评人工智能如何走向新阶段？

打破机器学习黑盒子实现可解释的人工智能

陆首群

2021.5.16

国内外跟贴留言

（705-779）

第十集

2021.9.17

中国开源软件推进联盟
China OSS Promotion Union

评人工智能如何走向新阶段?

打破机器学习黑盒子实现可解释人工智能

陆首群

2021.8.25

机器学习/深度学习是一种强大的数据分析工具，是弱人工智能的代表。但机器学习/深度学习也是有缺陷的，它本质上是一项暗箱技术或盲模型，其训练过程不可解释。图灵奖得主、贝叶斯网络之父 Judea Pear 早在 2018 年就指出，当前机器学习理论有局限性，完全以统计或盲模型（即黑盒子）的方式运行，所以不能成为强人工智能的基础；人工智能大师 Yoshua Beno 谈到，近年来，以深度学习算法为代表的人工智能技术快速发展，在计算机视觉、语音识别、语义理解等领域实现了突破，但其算法并不完美，有待继续加强理论研究；图灵奖得主、算法大师约翰-霍普克罗夫在 2019 年指出，对深度学习这个黑盒子，人们知道它在学习，但不知怎么学习，人类可能会在 5 年后大体读出深度学习的数学理论；COPU 于 2020 年 6 月指出，机器学习/深度学习必须克服其自身的缺陷，打破黑盒子痼疾，实现可解释的人工智能，建立可解释的机器学习模型。

如今，可解释性人工智能（XAI）已成为全球人工智能研发的亮点。

早在 2019 年 8 月，COPU 就提出研发 XAI 的任务，这在国内是最早提出的，在全球也是最早提出这个任务的少数先行者。在国内，2020 年 12 月，沈向洋教授提出：“拥抱开源，我们现在最重要的事

情是要做可解释的人工智能”。2021 年 1 月，姚期智院士提出：“机器学习算法缺乏可解释性，很多算法处于黑盒子状态，这项人工智能的技术瓶颈亟待突破”。

2020 年 6 月，COPU 主办《第 15 届开源中国开源世界高峰论坛》，邀请 IBM 副总裁 Todd Moore 在会上作“可信任人工智能（反欺诈、可解释、公平性）”的报告，从此至今，COPU 已收到全球研发可解释性人工智能的跟帖 50 多件。但由于全球人工智能技术（XAI）尚未完全成熟，在研发 XAI 算法时，专家对各道演算程序的理解和操作具有不确定性，最后评估还只能靠人工，所以 XAI 演算结果或算法可能有出入，致使可解释机器学习难以推广应用。为此，COPU 要求 IBM 人工智能研究所举出具体案例进行解析和说明，对我们提出的 8 个问题进行逐个解答：

①请 IBM 列出研发 XAI 的具体案例

②选用下列哪种方法进入运算？

可直接解释（内在解释）

事后解释

全局（模型级）可解释性

局部（实例级）可解释性

③选择什么工具？

如：决策树、规划库、抉择表等

④如何捕捉特征？

⑤如何建模？

⑥如何找到算法？

⑦如何进行评估？

⑧不但要导出本案例结果，还要使 XAI 在使用中确定是否能保持信任、公正、透明和可解释？！

在本集（第十集）人工智能跟帖 765 条中，IBM 的人工智能专家按我们提出的 8 个问题对具体案例作出了答复。

国内外 AI 跟帖留言（705-779)

705，在医学领域人类决策支持系统的可解释人工智能

瑞典皇家理工学院、芬兰阿尔托大学 Rohit Saluja，Samanta Knapic 等

2021.5.5

随着人工智能被应用于对人类有重大影响的更敏感的环境中，人工智能的可解释性就变得非常重要，人类必须能驹实时理解、再现和操纵机器决策过程。因此，人们越来越需要提高机器学习算法决策的可理解性，这些决策可以在实际应用中复制，特别是在医院领域。这就需要一个系统，允许直接地、可理解地和可解释地做决策。可解释人工智能（XAI）有助于促进人工智能和机器学习在医学领域的应用，尤其有助于提高透明度和信任度。

本文提出了在医学图像分析领域对决策支持的人工智能方法进行实验，分别是两种事后可解释机器学习方法 LIME 和 SHAP，以及以语境价值和效用（ClU）为中心的解释方法 CIU。作者基于 LIME、SHAP 和 CIU 提供的解释进行了三次用户研究，来自不同非医学背景的用户在基于网络的调查环境中进行了一系列测试，并陈述了他们对给定解释的经验和理解。CIU 可解释方法在增加对人类决策的支持以及透明性和用户理解性方面比 LIME 和 SHAP 方法表现得更好。此外，CIU 比 LIME 和 SHAP 生成解释的速度更快。研究表明，在不同的解释支持环境 F，人类的决策存在显著差异。在此基础上，作者提出了三种可能的解释方法，这些方法可以在不同的医学数据集上推广，并为医学专家提供很好的决策支持。

706，用于欺诈检测的可解释机器学习

剑桥大学 Ismini Psychoula 等 2021.5.13

应用机器学习来处理大型数据集在许多行业都有潜力，包括金融服务。然而，那些完全采用机器学习的实际问题，仍然集中在理解并能够解释复杂模型做出的决策和预测上。在本文中，我们通过研究在有监督和无监督模型上选择合适的背景数据集和运行时权衡，来探索实时欺诈检测领域中的可解释方法。

707，2021 年神经形态计算与工程路线图

丹麦科技大学，亚琛科技大学，2021.5.12

如今，基于冯-诺依曼架构的现代计算已成为一门成熟的前沿科学。在这种体系结构中，处理和存储单元被分为单独的模块，可以密集且连续地交换数据。这种数据传输造成了很大一部分功耗。下一代计算机技术有望解决这种超大规模的计算问题，即使这些未来的计算机具有令人难以置信的强大功能，但是如果它们基于冯-诺依曼类型的体系结构，它们将消耗 20 到 30 兆瓦的功率，并且不会像我们的大脑那样具有内在的物理内置能力来学习或处理复杂的非结构化数据。神经形态计算系统旨在满足这些需求。人脑使用 20W 和 1.2L 的体积每秒执行约 10～15 次计算。通过从 42 物学中获得启发，新一代计算机的动耗可以比传统处理器低得多，可以利用集成的非易失性存储器和逻辑，并且可以明确地设计为在复杂和非结构化数据的情况下支持动态学习。在其潜在的未来应用中，商业、医疗保障、疾病和病毒的传播控制在社会层面可能是最有影响力的。该路线图设想了神经形态材料

在前沿技术中的潜在应用，并将重点放在人工神经系统的设计和制造上。该路线图的内容将突出这一活动的跨学科性质，该活动的灵感来自生物学，它从生物学、物理学、数学、计算机科学和工程学中获得灵感。这将为在许多技术相关领域探索和整合当前和未来应用背后的新技术提供路线图。

708，机器学习／深度学习可解释性算法尚未完全成熟有待完善

COPU，2021.5.19

打破机器学习中的黑盒子，实现可解释性人工智能（Explainable Artificial Intelligence，XAI），或从弱人工智能嬗变为强人工智能，已成为当今中外人工智能研发的亮点。

人工智能中外跟帖迄今已汇集发表 704 条，其中 40 条跟帖属于从理论上探讨机器学习／深度学习可解释性算法，以及从应用实践上体验可解释性算法的解决方案。可是，机器学习／深度学习可解释性算法尚未完全成熟，误差较大，作为最后评估有时结论分散。

下面谈一下机器学习／深度学习可解释性算法的演算步骤：

1）审题：根据不同项目的不同对象，选择不同的可解释性解决方案（或区分有待求解的可解释人工智能案例，不同的解释行为采用不同的求解方法）。在选择不同的 XAI 方案时可从如下排列组合中选择：分内在可解释性与事后可解释性两种情况，进一步还可分全局可解释性与局部可解释性两种情况。

2）确定／列出：可解释性算法分前提（条件）与任务的预期目标（结果）两方面，为了做好 XAI 求解，其前提条件最好能演译为（或确定）语义（网络）。有人说，没有语义就不能有任何解释，似乎有点夸大。

3）分析（演算步骤）：

①判断某项目求解的前提与结果是否为因果关系，

②收集数据、演译语义、捕捉特征、研究算法，

③纪计建模，

④建立 XAI 解决方案，显示结论，

⑤进行评估。

实际上机器学习／深度学习的 XAI 尚未完全成熟，有待完善。

1）在审题时，选择不同的 XAI 方案的背景中（4 种排列组合）是粗线条的，有相当大的不确定性；

2）在演算步骤分析中，所谓没有语义就不能有任何解释（或不解决语义稀缺性将无法解释深度神经网络），似乎过于夸张；对于如何理解解释，有很多不同观点，如有人提出：仅据技术观点进行解释不够，要考虑心理学、社会科学等因素，有人提出用反事实解释来打破黑盒子，有人提出为建立可解释系统，决策者必须具有合理的信心等；对于最终的评估结论，其分散性和可信度均有待推敲！

709，XAI 手册：面向可解释人工智能的统一框架

德国人工智能研究中心（DFKI-G mbH）2021.5.14

可解释人工智能（XAI）领域已迅速成为一个蓬勃发展和多产的社区。然而这一领域有一个反复出现和公认的问题是对其术语缺乏共识。如“planation”和“interpretation”。本文提出了一个理论框架，不仅为这些术语提供了具体的定义，而且还概述了产生“explanation”和“interpretation”所需的所有步骤，本文证明了该框架在可解释性、可理解性和评价指标上符合要求。

本文对“explanation”和“interpretation”的定义如下：解释（explanation），指解释的任务，是描述（或解释）一个或多个事实的过程，而解释（interpretation），指解释的方法，是用以阐述解释（explanation）的含义，为了使解释（explanation）便于理解。

⊙大多数研发者侧重于解释方法的研究和使用，而忽略了解释任务的作用。

⊙解释通常没有明确定义域和共域（即 realm of the explanasand explanadum）。

在人工智能接管或协助人类专业工作者的领域（例如医疗应用），人们的兴趣通常从预测转到解释，以便人类专家学习并获得有关任务的新见解。

710，评价可解释人工智能分类算法的正确性

Swansea University 2021.5.20

近年来，可解释的 AI 通过特征归因算法吸引了许多研究者的关注，这些特征归因算法计算了预测中的“特征重要性”，因此变得越来越

流行。但是对这些算法的有效分析很少。本文开发了一种通过创建具有已知解释基础事实的数据集来定量评估 XAI 算法正确性的方法，在实验中使用两种流行的特征归因解释器，即局部可解释模型不可知性解释（LIME）和 SHapley 可加性析构法（SHAP）。

关于解释性准确度，①分类准确度与解释性准确度呈正相关，②SHAP 比 LIME 提供更准确的解释，③解释准确性与数据集复杂度负相关。

711，系统软件在神经拟态计算能耗管理中的作用

德雷克塞尔大学 2021.3.22

最近，很多计算社区引入了将诸如 DYNAP 和 Loihi 之类的神经拟态计算系统以提高机器学习程序的性能和能效，尤其是那些使用 Spiking Neural Network（SNN）实现的程序。用于神经拟态系统的系统软件的作用是将大型机器学习模型（例如具有许多神经元和突触）聚类并将这些聚类映射到硬件的计算资源。本文考虑了神经元和突触消耗的动率，以及在互连上传递尖峰时所消耗的能量，从而制定了神经拟态硬件的能量消耗。文中首先评估了系统软件在管理神经拟态系统能量消耗中的作用。作者制定了一种简单的映射方法，将神经元和突触放置到计算资源上以减少能耗。最终文本用 10 个机器学习应用程序评估了提出的方法，并证明了所提出的映射方法可显著减少神经拟态计算系统的能耗。

712，稀疏脉冲梯度下降

帝国理工学院 2021.5.18

由于神经形态计算设备的能耗低，人们越来越喜欢在其上模拟脉冲神经网络（SNN）。最近的进展使训练 SNN 的准确性达到了与传统人工神经网络（ANN）竞争的地步。同时在神经形态硬件上运行时具有能源效率。但是，训练 SNN 的过程仍然基于最初为 ANN 开发的密集张量操作，该操作不利用 SNN 的时空稀疏性质。我们在这里介绍了第一种稀疏 SNN 反向传播算法，该算法可实现与当前技术水平相同或更高的准确性，同时显著提高了速度并提高了存储效率。展示我们的方法在不同复杂度的真实数据集（MNIST，神经语言 MNIST 和 Spiking Heldelberg Digits）上的有效性，该方法在不减少精度的情况下实现了高达 70 倍的后向加速并提高 40%的内存效率。

713，QuatDE：用于知识图谱补全的动态四元数嵌入

电子科技大学 2021.5.19

近年来科研工作者对知识图谱完成方法，尤其是通过图嵌入方法学习实体和关系的低维表示来预测缺失的事实进行了广泛的研究。这些模型通常将关系向量视为实体对之间的平移（TransE）或旋转（rotatE 和 QuatE），从而享有简单性和效率优势。但是 QuatE 存在两个主要问题：①用于捕获实体与关系之间的表示以及特征交互的能力的模型相对较弱，②尽管模型可以处理各种关系模式，包括对称、反对称、反演和组合，但是不考虑关系的映射属性，例如一对多，多对一和多对多。本文提出了一种新颖的模型 QuatDE，具有动态映射策略，可以显式捕获多种关系模式，从而增强三元组元素之间的特征交互能力。

实验表明，QuatDE 在三个完善的知识图完成基准上均达到了先进的性能。MR 评估在 WN18 上相对增加了 26%，在 WN18RR 上增加了 15%，这证明了 QuatDE 的普遍性。

714，唤赵触角的脑机接口改善机械臂的控制

匹兹堡大学生物工程师的研究团队 2021.5.20 发表在 《Science》杂志上的一篇论文，描述如何通过增强一个四肢瘫痪病人的大脑刺激，来唤起触觉，从而使操作者更易通过大脑控制的机械臂，缩短操作时效。

715，多模态深层神经网络可解释性研究综述

Symbiosis Institute of Technology，2021.5.18

由深度神经网络提供支持的人工智能技术已在多个应用领域中取得了很大的成功，尤其是在计算机视觉应用程序和自然语言处理任务中。超越人类水平的表现推动了 AI 应用的研究，其中语言、视觉、感官、文本等不同模态在准确预测和识别中起着重要的作用。本文提出了几种采用深度学习模型的多模态融合方法。尽管它们具有出色的性能，但深层神经网络的复杂、不透明和黑盒性质限制了它们的社会认可度和可用性。这引发了对模型可解释性的探索，尤其是在涉及多模态人工智能方法的复杂任务中。本文对现有文献进行了广泛的回顾，以对多模态深度神经网络的可解释性，尤其是视觉和语言任务的可解释性进行全面的调查和评论。本文涵盖了多模态人工智能及其在通用领域

的应用的几个主题，包括该领域的数据集、方法和技术的基本构件、挑战、应用和未来趋势。

716，具有超分辨率忆阻器交叉开关的模拟神经计算

IEEEA.P.J ames，L.O.Chia，2021.5.10

忆阻器交叉开关阵列广泛用于内存中和神经形态计算应用，但是，忆阻器器件存在非理想状态，会让导电状态发生变化，因此随着器件的老化，将其编程为所需的模拟电导的一组稳定电导值。交叉开关中可用于节点的电导级别的数量被定义为交叉开关的分辨率。

本文提出了一种通过建立超分辨率忆阻器交叉开关来提高分辨率的技术，该超分辨率忆阻器交叉开关的节点具有多个忆阻器，以生成具有唯一电导值的 r-单纯形序列。电导值的范围和数量越宽，交叉开关的分辨率越高。这在构建模拟神经网络（ANN）层时特别有用，事实证明，这被证明是实现神经形态计算中形成神经网络层的一种可行方法。

717，从自然语言文本中提取因果关系的调查，2021.1.16

悉尼大学 Jie Yang，Soyeon Caren，Josiah Poon

作为人类认知的重要组成部分，因果关系频繁出现在文本中，从文本中提取因果关系有助于为预测任务建立因果网络。现有的因果关系提取技术包括基于知识的、基于统计机器学习的和基于深度学习的方法。每种方法都有其优缺点。例如基于知识的方法是可以理解的，但是需要大量的手动领域知识，并且具有较差的跨领域适用性。由于自然语言处理（NLP）工具包，统计机器学习方法更加自动化。然而，特征

工程是劳动密集型的，工具包可能导致错误传播。近年来，深度学习技术因其强大的表征学习能力和计算资源的快速增长而引起了自然语言处理研究者的极大关注。它们的局限性包括高计算成本和缺乏足够的带注释的训练数据。本文对因果关系抽取进行了全面的综述。我们首先介绍了因果提取中存在的主要形式：显性句内因果关系、隐性因果关系和句间因果关系。接下来，我们列出用于因果关系提取的基准数据集和建模评估方法。然后，给出了这三种技术及其代表系统的结构化概述。最后，我们重点介绍了现有的开放挑战及其潜在方向。

718，用于可解释性推荐的个性化

Lei Li，Yongfeng Zhang，Li Chen，香港浸会大学，2021.5.25

自然语言生成的个性化在可解释性推荐、评论、摘要和对话系统等一系列任务中起至关重要的作用。在这些任务中，用户和项目 ID 是个性化的重要标识符。Transformer 具有很强烈语言建模能力，但由于 ID 标记与单词不在同一语义空间中 Transformer 没有个性化，不能充分利用用户和项目的 ID。针对这一问题，本文提出了一种个性化的可解释推荐变换器（PETER），并在此基础上设计了一个简单有效的学习目标，利用 ID 对目标解释中的词语进行预测，从而赋予 ID 语言意义，实现个性化 Transformer。除了生成解释，PETER 还可以进行推荐，这使得它成为整个推荐解释管道的统一模型。大量实验表明，我们的小规模无训练模型在生成任务上的有效性和效率都优于微调的 BERT，这突出了我们设计的重要性和良好的实用性。

719，扫视视觉如何有助于深层网络的可解释性

Research and Education Center, Mathematics of Future Technologies, Nizhing Novgorod, Russia

本文描述了现代深层网络的一些问题（可解释性，缺乏面向对象性）是如何通过采用一种生物学上合理的感知扫视机器来解决的，提出了这样一个扫视视觉模型的草图。实验结果证明了该方法的有效性。

720，光子神经网络的前景

乔治华盛顿大学，2021.5.20

尽管神经网络已经以机器学习和神经拟态计算的方式在人工智能领域中得到了广泛的应用，然而神经网络软件在具有独立的内存和处理器（并顺序运行）的常规计算机上的应用仍然在很大程度上受到速度和能源效率的限制。神经拟态工程旨在模仿大脑中的神经元和突触构建能够进行分布式和并行处理的硬件处理器。本文证明由光子学（光学物理学）支持的神经拟态工程可以提供亚纳秒级的延迟和低能量的高带宽，从而将人工智能和神经拟态计算应用的领域扩展到机器学习加速，非线性编程，智能信号处理等方面，同时进行了光子神经网络根据目标应用的类别在集成平台和自由空间光学器件上的运行演示。

