实应用层面的安全要求，例如遵守法规和确保可解释性。本文是《神经符号人工智能》专栏的首篇文章，将介绍并概述这一迅速兴起的范式——神经符号人工智能。该范式结合神经网络与知识引导的符号方法，打造能力更强、灵活性更高的人工智能系统。这些系统有望大幅提升人工智能系统在算法层面（如抽象、类比、推理）和应用层面（如可解释且受安全约束的决策）的能力。**

I. 为什么需要神经符号人工智能?

神经符号人工智能是指致力于将基于神经网络的方法与基于符号知识的方法相结合的人工智能系统。我们从两个角度来更好地理解这种结合的必要性：（1）算法层面的考量，例如支持抽象、类比和长期规划的能力。（2）人工智能系统的应用层面考量，例如确保可解释性和安全性。

算法层面的考量

研究人员已经在人类大脑中发现了不同的系统，分别专门处理与感知和认知相关的信息。这些系统协同工作以支持人类智能，并使个体能够理解周围的世界并与之互动。丹尼尔·卡尼曼普及了系统1和系统 2 [1]的目标与功能之间的区别。系统1对于帮助个体理解其周围环境中遇到的大量原始数据，并将其转换为有意义的符号（例如词语、数字和颜色），以供进一步认知处理，至关重要。系统2则执行更有意识、更审慎的高阶功能（例如推理和规划）。它利用背景知识准确定位感知模块的输出，从而支持类比、推理和长期规划等复杂任务。尽管功能不同，系统1和系统2彼此相连、协同合作，共同构成人类的体验。这两个系统使人们能够根据自己对环境的了解进行观察、理解和行动。在过去十年中，基于海量数据训练的神经网络算法展现出卓越的机器感知能力，例如在预测下一个词和识别数字等自监督任务中取得了优异表现。值得注意的是，仅通过这类简单的自监督任务进行训练，就已在蛋白质折叠、高效矩阵乘法和复杂谜题求解等具有挑战性的问题上取得令人瞩目的成果 [2], [3]。然而，知识使人类能够进行超越现有数据明确表述内容的认知过程。例如，人类会通过映射到阐明这些映射关系的知识结构，在抽象背景相似的概念之间建立类比联系 [4]。或许，GPT-4等当前的生成式人工智能系统能够仅凭数据习得支持认知功能的知识结构 [5]。这一假设认为，从互联网上的大量文本中预测下一个词，能够促使神经网络涌现出关于世界的“认知模型”，并以此支持认知。然而，人们对其黑箱特性及由此造成的难以理解性存在重大担忧，这阻碍了对其认知能力进行可靠评估。另一方面，符号模型虽然不适合处理海量数据，却非常适合利用知识结构（例如知识图谱）支持类人认知。因此，与其依赖其中一种系统，不如将两类系统整合起来：擅长大数据驱动处理的基于神经网络的系统1，以及擅长处理依赖知识的

认知。

应用层面的考量

神经符号人工智能结合系统1和系统2，可以实现重要的应用层特征，例如人工智能中的可解释性、可解释性、安全性和信任。近期关于解释神经网络决策的可解释人工智能（XAI）方法的研究，主要涉及显著性图、特征归因和基于原型的解释等事后技术。这类解释对开发人员很有用，但终端用户却不易理解。此外，神经网络可能会因数据伪影、对抗性攻击、分布偏移和系统故障等无法控制的训练时因素而失效。为确保严格的安全标准，有必要在训练期间引入适当的背景知识来设置护栏，而不是将其作为事后措施。符号知识结构能够提供一种有效的机制，用于施加领域约束以确保安全，并为可解释性提供明确的推理轨迹。这些结构可以为终端用户构建透明且具备可解释性的系统，从而带来更值得信赖、更可靠的人工智能系统，尤其适用于安全关键型应用 [6]。

为什么选择神经符号人工智能?

