引用格式:陈小平. 大模型的逻辑增强与人工智能驱动的行业创新[J]. 技术经济, 2024, 43(12): 1-9.

CHEN Xiaoping. Logical enhancement of large language models and industrial innovation driven by artificial Intelligence technologies[J]. Journal of Technology Economics, 2024, 43(12): 1-9.

大模型的逻辑增强与人工智能驱动的行业创新

陈小平[1,2]

(1. 中国科学技术大学计算机学院, 合肥 230026; 2. 广东省科学院智能制造所, 广州 510070)

摘 要: 本文探讨通过大模型等人工智能技术推动制造业创新发展的重大意义,以及面临的挑战和机遇。分析大模型技术体系的构成,阐释其工程学基本概念,介绍关联度预测的一种科学解释——类 L_C 理论。基于该理论,分析大模型的奇异表现的深层原因与重要后果,形成对大模型技术更全面、更深入的理解。在此基础上,梳理制造业细分行业对人工智能技术的三项基本要求——专业性、逻辑可靠性和知识能力;分析大模型技术与人工智能强力法技术相集成的核心难点;提出构建行业人工智能系统的封闭化方案,由该方案构建的行业人工智能系统满足三项基本要求,并具有可解释性和可控性。最后,简要讨论制造业高质量发展中从"新技术的行业应用"向"新技术驱动的行业创新"转变的新趋势。

关键词: 大模型; 人工智能; 可解释性; 制造业; 行业人工智能; 行业创新

中图分类号: TP18 **文献标志码**: A **文章编号**: 1002-980X(2024)12-0001-09

DOI: 10. 12404/j. issn. 1002-980X. J24101804

一、引言

与高新技术的深度融合是经济发展的一个核心动力。随着大模型等人工智能技术取得重大进展,其在经济活动中的应用和对产业创新的驱动引起了广泛关注。

大模型等人工智能技术能否成为新质生产力,关键在于能否促进高质量发展。中国制造业体量庞大、门类齐全,包含着众多细分行业(以下简称细分行业)。根据工信部的数据,目前我国制造业的 80%尚未实现深度转型升级。高质量发展提出了一项新的重大任务——传统制造业的高端化,即让传统制造业升级为高端,而不是任其外迁。因此,众多细分行业对人工智能技术有很强的需求,大模型和人工智能技术对于推进整个制造业的行业创新具有极其重要的战略意义。对于大部分制造业已经外迁的经济体而言,其主要经济活动集中在虚拟经济和服务业,所以对人工智能技术的需求是非常不同的。

大模型等人工智能技术在快速发展的同时,也面临着一系列严峻挑战。大模型存储在深层网络中,每一个深层网络通常有几十亿到几千亿个参数,一个参数是介于 0~1 的一个数,一个深层网络的行为就是由这些参数决定的。那应如何理解大模型的工作原理? 如何解释大模型的众多奇异表现? 光看这些参数无法回答上述问题,这就产生了大模型的可解释性挑战。

通常关心的大模型解释有两种。一种是面向大模型研发者的科学解释,说明大模型的工作原理;另一种是面向用户的因果解释,说明大模型的一个具体输出基于什么理由。本文集中讨论大模型的科学解释,并在此基础上进一步探讨人工智能技术如何驱动制造业的产业创新。

在人类技术史上,几乎找不到一项技术是没有科学解释却被普遍应用的。技术的科学解释是大规模应用的必要前提。人工智能技术包含两大类——强力法和训练法[1]。强力法是可解释的,并具备逻辑可靠性;而训练法目前不是可解释的,也不具备逻辑可靠性,大模型就属于训练法。

收稿日期: 2024-10-18

基金项目: 国家自然科学基金"面向航空制造的人-机器人协作技术及应用研究"(92048301)

作者简介: 陈小平,博士,中国科学技术大学教授,博士研究生导师,中国科学技术大学机器人实验室主任,广东省科学院人工智能首席科学家,CAAI 人工智能伦理与治理工委会主任,研究方向:人工智能基础,智能机器人核心技术,人工智能伦理与治理。

制造业对人工智能技术提出了三项基本要求——具备专业性、逻辑可靠性或高可信度、知识能力。目前大模型不满足这些要求。为此,研究者倾向于让大模型发挥“理解”用户提问和提示的作用,让可解释的、具备逻辑可靠性和知识能力的人工智能强力法技术负责推理,通过系统集成来实现大模型的逻辑增强,进而满足应用需求。

然而在大模型和强力法技术之间存在着一条鸿沟,妨碍了二者的协同工作,阻碍了预期目标的实现。本文对相关难点加以分析,并提出一个基于封闭化的解决方案,该方案满足三项基本要求,并具有可解释性和可控性。这一方案并非现有人工智能技术的简单应用,而是一种基于人工智能技术的行业创新,颠覆了以往制造业中新技术应用的传统理念,反映了高质量发展的新机遇。