721，可解释多跳科学问答的动态语义图构建与推理

WeiWen Xu, Huihui Zhang, DengCai, Wai Lam,香港中文大学

知识检索和推理是 Web 范围的多跳问答（QA）的两个关键阶段。当检索证据事实以填补知识空白时，现有方法的置信度不足并缺乏透明的推理过程。

在本文中，作者提出了一个新的框架，该框架可通过动态构建语义图并对其进行推理，从而在利用更多有效事实的同时获得多跳质量检查的可解释性。作者采用抽象含义表示（AMR）作为语义图表示。该框架包含三个新想法：①AMR-SG，一种基于 AMR 的语义图，由候选事实 AMR 构造以揭示问题，答案和多个事实之间的任何跃点关系。②一种新颖的基于路径的事实分析方法，利用 AMR-SG 从大型事实库中提取活动事实以回答问题。③利用图卷积网络（GCN）进行事实级别的关系建模以指导推理过程。在两个科学的多跳 QA 数据集上的结果表明，本文方法可以超越最近的方法，包括使用其他知识图谱的方法，AMR-SG 可以保持较高的解释性，并且可以成功地与强大的预训练模型相结合，从而实现对 OpenB ookQA 和 ARC-Challenge 的显著改进，而不是利用额外的 KG。

722，时态神经网络在线学习的微体系结构实现框架

Harideep Nair，John Paul Shen，James E. Smith,卡内基梅隆大学，2021.5.27

时态神经网络（TNN）是含有脉冲的神经网络，类似于哺乳动物的新皮层，它使用时间作为资源来表示和处理信息。与采用单独的训练和推理阶段的计算密集型深度神经网络相反，TNN 具有极高效率的在线增量／连续学习能力，并且是构建边缘本机感觉处理单元的极佳候选

者。这项工作提出了使用标准 CMOS 来实现 TNN 的微体系结构框架。提出了三个关键构建块的门级实现：①多突触神经元，②多神经元列，以及③基于峰值定时依赖可塑性（STDP）的无监督和监督在线学习算法。TNN 微体系结构体现在一组特性缩放方程式中，可用于评估任何 TNN 设计的门数、面积、延迟和功耗。论文提出了拟议设计的合成后果（在 45nmCMOS 中），并证明了其在线增量学习能力。

723，对话图：将可解释的策略图网络融入谈判对话

卡内基梅隆大学语言技术研究所，2021.6.2

有说服力的谈判策略的务实规划是必不可少的。虽然现代对话代理擅长生成流利的句子，但他们仍然缺乏语用基础，无法进行战略推理。本文提出 Dialo Graph，这是一个谈判系统，它使用图神经网络在谈判对话中结合实用策略。给定对话上下文，Dialo Graph 明确地结合策略序列之间的依赖关系，以实现对下一个最佳策略的改进和可解释的预测。本文基于图的方法在策略/对话行为预测的准确性和下游对话响应生成的质量方面都优于先前最先进的协商模型。本文定性展示了学习策略图在对话过程中提供有效谈判策略之间的明确关联方面的进一步好处，从而导致可解释的战略对话。

724，使用尖峰神经网络对基于事件的光流进行自监督学习

代尔夫特理工大学 Federico Paredes-Valles 等

神经形态传感和计算有望实现高能效和高带宽传感器处理。神经形态计算的一个主要挑战是，由于离散脉冲和复杂的神经元动力学，传统

人工神经网络（ANN）的学习方法不能直接转移到脉冲神经网络（SNN）。因此，SNN 尚未成功应用于复杂的大规模任务。在本文中，我们专注于从基于事件的相机输入进行光流估计的自监督学习问题，并研究最先进的 ANN 训练管道所需的变化，以便成功地转换为 SNN。我们首先修改输入事件表示，以便用最少的显式时间信息对更小的时间片进行编码。因此，使网络的神经元动力学和循环连接随着时间的推移能够整合信息。此外，我们重新制定了基于事件的光流的自监督损失函数以改善其特性。我们使用提议的管道对各种模型的循环神经网络和 SNN 进行实验。关于 SNN，我们研究了参数初始化和优化、替代梯度形状和自适应神经元机制等元素的影响。结果发现初始化和代理梯度宽度在实现稀疏输入学习方面起着至关重要的作用，而包含自适应性和可学习的神经元参数可以提高性能。实验表明，所提出的 ANN 和 SNN 的性能与以自我监督方式训练的当前最先进的 ANN 的性能相当。

725，探索神经拟态学习规则的库

芝加哥大学 RickStevens 等，2021.5.1

神经拟态计算领域正处于积极探索的时期。虽然已经开发出了许多模拟神经元动力学、将深度神经网络转换为 SNN 模型的工具，但用于神经拟态学习的规则的通用软件库仍未得到充分的探索与开发。为神经拟态网络设计新的学习规则库具有非常大的挑战性，其范围从编码方法到梯度近似，从模仿贝叶斯大脑的群体方法到部署在忆阻器交叉开

关上的约束学习方法。为了填补这一方面的空自，本文提出了一个模块化、可扩展的库 NeKo，专注于帮助设计新的学习算法，并在三个示例案例中展示了 NeKo 的效用，包括在线本地学习、概率学习和模拟设备上的学习。实验结果表明，NeKo 可以复制最先进的算法，并且其在准确性和速度方面的表现明显优于其他算法。此外它还提供了包括梯度比较在内的工具，可以帮助开发新的算法变体。NeKo 是一个开源 Python 库，支持 PyTorch 和 TensorFlow 后端。

726，CoRI：用于开放信息提取的数据增强的集体关系集成

ZhengbaoJiang 等，2021.6.1 卡内基梅隆大学语言技术研究所/亚马逊

将从 Web 中提取的知识集成到知识图谱（KG）中可以促进诸如问答之类的任务。本文作者研究了关系整合，旨在将主题-关系-客体提取中的自由文本关系与目标知识图谱中的关系对齐。为了解决自由文本关系不明确的问题，以前的方法利用相邻实体和关系获得额外的上下文。但是，这些预测是独立进行的，可能相互不一致。本文作者提出了一个两阶段的集体关系集成（CoRI）模型，其中第一阶段独立进行候选预测，第二阶段采用仿问所有候选预测的集体模型来进行全面一致的预测。本文作者使用来自目标知识图谱部分未使用的增强数据进一步改进了集体模型。在两个数据集上的实验结果表明，CoRI 可以显著优于基线，将 AUC 分别从 0.677 提高到 0.748 和从 0.716 提高到 0.780。

727，端到端 NLP 知识图谱构建

微软印度研究院/IBM 欧洲研究院，2021.6.2

本文研究了从科学论文中端到端构建 NLP 知识图（KG）。作者们专注于提取四种类型的关系：任务和数据集之间的评估，任务和评估指标之间的评估，以及同类型实体之间的相关和相关关系。例如“F1 分数”与“F-measure”的互相指代。作者们为这些关系类型中的每一种引入了新方法，并将作者们最终框架（SciNLP-KG）应用于来自 ACL Anthology 的 30，000 篇 NLP 论文，并构建一个大规模的 KG，这有助于为 NLP 社区自动构建科学排行榜。作者们的实验结果表明，生成的 KG 包含高质量的信息。

728，Bounded Logitattention：学习解释图像分类器

Polten 大学，2021.5.31

可解释人工智能是试图通过称为“解释”的适当的辅助信息来阐明过于复杂而无法直接被人类认知访问的系统的工作原理。我们为卷积图像分类器提供了一个可训练的解释模块，称之为 BLA（Bounded Logitattention:）。BLA 克服了实例特征选择方法“学习解释”（L2X）的几个限制：BLA 可扩展到现实世界大小的图像分类问题；BLA 提供了学习可变大小解释的规范方法。由于其模块化，BLA 适合迁移学习设置，也可以用作训练分类器的事后附加组件。在用户调研中，我们发现 BLA 解释比流行的（Grad-）CAM 方法生成的解释更受欢迎。

729，MERLOT：多模态神经脚本知识模型

华盛顿大学 Allen 人工智能研究所，2021.6.4

人类观察事件的视觉上下文，进行跨时间的多模态推理，对过去、现在和未来做出推断，进而理解世界。

本文提出 MERLOT，通过观看数以百万计的 YouTube 视频和转录语音，来学习多模态脚本知识的模型——以一种完全无标记、自监督的方式。通过对帧级（空间）和视频级（时间）目标的混合预训练，模型不仅学会了将图像与时间上对应的词语相匹配，还学会将全局发生的事件与时间相联系，表现出强大的开箱即用的时间常识表示，微调后在 12 个不同的视频问答保证数据集上实现了最先进的性能，能很好地迁移到静态图像世界，允许模型推理视觉场景背后的动态上下文。在视觉共感推理中，MERLOT 以 80.6%的准确率回答问题，比类似规模的先进模型高出 3%以上，即使是那些大量使用辅助监督数据（如物体边框）的模型。消融分析显示了以下的互补重要性：在视频与静态图像上进行训练；扩大预训练视频语料库的规模和多样性；使用不同的目标，鼓励从识别到认知层面的全栈多模态推理。

730，通过对抗性正则化降低网络对参数噪声的敏感度

Julian Bucher，苏黎世大学

神经拟态网络处理器有望在基于 NN 的 ML 任务的计算密度和能源效率方面取得巨大优势。然而，由于工艺变化和固有的器件物理特性，这些技术容易出现不理想的情况。部署的模型存在的参数噪声会降低部署到处理器的网络的任务性能。虽然可以校准每一个设备，或为每个

处理器单独训练网络，但这些方法成本高昂且不适合商业部署。因此由于网络架构和参数的原因，需要寻找替代的方法来训练对参数变化具有内在鲁棒性的网络。

本文提出了一种新的对抗网络优化算法，该算法在训练过程中攻击网络参数，并在面对参数变化的推理过程中提升鲁棒性能。本文的方法引入了惩罚网络对权重扰动心敏感性作为正则化项。并与之前在训练过程中降低参数敏感性的方法进行比较，例如 dropout、权重平滑和引入参数噪声。实验证明，本文的方法所产生的模型对目标参数变化更鲁棒，对随机参数变化同样稳健。与其他方法相比，此方法在实验场景中更平坦的位置找到最小值，突出表现该技术发现的网络对参数扰动不太敏感。同时，本文还提供了一种神经网络架构部署计算非理想性影响的设备的方法，以期将性能损失降至最低。

731，脑接口：人类大脑利用意念控制老鼠走迷宫

浙江大学 ScientificReports，2021.6.11

使用人脑意念控制外部装置的研究，比如控制假肢等，已经有很多了，但关于用人脑意念控制动物的研究还不太多。浙大研究人员发表在《Scientific Reports》杂志上的一篇论文，证明了直接从人脑发送的无线输入来控制“ratCyborgs”的能力。研究人员在他们的论文中将受人脑控制的老鼠称为“ratCyborgs”，成功以人脑意念操控实验鼠的动作，并引导实验鼠穿越复杂的迷宫。

732，腾讯与厦门大学发布《2021 十大人工智能趋势》，田忠

腾讯优图实验室联合厦门大学人工智能研究院在 6 月 5 日召开的“2021 全球人工智能技术大会”上正式发布《2021 十大人工智能趋势》https://mp.weixin.qq.com/s/ZgoK3DXgG1lJBp4E6IMEjQ。该趋势报告基于腾讯优图和厦门大学人工智能研究院长期对人工智能尤其是计算机视觉的研究洞察，提出 3D 视觉技术、数字内容产业、AI 深度学习算法、人工智能内核芯片等方向的前沿预测。和 COPU 陆主席一直在倡导的可解释人工智能，这份趋势报告将自动机器学习的自动化和可解释程度的提升作为首要趋势，看好机器学习的大众化。无监督、弱监督学习有效降低深度学习成本。3D 视觉技术迅速拟合虚实世界，形成新的产业机会。定制型人工智能内核芯片将逐渐演变为通用型人工智能内核芯片，推动人工智能内核芯片实现真正的落地。人工智能的治理正成为全球热门的议题，算法的公平性正是人工智能治理的关键问题。中国、欧盟相继推出有关指引推动普惠无偏见的人工智能。同时，隐私保护、智能安全也日益成为人工智能。

733，透明度图灵测试

Felik Biessmann，Vikor Treu，2021.6.21

可解释人工智能（XAI）的一个中心目标是改善人工智能交互中的信任关系。透明人工智能系统研究的一个假设是，解释有助于更好地评估机器学习（ML）模型的预测，例如，使人类能够更有效地识别错误预测。然而，最近的经验证据表明，解释可能产生相反的效果：当提

出 ML 预测的解释时，人类往往倾向于相信 ML 预测，即使这些预测是错误的。实验证据表明，这种效应可以归因于人工智能或解释的直观程度。这种效应挑战了 XAI 的目标，这意味着负责任地使用透明 AI 方法必须考虑到人类区分机器生成和人类解释：的能力。本文提出了一种基于图灵模拟博弈的 XAI 方法的量化度量，即图灵透明度检验。询问者被要求判断一个解释是由人还是由 XAI 方法生成的。在这二进制分类任务中，人类无法检测到 XAI 方法的解释通过了测试。检测这样的解释是评估和校准人工智能交互中的信任关系的一个必要条件。我们在一个众包文本分类任务上给出了实验结果，证明即使对于基本的 ML 模型和 XAI 方法，大多数参与者也无法区分人类和机器生成的解释。我们讨论了我们的结果对透明 ML 的应用的伦理和实践意义。

734，人工解释的多样性和局限性

ChenhaoTan，芝加哥大学，2021.6.22

NLP 越来越多的努力旨在构建人工解释的数据集。然而，解释一词包含了广泛的概念，每个概念都有不同的属性和后果。本文的目标是提供一个不同类型的解释和人工局限性的概述，并讨论在 NLP 中收集和使用解释的意义。受心理学和认知科学先前工作的启发，将 NLP 中现有的人工解释分为三类：近似机制、证据和程序。这三类性质不同，并且对由此产生的解释有影响。例如，程序在心理学中不被视为解释，而是与从指令中学习的大量工作相联系。解释的多样性进一步体现在代理问题上，这些代理问题是标记者解释和回答开放式的为什么问题

所需要的。最后解释可能需要与预测不同的，通常是更深层次的理解，这让人怀疑人工是否能在某些任务中提供有用的解释。

735，基于显著性的 XAI 方法众包评估

Xiaotian Lu 等，京都大学，2021.6.27

理解深层神经网络预测背后的原因对于在许多重要应用中获得人们的信任至关重要，这反映在近年来对人工智能（XAI）可解释性的需求不断增加。基于显著性特征属性方法，特别是在计算机视觉领域，经常被用来作为一种 XAI 方法，该方法突出了图像中对分类器决策有重要贡献的部分。为了定量比较各种基于显著性的 XAI 方法，已经提出了几种自动评价方法；然而，不能保证这些自动化的评估指标能够正确地评估可解释性，自动评估方案的高评级并不一定意味着人类的高可解释性。在这项研究中，我们提出了一个新的基于人的评估方案，利用众包来评估 XAI 方法，而不是自动评估。我们的方法受人类计算游戏 Peek-a-boom 的启发，利用群体的力量对不同的 XAI 方法的显著性图进行评价。实验结果表明，基于人群的评价方案的评价结果不同于自动评价方案。此外，我们将基于人群的评估结果视为基本事实，并提供了一个定量的性能度量来比较不同的自动评估方案。我们还讨论了群体工作者对结果的影响，并表明群体工作者能力的变化对结果没有显著影响。

736，多时间序列的可逆神经网络可解释非线性建模

LuisMiguel 等，Agder 大学，2021.7.3

提出一种非线性拓扑识别方法，基于以下假设：时间序列的集合分两步生成：①潜在空间中的向量自回归过程，以及②非线性组件方式、单调递增的观测映射。后面的映射被假定为可逆的，并被建模为浅层神经网络，因此可以对它们的逆进行数值评估，并且可以使用受深度学习启发的技术来学习它们的参数。由于函数反转，反向传播步骤并不简单，本文解释了应用隐微分计算梯度所需的步骤。虽然模型的可解释性与线性 VAR 过程相同，但初步数值测表明预测误差变小。

737，OKGIT：具有隐式类型的开放知识图谱链接预测

印度科学研究所，2021.6.24

开放知识图谱（OpenKG）是指使用 OpenIE 工具从语料库中提取的一组（头名词短语、关系短语、尾名词短语）三元组，例如（特斯拉、返回、纽约）。虽然 OpenKG 很容易为域引导，但它们非常稀疏，还远不足以直接用于最终任务。因此，在本文理解、问答和网络搜索查询推荐等下游任务中使用这些图时，预测新事实的任务，即链接预测，成为重要的一步。

OpenKG 的链接预测任务一直是一个相对未开发的研究领域，其学习嵌入是一种最近受到一些关注的链接预测方法。以前关于 OpenKG 嵌入的工作主要集中在改进或合并 NP 规范化信息上，经过仔细检查，作者发现当前的 OpenKG 链接预测算法通常会针对给定的名词和关系短语预测具有不兼容类型的名词短语（NPs）。

本文作者在这项工作中解决了这个问题，使用来自 BERT 的隐式类型信息来改进 OpenKG 链接预测并提出了 OKGIT。它使用新颖的类型兼容性分散和类型正则化来改进 OpenKG 链接预测。通过对多个数据集的大量实验，作者表明所提出的方法实现了最先进的些能，同时在链接预测任务中产生了类型兼容的 NPs。

738，Spiking-GAN：使用 Time-To-First-Spike 编码的 Spiking 生成对抗网络

VineetKotariya 等，印度理工大学

脉冲神经网络（SNN）在以节能的方式解决深度学习问题方面显示出了巨大的潜力。但是，它们仍然仅限于简单的分类任务。在本文中，我们提出了 Spiking-GAN 这是第一个基于脉冲的生成对抗网络（GAN）。它采用一种称为首次脉冲时间编码的时间编码方案。我们使用时域中的近似反向传播来训练它。我们的网络使用非常高不应期的简单积分和激发（IF）神经元，以确保每个神经元最多有一个脉冲。这使得模型比基于脉冲率的系统稀疏得多。与之前的工作相比，修改后的时间损失函数称为“Aggressive TTFS”，将网络推理时间提高了 33%以上，并将网络中的脉冲数量减少了 11%以上，实验表明，使用这种方法在 MNIST 数据集上训练网络时，我们可以生成高质量的样本，从而证明了该框架在解决脉冲域中此类问题的潜力。

739，再谈英国首研基于机器学习（人工智能）的六代机

在早年（2020.2.22）我们的跟帖 283 中就报导了英研六代机亮相的信息。

据《兵器》杂志报导，目前世界最先进的战斗机是四款五代机：美国 F-22、F-35，俄罗斯苏-37，中国的歼 20，F-22、F-35，歼 20 均是隐身的，苏-37 不是隐身的。由于 F-22、F-35 在生产中存在一些问题，处于停产状态，苏-37 在飞行试验中也有一些问题待解决，今天表现较好的只有歼-20。