在 *AI* 系统中实现智能行为，必须同时涉及感知（即处理原始数据）和认知（即利用背景知识支持抽象、类比、推理和规划）。符号结构能够显式地表示这些背景知识。尽管神经网络是处理数据并从中提取模式的强大工具，但它们缺乏对背景知识的显式表示，这阻碍了对其认知能力进行可靠评估。此外，在提供由背景知识中的概念所引导的可解释结果的同时，应用适当的安全标准，对于建立值得信赖的认知模型以支持决策至关重要。

II. 什么是神经符号人工智能？我们如何实现它?

神经符号人工智能是一个用于描述相关技术的术语，这些技术旨在将基于知识的符号方法与神经网络方法相结合，以提升人工智能系统的整体性能。这些系统能够将神经网络强大的逼近能力与符号推理能力相融合，使其能够对抽象概念进行推理、从有限数据中外推并生成可解释结果 [7]。这些组件共同支持前文介绍的算法层面和应用层面的考量。神经符号人工智能方法可分为两大类：（1）压缩结构化符号知识，将其与神经模式集成，并利用集成后的神经模式进行推理的方法；（2）从神经模式中提取信息，以便映射到结构化符号知识（提升）并执行符号推理的方法。此外，我们将（1）进一步细分为采用以下方法：(a) 使用压缩的知识图谱表示与神经模式集成；(b) 使用压缩的基于形式逻辑的表示与神经模式集成。我们还将（2）进一步细分为采用以下方法：(a) 神经组件与符号组件之间的解耦集成；(b) 神经组件与符号组件之间的交织集成。这些方法能够以不同程度的有效性支持算法层面和应用层面的功能，其水平范围从低（**L**）、中（**M**）到高（**H**）。图1 详细介绍了我们对神经符号人工智能方法的分类。

类别1方法的算法层面分析。

对于类别1(a)，以往的工作使用两种方法压缩知识图谱。一种方法是使用知识图谱嵌入方法，通过图神经网络等技术将知识图谱嵌入高维实值向量空间，从而实现压缩。这使其能够与神经网络的隐藏表征相融合。另一种方法是使用知识图谱掩码方法，以适合与神经网络的归纳偏置相融合的方式对知识图谱进行编码。图2 展示了这两种方法。神经网络处理大量原始数据的能力，也使其在处理数百万乃至数十亿个节点和边时，能够应用于知识图谱压缩，即实现大规模感知（图**H**中的图1）。在神经推理流程中利用压缩表示，能够提升系统的认知能力，例如抽象、类比和规划能力。然而，由于知识图谱中的完整语义遭到有损压缩（例如，压缩表示无法有效建模关系），这种提升较为有限（图 **M**中的图 1）。类别1(b)方法使用矩阵和高阶张量分解方法，获取对象及描述对象间关系的形式逻辑陈述的压缩表示（例如命题逻辑、一阶逻辑和二阶情境演算）。认知能力的提升趋势与类别1(a)类似。然而，形式逻辑的压缩技术计算效率低下，也无法实现大规模感知（图 **L**中的图 1）。

类别1方法的应用层面分析。对于类别1(a)，将知识图谱压缩后集成到神经处理流水线时，其完整语义不再被显式保留。显著性图、特征归因和基于原型的解释等事后解释技术，只能解释神经网络的输出。这些解释主要用于协助系统开发者诊断和排查神经网络决策过程中的算法变更。遗憾的是，这些解释并非以领域或应用层面的术语来表述，因此对终端用户的价值有限（（**L**），图 1中可解释性较低）。知识图谱压缩方法仍可用于施加领域约束，例如指定对神经网络中模式相关性的修改，如