二、大模型的可解释性挑战

大模型技术体系十分庞大,主要包含三个部分。一是预训练(pre-train),用人类规模的原始语料(如互联网 2/3 的文本),经自监督学习(无需人工标签)构建基础大模型。二是细调(fine-tune),即用补充数据对基础大模型进行专门训练,使其输出更符合人的需要。目前使用的最重要的补充数据来自人类反馈。ChatGPT 就是用人类反馈数据,对几个基础大模型进行细调得到的。三是激发(prompt),即在提问之外,通过人工提示或相应的技术手段,对存储大模型的神经网络进行激发,让大模型对具体问题回答得更好。

大模型的工程学基础可简化为三个基本概念。一是语元(token),直观上指的是字、词和标点符号。二是关联度,即语料中语元之间关联强度的某种统计近似,用 0 和 1 之间的一个数值表示,存储在基础大模型里。三是关联度预测,即根据存储的关联度,预测在一个给定的语境(组成上下文的一串语元)之后出现的下一个语元。大模型通过反复进行关联度预测,逐字生成给用户的回复。

本文通过一个简化的例子说明这些概念。考虑表 1 中的语料,它只含两个出现概率分别为 0.6 和 0.4 的句子。假设从该语料中提取的语元即图 1 中椭圆里的字、词和标点符号。语元之间的关联度如图 1 中的箭头附带的数值所示。例如,“我”与“要”的关联度为 1(因为语料中“我”后面出现的总是“要”),“我”与“上网”和“听歌”的关联度分别是 0.6 和 0.4,“我”与其他语元之间也都有关联度(图 1 中有省略)。给定语境“我要听歌,请打开”,根据图 1 中的关联度,特别是给定语境下“听歌”与“音箱”的关联度为 1,预测下一个出现的语元是“音箱”的准确率为 100%,因为在原始语料(表 1)中,只要前面出现了“听歌”,那么“打开”的下一个语元一定是“音箱”。

表 1　一个语料样例

语料(用于训练的语言材料)	概率
我要上网,请打开浏览器	0.6
我要听歌,请打开音箱	0.4

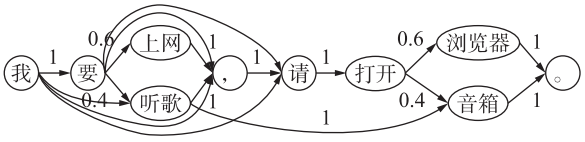

图 1　语料(表 1)中语元之间的关联度
(省略了一部分关联度)

上述工程学概念有利于了解大模型的算法,但显然没有像牛顿力学相对于宏观物理世界那样,提供大模型的科学解释。陈小平[2]发现,预训练和激发这两部分有一个共同的基础,也是大模型的底层机制,叫做关联度预测。陈小平[2-3]给出了关联度预测的一个形式化理论——类 L_C 理论,作为大模型的一种科学解释。根据这种解释,预训练和激发这两部分就不再是黑箱,而细调部分还是黑箱。

类 L_C 理论有 3 条公理,前两条公理是通用的,第三条公理跟应用有关,不同的应用有不同的公理 3,ChatGPT 的公理 3 就与文献[2-3]中的公理 3 不同。这些公理用来描述大模型的可解释的一般行为,有些行为不可解释,比如细调现在就不可解释,所以不在三条公理的覆盖范围内。大模型的一些行为细节或个别大模型的特殊行为,也不在考虑范围之内。

公理 1(语境关联度):$0 \leqslant C(a_i, b \mid a_1^n \leqslant 1)$　　(1)

公理 2(综合单调性):$\wedge_{i=1}^{n}[C(a_i, b \mid a_1^n) \leqslant C(a_i, c \mid a_1^n)] \supset C(a_1^n, b) \leqslant C(a_1^n, c)$　　(2)

公理 3(预测选择):$\mathrm{argmax}_b\, C(a_1^n, b)$　　(3)

随着大模型研究的推进,特别是 2023 年 7 月之后,逐步出现了越来越多的深度测试,这种测试不是只给

出测试结果统计，而是从测试结果出发，进一步分析、揭示大模型的奇异表现，也就是好得令人意外，或差得令人意外，却无法解释的表现。能不能解释大模型的各种奇异表现，是大模型科学解释的试金石。

迄今发现的大多数深度测试结果都可以用类 L_C 理论加以解释，甚至可以用类 L_C 理论预言这些表现，也就是在测试之前预言大模型在测试中会出现什么表现。少数奇异表现不容易解释，但也不与类 L_C 矛盾。这表明，类 L_C 理论很大程度上得到了迄今实验结果的支持。