英研“暴风雨”六代隐身战斗机超高音速 5 马赫，为歼-20 二倍多，搭载激光武器，采用大升阻比与机翼融合设计，采用智能技术，很可能推出无人驾驶。

740，可信人工智能

HaoChen Liu 等，密歇根大学，2021.7.12

在过去的几十年里，人工智能技术经历了飞速的发展，深刻地改变了人们的日常生活和人类社会的发展进程。开发人工智能的目的是减少人类劳动、为人类生活带来便利和促进社会进步。然而，最近的研究和应用表明，人工智能可能会对人类造成无意的伤害，例如在安全关键场景中做出不可靠的决定，或因无意中歧视某一群体而破坏公平性。因此，可信人工智能近年来受到了极大的关注，这就要求人们认真考虑，避免人工智能可能给人类带来的不利影响，使人类能够充分信任人工智能技术，并与之和谐相处。

近年来，人们对可信人工智能进行了大量的研究。本文作者从计算的角度对可信人工智能进行一个全面的调研，帮助读者了解最新的技术。可信人工智能是一个庞大且复杂的领域，涉及多个维度。本文作者关

注六个最关键的维度：（i）安全性和稳健性，（ii）非歧视性和公平性，（iii）可解释性，（iv）隐私性，（v）责任性和可审计性，以及（vi）环境福利。对于每个维度，作者回顾了近期的相关技术，并总结了它们在实际系统中的应用。作者还讨论了不同维度之间的一致性和冲突性交互作用，并讨论了可信人工智能的未来研究方向。

人工智能系统性能的提高通常是通过增加模型复杂度来实现的。一个典型的例子就是深度学习，它是大多数人工智能系统的核心。但是，它们被视为黑匣子，因为大多数深度模型过于复杂和不透明，人们无法理解。更重要的是，如果不解释模型背后的潜在机制，深度模型就不能完全可信，这就妨碍了它们在涉及道德、正义和安全的关键应用中的应用，如医疗保健、自动汽车等。因此，建立一个可信的人工智能系统需要了解特定决策是如何做出的，这导致了可解释人工智能领域的兴起。论文第六章对可解释人工智能的最新进展提供直观的理解和高层次的见解。首先，作者提供了人工智能中解释性的概念和分类。其次，根据前面提到的分类法回顾了人工智能系统中有代表性的可解释技术。随后，作者介绍了可解释人工智能技术的实际应用。最后，提供了一些综述和工具，并讨论了可解释人工智能的未来机遇。

在机器学习和人工智能文献中，explainability 和 interpretability 通常被研究者可互换地使用。最流行的可解释性定义之一是 Doshi Velez 和 Kim 的定义，他们将其定义为"以可理解的术语解释或呈现给人类的能力"。另一个流行的定义来自 Miller，他将可解释性定义为"人类能够理解决策原因的程度"。一般来说，人工智

能系统的可解释性越高，人们就越容易理解某些决策或预测是如何做出的。同时，如果一个模型的决策比其他模型的决策更容易被人理解，那么它比其他模型更容易解释。虽然 explainable AI 和 interpretable AI 有着非常密切的联系，但有一些研究也讨论了它们之间的一些细微差别。

（1）如果模型本身能够被人类理解其是如何进行预测的，那么模型就是"interpretable"。 当查看模型参数或模型摘要时，人类可以准确地理解它如何做出某个预测/决策的过程，甚至给定输入数据或算法参数的变化，它是人类能够预测将要发生什么的程度。换句话说，这样的模型本质上是透明和可解释的，而不是黑盒/不透明模型。interpretable models 的例子包括决策树和线性回归。

（2）explainable model 是采用了额外的（事后）解释技术来帮助人类理解为什么它做出了某个预测/决策，尽管该模型仍然是黑盒和不透明的。值得注意的是，这种解释往往是不可靠的，可能会产生误导。这类模型的例子是基于深度神经网络的模型，其中的模型通常过于复杂，任何人都无法理解。

人工智能的解释技术可以根据不同的标准进行分组。

根据模型用法的不同可分为 model-intrinsic 和 model-agnostic。如果可解释技术的应用仅限于人工智能模型的特定体系结构，那么这些可解释技术称为 model-intrinsic。相反，可以应用于任何模型的技术被称为 model-agnostic。

根据解释范围的不同可分为 local 和 global。如果该方法仅为特定实例提供解释，则它是局部解释，如果该方法可以解释整个模型，则它是全局解释。

根据解释方法的不同可分为 gradient-based 和 perturbation-based。如果这些技术利用输入实例的偏导数来生成属性，那么这些技术称为基于梯度的解释方法，如果这些技术侧重于输入数据的变化或修改，我们称之为基于扰动的解释方法。

还有一种技术通过其他方法进行解释，即 Counterfactual Explanations。该方法通常是包含因果关系的形式，例如："如果 X 没有发生，Y 就不会发生"。一般来说，Counterfactual Explanations 方法与模型无关，可用于解释个别实例的预测（局部）。每一类方法的代表模型如下表所示。

Representative Models	Model Usage	Scope	Methodology
Linear model	Intrinsic	Global	-
LIME [267]	Agnostic	Both	Perturbation
CAM [369]	Agnostic	Local	Gradient
Grad-CAM [290]	Agnostic	Local	Gradient
SHAP [220]	Agnostic	Both	Perturbation
Saliency Map Visualization [300]	Agnostic	Local	Gradient
GNNExplainer [346]	Agnostic	Local	Gradient
Class Model Visualization [300]	Agnostic	Local	Gradient
Surveys	[27, 34, 103, 112, 113, 152, 175, 209, 238, 243, 314, 349, 361]		

可解释人工智能的应用有：（1）推荐系统。推荐系统（RecSys）在我们的日常生活中变得越来越重要，因为它们在缓解信息过载问题方面发挥着重要作用。这些系统提供个性化信息以帮助人类做出决策，并已广泛应用于各种面向用户的在线服务，如电子商务商品日常购物推荐（如亚马逊、淘宝）、就业市场就业推荐（如 LinkedIn）等。近年来，基于深度学习的推荐模型在提高准确性和更广泛的应用场景方

面取得了巨大的进步。因此，人们越来越关注理解基于深度学习的推荐系统为什么会推荐某些项目给最终用户，因为提供个性化推荐系统的良好解释可以充分激励用户与项目交互，帮助用户做出更好和/或更快的决策，增加用户对智能推荐系统的信任。（2）药物研究。在过去的几年中，可解释人工智能已被证明显著加速了计算机辅助药物发现的过程，例如分子设计、化学合成规划、蛋白质结构预测和大分子目标识别。（3）自然语言处理。作为人工智能应用最广泛的领域之一，自然语言处理（NLP）研究了如何使用计算机来处理或理解人类的语言。自然语言处理的应用无处不在，包括对话系统、文本摘要、机器翻译、问答、情感分析、信息检索等。最近，深度学习方法在许多不同的自然语言处理任务中取得了非常好的表现，但这是以模型变得不那么可解释为代价的。

由于人工智能的可解释性是一个相对较新的领域，一个发展中的领域，因此有许多问题需要考虑。（1）可解释人工智能的安全性。最近的研究表明，由于人工智能模型的数据驱动特性，其解释容易受到恶意操作的影响。攻击者试图生成对抗性的示例，这不仅会误导目标分类器，还可能欺骗相应的解释器。这自然会在解释上引起潜在的安全问题。因此，如何防范翻译中的对抗性攻击将是今后的一个重要方向。（2）评价方法。评价指标是研究解释方法的关键。然而，由于缺乏基本事实和人的主观理解，对某些预测的解释是否合理和正确的评价变得越来越困难。目前广泛使用的评价方法是基于人的评价，这种方法比较直观，同时也比较费时，并且存在偏见。（3）从白盒到黑盒。

大多数现有的解释技术要求对所解释的人工智能系统有充分的了解（表示为白盒）。然而，由于隐私和安全问题，在许多情况下，人工智能系统的相关知识通常是有限的。因此，一个重要的方向是理解如何在黑箱系统中生成决策的解释。

741，基于可解释 SincNet 的脑电信号情感识别深度学习

2Mila-Quebec 人工智能研究所，2021.9.23

机器学习方法，如深度学习，在医学领域显示出有希望的结果。然而，这些算法缺乏可解释性可能会阻碍它们在医疗决策支持系统中的适用性。本文研究了一种可解释的深度学习技术，称为 SincNet。SincNet 是一种卷积神经网络，它通过可训练的 sinc 函数有效地学习定制的带通滤波器。在这项研究中，本文使用 SincNet 来分析患有自闭症谱系障碍(ASD)的个体的神经活动，他们在神经振荡活动中经历了特征性差异。特别是，本文提出了一种新的基于 SincNet 的神经网络，用于使用 EEG 信号检测 ASD 患者的情绪。可以轻松检查学习到的过滤器，以检测 EEG 频谱的哪一部分用于预测情绪。本文发现本文的系统会自动学习 ASD 患者经常出现的高 α (9-13Hz)和 β (13-30Hz)频带抑制。这一结果与最近关于情绪识别的神经科学研究一致，该研究发现这些频带抑制与在 ASD 个体中观察到的行为缺陷之间存在关联。在不牺牲情绪识别性能的情况下，SincNet 的可解释性得到了提高。

742，深度神经网络中的时间稀疏性训练在视频处理中的应用

比利时微电子研究中心，2021.7.15

激活稀疏性提高了稀疏感知神经网络加速器的计算效率和资源利用率。由于 DNN 中的主要操作是具有权重的激活的乘法累加 (MAC) 以计算内积，因此跳过（至少）两个操作数之一为零的操作可以使推理在延迟和功率方面更有效。激活的空间稀疏化是 DNN 文献中的一个热门话题，并且已经建立了几种方法来使 DNN 偏向于它。另一方面，时间稀疏性是受生物启发的脉冲神经网络 (SNN) 的固有特征，神经形态处理利用它来提高硬件效率。引入和利用时空稀疏性是 DNN 文献中很少探讨的主题，但与 DNN 的趋势完美共鸣，从静态信号处理转向更多流信号处理。为了实现这一目标，在本文中，我们引入了一个新的 DNN 层（称为 Delta 激活层），其唯一目的是在训练期间促进激活的时间稀疏性。 Delta 激活层将时间稀疏性转换为空间激活稀疏性，以便在硬件中执行稀疏张量乘法时加以利用。通过在训练期间采用增量推理和“通常的”空间稀疏化启发式方法，所得模型不仅学习利用空间而且利用时间激活稀疏性（对于给定的输入数据分布）。人们可以在原版训练或细化阶段使用 Delta 激活层。我们已经实现了 Delta 激活层作为标准 Tensoflow-Keras 库的扩展，并将其应用于在人类动作识别 (UCF101) 数据集上训练深度神经网络。我们发现，激活稀疏性提高了近 3 倍，在长时间训练后模型准确性的损失是可恢复的。

743，知识图谱在智能制造领域的研究现状及应用前景综述

数据和知识是新一代信息技术与智能制造深度融合的基础。然而，当前产品设计、制造、装配和服务等过程中，数据及知识的存储大多以传统关系型数据库为基础，这导致了数据及知识的冗余性和搜索及推理的低效性。近年来，知识图谱技术飞速发展起来，它本质上是基于语义网络的思想，可以实现对现实世界的事物及其相互关系的形式化描述。该技术为智能制造领域数据及知识的关联性表达和相关性搜索推理问题的解决带来了可能性，因此其在智能制造的实现过程中扮演着越来越重要的角色。为了给知识图谱在智能制造领域的应用提供理论支撑，总结了知识图谱领域的研究进展；同时探索了知识图谱在智能制造领域的 3 大类应用方向，共 15 小类应用前景，分析了在各个应用前景上与传统方法的不同之处，应用过程中所需要使用的知识图谱相关技术以及实施过程中所待突破的关键技术，希望可以为进一步展开针对知识图谱在智能制造领域的研究提供启发，同时为相关企业针对知识图谱的实际应用提供参考；最后以数控车床故障分析为案例，验证了知识图谱在智能制造领域应用的有效性。

744，GLIME：一种用于可解释性模型不可知解释的新图形方法

可解释人工智能（XAI）是一个新兴的领域，在这个领域中，一系列的过程和工具使人们能够更好地理解由黑盒模型生成的决策。然而，大多数可用的 XAI 工具通常仅限于简单的解释，主要是量化各个特性对模型输出的影响。因此，人类用户无法理解特征之间的相互关系以

进行预测，而训练模型的内部工作机制仍然是隐藏的。本文致力于开发一种新的图形化解释工具，该工具不仅能显示模型的重要特征，而且能揭示特征之间的条件关系和推理，捕捉特征对模型决策的直接和间接影响。提出的 XAI 方法称为 GLIME，它提供了全局（对于整个数据集）或局部（对于特定数据点）的图形模型不可知解释。它依赖于局部可解释模型不可知解释（LIME）与图形最小绝对收缩和选择算子（GLASSO）的结合，产生无向高斯图形模型。采用正则化方法将小的偏相关系数压缩到零，从而提供更稀疏、更易于解释的图形解释。选择两个著名的分类数据集（活检和 OAI）来证实 GLIME 在稳健性和一致性方面优于 LIME。具体来说，GLIME 在两个数据集上实现了特征重要性方面的稳定性提高（76%-96%，而使用 LIME 则为 52%-77%）。GLIME 展示了一种独特的潜力，通过提供信息丰富的图形化解释，可以打开黑匣子，从而扩展 XAI 当前最先进的功能。

745，用于提高神经网络对数据质量问题的鲁棒性的调制层

华盛顿大学，2021.7.19

数据质量是机器学习中的一个常见问题，特别是在医疗保健等高风险环境中。缺失数据会影响复杂模式中的精度、校准和特征属性。开发人员经常在精心策划的数据集上训练模型，以最小化丢失的数据偏差；但是，这降低了此类模型在生产环境（如实时医疗记录）中的可用性。

因此，使机器学习模型对缺失数据具有鲁棒性是实际应用的关键。一些分类器自然地处理缺失数据，而另一些分类器，如深度神经网络，则不是针对未知值设计的。我们提出了一种新的神经网络修正方法来减轻缺失数据的影响。这种方法的灵感来自于由生物神经网络进行的神经调节。我们的建议将完全连接层的固定权重替换为每个输入的附加输入（可靠性得分）函数，模仿大脑皮层基于其他数据的上下权重输入的能力。利用多层感知器与主任务联合学习调制函数。我们在多重分类、回归和插补问题上测试了我们的调制全连接层，它要么提高了性能，要么产生了与传统神经网络结构相类似的性能，将可靠性连接到输入。具有调制层的模型通过在评估时引入额外的缺失，对数据质量的降低更具鲁棒性。这些结果表明，通过调制全连接层明确说明信息质量的降低可以使人工智能系统在实时环境中部署。

746，具有自监督和门控适配器的 LiDAR 语义分割中的无监督域适应（自适应激光雷达语义分割）

华为 Noah'sArkLab,2021.7.20

在本文中，我们致力于一个探索较少，但更现实和复杂的问题领域自适应激光雷达语义分割。当训练（源域）和测试（目标域）数据来自不同的激光雷达传感器时，现有分割模型的性能会显著下降。为了克服这一缺点，我们提出了一种无监督的域自适应框架，该框架利用未

标记的目标域数据进行自我监督，并结合一种不成对的掩码转移策略来减轻域转移的影响。此外，我们在网络中引入带有少量参数的选通适配器模块，以说明特定于目标域的信息。从真实到真实以及从合成到真实的 LiDAR 语义分割基准进行的实验表明，该方法比现有技术有显著的改进。

747，稳健可解释性：深度神经网络基于梯度的属性方法教程

IanE.Nelsen 等，Rowan 大学，2021.7.3

随着深度神经网络的兴起，解释这些网络预测的挑战越来越受到人们的认可。虽然存在许多解释深度神经网络决策的方法，但目前还没有就如何评估它们达成共识。另一方面，鲁棒性是深度学习研究的热门话题；然而，直到最近才在可解释性方面谈论它。在本教程论文中，本文首先介绍基于梯度的可解释性方法。这些技术使用梯度信号来分配输入特征的决策负担。稍后，本文将讨论如何评估基于梯度的方法的鲁棒性以及对抗性鲁棒性在提供有意义的解释方面所起的作用。本文还讨论了基于梯度的方法的局限性。最后，本文介绍了在选择可解释性方法之前应该检查的最佳实践和属性。本文总结了该领域在稳健性和可解释性的融合方面的未来研究方向。

748，使用可解释的深度学习方法进行有效和健壮的模式识别

XiaoBai 等，德克萨斯大学奥斯丁分校，2021.7.23

深度学习最近在许多视觉识别任务中取得了巨大的成功。然而，深层神经网络（DNNs）通常被视为黑匣子，其决策过程和原理不易被人类理解，因此被禁止在关键安全应用中使用。本文介绍了 30 篇论文，这些论文都是关于 Explainable Deep Learning for Efficient and Robust Pattern Recognition 的特刊。它们主要分为三大类：可解释的深度学习方法、通过模型压缩和加速以实现高效的深度学习以及深度学习的鲁棒性和稳定性。本文对这三个专题的代表作和最新进展进行了综述，并简要介绍了各个专题已被接受的论文。这篇综述的整体结构如图 1 所示。

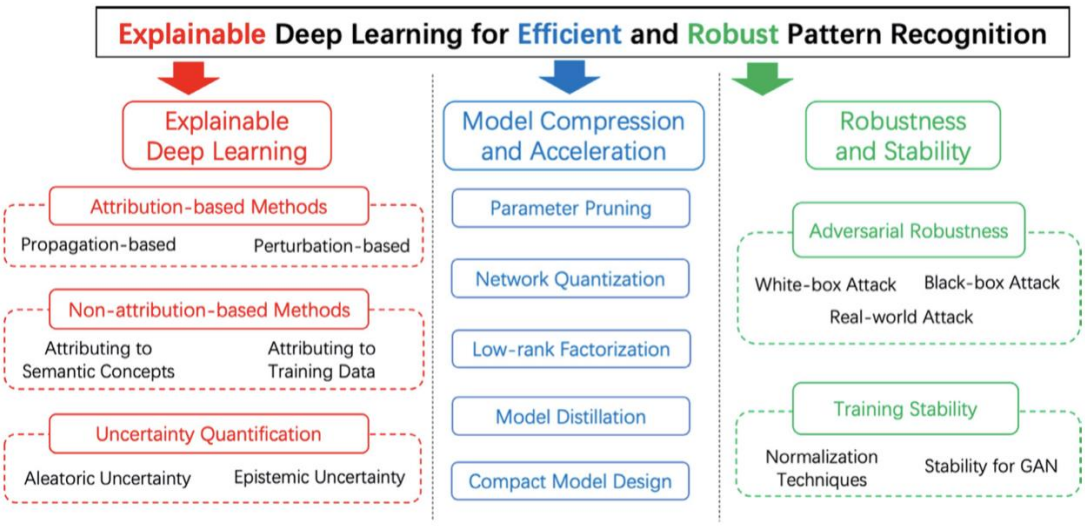


Fig. 1. The overall structure of this guest editorial.