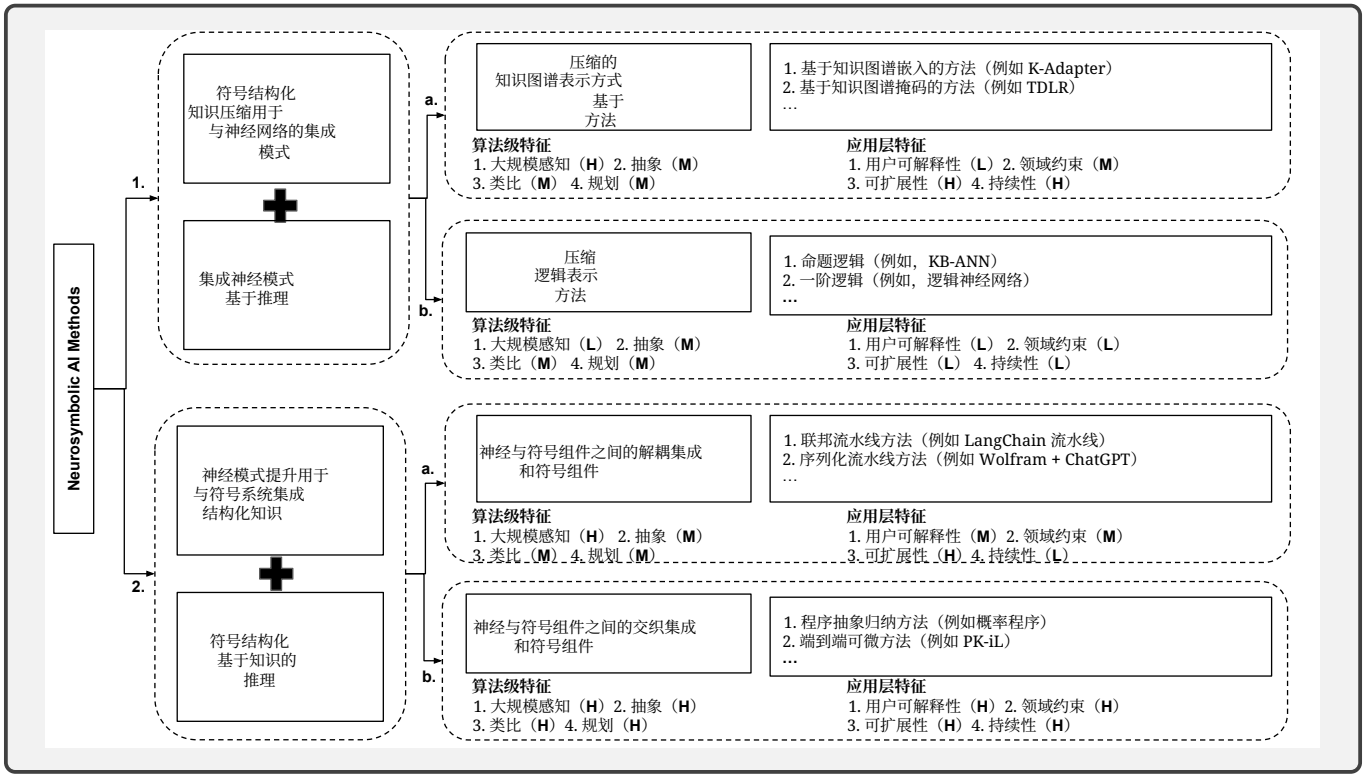


图 1。神经符号技术的两种主要类型——下沉和提升——还可进一步划分为四个子类别。在低（**L**）、中（**M**）和高（**H**）尺度上，这些方法可用于在算法层面和应用层面提供多种功能。

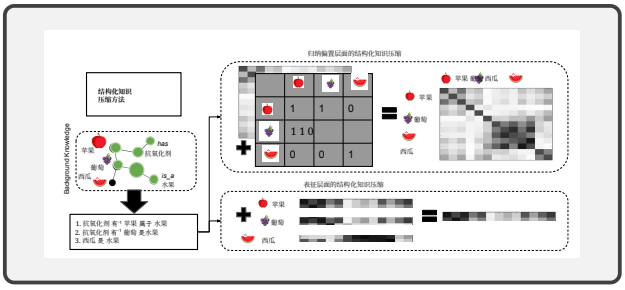


图 2。该图展示了压缩知识图谱并将其与神经处理流水线相集成的两种方法。一种方法是将知识图谱路径嵌入向量空间，从而实现与神经网络隐藏表征的集成。另一种方法是将知识图谱编码为掩码，以修改神经网络的归纳偏置。归纳偏置的一个例子是存储在 Transformer 神经网络自注意力矩阵中的相关性信息 [8], [9]。

图2所示。不过，这一过程的约束规范能力有限，因为大型神经网络包含多个处理层和众多活动组件（（**M**），图1）。要确定对网络所做的修改是否在各个处理层中得以保留，颇具挑战。神经处理流水线确实能实现高度自动化，使系统更容易扩展到各种用例（例如接入特定用例的知识图谱），并支持持续适应贯穿系统的生命周期（例如持续修改知识图谱）。该能力在图 **1** 中以字母（H）表示。对于类别1(b)，当压缩后的形式逻辑表征与神经处理流水线相集成时，在用户可解释性、领域约束、可扩展性和持续适应等应用层面的所有方面，系统得分往往都较低，如图 **1** 中的字母（L）所示。这主要是由于显著的用户与技术之间的障碍所致。最终用户必须熟悉形式逻辑语义的严谨性和细节，才能与系统进行交流（例如，提供领域约束规范）。

类别 2 中方法的算法层面分析。

对于类别 2(a)，大语言模型及其相应插件的兴起推动了联邦流水线方法的发展。这些方法利用神经网络，根据使用自然语言和图像等适当模态描述的任务，识别符号函数。识别出符号函数后，方法会将任务交给相应的符号推理器，例如数学工具或基于事实的搜索工具。图3 展示了一个使用 Langchain 库的联邦流水线方法。这些方法能够通过大语言模型（图**H1** 中的 ）支持大规模感知。然而，它们促进与认知相关的算法级功能（例如抽象、类比、推理和规划）的能力，受限于语言模型对输入查询（图1中的 M）的理解