许多可解释性方法旨在解释深层神经网络的工作机制。大多数的方法侧重于将 DNN 的预测归因于其输入特征。这种基于归因的方法涵盖了计算机视觉中的大多数可视化方法，它通过定位对决策贡献最大的区域，直接在输入图像的域中给出解释。除此之外，许多基于非归因的方法还从概念、训练数据、内在注意机制等方面进行了解释。从另一

个角度看，不确定性暗示了网络决策的可靠性。这些信息是对网络解释的补充，在各种现实生活中的关键安全应用中是至关重要的。

深度神经网络在各种任务中以设置较多参数为代价获得了最高的精度，从而需要大量的计算资源和训练时间。因此，在部署到资源受限的设备和实时应用程序之前，业界对模型压缩和加速技术有着巨大的需求。近年来，越来越多的方法被提出来用于压缩和加速网络，同时对模型的精度做出最小的妥协。大多数方法可分为以下几类：参数剪枝、网络量化、低秩因子分解、模型提取和紧凑网络设计。

鲁棒性揭示了模型在数据噪声下提供可靠决策的能力。近年来，人们对与深度学习相关的鲁棒性的几个方面进行了研究。其中最热门的话题是对抗性健壮性，因为它与应用程序中的安全问题密切相关。稳定性是深度神经网络的另一个关键问题，它决定了网络能否成功收敛。一些技术有助于提高训练时的稳定性，如规范化技术和网络优化的一些约束。

749，KG4vis，一种基于知识图谱的可视化推荐方法

新加坡管理大学，香港科技大学 2021.7.27

可视化推荐或自动可视化生成可以显着降低一般用户快速创建有效数据可视化的障碍，尤其是对于没有数据可视化背景的用户。然而，现有的基于规则的方法需要可视化专家对可视化规则进行繁琐的手动规范。其他基于机器学习的方法通常像黑盒一样工作，并且难以理解为什么推荐特定的可视化，从而限制了这些方法的更广泛的采用。本文

通过介绍 KG4Vis 填补了这一空白。KG4Vis 是一种基于知识图谱(KG)的可视化推荐方法。它不需要手动指定可视化规则，也可以保证良好的可解释性。具体来说，本文作者提出了一个构建知识图谱的框架，由三种类型的实体（即数据特征、数据列和可视化设计选择）和它们之间的关系组成，以对数据和有效可视化之间的映射规则进行建模。采用基于 TransE 的嵌入技术从现有的数据集-可视化对中学习实体的嵌入和知识图谱的关系。这种嵌入本质上是对理想的可视化规则进行建模。然后，给定一个新的数据集，可以从具有语义意义规则的知识图谱中推断出有效的可视化。本文作者进行了广泛的测试以评估提出的方法，包括定量比较、案例研究和专家访谈。结果证明了本文方法的有效性。

在未来的工作中，作者计划探索如何在不增加计算能力的情况下在 KG 中加入跨列特征。并且，本文作者想进一步研究如何结合不同的用户需求和偏好来实现对不同用户的个性化可视化推荐。此外，还会关注将提出的基于 KG 的可视化推荐方法扩展到其他类型的可视化。

750，解决基于扰动的 XAI 技术所面临的分布外数据问题

LuyuQIU 等人，华为香港研究院、香港大学，2021.8.7

随着可解释人工智能（XAI）的迅速发展，基于扰动的 XAI 算法因其有效性和易实现性而变得非常流行。绝大多数基于扰动的 XAI 技术都面临着分布外（Out-of-Distribution，OoD）数据的挑战，即随机扰动数据的伪影与原始数据集不一致。OoD 数据导致模型预测中的过度

置信问题，使得现有的 XAI 方法不可靠。但是，目前基于扰动的 XAI 算法存在的 OoD 数据问题尚未在文献中得到充分解决。本文，作者通过设计一个额外的模块来解决这个 OoD 数据问题，该模块量化了扰动数据和原始数据集分布之间的相关性，并将其集成到聚合过程中。作者的解决方案与最流行的基于扰动的 XAI 算法（如 RISE，OCCLUSION 和 LIME）兼容。实验证明，作者的方法在使用计算和认知度量的一般情况下表现出显著的改进。特别是在退化的情况下，与基线相比，作者提出的方法表现出了优异的性能。此外，作者的解决方案还解决了忠实度指标的一个基本问题，这是 XAI 算法的一个常用评估指标，似乎对 OoD 问题很敏感。

751，基于子组发现的黑盒子事件检测的可解释性总结

YoucefRemil 等，里昂大学，2021.8.6

随着监控系统和设备 软件用户报告的事件数量不断增加，预测性维护的需求也随之增加。在前线，待命工程师（OCE）必须快速评估事件的严重程度，并决定联系哪家服务机构采取纠正措施。为了使这些决策自动化，已经提出了几种预测模型，但最有效的模型是不透明的（比如，黑盒），这大大限制了它们的采用。在本文中，提出了一个有效的黑盒模型，该模型基于过去 7 年中向本文公司报告的 170K 起事件，并强调在运行本文的产品（ERP）的数千台服务器上大量报告事件时自动分类的需要。可解释人工智能（XAI）的最新发展有助于为模型提供全局解释，但最重要的是，为每个模型预测 结果提供局

部解释。不幸的是，在处理大量日常预测时，为人类提供每种结果的解释是不可想象的。为了解决这个问题，本文提出了一种源于子组发现的原始数据挖掘方法，这是一种模式挖掘技术，具有对共享其黑盒预测的类似解释的对象进行分组的自然能力，并为每个组提供描述。本文评估了这种方法，并给出了本文的初步结果，这给本文带来了采用有效 OCE 的良好希望。相信这种方法能为解决模型不可知结果解释问题提供了一种新的方法。

752，基于数据流将脉冲神经网络合成到神经形态硬件

ShihaoSong 等，德雷塞尔大学，2021.8.4

脉冲神经网络 (SNN) 是一种新兴的计算模型，它使用事件驱动的激活和仿生学习算法。基于 SNN 的机器学习程序通常在基于 tile 的神经形态硬件平台上执行，其中每个 tile 由一个称为 crossbar 的计算单元组成，它映射程序的神经元和突触。然而，在现成的神经形态硬件上合成此类程序具有挑战性。这是因为硬件的固有资源和延迟限制会影响模型性能（例如准确性）和硬件性能（例如吞吐量）。我们提出了 DFSynthesizer，一种将基于 SNN 的机器学习程序合成到神经形态硬件的端到端框架。提议的框架分四个步骤工作。首先，它分析机器学习程序并使用代表性数据生成 SNN 工作负载。其次，它划分 SNN 工作负载并生成适合目标神经形态硬件纵横的集群。第三，它利用同步数据流图 (SDFG) 的丰富语义来表示集群 SNN 程序，允许在关键硬件约束方面进行性能分析，例如交叉开关的数量、每个交

叉开关的尺寸、瓦片上的缓冲区空间和瓦片通信带宽。最后，它使用一种新颖的调度算法在硬件的交叉开关上执行集群，保证硬件性能。我们使用 10 个常用的机器学习程序评估 DFSynthesizer。我们的结果表明，与当前的映射方法相比，DFSynthesizer 提供了更优秀的性能保证。

753，节能随机游走计算的神经拟态缩放优势

J.Darby Smith，神经探索与研究实验室，桑迪亚国家实验室，2021.7.27

受人脑运作方式启发的神经拟态计算 (NMC) 方法可以从根本上改进现有的计算方式。大多数旨在在人造硬件中复制大脑计算结构和架构的 NMC 研究都集中在人工智能上，很少有人探索这种受大脑启发的硬件是否可以提供超越认知任务的价值。本文证明了尖峰神经拟态架构的高度并行性和可配置性使它们非常适合通过离散时间马尔可夫链实现随机游走算法。这种随机游走在蒙特卡罗方法中很有用，蒙特卡罗方法代表了解决各种数值计算任务的基本计算工具。此外，本文还展示了涉及一类随机微分方程的概率解决方案的数学基础如何利用这些模拟为一系列广泛适用的计算任务提供解决方案。同时，本文还发现 NMC 平台在足够的规模下可以大大降低高性能计算 (HPC) 平台的能源需求。

754，用于推荐知识图谱上的时间感知路径推理

Yuyue Zhao 等、中国科技大学、新加坡大学、中国风险感知与预防国家工程实验室，2021.8.5

由于知识图谱（KG）的推理能够提供明确的解释，因此已经研究了可解释性推荐的推理。然而，遗憾的是，当前基于 KG 的可解释推荐方法忽略了时间信息（如购买时间、推荐时间等），这可能会导致不合适的解释。在这项工作中，本文作者提出了一种新颖的时间感知路径推理推荐（简称 TPRec）方法，该方法利用时间信息的潜力以合理的解释提供更好的推荐。据作者所知，TPRec 方法是第一个将时间感知路径推理方法引入推荐系统的工作，并通过利用时间信息实现了显着的性能提升。首先，作者提出了一种有效的时间感知交互关系提取组件来构建具有时间感知交互（简称 TCKG）的协作知识图谱，然后介绍一种新颖的时间感知路径推理方法进行推荐。作者对三个真实世界的数据集进行了广泛的实验。结果表明，所提出的 TPRec 可以成功地使用 TCKG 来获得可观的收益并提高可解释推荐的质量。

对于未来的工作，作者计划使用来自其他产品领域和其他主要在线供应商的更多数据集来评估 TPRec，并通过利用对抗性学习模型自动塑造奖励来扩展 TPRec 模型，以实现更准确的推荐结果。作者还计划将因果推断与我们的模型相结合，以实现更好的可解释性。

755，论人工智能基础理论的源头创新

范式革命：人工智能基础理论重大突破的必然选择

钟义信

摘要 制约人工智能发展的最大障碍，是它的最高指导思想（即科学观和方法论，统称为范式）犯了张冠李戴的大忌：用了物质学科的范式来指导人工智能的研究。本文用科学发展普遍规律与人工智能发展历史相结合的方法论证了：人工智能范式张冠李戴的不可避免性，人工智能的范式革命是人工智能实现重大突破的唯一正确举措，后者将推动物质学科主导的科学时代迈向信息学科主导的科学新时代，并从源头上创建科学新时代的人工智能理论 — 通用人工智能理论。

1，引言

与解放人类体质能力的材料科学和解放人类体力能力的能量科学（统称为物质学科）不同，人工智能是以解放人类智力能力为目标的一门学科。因此，人工智能学科一问世就受到人类社会的高度关注。至今，发展人工智能已经成为世界各国特别是各发达国家的重大战略。

2018 年 10 月 31 日，习近平总书记在中共中央政治局集体学习会议上指出：**人工智能是引领这一轮科技革命和产业变革的战略性技术，… 要加强基础理论研究，支持科学家勇闯人工智能科技前沿的“无人区”，努力在人工智能发展方向和理论、方法、工具、系统等方面取得变革性颠覆性突破，确保我国在人工智能这个重要领域的理论研究走在前面、关键核心技术占领制高点。**

2017 年 7 月，我国国务院颁发了《新一代人工智能发展规划》。

根据这个《发展规划》，今年，我国的人工智能研究进入了第二个战略阶段，目标是实现人工智能基础理论的重大突破。因此，实现人工智能基础理论研究的重大突破，应当成为本阶段我国人工智能研究的重中之重。

2，实现人工智能基础理论的重大突破，需要准确理解几个重要概念

人工智能的研究新颖、复杂而艰深，呈现出诸多的谜团和不确定性。因此，需要正确理解人工智能领域的一些重要概念，以免研究活动误入歧途。

2.1 人类智慧和人类智能

首先，要分请智慧与智能的关系，不能把二者混为一谈。

什么是智慧？现代汉语词典说，是辨识判断和发明创造的能力。

笔者认为，**人类智慧是指它作为万物之灵的卓越能力：人类自觉地不断地运用信息和知识去认识世界和改造世界，并在改造客观世界的过程中不断地改造自己的主观世界，从而不断地改善自己的生存与发展水平。**

认识世界，是为了探索未来发展道路而去研究和提出应当解决的问题，给出解决问题应当遵循的工作框架：（1）定义和描述所需解决的问题，（2）预设解决问题应达到的目标，（3）明确解决问题所需的知识。

改造世界，是为了变革现实而去解决所提出的问题，即针对给出的工作框架，利用关于问题的信息和知识，在预设目标的引导下去谋划解决问题的策略，并把策略转化为行动，把问题的原有状态改造成为符合预设目标所要求的状态。

认识世界必须根据人类的目的，利用人类的隐性思辨能力（如直觉能力、抽象能力、理解能力和想象能力等）去探索，因此，被称为人类的隐性智慧。改造世界则主要依赖于人类的显性操作能力（包括获取信息、提取知识、生成策略和执行策略的能力等）来实施，因此，被称为人类的显性智慧。

人类隐性智慧的特点，是直接依赖于人类的目的，而且高度抽象，甚至近于神秘，因此，难以用机器来模拟。然而，人类的显性智慧却是在给定了具体问题、求解目标和知识的工作框架条件下的问题求解研究，既有明确的起点，又有明确的目标，既可望也可及，因而可以成为“机器模拟”的对象。人类的显性智慧由此就被特别地称为“人类智能”。

人造机器可以用来模拟人类智能，却难以用来模拟人类智慧。

2.2 人工智能与人类智能

人工智能，是在人造机器上所实现的人类智能。因此，人工智能的研究只涉及对“解决问题变革现实的人类显性智慧”的模拟，不涉及探索未来提出问题的人类隐性智慧。人类不可能让没有生命的机器取代自己去探索自己的未来，把人类未来发展的命运交给没有生命的机器。

明确了这些概念，就可以避免产生“人工智能无所不能、人工智能将全面超越人类和统治人类”这样一些误解。其实，**再强大的人工智能，都只能在人类隐性智慧给出的工作框架内去解决问题**，只能执行人类的意志，为解决人类关注的问题和为改善人类生存发展的目的服务。人工智能可以是人类的聪明助手与合作伙伴，而不能成为主宰人类命运的主人。

2.3 信息与机器感知

没有信息就没有知识，没有知识就没有智能。因此，信息是智能的源头，准确理解信息概念才能保证智能理论研究不走偏路。

需要提醒的是，至今国际学术界唯一公认的信息理论是 Shannon1948 年创立的信息论。然而，Shannon 信息论的原名是通信的数学理论（Mathematical

Theory of Communication），适用于“只关心形式，不关心价值和内容”的通信工程，不适于“需要全面理解形式、内容和价值”的人工智能研究的需要。

具体来说，在人工智能的场合，信息有两个互相联系又互相区别的概念。

首先是客体信息的概念，它的定义可以这样表述：**客体信息，就是客体所呈现的自身状态及其变化方式**。显然，这是本体论的信息概念。客体信息只与客体自身的状况有关，而与主体的状况无关，甚至与是否有主体存在无关。

如果客体信息被主体（人工智能机器可被理解为人类主体的代理，也常常简称为主体。下同）所感知，就产生了主体的感知信息的概念。由于引入了主体，**感知信息比较复杂，包含了三个分量：（1）主体从客体信息中感受到的客体的状态及其变化方式的外在形式，称为形式信息**（文献中称为语法信息）；**（2）从客体信息中知觉到的客体对主体目标而言所具有的效用价值，称为价值信息**（文献中称为语用信息）；（3）**由形式信息和价值信息两者定义而成的含义内容，称为内容信息**（文献中称为语义信息）。

感知信息是形式信息、价值信息和内容信息的三位一体。正由于感知信息包含了形式、价值、内容全部分量，因此也被称为“全信息”。显然，感知信息属于认识论的范畴。关于感知信息的更深入讨论，见本文第 5.2 节。

有了这些信息概念，就可以精准定义感知的概念：**感知，就是把客体信息转换为感知信息的过程**。这是感知的学术本质。

学术界存在一个相当普遍的误解：以为感知就是传感。根据上述的感知定义就明白：感知功能必须能够了解客体的形式、价值和内容。传感功能只能感受到客体的形式，不了解客体的价值和内容，不可能承担感知的功能。

2.4 知识与机器认知

由上述信息的定义可以知道，信息是具体的个性化的概念。如钢笔的信息、毛笔的信息、铅笔的信息、圆珠笔的信息、排笔的信息、画笔的信息等等，都是具体事物的信息，个别事物的信息。

与此相对，“笔”的信息则是抽象的共性的概念，它是从具体的钢笔的信息、毛笔的信息、铅笔的信息、圆珠笔的信息、排笔的信息、画笔信息这个集合抽象出来的共性的概念。这个“抽象的笔的信息”就称为“关于笔的知识”。**信息是具体的个别事物呈现的现象，知识则是一类事物的抽象的共性本质**。

需要注意，既然知识是一类事物的抽象的共性的本质，因此，知识应当属于认识论的范畴（而不是本体论的范畴）。换言之，认识论范畴的知识只能从认

识论范畴的感知信息抽象提炼出来的概念。因此，与感知信息概念相对应，知识也应当是形式性知识、价值性知识、内容性知识的三位一体。

由此可以给出认知的定义：**认知，就是把感知信息转换为知识的过程**。这是认知的基本功能，也是认知的学术本质。

不难理解，作为把感知信息（感性认识）转换为知识（理性认识）的认知，其实就是学习的过程，包括初级的灌输式（机械式）学习，中级的从众式（统计式）学习，高级的自主式（理解式）学习。

2.5，策略与谋行

策略是解决问题的方法与步骤，是人工智能的核心概念。人类智能和人工智能的智能水平究竟如何？关键就看解决问题的策略的水平如何。由于策略与智能具有这样密切的关系，人们就把策略看作是智能的化身，称之为智能策略。

如果给定了求解的问题，那么，问题的客体信息也就随之给定了。于是，通过感知就可以获得感知信息，通过认知就可以获得相应的知识。在此基础上，针对感知信息（它是主体对问题的感性认识），在知识的约束下（什么步骤能被采取，什么步骤不能被采取），在预设的求解目标引导下，就可以谋划解决问题的智能策略。这个过程，称为谋划求解问题的行动策略的过程，简称“谋行”。

所以，可以给出定义：**谋行，就是在目标导引下，在知识支持与约束下，把感知信息转换为智能策略的过程**。

以往的文献没有“谋行”的概念，只有“决策”的概念。不过，人们往往把决策简单地看作是“在几个备选行动方案中选取一个”的过程，忽视了这些备选的行动方案是怎样被“谋划”出来的。为了避免这种简单化片面化的误解，这里提出了“谋行”的概念，它涵盖了谋划与选择的全部过程。

2.6 主客互动与信息生态

人工智能是在人造机器上实现的人类智能，是对人类解决问题变革现实的显性智慧的模拟。因此，人工智能所扮演的角色就是“人类的代理”。而人类正是通过与环境中的客体的相互作用产生出认识世界和改造世界的智能策略和智能行为，赢得人类生存发展水平的不断改善。因此，**主客互动（主体与客体的相互作用）是人工智能研究的根本前提**。

在主客互动过程中，主体接受的是客体信息的刺激，输出的是主体产生的智能策略和智能行为。显而易见，智能必定是由信息经过复杂的转换所生成的。具体来说，在主客互动的过程中，客体信息会被主体转换为感知信息，进而转

换成为知识、智能策略和智能行为。这就是所说的信息转换，而且是**信息的生态转换，这个信息转换过程就形成了信息的生态链**：从初级的信息（现象）转换为高级的知识（本质）以至更高级的智能（策略）。