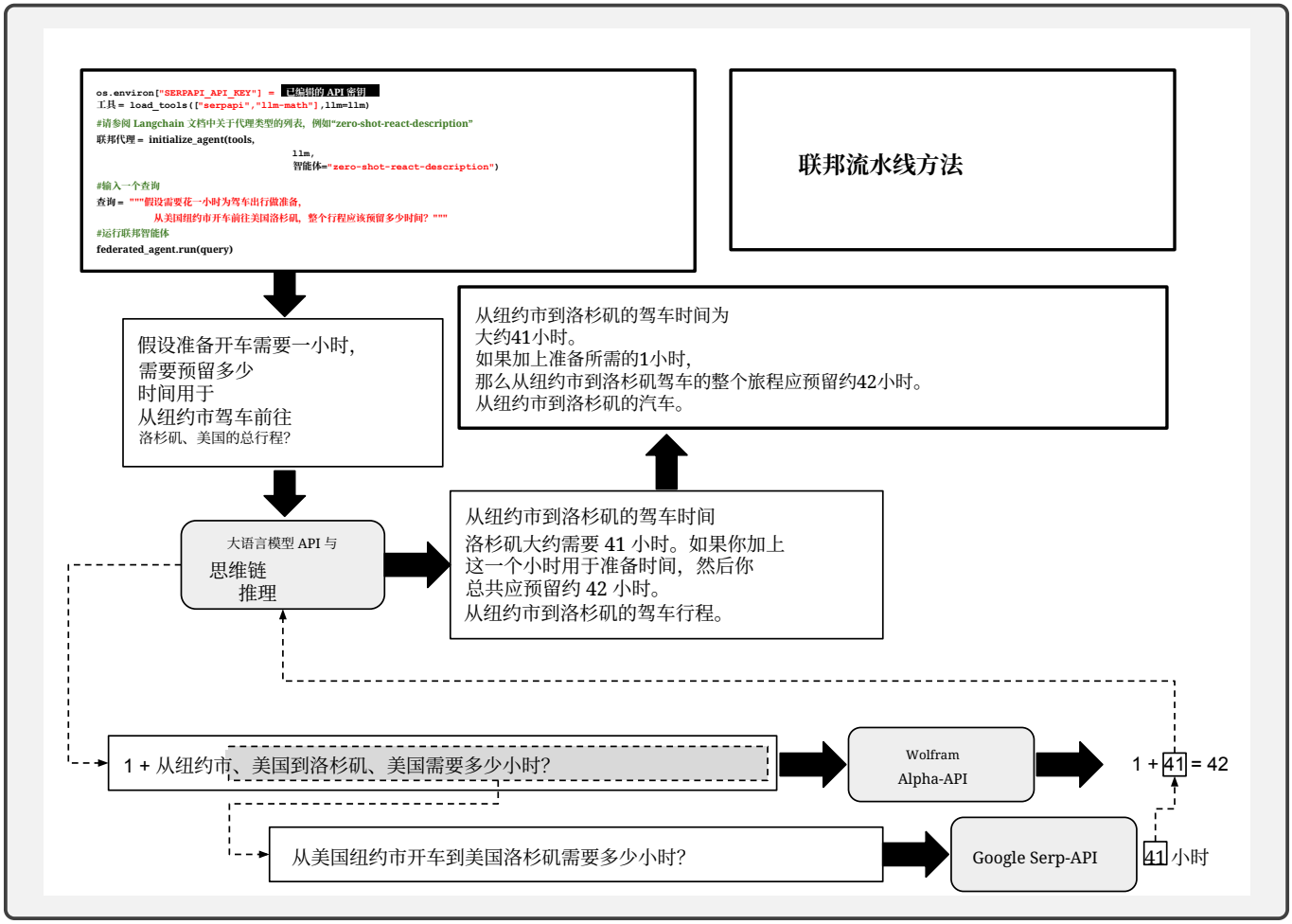


图3。展示了使用 Langchain 库的联邦管道方法。该方法采用经过思维链推理训练的语言模型，将输入查询分解为多个任务。随后，语言模型利用特定于任务的符号求解器推导解决方案。具体而言，语言模型识别出，对于给定查询，需要使用搜索和科学计算（数学）符号求解器。之后，将所得解决方案组合起来，并转换为自然语言呈现给用户。

。类别2(b)方法采用与类别 2(a) 联邦式流水线相似的流水线，但还具备一项额外能力：通过与各组件相对应的端到端函数微分组合，全面控制所有流水线组件的学习。这种控制程度使我们能够针对给定应用，在抽象、类比和规划等方面达到所需的认知水平（图**H1** 中的1），同时保留大规模感知能力。图4展示了这种方法在心理健康诊断辅助中的应用示例。

类别 2 方法的应用层面分析。

对于属于类别 2(a) 的系统，在处理过程中追踪其思维链，能够极大地提升用户可解释性的应用层面效果。然而，语言模型在响应生成过程中解析输入查询并将其与领域模型概念关联起来的能力，限制了这种能力（（**M**）（图1）。此外，使用提示模板以自然语言指定领域约束也限制了约束建模能力，而这取决于语言模型理解应用或特定领域概念的能力（（**M**）见图1）。联邦式流水线具有出色的可扩展性，因为用于促进特定领域用例的语言模型和应用插件正变得越来越普及、越来越容易获取（（**H**）见图1）。遗憾的是，训练语言模型需要耗费大量时间和空间资源，因此，利用联邦式流水线进行持续领域适应仍然面临挑战（（**L**）见图1）。尽管如此，支持持续学习目标的语言建模架构正迅速受到关注。类别2(b)方法前景显著，因为它们在所有应用层面的方面得分都很高，包括用户可解释性、领域约束、跨用例的可扩展性，以及对持续适应特定应用的变化的支持（（**H**）见图1）。这得益于其高度的建模灵活性，以及系统组件之间紧密交织的耦合。因此，任何特定组件的变化都会带来所有组件的积极变化