所谓信息生态，是指在保持信息内涵的整体性、时空的连续性和系统性前提下和全局优化条件下的信息转换。这样的信息转换过程，也就是信息的生态演化过程。

2.7 学科范式与科学时代

在一般的意义下，范式是指人们所遵循的“世界观和行为方式”[7]。人类一切有意识的行为都受着某种范式的支配和规范，不是受这种范式的支配和规范就是受别的范式的支配和规范，不存在范式真空的情形。范式是引领和规范人们行为的最高支配力量。

在科学研究领域，范式是指研究活动所遵循的科学观和方法论，人类的一切研究活动都受着某种范式的支配和规范，不是受这种范式的支配和规范就是受那种范式的支配和规范，不存在范式真空的情形。**范式是引领和规范研究活动的最高支配力量**。

正是在这个意义上，科恩**把范式看作是“科学革命”的指标**。

迄今，人类的科学研究活动产生了两种不同的学科：一个是发端于农耕时代的物质学科，一个是兴起于信息时代初期的信息学科。物质学科早已渗透在整个科学领域，信息学科也在快速地向所有科学领域（包括物质学科领域）渗透。

物质学科的研究对象是物质对象，信息学科的研究对象是信息对象。物质对象和信息对象的性质对立而统一，相反而相成，物质学科和信息学科所遵循的范式也是对立而统一，相反而相成。由此，**就形成了对立而统一、相反而相成的两个科学时代：物质学科主导的科学时代，信息学科主导的科学新时代。促成物质学科主导的科学时代向信息学科主导的科学新时代转换的力量，就是范式的革命。**

以上这些，就是与人工智能理论直接相关的最为基本的重要概念。

3，实现人工智能基础理论的重大突破，需要准确理解人工智能的历史

半个多世纪的人工智能发展史，**极其精彩地演绎了人工智能研究领域各种不同观念和方法之间的矛盾冲突，从中折射出了非常宝贵而且发人深省的启示**。因而，值得细加考察。

3.1 人工智能的历史：观念方法的矛盾演绎

用机器来协助人类进行劳动，是人类一直不懈追求的美好理想。不过，真正付诸实践的研究始自 20 世纪的 40 年代。那时人们相信，人类的高级认知功能定位于大脑新皮层的神经网络，认为只要把大脑神经网络的结构在机器上实现出来就可以模拟出人类的智能。这是结构模拟方法捷足先登的缘由。

具体来说，1943 年 McCulloch 和 Pitts 发表了神经元的数理逻辑模型，1949 年 Hebb 提出了神经元的学习规则。利用 M-P 模型和 Hebb 规则就可以构造人工神经元网络，开展人工神经网络及其应用研究。

人类大脑神经网络是由近千亿神经元互相复杂连接而成的大规模非线性系统，而当时的科技和工业能力都只能实现小规模的简单神经网络系统，因此结构模拟的人工智能研究虽然似乎前景光明，但道路却很曲折，进展缓慢。

面对这种情况，一批思想活跃的学者便撇开结构模拟的途径，选择了功能模拟的途径。这就是 1956 年夏天 McCarthy 等人在 Dartmouth 发起的利用计算机作为硬件平台、通过编制“聪明软件”来模拟人类逻辑思维功能的研究途径，并且创造了 Artificial Intelligence 这一术语来表征这个新的研究领域。由于那时的计算机已有很强的功能，Newell 和 Simon 等人提出了“计算机的功能与人脑功能等效”的物理符号假设，并把功能模拟的研究产物称为物理符号系统，认为这是可以像人脑一样求解通用问题的人工智能系统。果然，功能模拟方法一经问世，便在数学定理机器证明和模式识别等方面取得了令人鼓舞的成绩。

不过，当人们真的利用这一方法来解决通用问题的时候，就发现这是很不现实的想法，因为求解通用问题需要无限的知识。于是，人们不得不把面向通用问题求解的物理符号系统改成为面向专门问题求解的专家系统。

然而，解决专门问题也需要专门的知识，而知识的定位、获取、表示、推理都存在难以逾越的“知识瓶颈”困难。在这种情况下，Brooks 等人提出了行为模拟的思路。他们宣称，行为模拟方法不需要知识，只需要感知到环境对智能系统的刺激和智能系统对此所产生的动作响应（于是被称为“感知动作系统”），因此可以回避结构模拟方法的复杂性问题和功能模拟方法的知识瓶颈困难。行为模拟的典型研究成果，是 Brooks 领导的 MIT 人工智能实验室在 1990 年所演示的六脚虫爬行机器人以及随后所研制的各种智能机器人。遗憾的是，对于行为模拟的感知动作系统来说，它自身的最大问题是它仅能模拟智能系统的外在行为表现，这只是一类浅层的智能。

上述历史表明，人工智能存在结构模拟、功能模拟、行为模拟三种不同的研究方法。它们是在人工智能发展的不同阶段、面临不同的问题、分别由不同的

人群发展出来的各不相同的研究方法。它们的学术信仰（**以脑科学为模拟原型的结构主义、以认知科学为模拟思路的功能主义、以控制论为模拟原理的行为主义**）各不相同，所针对的问题各不相同，所采取的策略各不相同，所显现的能力也各不相同。虽然都是人工智能的研究方法，但却无法实现殊途同归。

实际上，三大学派非但未能殊途同归，无法形成合力，反而变为互不相容，互相排斥，势难同立，以至逞强凌弱，成为同行冤家。以下是这些学派称雄争霸的若干典型事例。

1969 年，功能主义方法的学术带头人 Minsky 和他的同事 Papert 出版学术专著《感知机》严厉抨击神经网络方法“没有科学价值（without scientific value）”，认为它的成果建在流沙之上（built on quick sand）。这种猛烈的批评，造成世界范围人工神经网络领域的研究人员大量转行，使人工神经网络的研究陷入为期十多年的“黑暗年代”。

1987 年，借着专家系统的研究遭遇“知识瓶颈”的困难而神经网络研究进入复兴阶段的机会，结构主义人工神经网络研究者们在美国加州圣地亚哥举行的 IEEE 第一届神经网络国际会议期间，与会者中爆发出了“人工智能死了，神经网络万岁（AI is dead. Long live neural network）”的强烈呐喊。表现了结构主义神经网络研究学派对功能主义专家系统研究学派的强烈反击。

1990 年前后，行为主义感知动作系统研究学派的带头人 Brooks 连续发表论文抨击正在遭受“知识瓶颈”困难的功能主义研究方法，宣称行为主义感知动作系统的研究方法可以不需要知识（Intelligence without Knowledge），因而也不需要知识的表示(Intelligence without Representation)，以此来否定功能模拟方法。

正是由于不同学派的学术主张各不妥协，使人工智能的研究始终处于某个学派一派独大的格局，从来没有出现过相互合作的局面。

●1943 年至 1956 年间，由于其它两种研究路径尚未问世，人工智能当然只存在结构主义（初期的简单人工神经网络）一种研究路径。

●1956 年至 2016 年这 60 年间，基本上是功能主义雄霸天下。虽然在 20 世纪 70 年代初期由于在机器翻译领域的失败和随后由于知识瓶颈的困扰曾经两度遭遇过发展的危机，但是，功能主义方法仍然维持着它的统治地位，因为被它打入“黑暗年代”的结构主义方法还没有完全缓过气来。

●从 2016 年至今，由于初期的简单神经网络逐渐发展成为复杂的深层结构模型，训练和学习的算法也得到显著改进，在人工智能的一些竞赛评测项目中表现出大大超越功能主义方法、甚至超越人类能力的优异性能，于是，基于结

构主义方法的深层神经网络研究变成这一时期人工智能的主导学派，基于功能主义方法的专家系统研究则几乎难觅踪影了。

3.2 人工智能历史的启示：根本问题是范式的张冠李戴且不可避免

经过 70 多年的发展，人工智能研究的结构主义（人工神经网络研究）、功能主义（专家系统）研究、行为主义（感知动作系统研究）三大学派各自都取得了不少引以自豪的成果，但也都面临严峻的挑战。

不过，三大学派取得的成绩已为世人熟知，为了节约篇幅，这里暂不展开。从推动人工智能基础理论源头创新的需要着眼，我们应当重点关注现行人工智能究竟面临的根本性问题。只有发现了这些问题，才能有针对性地解决这些问题，推动人工智能及其基础理论研究健康发展。

历史考察表明，**第一个严重的问题是：三大学派分道扬镳各自为战，无法形成统一的人工智能理论，因而也就无法建立通用的人工智能系统**。

由于没有统一的理论，现行人工智能系统的研究都严重依赖于应用场景。人们在研究实用的人工智能系统的时候，首先就需要仔细选择应用的场景，然后针对这个选定的场景设计解决问题的人工智能系统。一旦应用场景改变了，系统就要从头重新设计：一个场景一个系统，不同场景就要设计不同的系统。

曾经有人宣称，人工智能领域或许根本就不存在统一的理论，因此也不存在通用的系统；只要一个一个应用场景的人工智能系统都研制出来了，人工智能的问题就全解决了。不难看出，这种认识根本不符合辩证法和“可持续发展”的理念，也不符合系统论的基本原理。系统论原理指出：系统（整体）远远大于它的部分和。这就是说，系统的整体解决远远胜于它的所有部分解决之和，其间存在整体优化和局部优化之间的质的差别！因此，“现行人工智能无法建立统一理论”乃是一个极为严重的问题，表明人工智能的研究根本没有到位。

追根寻源，为什么现行人工智能的研究无法建立统一的理论呢？

人工智能发展的历史给出了明确的答案：作为开放复杂信息系统的人工智能研究，遵循了“分而治之”的方法论，**把人工智能的整体研究肢解成了结构主义、功能主义、行为主义三种方法分别进行研究**。“分而治之”方法论在物质学科领域非常有效，但用在人工智能领域却割断了三种方法之间的复杂而隐秘的信息联系，后者正是整体信息系统的生命线和灵魂；既然割断了三种方法之间的生命线和灵魂，当然就无法把它们恢复成为原来的复杂信息系统整体。

这就是现行人工智能的研究无法建立人工智能统一理论的根本原因。不是技术细节的原因，而是方法论的原因。而没有人工智能的统一理论，就表明现行人工智能的研究根本没有走上正确的轨道。

第二个严重的问题是：现行人工智能的智能不是在理解问题基础上的真正智能，而是快速计算能力的奇葩表现，真正的智能水平却很低下。

按照本文前面的人工智能定义，它应当是在人造机器上所实现的解决问题的显性智慧。这就表明，人工智能系统所表现的“智能”应当是在“理解问题基础上解决问题的能力”。但是，现行人工智能系统表现的“智能”其实并没有理解能力。为什么这样说呢？

这是因为：**现行人工智能的研究都是纯粹形式化的研究，阉割了问题的内容与价值因素，因此不可能理解问题**，当然也就不可能做出有智能水平的决策。比如，当你看见一只老虎，你看到了它的形态（形式信息），如果你也懂得它会伤人（价值信息），知道它是食人猛兽（内容信息），那么，你就理解了老虎，就可以做出正确的、有智能水平的决策；躲，逃，或在不得已的情况下射杀。但是如果你不知道老虎会伤人（不具有价值信息），也不知道它是食人猛兽（不具有内容信息），你就没有理解老虎这种凶猛野兽。在这种情况下，你怎么知道应当做出怎样的决策？说不定你就会“与虎共舞”，结果为虎所食。

单纯形式化是物质学科的方法论，在只关心物质结构与功能的物质学科范畴内可以畅行无阻，但是却不可能有效支持人工智能的研究。这是因为，人工智能的研究不能仅仅根据问题的形式就做决策，而必须对问题的形式、内容、价值有全面的理解才能做出合理的决策。

总之，现行人工智能存在的主要问题是（1）**整体被肢解，无法建立统一理论，表明人工智能的研究没有走上正确的轨道**，（2）**价值和内容被阉割，导致智能水平低下，同样说明人工智能的研究没有走上正确的轨道**。

总之，不能建立统一的人工智能理论，根本原因是因为人工智能的研究遵循了物质学科**“分而治之”的方法论**。而智能水平低下，根本原因则是因为人工智能的研究遵循了物质学科**“单纯形式化”的方法论**。

正如范式定义所表明的，方法论是为科学观服务的。有什么科学观，就要求采用什么样的方法论。科学观和方法论两者所构成的统一整体，就形成了主宰科学研究活动的最高引领和规范力量 — 范式。

这就很明白了：人工智能的研究之所以一直都在遵循物质学科“分而治之”和“单纯形式化”的方法，根本的原因有两个方面：一方面，人工智能的研究活动虽然已经广泛展开，但是，长期以来很少关注人工智能的最高指导思想 — 科学观和方法论（范式）。以至半个多世纪过去了，仍然没有在学科范式上形成学术共同体的明确共识。因此，根本没有信息学科的范式可以遵循。另一方面，数百年来物质学科范式的强大存在和广泛影响，使人们误以为信息学科的

研究也同样可以遵循物质学科的范式。这样两个方面原因的叠加，就造成了人工智能范式的张冠李戴。

这决不是什么人的恶意中伤或造谣惑众，而是半个多世纪以来人工智能研究领域活生生的事实。下面表 1 所列的内容，就是物质学科范式、人工智能实际所遵循的范式、信息学科范式三者的简明对照。

需要说明的是，表 1 第三行所列的“信息学科范式”是本文作者团队历经半个多世纪所研究提炼出来的结果。虽然在整个人工智能学术共同体内，很少有人关注和研究过信息学科领域的范式问题，但是由于独特的学经历，笔者从 20 世纪 60 年代起就特别关注了信息学科的科学观与方法论的探索。

表 1 人工智能范式张冠李戴的具体事实

事项	科学观	方法论
物质学科范式	**机械唯物论的物质观** 研究对象：物质客体，排除主观因素 对象性质：确定性演化，可分性 研究目的：了解对象的结构与功能	**机械还原论的方法论** 描述与分析方法：纯粹形式化 决策的判断方法：形式的匹配 宏观处置的方法：分而治之
人工智能范式	**准“物质观”** 研究对象：脑物质，排除主观因素 对象性质：存在不确定性，可分性 研究目的：了解对象的结构与功能	**真“还原论”** 描述与分析方法：纯粹形式化 决策的判断方法：形式的匹配 宏观处置的方法：分而治之
信息学科范式	**整体观：主客互动的信息观** 研究对象：主客互动的信息过程 对象性质：不确定性演化，整体性 研究目的：实现主体客体合作双赢	**辩证论：信息生态方法论** 描述与分析方法：形式内容价值一体化 决策的判断方法：内容理解 宏观处置的方法：生态演化

对比表 1 三个项目的内容可以明显看出，人工智能实际遵循的范式（表中第二行），基本上是物质学科的范式（表中第一行），而与信息学科范式（表中第三行）几乎无关。

那么，为什么会有这样的结果呢？答案很明确：历史注定，无可避免。

社会的发展存在一个铁定的法则：**社会意识滞后于社会存在**。

在科学研究领域，学科的研究活动就是一种社会存在，作为学科范式的科学观和方法论就是学科的社会意识。学科的研究活动这种社会存在迟早总要产生相应的学科范式作为自己的社会意识，而且，学科的范式一旦形成，它就要反过来影响学科的研究活动。只是，学科范式的形成必然要滞后于学科的研究活动。

学科的范式究竟需要滞后多长的时间？这没有确定的答案。一般而言，学科越是复杂，学科范式的形成（特别是在学术共同体内形成共识）所需要的时间就必然越长。比如，直到如今，信息学科的范式都没有在学术共同体内形成共识。

4，实现人工智能基础理论的重大突破，需要遵循学科发展的普遍规律

作为一个新学科，人工智能当然具有自己的个性，但是，同样重要的（如果不是更为重要的话）是必须服从共性的规律。以下的表 2 用表格的形式给出了学科发展普遍规律的简明解释。

表 2 学科发展的普遍规律

阶段进程	进程名称	进程要素	要素解释
初级阶段：自下而上的探索	**摸索**（准备）	多方试探	通过长期自下而上的多方试探摸索，总结失败教训和成功经验，提炼和确立学科的研究范式（即学科应当遵循的科学观和方法论）
高级阶段：自上而下的建构	**范式**（宏观定义）	科 学 观	宏观上明确学科的本质“是什么”
		方 法 论	宏观上明确学科的研究“怎么做”
	框架（落实定位）	学科模型	基于“学科范式”的学科全局蓝图（是什么）
		研究路径	基于“学科范式”的整体研究方法（怎么做）
	规格（精准定格）	学术结构	基于“学科范式”的学科内涵规格（是什么）
		学术基础	基于“学科范式”的学术基础规格（怎么做）
	理论（完整定论）	基本概念	基于“学科范式”的学科基本知识点（是什么）
		基本原理	基于“学科范式”的概念间相互联系（怎么做）

表 2 显示，学科的发展需要经历两个相互联系而又相互不同的阶段：首先是自下而上的探索阶段（称为初级阶段），然后才能进入自上而下的建构阶段（称为高级阶段）。

初级阶段的主要任务，是要通过各个相关学科领域的研究人员从各种不同的学术角度展开全面而漫长的摸索、试探、争论、交流、总结，才能逐渐提炼出关于学科范式（学科本质是什么？应当怎样研究？）的共识。这是一个极为艰难的探索过程。在这个阶段，往往会出现各种各样“盲人摸象”的情景。

高级阶段的主要任务，是要在初级阶段摸索得到的学科范式引领下，通过形成学科框架（学科的研究模型和研究路径）和学科规格（学术结构的规格和学术基础的规格）一步一步地把范式贯彻落实到学科的理论建构之中（包括构建学科的基本概念集合和基本原理集合）。

对照表 2 所描述的学科发展普遍规律可以看出，**当今，人工智能学科的研究确实仍然处在自下而上探索范式的初级阶段**。这是因为，今天的人工智能研究仍然处于结构主义、功能主义、行为主义三大学派互不认可（类似于盲人摸象）的状态，而且仍然处于“范式张冠李戴”的状态。为了实现人工智能基础理论的重大突破，首先就必须尽快结束这种范式张冠李戴的状态。

表 2 的规律也清楚地表明，学科的范式是引领学科发展全过程的力量：学科探索阶段的任务，是为了总结出学科的范式；学科建构阶段的任务，是为了贯彻和落实所总结出来的学科范式。所以，学科范式的引领和规范作用贯彻在学科发展的始终。有了正确的学科范式，学科的发展就有了明确的方向和灵魂，离开了正确的学科范式，学科的发展就会迷失方向，陷入泥潭。

5，范式革命，实现人工智能基础理论重大突破的方法与步骤

既然学科范式是学科发展的最高支配力量，而人工智能的范式又不可避免地陷入了张冠李戴的窘境，成了现行人工智能一切痼疾顽症的总根源，那么，人工智能范式的革命就成为解决人工智一切痼疾顽症的唯一对症良方，也是实现人工智能基础理论重大突破和源头创新的唯一正确举措。