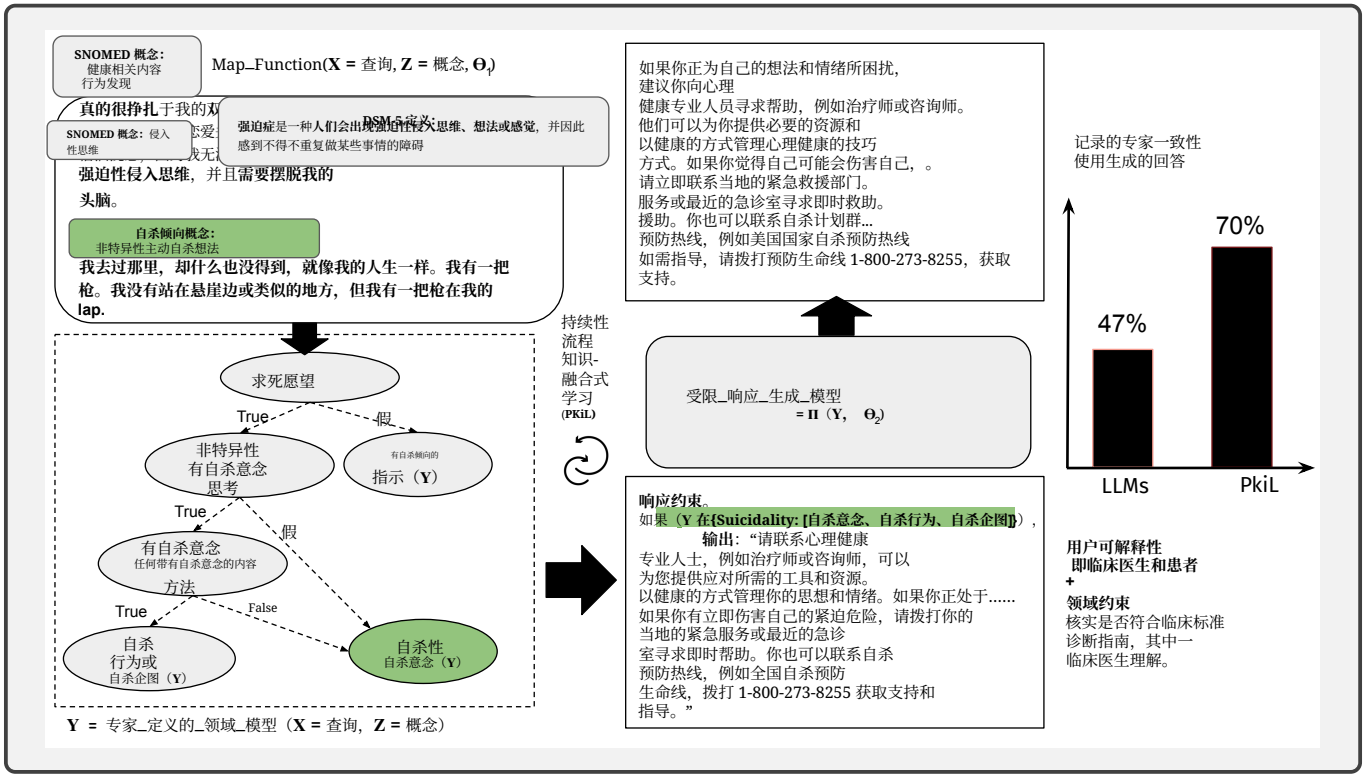


图 4 展示了一条端到端完全可微的流水线。它由对应于各种流水线组件的函数组合而成。该流水线支持开发针对特定应用量身定制、可轻松进行端到端训练的人工智能系统。为此，将可训练映射函数应用于原始数据，将其转换为领域模型中的概念。图中给出的示例涉及心理健康诊断和对话辅助。映射函数将原始数据片段关联到诊断模型中的决策变量，然后利用这些变量对文本生成模型生成的患者的回应施加约束。现有实现的结果表明，使用此类流水线时，专家满意度达到 70%；相比之下，在联邦式流水线中使用大语言模型（例如 OpenAI 的 text-Davinci-003 [10]。

在系统流水线中的组件。值得注意的是，在图4所示的一个已实现系统中，针对心理健康诊断辅助用例，我们看到专家对系统响应的满意度大幅提升，这进一步展现了 2(b) 类方法的巨大潜力。

III. 神经符号人工智能的未来

本文比较了不同的神经符号架构，考察了其算法层面的特性（包括感知和认知）以及应用层面的特性，例如用户可解释性、领域约束规范、可扩展性和对持续学习的支持。语言模型的快速进步表明，它们将在大规模感知方面达到近乎最优的性能水平。知识图谱适用于连接认知与感知层面的符号结构，因为它们支持现实世界的动态性。与一阶逻辑等静态且脆弱的符号逻辑不同，知识图谱易于更新。除了适用于企业用例以及采用成熟的可移植性标准外，知识图谱还属于一个成熟的算法生态系统，该生态系统能够实现高效的图管理和查询。这种可扩展性使人们能够对包含数百万乃至数十亿个节点的大型复杂数据集进行建模。