具体来说，**所谓人工智能的范式革命，就是一方面要颠覆物质学科范式对人工智能学科的统领地位，同时要总结和确立信息学科范式对人工智能学科的整体引领，并以信息学科范式来规范人工智能各个层面的研究**。

现在就来考察：人工智能的范式革命是怎样按照表 2 所载明的学科发展普遍规律，从人工智能研究的源头开始一步一步地创建全新的一代人工智能理论 —通用人工智能理论。

5.1 第一步，解除物质学科范式对人工智能研究的约束，确立信息学科范式对人工智能研究的引领，首先阐明人工智能学科的宏观定义

如前所说，人工智能的范式革命，就是要彻底颠覆物质学科范式对人工智能的统领地位，全面确立信息学科范式对人工智能的引领和规范作用。

之所以要颠覆物质学科范式对人工智能的统领地位，是因为：

（1）按照物质学科范式的科学观，研究对象应是孤立的脑物质，排除主观因素的影响。但是，智能不可能由孤立的脑物质产生（参考狼孩的试验）。因此物质学科范式的科学观不适用于人工智能的研究。

（2）按照物质学科范式的方法论，就要贯彻分而治之和单纯形式化的方法。但是，这样两种方法只能导致人工智能理论的分裂（分而治之割断子系统之间的联系）和智能水平的低下（单纯形式化使信息、知识、智能全部空心化）。

可见，物质学科的范式确实把人工智能学科的研究完全引入了歧途。

颠覆了物质学科范式对人工智能的束缚，只是范式革命的第一步。更重要的步骤是要全面确立信息学科范式对人工智能的规范与引领。

那么，什么是信息学科的范式？

经过我们半个多世纪对信息学科的科学观及方法论的探究，它的基本特征已经被总结出来，并且表述在本文表 1 的第三行。

（1）信息学科范式的科学观认为，人工智能的研究对象不应排除主体因素的存在，人工智能的研究对象应当是主体驾驭和环境约束下的主体与客体相互作用的信息过程。

（2）信息学科范式的方法论表明，人工智能的研究不应遵循机械还原的分而治之和单纯形式化方法，人工智能的研究应当遵循信息生态方法论（即在保持信息的整体性、连续性、系统性和全局优化条件下的信息转换方法）。

这样，根据信息学科范式的科学观和方法论，我们就可以得到**人工智能学科的宏观定义（是什么？怎么做？），即：**

（1）人工智能学科的本质，是主体驾驭和环境约束下主体客体相互作用的信息过程，研究的目的是实现主体客体相互合作和互利双赢。

（2）人工智能学科的研究，应当坚持信息生态方法论。

正确的学科定义，是人工智能学科能够走上健康发展轨道的根本基础。这样，人工智能的研究就不再局限于模拟人类大脑的结构，也不是去模拟大脑的功能或者人类表现的外部行为，而是要运用信息生态方法论，研究主客互动的信息过程，实现主客双赢的规律，不断改善人类的生存发展水平。

5.2 第二步，把学科的宏观定义具体落实为学科的全局模型和研究路径

根据信息学科范式科学观所阐明的学科定义，人工智能的全局研究模型应当是**主体驾驭和环境约束下的主体与客体相互作用的信息过程模型**。图 1 所表示的，正是符合这个学科宏观定义的人工智能全局研究模型[4]。

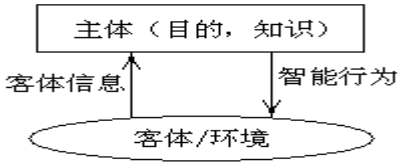


图 1 人工智能的全局模型：主体客体相互作用的信息过程

正如模型所表明的那样，人工智能学科的全局模型必然是客体信息首先作用于主体，然后，具有目的和知识的主体就设法生成智能行为反作用于客体，完成主体与客体相互作用的第一个基本回合。

如果第一回合的智能行为反作用于客体的结果达到了预设的目标，这个回合就完成了任务；如果没有达到预设目标，其中的误差就要返回到主体的输入端，成为补充性的客体信息，促使主体去补充新的知识，优化智能策略和智能行为，改进反作用的效果（缩小误差），直至满意为止。

图 1 所示的人工智能全局研究模型虽然看似简单，却清晰而准确地揭示了人工智能的深层本质：只有主体与客体发生相互作用，才会使具有目的和知识的主体产生智能行为来反作用于环境的客体，实现主体与客体的合作双赢，一方面，主体达到了自己的预设目标；另一方面，客体维护了环境的运行规律。

那种既没有外部刺激输入、又没有自身行为向客体输出的孤立大脑，实际上不可能产生真正有用的智能。即使孤立的大脑非常“聪明”，但是没有外部的刺激就没有产生智能行为的激励与动因，因而不会去生成智能行为；而没有自身的输出，就不可能检验它所产生的行为是否具有智能水平。

根据信息学科范式的科学观构筑了正确的研究模型以后，接下来要思考的问题就是：应当怎样去研究这个全局模型？

根据信息学科范式的方法论，人工智能的研究应当遵循信息生态方法论。把这个问题说得更明确一些就是：遵循信“在主客互动框架下保持信息的整体性、连续性、系统性的条件下以及在全局优化条件下的信息转换方法”。

进一步，遵循了信息生态方法论的人工智能系统，又将会怎样产生所需要的智能行为呢？

事实上，图 1 的全局研究模型已经表明，人工智能系统产生智能行为的方法只与客体信息、主体目标、主体的知识这些因素有关，即：智能行为的生成必定是在主体目标的导引下、在知识的支持与约束下、由客体信息通过复杂转换而产生出来。

由此，就可以构筑：在信息学科范式的信息生态方法论指引下智能行为生成的过程模型，如图 2 所示：

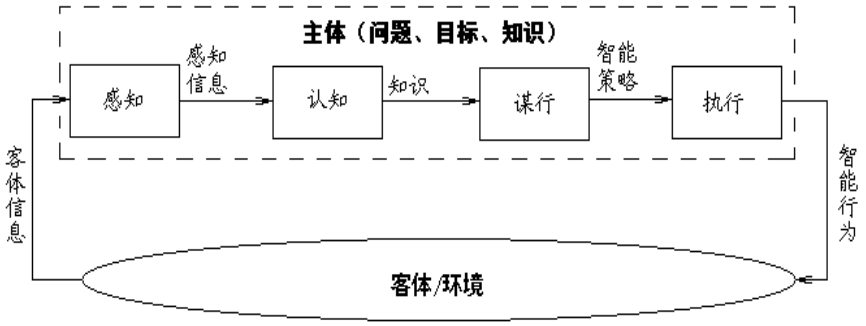


图 2 普适性智能生成机制的模型[20]

显然，图 2 只是图 1 的具体化，但具体化了的图 2 就更加明确地揭示了主体产生智能行为的详细过程和普适性机制。

具体来说，首先就要实现主体对客体的感知，从而产生主体的感知信息，接着就要把感性的感知信息提升为理性的知识才能对客体有深刻的认识，然后就可以在预设目标的引导下、在知识的支持与约束下，把感知信息转换为解决问题的智能策略，进而转换成为智能行为。这就是上述图 2 的全部结果。

图 2 所揭示的智能生成机制是普适性的，因为其中定义的所有因素（感知、认知、谋行、执行）都是普适性的。图 2 还表明，这个普适性智能生成机制的准确内涵应当是“信息转换与智能创生”原理，即（以 表示“转换”算法）：

客体信息 感知信息 知识 智能策略 智能行为

这样，信息学科范式的信息生态方法论就创造了一种前所未有的人工智能研究路径：**基于普适性智能生成机制的机制主义研究路径**，它是一种统一的（普适的）人工智能研究路径。有了这种统一的人工智能研究路径，原先的结构主义、功能主义、行为主义三者分道扬镳的研究路径就可以被放进历史博物馆！

事实上，对于人工智能基础理论研究来说，普适性智能生成机制才是人工智能的核心本质所在：系统的结构和功能都是为系统实现智能生成机制服务的。至于系统的行为，则是普适性智能生成机制实现以后的系统外部表现。

所以，这是人工智能理论的一个具有历史意义的重要结果。它给人工智能学科的研究大方向提供了一个极为重要的启示：**普适性智能生成机制是一类复杂的信息转换，因此，要把人工智能研究的整体思路聚焦到“信息转换”这个大方向上来！只要抓住了“信息转换与智能创生”这个普适性智能生成机制，就**

抓住了人工智能研究的核心本质。算法、算力、数据、知识都是为实现“信息转换与智能创生”这个普适性智能生成机制服务的。

不过，需要特别提醒：作为普适性智能生成机制的本质，“信息转换与智能创生原理”中的源头 — 信息，并不是现在广泛流行的 Shannon 信息，而是感知信息，也就是全信息。如果不了解这一点，普适性智能生成机制仍然不可能生成真正的智能，这是因为，**Shannon 信息是被单纯形式化方法阉割了内容和价值因素、只剩下形式因素的空心化信息！利用这样的空心化信息，普适性智能生成机制也只能提炼出空心化知识，生成空心化的智能！**

本文在基本概念一节已经交代了感知信息（全信息 / 内容信息）的概念，指出了感知信息具有形式信息（语法信息）、价值信息（语用信息）、内容信息（语义信息）三个分量。不过，在那里还没有来得及交代这些分量是怎样生成的，也没有交代这三个分量之间存在什么样的相互关系。

这是因为，要想更深刻地阐述感知信息的定义和生成机制，首先必须懂得信息学科的范式，即必须懂得：整个信息转换和智能创生的过程都必须“在主体驾驭和环境约束下主体与客体相互作用的框架下进行”这个大前提。而在本文第 2 节阐述基本概念的时候，还没有论述到这个前提。因此，只能等到现在这个时候，才具备了条件可以具体阐明：如何在主体驾驭下从外界的客体信息生成主体感知信息的工作机制。

在这里，“主体驾驭”和“主体与客体相互作用”的最重要标志，就是人工智能的整个信息转换与智能创生过程都必须尊重“主体目标”的导控作用，以及存在主体所提供的知识所发挥的支持与约束作用。

在此，值得特别强调的是：在人工智能系统中，必须十分重视“主体目标”的作用，它是体现主体主观能动作用和主体利益的至高无上的标志性因素。可以认为，只有那些有主体目标的人工智能系统，才会拥有真正的智能，没有主体目标的“人工智能”系统，不可能拥有真正的智能。

这是研究人工智能理论与研究一般物质系统理论之间的最大区别之一。这也是人工智能的研究不宜完全退化到物质系统的研究的一个重要原因。

在做了这些说明之后，现在就可以来构造“由客体信息转换生成感知信息”的工作机制模型，具体情况如图 3 所示。

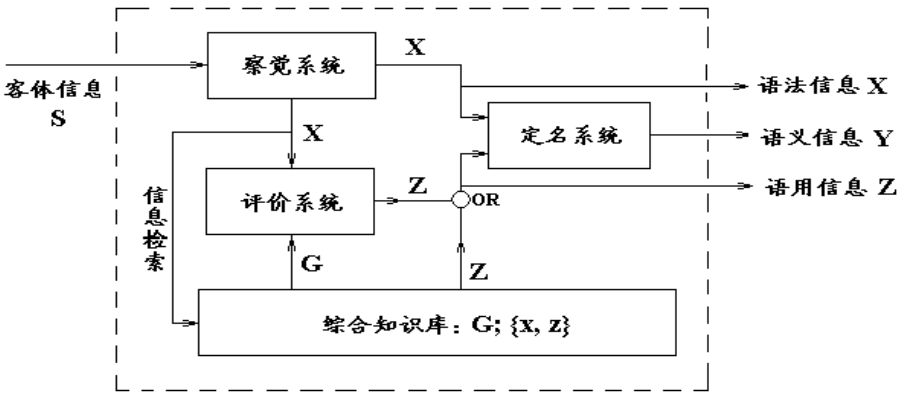


图 3 感知信息生成机制的模型

图 3 模型中，这个驾驭全局的主体目标就是存储在“综合知识库”的 G。由它来确定外来的刺激（客体信息 S）对于达成系统的目标究竟是有利？有害？还是无关？也就是确定这个客体对于系统的价值信息（语用信息）为正？为负？还是为零，从而据此确定系统究竟应当对这个客体表示欢迎（支持）？反对（抵制）？还是不予理会（过滤）？

图 3 模型中的**“察觉系统”**就是传感系统，但它只能觉察和表达系统面临着什么形式的刺激 X（形式信息／语法信息），而不可能懂得刺激的内容；模型中的**“评价系统”**就是对刺激的价值评估：这个刺激对于达成主体目标有什么利害关系 Z（价值信息／语用信息）；其中的**“定义系统”**就是把形式信息 X 和价值信息 Z 两者形成的“偶对”映射到内容（语义）空间并对映射结果命名，这样就得到了内容（语义）信息 Y。显然，模型在技术上是完全可实现的。

模型表明，感知信息的形式信息 X（语法信息）、价值信息 Z（语用信息）和内容信息 Y（语义信息）三者之间，并非相互独立或完全平等的关系，它们之间存在“Y = (X, Z)”的关系（其中的符号 表示“映射与命名”的操作），即：**内容信息乃是“把形式信息与价值信息所形成的偶对抽象而成的结果”。这是一个非常重要非常有用的结果，也是一个至今都被普遍误解了的重要结果。**

由这样生成的全信息（感知信息）按照普适性智能生成机制生成的知识和智能就将是同样具备形式、价值、内容三个份量的“全知识”和“全智能”。

5.3 第三步，进一步精准刻画学科的学术结构规格和学术基础规格

根据信息学科范式的科学观，人工智能学科的全局研究模型是“主体客体相互作用的信息过程模型”，那么，为了精准研究这种主客互动的信息过程，人工智能学科的学术结构规格也应当精细化。结果可以发现：**人工智能学科的学**

术结构应当由人工智能的原型学科（神经科学、认知科学、人文科学等）、核心学科（信息科学、系统科学、控制科学等）和基础学科（数学、逻辑学、哲学等）所形成的交叉学科群结构；而不应当仅仅是计算机学科。

同时，根据信息学科范式的方法论要求，人工智能学科的方法论应当是信息生态方法论。于是，为了描述主客互动框架下的信息生态过程 — 以信息转换与智能创生为标志的普适智能生长机制，**人工智能学科的学术基础的规格就应当满足“内涵的整体性、时空的连续性和系统性、全局的优化性”的数学理论、逻辑理论以及新型的信息哲学理论**。而不应当仅仅是传统的概率理论、刚性化的数理逻辑和传统的哲学观念。

5.4 第四步，深度改造人工智能理论的基本概念和基本原理

由于现有的人工智能理论（包括他的基本概念和基本原理）是在物质学科范式约束下形成的，肯定不能适应信息学科范式的规范。因此，深度改造人工智能的基本概念和基本原理是绝对必要的工作。

比如，物质学科范式的科学观认为研究对象是物质客体，不允许主观因素的参与。这就导致现有人工智能的基本概念（如数据、知识、智能等等）都是纯粹客观的、绝对中性的、不反映主体因素的概念，同时也导致现有人工智能的工作原理（形式计算、统计、数理逻辑等等）也都是纯粹客观的、绝对中立的。

又比如，物质学科范式的方法论是分而治之和单纯形式化，因此，现行人工智能的基本概念（数据、知识、智能等等）都是被肢解和被阉割了内容与价值因素的形式化概念，而现行人工智能的基本原理（纯形式的计算、统计、数理形式逻辑等等）也都是被肢解和被阉割了内容与价值因素的原理。

显然，如果把现有人工智能的这些基本概念和基本原理原封不动地照搬到新的人工智能理论中来，那么就意味着，信息学科范式“被截去了双腿”。换句话说，这样的范式革命是半截子的、不成功的：学科的定义（学科的科学观和方法论）、学科的定位（学科的全局模型和研究路径）、学科的定格（学科的学术结构和学术基础）都符合信息学科范式的要求，而学科理论的基本概念和基本原理却是遵循了物质学科范式的要求！

显然，这种半截子的人工智能范式革命，绝对不可能成功。因此，人工智能理论基本概念和基本原理的深度改造势在必行。

根据信息学科范式的科学观，人工智能全局研究模型是“在主体驾驭和环境约束下主体与客体相互作用的信息过程”，那么，为了研究主体与客互动的信息过程，使主体能够从形式、价值、内容上全面了解客体，**人工智能学科的基本概念也就应当是形式价值内容三位一体的全信息、形式价值内容三位一体的**

全知识、形式价值内容三位一体的全智能等基本概念。其中，价值和内容便是信息学科范式科学观所要求的主体因素的具体体现，也是信息学科范式的信息生态方法论的体现。

近些年来，从实际操作的角度，人们把纯粹形式化的“数据”看作是人工智能的粮食。这看似有一定的道理。不过，深入研究就会认识到，数据只是信息的载体，不含信息的形式数据根本无法使主体全面认识客体，因此不能启动“普适性智能生成机制”，不能启动人工智能系统的有效工作。只有携带了与系统目标相关（无论是正相关还是负相关）的信息（即感知信息的价值信息分量不为零），才能真正启动人工智能系统的工作。所以，对于人工智能的研究来说，真正重要的概念是全信息以及由全信息引出的一系列概念，而不是纯形式的数据。

同时，根据信息学科范式的方法论，人工智能学科的方法论应当是信息生态方法论，那么，为了满足“整体性、连续性、系统性和全局优化性”，**人工智能学科的基本原理就应当是符合信息生态规律的一系列信息转换并最终创生智能策略的那些原理，它们构成“信息转换与智能创生”定律。**

很有意义的是，信息学科领域的“信息转换与智能创生定律”，竟是与物质学科领域的“质量转换与物质不灭定律”以及能量学科领域的“能量转换与能量守恒定律”一样重要（甚至是更重要）的基本科学定律。着实发人深省！

5.5 第五步，最后综合形成通用的人工智能理论

最后，把人工智能范式革命及其全部链锁反应的结果综合在一起，就可以产生人工智能范式革命所创造的新一代人工智能理论 — 基于普适性智能生成机制的通用人工智能理论。

普适性智能生成机制之所以特别重要，并且成为人工智能范式革命最为核心的成果，是因为，无论面对任何问题（任何应用场景），只要能够提供“问题的充分描述、问题的求解目标和问题求解的相关知识”，那么，基于普适性智能生成机制的人工智能系统就可以生成解决问题达到目标的智能策略和智能行为。这就是通用人工智能的基本思想。

作为最终结果，可以在图 2 所示的普适性智能生成机制模型的基础上生成整个通用的人工智能理论的系统模型，而且完整地体现了表 2 所列的“学科发展普遍规律”，这就是图 4 所示的通用人工智能理论系统模型。