总而言之，本文强调了在当前实现中结合语言模型与知识图谱的有效性。然而，本文也指出，未来的知识图谱有潜力对超越模式的异构应用类型和领域级知识进行建模。这包括工作流、约束规范和流程结构，从而进一步增强神经符号架构的能力和实用性。将此类增强型知识图谱与高容量神经网络相结合，将为终端用户提供极高程度的算法和应用层面效用。出于对安全性的担忧，近期人们推动暂停进一步部署 GPT* 等生成式人工智能系统，因为在没有额外护栏的情况下，当前系统可能会对个人和社会造成重大危害。我们相信，如图4 所示，可以通过知识图谱的扩展形式对指南、政策和法规进行编码（因而采用符号手段），进而提供可解释性、问责性、严格的审计能力和安全性。令人鼓舞的是，所有这些方面都在迅速取得进展，未来前景十分光明。

致谢

本研究部分由美国国家科学基金会资助，资助编号 2133842，“EAGER：推进

深度知识融合学习与神经符号人工智能。”

作者

Amit Sheth 是南卡罗来纳州人工智能研究所（AIISC）的创始主任、NCR讲席教授，也是南加州大学计算机科学与工程教授。他荣获2023年 IEEE-CS Wallace McDowell奖，并且是 IEEE、AAAI、AAIA、AAAS 和 ACM 会士。联系邮箱：amit@sc.edu

Kaushik Roy 是一名博士生，在本文相关领域发表过多篇论文。联系方式：kaushikr@email.sc.edu

Manas Gaur 是马里兰大学巴尔的摩县分校的助理教授。他的博士论文研究知识融合学习，目前的研究重点是本文所讨论系统的可解释性和安全性。他曾获EPSRC-UKRI奖学金、社会公益数据科学奖学金和社会公益人工智能奖学金，并于近期获评2023年AAAI新晋教员。联系方式：manas@umbc.edu

参考文献

[1] D. Kahneman，《思考，快与慢》，*Farrar, Straus and Giroux*，2011年。[2] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. ˇZ ́ıdek, A. Potapenko 等，“借助 AlphaFold 实现高精度蛋白质结构预测”，《自然》，第596卷，第7873期，583–589页，2021年。[3] A. Fawzi, M. Balog, A. Huang, T. Hubert, B. Romera-Paredes, M. Barekatain, A. Novikov, F. J. R Ruiz, J. Schrittwieser, G. Swirszcz 等，“利用强化学习发现更快的矩阵乘法算法”，《自然》，第610卷，第7930期，47–53页，2022年。[4] D. Gentner，“结构映射：类比的理论框架”，认知科学，第7卷，第2期，155–170页，1983年。[5] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Ka-mar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg 等，“通用人工智能的曙光：GPT-4的早期实验”，*arXiv*预印本 *arXiv:2303.12712*，2023年。[6] A. Sheth, M. Gaur, K. Roy, R. Venkataraman, 和 V. Khandelwal，“过程知识融合人工智能：迈向用户层面的可解释性、可理解性与安全性”，《*IEEE*互联网计算》，第26卷，第5期，76–84页，2022年。[7] A. d. Garcez 和 L. C. Lamb，“神经符号人工智能：第三次浪潮”，《人工智能评论》，第1–20页，2023年。[8] V. Rawte, M. Chakraborty, K. Roy, M. Gaur, K. Faldu, P. Kikani, H. Akbari, 和 A. P. Sheth，“TDLR：自上而下的语义—自下而上的句法语言表示”，载于*NeurIPS'22*关于注意力的各个方面：弥合注意力不同视角的研讨会。[9] R. Wang, D. Tang, N. Duan, Z. Wei, X. Huang, G. Cao, D. Jiang, M. Zhou 等，“K-Adapter：借助适配器将知识融入预训练模型”，*arXiv*预印本 *arXiv:2002.01808*，2020年。[10] K. Roy, M. Gaur, Q. Zhang, 和 A. Sheth，“用于社交媒体自杀倾向评估的过程知识融合学习”，*arXiv*预印本 *arXiv:2204.12560*，2022年。