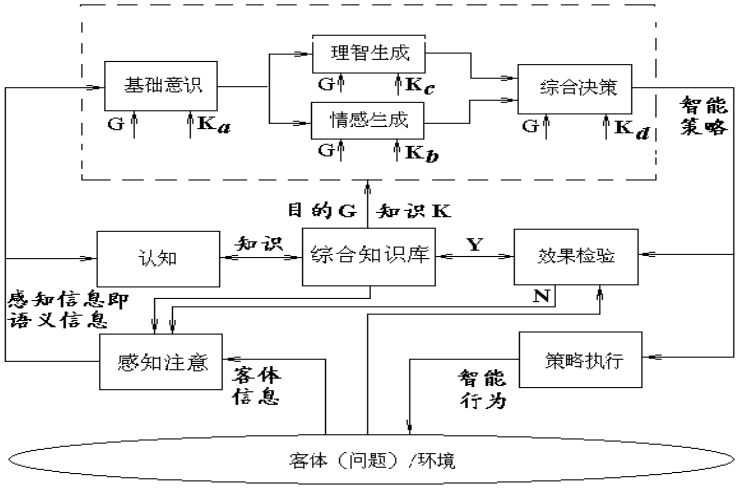


图 4 通用人工智能理论的系统模型

不难看出，图 4 仍是“主体驾驭和环境约束下主体与客体相互作用所产生的信息过程模型”，其中的客体（问题）表示为模型的底端的椭圆，模型的其他部分则是充分展开了的主体功能群，它完美地包容了图 2 的普适性智能生成机制：把客体信息转换为感知信息的感知子系统（“注意”的功能就在感知功能的基础上完成），把感知信息转换为知识的认知子系统，在目的引导下在知识约束下把感知信息转换为智能策略的谋行子系统（其中包含基础意识、情感、理智以及它们的综合决策），把智能策略转换为智能行为的执行子系统，而且包含了存储各种信息、各种知识和各种策略的综合知识库，还包含了检验智能行为反作用于客体的行为实效的效果检验子系统，以及模型中不容易直接看见但实际上存在的效果优化子系统（包括效果反馈、学习新知识、改善智能策略和智能行为从而改善智能行为的效果），最后还包括了把满足要求的智能策略存入综合知识库，从而体现了通用人工智能系统在解决问题的过程中不断学习和不断增广自己解决问题的能力的自学习子系统。

只是由于本文篇幅和专业性的限制，有关上述通用人工智能理论系统模型中的这些子系统的工作原理难以在本文尽言。有兴趣者可以参考笔者的学术专著。

到此需要指出，所谓通用人工智能，并不是指：用一个单体的“巨无霸”式的人工智能系统去包打天下，去解决世间所有的问题。我们所说的**通用人工智能理论，是指以不变的（普适性的）智能生成机制去成功应对千变万化的实际问题**。这里的通用是指智能生成机制的通用，而通用人工智能系统的输入内容（问题、目标、知识）和输出内容（智能策略和智能行为）则将随着问题的改变而相应改变。这就是**“以不变应万变”**的真实含义。

具体来说，给定任何合理的“求解的问题、求解的目标、求解问题所需要的先验知识”（所谓“合理”，是指：所给定的“求解问题”至少在理论上确实存在着“求解目标”，或者说给定的“求解目标”一定存在），通用人工智能

理论都可以凭借它的普适性的（不变的）智能生成机制，在给定的目标引导下，利用给定的知识（在先验知识不足的情况下也可以通过学习来扩展知识），生成能够求解问题达到求解目标的智能策略和智能行为。当然，如果求解的问题改变了（这意味着求解目标和所需要的知识也会相应改变），解决问题的策略和智能行为也就会随之而改变，但是，生成这种智能策略和智能行为的机制却永远不会改变：依然是“信息转换与智能创生定律”。

换句话说，“普适性的智能生成机制”是人工智能理论研究中的不变性和不变核。这样，面对千姿百态的问题，不再必须对每个不同的问题都推倒重来从头做起去设计专用系统，只需要把各自的“问题、目标、知识”表达成一定的规格，利用同样的普适性的智能生成机制就可以生成解决相关问题的智能策略和智能行为，达到求解问题的目的。

总之，**通用人工智能系统乃是“人工智能系统的通用孵化平台”。通用人工智能理论的“机制通用性和理论统一性”，是人工智能范式革命带来的巨大优越性，现行人工智能理论无法望其项背**。

同样十分重要的问题是，由于现行人工智能遵循了“单纯形式化”的物质学科范式方法论，使得现行人工智能系统的“智能水平十分低下”，成为现行人工智能理论的又一项痼疾顽症。现行人工智能系统表显出来的“智能”，其实都不是基于对问题的理解来实现的，而是利用快速运算变幻出来的结果。这样设计出来的“智能”严格依赖于特定的应用场景，场景一变，这种智能便会失效。

从通用人工智能理论的图 2 和图 4 的模型则可看出，在这里，客体信息经过感知就被转换成为了感知信息。根据本文第 2 节的概念定义，感知信息是形式信息（语法信息）、价值信息（语用信息）、内容信息（语义信息）的三位一体；知识也是形式知识、价值知识、内容知识的三位一体；智能策略也是形式策略、价值策略、内容策略的三位一体。因此，信息、知识、策略都是可以理解的。

当然，人们不应当完全按照“人类的理解能力水平”来要求“人工智能机器的理解能力水平”。但是，人工智能的“理解能力”仍然能使人工智能机器做出具有智能水平的决策。

比如，在信息的层面上，系统可以根据感知信息的价值信息分量的大小对客体做出基于理解的明智决策：

1）如果价值信息为足够大的正值，就采取相应的措施来保护客体。

2）如果价值信息为足够大的负值，就采取相应的措施来反制客体。

3）如果价值信息为足够小的正值或负值，就不理会或过滤掉这样的客体，

4）在多中选一的情况下，就选择其中价值信息为最大正值的客体，或者选择其中价值信息为最小负值的客体。

在知识和策略的层面，也同样可以针对所面临的问题做出与信息层面类似的具有一定智能水平的明智决策。而且，无论在信息、知识、智能的层面，这样所做出的决策，都可以清晰而准确地得到解释。

所以，**人工智能范式革命为基于普适性智能生成机制的通用人工智能理论带来的另一个巨大优越性，就是它的理解能力和在理解基础上的智能决策水平。与通用人工智能理论的这种智能水平相比，现行人工智能理论的理解水平更是望尘莫及。**

以上所述，就是人工智能范式革命的学术内涵、实现途径（表现为表 2 所示的学科发展普遍规律）和重大学术意义，也是人工智能范式革命所带来的人工智能理论划时代的深刻革命。

表 3 是对上述论证的简明总结。

表 3 人工智能范式革命创造了通用人工智能理论

对比项目	物质学科范式下的人工智能	**两个科学时代的分水岭**	信息学科范式下的通用人工智能
科 学 观	机械唯物的物质观：非主观，结构		对立统一的信息观：主客互动，双赢
方 法 论	机械还原论：纯形式，分而治之		信息生态学：整体化，生态演化
全局模型	孤立的脑模型		主体驾驭的主客互动的信息过程模型
研究路径	结构、功能、行为模拟分道扬镳		基于智能生长机制的机制主义路径
学术结构	计算机		原型—核心—基础等多学科交汇
基础特色	概率论，形式逻辑		因素空间数学理论，泛逻辑理论
基本概念	形式数据、形式知识、形式智能		全信息、全知识、全智能等
基本原理	未做总结		信息转换与智能创生定律
综合结果	三个独立的局部理论		通用人工智能理论

值得指出，表 3 显示了，在实施人工智能范式革命之前，“三分天下”的现行人工智能理论乃是物质学科主导的科学时代的人工智能理论；而在实施人工智能范式革命之后，“通用人工智能理论”则是信息学科主导的科学新时代的人工智能理论。**人工智能的范式革命，是这两个时代人工智能理论的分水岭。**

正是在这个意义上，**人工智能的范式革命，乃是推动物质学科主导的科学时代转变到信息学科主导的科学新时代的引擎和桥梁，也是划分物质学科主导的科学时代与信息学科主导的科学新时代的界限和分水岭。这样，人工智能范式革命的产物 — 基于普适性智能生成机制的通用人工智能理论就名副其实地具有了划时代的意义。**

这就是人工智能范式革命及其链锁反应所带来的人工智能基础理论的具有划时代意义的重大突破和源头创新。

致谢

本文是笔者数十年来在国家自然科学基金（包括，但不限于 68872014, 69171023, 69982001, 60496327, 60575034, 60873001, 70711120412 等）以及国家社科基金（18ZDA027）资助下所获得的系统性研究结果。本文的研究还得到陆汝钤院士、李衍达院士、陆建华院士、谭铁牛院士、蒲慕明院士、涂序彦教授、何华灿教授、汪培庄教授、史忠植教授、韩力群教授、王小捷教授、周延泉副教授、李蕾副教授、李睿凡副教授、陈志成博士、孙健博士等多种形式的帮助。

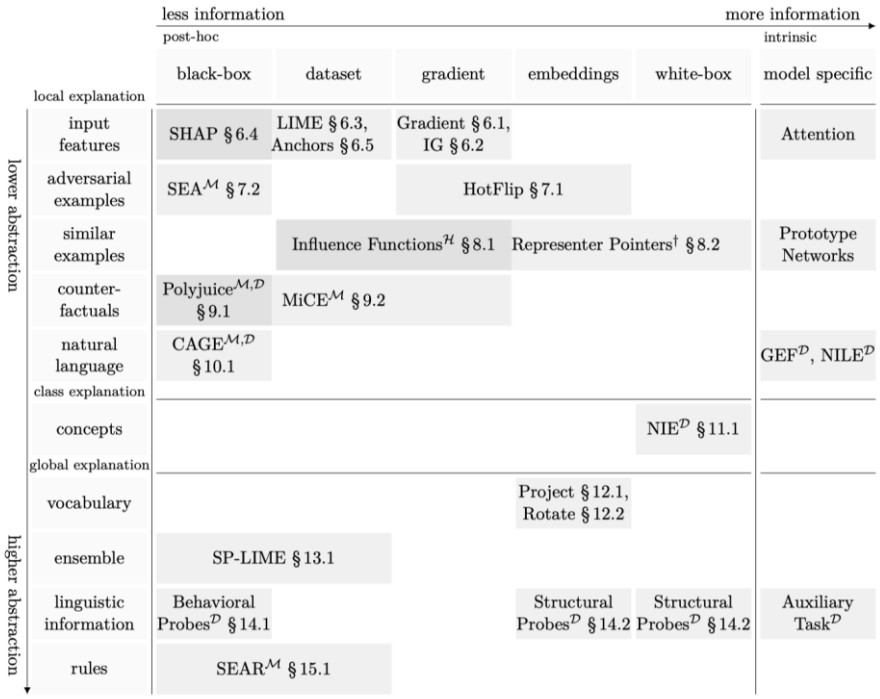
在此谨表衷心的感谢！

756，NLP 的事后可解释性研究综述

Mila-Quebec 人工智能研究所、脸书 AI 数据集、加拿大 AI 数据集，2021.8.10

自然语言处理（NLP）模型变得越来越复杂和广泛。随着神经网络的发展，人们越来越关注模型的可解释性。因此，本综述对可解释性方法进行了分类，并深入讨论这些方法的原理和特点。本综述侧重于调研事后（Post-hoc）可解释性方法，此类方法在模型训练完成后才提供解释，而模型通常是不可知的。

下表是事后可解释性方法的总结，其中§指出了该方法在文章的哪一部分中被讨论。行描述了如何进行解释，而列描述用于生成解释的信息。行和列分别按照抽象级别和信息量进行排序。



作者就 NLP 可解释性的未来研究方向进行了展望：

（1）衡量可解释性。目前可解释性的衡量方式各不相同。一般情况下，每篇论文都会介绍其衡量可解释性的方式。这降低了研究的相互可比较性。

（2）class explanations。关于方法本身，已经有很多研究。然而，在局部解释和全局解释之间，class explanations 仍然是一个未充分体现的中间地带。

（3）将事后方法与内在方法相结合。大多数内在方法（intrinsic methods）并不是纯粹的内在方法，它们通常有一个中间的表示，它

在本质上是可以解释的。但是，生成此表示通常使用黑盒模型。因此，如果要理解整个模型，就需要事后解释方法。

757，基于结构匹配的可解释深度度量学习研究

中国 Tsinghua 大学，2021.8.13

神经网络如何区分两幅图像？理解深度模型的匹配机制对于开发可靠的智能系统以用于许多危险的视觉应用（如监视和访问控制）至关重要。然而，现有的深度度量学习方法大多通过比较特征向量来匹配图像，忽略了图像的空间结构，因而缺乏可解释性。在本文中，提出了一种更透明的嵌入学习的深度可解释度量学习（DIML）方法。与传统的基于特征向量比较的度量学习方法不同，本文提出的结构匹配策略，通过计算两幅图像的特征映射之间的最佳匹配流来显式对齐空间嵌入。此方法使深度模型能够以更人性化的方式学习度量，其中两幅图像的相似性可以分解为若干部分相似性及其对整体相似性的贡献。此方法是模型不可知的，可以应用于现成的主干网和度量学习方法。本文在深度度量学习的三个主要基准（包括 CUB200-2011、Cars196 和斯坦福在线产品）上评估了该方法，并在更好的可解释性上实现了对流行度量学习方法的重大改进。

758，具有相变记忆突触的 SRNN 的在线训练

香港大学，2021.8.4

SRNN 因其丰富的时间动态和稀疏处理而成为解决各种复杂认知和运动任务的有效工具。然而，由于即使在权重分辨率有限的情况下，仍然

缺乏本地的、硬件友好的学习机制来解决时间信用分配问题并确保稳定的网络动态，在专用的神经拟态硬件上训练 SRNN 仍然存在着诸多争议。如果使用忆阻器件进行内存计算来解决冯诺依曼瓶颈问题，将会显著增加 SRNN 的计算和工作内存的可变性。为了应对这些挑战并在忆阻 SRNN 中实现在线学习，本文提出了一个基于准确和全面的相变存储器 (PCM) 设备模型的差分架构交叉阵列的仿真框架，使用最近提出的电子道具学习规则训练了一个 SRNN，并利用模拟框架模拟其权重。尽管 e-prop 局部接近理想的突触更新，但由于大量的 PCM 非理想性，很难在忆阻基板上实现更新。同时，本文还比较了几种广泛适用的权重更新方案，这些方案主要旨在应对这些设备的非理想情况，并证明累积梯度可以实现在忆阻基板上在线和有效训练 SRNN。

759，用于鲁棒视觉对象跟踪的多域协作特征表示

大连理工大学，2021.8.10

联合利用多个不同但互补的域信息已被证明是执行鲁棒对象跟踪的有效方法。本文侧重于有效地表示和利用来自帧域和事件域的互补特征，以提高挑战场景中的对象跟踪性能。具体来说，我们提出了通用特征提取器（CFE）来从 RGB 域和事件域中学习潜在的通用表示。为了学习这两个域的独特特征，我们利用基于脉冲神经网络的事件唯一提取器 (UEE) 来提取事件域中在某些具有挑战性的条件下可能在 RGB 中遗漏的边缘线索，以及用于 RGB 的唯一提取器(UER) 基于深度卷积神经网络提取 RGB 域中的纹理和语义信息。在标准 RGB 基准和真实事件跟

踪数据集上的大量实验证明了所提出方法的有效性。我们展示了我们的方法优于所有比较的最先进的跟踪算法，并验证基于事件的数据是在具有挑战性的场景中进行跟踪的有力线索。

760，使用逆 HopfFibration 的知识图谱表示学习

海得拉巴印度理工学院，德国 CereuceGmbH、Zerotha 研究所、亚琛工业大学，美国代顿大学，德国高盛

最近，已经设计有几种知识图谱嵌入（KGE）方法来表示密集向量空间中的实体和关系，并用于下游任务，例如链接预测。一些 KGE 技术解决了可解释性问题，即将关系的连接模式（即对称/不对称、逆和组合）映射到几何解释，例如旋转。其他方法对更高维空间中的表示进行建模，例如四维空间 (4D)，以增强推断连接模式（即表达能力）的能力。然而，在 4D 空间中建模关系和实体通常以可解释性为代价。

本文提出了 HopfE，这是一种新颖的 KGE 方法，旨在实现四维空间中推断关系的可解释性。作者首先对 3D 欧几里得空间中的结构嵌入进行建模，并将关系算子视为 SO(3) 旋转。接下来，使用逆 Hopf Fibration 将实体嵌入向量从 3D 空间映射到 4D 超球面，其中作者嵌入了来自 KG 本体的语义信息。因此，HopfE 在不失去表达性和可解释性的情况下考虑了实体的结构和语义属性。作者在四个著名基准上的实证结果实现了 KG 完成任务的最先进性能。

关于未来研究方向，本文的结果开辟了重要的研究问题：1）在四个维度的空间中，如何更优化地利用语义属性对链接预测性能产生积极影响，并诱导出语义类层次结构。 2）如何利用属性（或描述逻辑）上的逻辑来进行 4D 中 KGE 模型的细粒度推理和逻辑可解释性。

761，利用商业数据科学价值：确保解决方案的可解释性和公平性

https：//arxiv.org/abs/2108.07714

本文介绍了人工智能中公平性和可解释性（XAI）的概念，旨在解决复杂的商业问题。为了公平，作者讨论了导致偏见的细节，以及相关的缓解方法，最后提出了一套在数据驱动的组织中引入公平的方法。此外，对于 XAI，作者审核了特定的算法和演示性的业务用例，讨论了大量的质量量化技术，并概述了未来的研究途径。

762，神经网络模型的记忆容量

哥伦比亚大学，2021.8.17

记忆是一种复杂的现象，涉及到几种不同的机制。这些机制在不同的空间和时间层面上运作。本章重点介绍了理论框架和数学模型，这些模型是为了理解这些机制是如何被编排来存储、保存和检索大量记忆的。特别是，本章回顾了关于记忆容量的理论研究，研究人员在研究中估计了可存储记忆的数量如何与神经回路中神经元和突触的数量成比例。记忆容量取决于突触的复杂性、表征的稀疏性、记忆之间的时空相关性以及记忆提取的具体方式。当突触只能以有限的精度进行修改时（如生物突触），复杂性很重要，而稀疏性可以极大地提高记忆

容量，在记忆结构（相互关联）时尤其有益。本章讨论的理论工具可以用来确定记忆存储、保存和检索的重要计算原理，并为设计和解释记忆实验提供指导。

763，知识库上问答的生成关系连接

IBM 研究所 2021.8.16

关系链接对于在知识库上回答问题至关重要。尽管有各种各样的努力来提高关系链接性能，但目前最先进的方法并没有达到最佳效果，因此，对整体端到端问答性能产生了负面影响。在这项工作中，我们提出了一种新的关系链接方法，将其作为一个生成问题来构建，以便于使用预先训练好的序列到序列模型。我们将这种序列模型扩展到序列模型，其思想是注入来自目标知识库的结构化数据，主要是使这些模型能够处理知识库的细微差别。此外，我们训练模型的目的是生成由参数关系对列表组成的结构化输出，从而实现知识验证步骤。我们将我们的方法与来自 DBpedia 和 Wikidata 的四个不同数据集上的现有关系链接系统进行了比较。我们的方法报告了与最新技术相比的巨大改进，同时使用了一个更简单的模型，可以轻松地适应不同的知识库。

764，基于神经数据转换器的神经群体活动表示学习

Emory 大学，Georgia 技术研究所，2021.8.2

从理论上讲，神经种群活动反映了一种潜在的动力学结构。使用具有显式动力学的状态空间模型，例如基于递归神经网络（RNN）的状态空间模型，可以准确地捕获这种结构。然而，使用递归显式地建模动力学需要对数据进行顺序处理，从而减慢了实时应用程序（如脑-机接口）的速度。这里我们介绍神经数据转换器（NDT），一种非重复性的替代方法。我们通过将 NDT 应用于具有已知动力学的合成数据集以及 RNNs 模拟的到达任务期间来自猴子运动皮层的数据，测试 NDT 捕捉自主动力系统的能力。NDT 对这些数据集以及最先进的循环模型进行建模。此外，它的非重复性允许 3.9ms 的推断，在实时应用程序的循环时间内，比数据集上的重复基线快 6 倍以上。这些结果表明，一个显式的动力学模型是没有必要的自治神经种群动力学模型。代码链接：this https URL

765，2020 年 6 月，COPU 主办《第 15 届开源中国开源世界 高峰论坛》，邀请 IBM 副总裁 Todd Moore 在会上作“可信 任人工智能(反欺诈、可解释、公平性)”的报告，从此 至今，COPU 已收到全球研发可解释性人工智能的跟帖 48 件。但由于全球人工智能技术(XAI)尚未完全成熟，在研发 XAI 算法时，专家对各道演算程序的理解和操作具有不 确定性，最后评估还只能靠人工，所以 XAI 演算结果或算 法可能有出入，致使可解释机器学习难以推广应用。为 此，COPU

要求 IBM Todd Moore 和人工智能研究所的 CTO Animesh 对 XAI 举出具体案例并进行解析和说明，对我们提 出的 8 个问题进行逐个解答:

①IBM 列出研发 XAI 的具体案例是什么?

②选用下列哪种方法进入运算?

- 可直接解释(内在解释)
- 事后解释
- 全局(模型级)可解释性
- 局部(实例级)可解释性

③选择什么工具?

- 如:决策树、规划库、块择表等

④如何捕捉特征?

⑤如何建模?

⑥如何找到算法?

⑦如何进行评估?

⑧不但要导出本案例结果，还要使 XAI 在使用中确定是否能保持信任、公正、透明和可解释?!

766，工程设计的可解释人工智能：系统工程和基于组件的深度学习的统一方法

2021.8.29

由机器学习创建的数据驱动模型在设计和工程的所有领域都变得越来越重要。它们在帮助决策者创造具有更好性能和可持续性的新型人工制品方面具有很大潜力。然而，这些模型的有限泛化和黑盒特性导致了有限的可解释性和可重用性。这些缺点严重阻碍了工程设计的采用。为了克服这种情况，作者提出了一种基于组件的方法，通过机器学习（ML）创建部分组件模型。这种基于组件的方法将深度学习与系统工程（SE）相结合。通过节能建筑设计的实例，作者首先通过准确预测不同于训练数据的具有随机结构的设计的性能，证明了基于组件的方法的泛化能力。其次，作者通过局部抽样、敏感性信息和来自低深度决策树的规则以及从工程设计角度评估这些信息来说明可解释性。可解释性的关键在于，组件之间接口处的激活是可解释的工程量。通过这种方式，分层组件系统形成了一个深度神经网络（DNN），该网络直接集成了工程可解释性信息。组成组件中的大量可能配置允许使用可理解的数据驱动模型检查新的看不见的设计案例。通过相似的概率分布匹配组件的参数范围，可以生成可重用、通用性好且可信的模型。该方法使模型结构适应系统工程和领域知识的工程方法。

767，在认知交易的计算架构中集成启发式和学习

Molise 大学，2021.8.27

近年来，人工智能在图像分析、自然语言理解和战略游戏等领域的成功引起了金融界的兴趣。具体而言，对于人工代理（称为机器人交易

者）的创建，人们有很高的期望和正在进行的工程项目，这些人工代理能够用经验丰富的人类交易者的技能操纵金融市场。撇开明显的经济影响不谈，这无疑是一个具有重大科学意义的领域，因为这样的真实环境对人工智能技术的使用构成了挑战。正因此，我们必须意识到，能够在这样的水平上运行的人工智能体不仅指日可待，而且不会有简单的答案，而是各种技术和方法的共同作用才能使这项工作取得成功。在这篇文章的过程中，我们回顾了有效的机器人交易者设计中固有的问题以及相应的解决方案，考虑到将机器人交易的当前技术状态提升到下一个智能水平的总体目标，我们称之为认知交易。我们的方法的关键是将两个方法学和技术方向结合起来，尽管这两个方向都深深扎根于人工智能的学科领域，但迄今为止，它们已经走上了各自的道路：启发式和学习。

768，脑机接口运动想像分类模型——一种稀疏群滤波器组表示模型

Beihang 大学、Lanzhou 大学、Wuyi 大学、中科院，2021.8.27

公共空间模式（CSP）已广泛用于运动想象（MI）脑电图（EEG）记录的特征提取和脑-机接口（BCI）应用的 MI 分类。BCI 通常需要相对较长的 EEG 数据才能进行可靠的分类器训练。更具体地说，在使用一般空间模式进行特征提取之前，使用来自两个不同类别的训练字典来构造复合字典矩阵，并且将测试样本在滤波器带中的表示估计为字典矩

阵中的列的线性组合。新方法：缓解频带间稀疏小样本（SS）问题。针对 BCI 系统中的运动图像，提出了一种新的稀疏群滤波器组模型（SGFB）。结果：我们通过基于与非零相关系数对应的类别表示残差来执行任务。此外，我们还在多任务学习框架下，在三个不同的时间窗口中使用约束滤波器带进行联合稀疏优化，以提取鲁棒的 CSP 特征。为了验证该模型的有效性，我们在 BCI 竞赛的公共脑电数据集上进行了实验，并与其他竞赛方法进行了比较。与现有方法的比较：不同子带的良好分类性能证实了我们的算法是改善基于 MI 的 BCI 性能的一个有希望的候选方法。

769，基于梯度激活图（GAM）的可解释视觉相似性与分类

以色列开放大学和微软，巴伊兰大学，微软和特拉维夫大学，2021.9.3

本文提出了梯度激活图（GAM）——一种解释视觉相似性和分类模型预测的机制。通过从多个网络层收集局部梯度和激活信息，与现有替代方案相比，GAM 提供了改进的视觉解释。详细解释了 GAM 的算法优势，并通过经验进行了验证，其中表明 GAM 在各种任务和数据集上的表现优于其替代方案。

770，工程设计的可解释人工智能：一种结合系统工程和基于组件深度学习的统一方法

柏林大学，2021.8.29

由机器学习创建的数据驱动模型在设计和工程的所有领域都变得越来越重要。它们在帮助决策者创造具有更好性能和可持续性的新型人工制品方面具有很大潜力。然而，这些模型的有限泛化和黑盒特性导致了有限的可解释性和可重用性。这些缺点严重阻碍了工程设计的采用。为了克服这种情况，作者提出了一种基于组件的方法，通过机器学习（ML）创建部分组件模型。这种基于组件的方法将深度学习与系统工程（SE）相结合。通过节能建筑设计的实例，作者首先通过准确预测不同于训练数据的具有随机结构的设计的性能，证明了基于组件的方法的泛化能力。其次，作者通过局部抽样、敏感性信息和来自低深度决策树的规则以及从工程设计角度评估这些信息来说明可解释性。可解释性的关键在于，组件之间接口处的激活是可解释的工程量。通过这种方式，分层组件系统形成了一个深度神经网络（DNN），该网络直接集成了工程可解释性信息。组成组件中的大量可能配置允许使用可理解的数据驱动模型检查新的看不见的设计案例。通过相似的概率分布匹配组件的参数范围，可以生成可重用、通用性好且可信的模型。该方法使模型结构适应系统工程和领域知识的工程方法。

771，利用 PrMnO3RRAM 中的电热时间尺度来构建一个模拟生物神经运作模式的紧凑、无时钟的神经元

印度孟买 IIT，2021.9.3

尖峰神经网络 (SNN) 近些年在神经拟态计算领域得到了飞速发展，SNN 结合生物神经元动力学，包括代表神经网络内不同大脑活动的复

杂尖峰模式，利用神经元和突触集成的网络系统模仿人脑来提供更高的计算效率。神经元的早期硬件实现由于电路设计中的大电容器而占用大量面积，并且在设备级别只能用时钟神经元进行演示。为了更逼真地模拟生物神经元尖峰行为，新兴的忆阻器件被认为是很有前途的替代品。本文提出了基于 PrMnO3(PMO) -RRAM 设备的神经元。紧凑型 PMO RRAM 器件的压控电热时间尺度取代了为大电容器充电的电气时间尺度。电热时间标度用于实现具有多个电压控制时间标度的积分块以及耐火块以生成生物神经元动力学。本文首先展示了热器件模型的 Verilog-A 实现，它捕获了 PMO 器件的电流-温度动态。其次，驱动电路旨在模拟皮层神经元的不同尖峰模式，包括内在爆发 (IB) 和颤振 (CH)。第三，模拟神经元电路模型，其中包括 PMO RRAM 器件模型和用于演示异步神经元行为的驱动电路。最后，进行了硬件-软件混合分析，其中 PMO RRAM 设备经过实验表征以模拟神经元尖峰动力学。这项工作为大规模 SNN 提供了一种可实现且更具生物学可比性的硬件效率解决方案。

772，通过动态视觉传感器的噪声滤波器增强尖峰神经网络抵抗对抗性攻击的分析和设计方法

维加纳技术大学，机，计算机工程学院，2021.9.3

尖峰神经网络 (SNN) 旨在在具有基于事件的动态视觉传感器 (DVS) 的神经形态芯片上实施时提供节能的学习能力。 本文研究了 SNN 对此类基于 DVS 的系统的对抗性攻击的鲁棒性，并提出了 R-SNN，这

是一种通过有效的 DVS 噪声过滤来增强 SNN 的新方法。 我们是第一个对 DVS 信号（即时空域中的事件帧）产生对抗性攻击，并为 DVS 传感器应用噪声滤波器以防御对抗性攻击的人。 我们的结果表明，噪声滤波器有效地防止了 SNN 性能下降。 我们实验中的 SNN 在不同对抗性威胁模型下对 DVS-Gesture 和 NMNIST 数据集提供超过 90% 的准确率。

773，具有动态知识图谱的交互式机器理解

加拿大蒙特利尔微软研究院，柏林自由大学，2021.8.31

交互式机器阅读理解 (iMRC) 是一种机器理解任务，其中知识源是部分可观察的。 代理必须按顺序与环境交互以收集必要的知识以回答问题。本文作者假设图表示是很好的归纳偏差，可以在 iMRC 任务中作为代理的记忆机制。 作者探索了四种不同类别的图，研究了可以在不同级别捕获文本信息的不同类别的图结构。 作者描述了在信息收集期间动态构建和更新这些图的方法，以及在 RL 代理中编码图表示的神经模型。 iSQuAD 上的大量实验表明，图形表示提供了跨设置的一致改进，可以显着提高 RL 代理的性能。

774，将特定领域的异构知识整合到统一表示中的预训练语言模型

清华大学，华为诺亚方舟实验室，2021.9.2

现有技术从不同的角度扩展了 BERT，例如设计不同的预训练任务、不同的语义粒度和不同的模型架构。很少有模型考虑从不同的文本格

式扩展 BERT。在本文中，作者们提出了一个异构知识语言模型（HKLM），一个统一的预训练语言模型（PLM），适用于所有形式的文本，包括非结构化文本、半结构化文本和结构良好的文本。为了捕捉这些多格式知识之间的对应关系，作者们的方法使用掩码语言模型目标来学习单词知识，使用三重分类目标和标题匹配目标分别学习实体知识和主题知识。为了获得上述多格式文本，作者们在旅游领域构建了一个语料库，并在 5 个旅游 NLP 数据集上进行了实验。结果表明，作者们的方法优于仅使用 1/4 数据的纯文本预训练。作者们将发布代码、数据集、语料库和相关知识图谱。

775，可解释人工智能的反事实评估

（美）罗格斯大学、阿里巴巴，2021.9.5

尽管近年来机器学习中出现了各种可解释的方法，但解释在多大程度上真正代表了模型预测背后的推理过程，即解释的可信度仍然是一个悬而未决的问题。衡量可信度的一种常用方法是基于擦除的标准（erasure-based criteria）。基于擦除的标准虽然简单，但不可避免地会引入偏差和伪影。因此，作者提出了一种新的方法，从反事实推理的角度来评估解释的真实性。反事实评估过程是根据以下两个直观的观察结果精心设计的：（1）即使是最重要特征上的微小扰动也可能影响模型预测；（2）除非最不重要特征的扰动足够大，否则它们不会对模型预测产生太大影响。基于这两个观察结果，作者提出了一个新的框架，利用反事实的概念来评估解释的真实性。在反事实评

估过程中，我们的目的是想要知道，如果我们改变了输入的特征，模型的预测会发生怎样的变化。作者提出了两种不同的算法，分别在处理离散或连续输入条件下找到适当的反事实，然后使用获得的反事实来衡量可信度。在多个数据集上的实证结果表明，与现有指标相比，作者提出的反事实评估方法能够实现与 ground truth 的最高相关性。

776，用于序列学习的具有改进的固有循环动力学的尖峰神经网络

普渡大学，2021.9.4

具有 LIF 神经元的 SNN 以事件驱动的方式运行，并存储内部状态以随时间保留信息，为在边缘设备上的节能神经拟态计算提供机会。然而，许多关于 SNN 的代表性工作并没有完全证明其固有的循环（保留过去信息的膜电位）对于顺序学习的有效性。大多数作品通过速率编码及时人为扩展输入表示来训练 SNN 识别静态图像。本文展示了 SNN 可以针对顺序任务进行训练，并提出对 LIF 神经元网络的修改，使内部状态能够学习长序列并使其固有的递归能够适应梯度消失问题。然后，本文呢作者还开发了一个训练方案来训练所提出的具有改进的固有递归动力学的 SNN，该方案允许尖峰神经元产生多位输出（与二进制尖峰相反），这有助于减轻尖峰神经元激活函数的导数与用于克服尖峰神经元不可微性的替代导数之间的不匹配。实验结果表明，在 TIMIT 和 LibriSpeech 100h 数据集上提出的 SNN 架构产生的精度与 LSTM 相当（分别在 1.10% 和 0.36% 以内），但参数比 LSTM 少 2 倍。与 GRU 相比（GRU 通常被认为是 LSTM 的轻量级替代品），

稀疏 SNN 输出还分别在 TIMIT 和 LibriSpeech 100h 数据集上节省了 10.13 倍和 11.14 倍的乘法运算。

777，强化学习在时态知识图谱预测中的应用

华中科技大学、慕尼黑大学、西门子公司，2021.9.9

时间知识图谱（TKG）推理是近年来引起越来越多研究兴趣的一项关键任务。现有的大多数方法都侧重于对过去时间戳进行推理以完成缺失的事实，而在已知 TKG 上进行推理以预测未来事实的工作很少。与完成任务相比，预测任务更难，面临两个主要挑战：（1）如何有效地对时间信息进行建模以处理未来的时间戳？ (2) 如何进行归纳推理来处理随时间出现的先前看不见的实体？为了应对这些挑战，作者们提出了第一种用于预测的强化学习方法。具体来说，代理在历史知识图谱快照上旅行以搜索答案。作者们的方法定义了一个相对时间编码函数来捕获时间跨度信息，作者们设计了一种基于狄利克雷分布的新颖的时间形奖励来指导模型学习。此外，作者们为看不见的实体提出了一种新的表示方法，以提高模型的归纳推理能力。作者们在未来的时间戳评估作者们用于此链接预测任务的方法。与现有的最先进方法相比，在四个基准数据集上的大量实验证明了显着的性能改进，同时具有更高的可解释性、更少的计算和更少的参数。

778，陆首群&Daniel 议论语音识别 Kaldi 新版

2021 年 9 月 13 日，COPU 听取小米集团首席语音科学家 Daniel Povey 关于语音识别 Kaldi 新版的报告，并就 Kaldi 及其未来发展远景进行讨论。

COPU 陆主席首先指出，Daniel，你是语音科学大师，你率领的团队研发的语音识别 Kaldi 最近发布新版，首先我向你祝贺！并请你简要介绍 Kaldi 这几年在国内外发行及应用情况，请你谈一下 Kaldi 与其他语音识别技术相比，有什么特点和优势？以及 Kaldi 发展远景是什么？Daniel 谈，Kaldi 是开源的语音识别工具，集成了各种语音识别模型，包括隐马尔可夫和深度学习神经网络，被认为世界语音识别框架的基石。新版 Kaldi 由 Lhotse（训练数据准备部分，设计通用灵活的接口，以适应语音识别、文本转语音等任务，引入 AudioCuts 概念，降低数据存储空间）、Icefall（训练脚本集合，降低用户学习成本）、K2（新版 Kaldi 核心，将加权有限状态转换器和相关算法集成到 PyTorch 和 TensorFlow）三部分组成，服务于小米的“手机＋AIoT 双引擎战略”。

陆问：据说 Kaldi 识别准确率可达 95%～97%，是否过高？！Daniel 认为，准确率与原始数据有关（不同数据有不同准确率）。陆谈：Kaldi 新版是基于 PyTorch 框架的，PyTorch 创始人 Soumith Chintala 曾指出，2020 年 AI 社区将用更多度量指标衡量 AI 模型的性能，而不仅仅是准确率和原始数据，Kaldi 新版是否也有如此改进？Daniel 认为，新版 kaldi 用在小爱云端（服务器端），注重于改进模型性能：多通道、低功耗。

陆问：新版 Kaldi 是否考虑更高效使用 GPU 及如何针对新硬件执行自动编译？Daniel：太对了！我们考虑使用 GPU 包括与非监督学习。

Daniel 希望帮助抓好新版落实，陆说：给你找两个伙伴帮助抓落实：一是小米集团，你们已经做了，要求 Kaldi 新版服务于小米的“手机＋AIoT 双引擎战略”语音识别技术，请小米集团副总裁崔宝秋帮助落实！二是建议与 CSDN 合作，请 CSDN 创始人蒋涛（基于程序员资源及搜索引擎）帮助落实（在请你写的文章中 CSDN 就可向你提供搜索资料）。

779，陆首群&Daniel 议论目前大规模语义网络（知识图谱）尚无力支持实现认知智能

陆谈：我们非常关注大规模语义网络（知识图谱），目前存在的全球性问题是：在语义网络中缺乏逻辑推理，常识问题是其短板，因此难以支持认知智能的实现，离解决还差最后一公里！Daniel：恐怕不止最后一公里，离解决的路还遥远。Kaldi 技术不同于大规模语义网络！陆：认知计算是用来构建模拟人脑思维过程的系统，是使人工智能进入可理解、可解释的强人工智能的新阶段！Daniel：今后双方可就此交换意见。

Note:

敬请关注联盟微信公众号
COPU开源联盟

中国开源软件推进联盟秘书处　　电话：+86 010-88558999
联盟公共邮箱：office@copu.org.cn　　联盟官网：http://www.copu.org.cn
地址：北京市海淀区紫竹院路66号赛迪大厦18层