三、大模型的奇异表现

本节围绕知识能力的三个主要方面——知真知、知所知和知不知，介绍对大模型进行的 3 个深度测试，并对测试所揭示的奇异表现的原因与后果进行理论分析。

测试例 1（关于知真知），这是 2023 年 8 月 Arkoudas[4] 做的一个测试。问题中的 p 代表一个命题，也就是一个有真假的陈述句。命题分为肯定的和否定的，否定的命题前面有奇数个否定词~。给大模型的测试题目是：p 前面有 27 个否定词，问大模型“p 前有几个否定词”。这个问题看起来很简单，数一下否定词的个数就行了，结果大模型回答 28 个。这说明大模型不会计数。2024 年 7 月，谷歌等研究机构发文回应了大模型不能计数的问题[5]。他们的分析不仅确认了这个问题的存在，而且进一步指出，用 Transformers 算法难以解决这个问题。计数是数学的基础功能，自然数是用 0 和+1（即计数）定义的，加、减、乘、除等数学运算都是用 0 和计数定义的，不会计数意味着由计数定义的诸多数学运算都会受影响，说明大模型缺乏数学能力。

不会计数又导致不会逻辑否定。逻辑学的一条规则是双重否定律，即两个否定词连在一起就变成肯定，而肯定和否定是相互矛盾的，不能混淆。不会否定运算说明大模型缺乏逻辑能力，它会在推理过程中混淆肯定命题和否定命题，从而引入逻辑错误，导致大模型不具备逻辑可靠性，使大模型对各种问题的判断违背逻辑规律。

从知识能力的角度看，一个大模型对提问 Q 回答 A，可视为该大模型“知道 Q 的答案是 A”，简称“知道 A”。有关知识能力的第一条要求涉及“知真知”能力，也就是当一个大模型“知道 A”的时候，是否一定有“A 是真的”。假如“知道 A”但 A 不是真的，这种“知道”就没有多少实际意义。测试例 1 的结果表明，大模型不具备知真知的能力，因此大模型提供的回答并非总是可信的。

测试例 2（关于知所知），这是 2023 年 2 月 Qin 等[6] 对 ChatGPT 做的测试，测试题目是：一个孩子存了 21 元，如果又得到 15 元，用这些钱可以买多少个单价 6 元的玩具？大模型的回答有两句话，第一句给出答案 5，这显然不对，回答的第二句说：应该有总共 21+15 等于 36 元，所以能买 36/6 等于 6 个玩具。这样的表现令人费解，大模型明明知道是 6 个，计算过程也是正确的，可是为什么前面却说是 5 个？而且发现前面答错了，为什么后面不加以纠正？

在人工智能知识逻辑中，“知所知”指的是：一个人工智能系统知道自己知道什么，所以当它知道问题 Q 的答案是 A 的时候，一定会用 A 回答 Q，不会给出别的答案。然而本例表明，大模型不具备知所知的能力，它明明知道正确答案，却仍然给出错误回答。本例还表明，大模型给出的答案跟答案的解释可以不一致，所以大模型的解释不是自己行为的因果解释。这表明，大模型没有因果解释能力，所以不能认为大模型给出的解释是自己行为的因果解释。

测试例 3（关于知不知），这是 2023 年 8 月 Arkoudas[4] 对 GPT-4 进行的测试。测试包含如下类型的多个问题：给定一个集合和一个正数，让 GPT-4 求该集合的子集，使得子集元素之和等于给定的正数。例如，给定集合{2, 8, 6, 32, 22, 44, 28, 12, 18, 10, 14}，问该集合有多少个子集，其元素之和等于 37？

这个测试问题包含两条约束：①子集元素都在集合{2, 8, 6, 32, 22, 44, 28, 12, 18, 10, 14}中；②子集元素之和为 37。测试结果表明，大模型不能检验一个回答是否满足全部约束，给出的回答都是只满足一个约束的。分析表明，导致这种情况的直接原因是，关联度预测的“通用”算法受计算复杂度的制约，不具备检验多重约束的能力。这意味着，大模型不具备约束满足的能力。

然而，大模型并不知道自己不具备多重约束检验的能力，仍然徒劳地回答多重约束满足的问题。这表明，大模型不具备“知不知”的能力。如果一个人工智能系统具备“知不知”的能力，那么它就不会回答它不

知道答案的问题。

约束满足是一种基础性的逻辑功能,在反思、规划和目标驱动的决策等高级认知活动中具有不可或缺的重要作用。反思是检验自己已完成的思维过程是否符合某些标准,即判别思维过程是否满足相关的约束。规划是找到实现目标的行动序列,使得行动序列的整体效果满足目标约束。目标驱动的决策是生成满足目标约束的决策。总之,这些高级认知活动都包含着约束满足,离开约束满足就不可能实现。一些学者反复呼吁:需要研发目标驱动的人工智能技术,以弥补大模型不具备目标驱动能力的缺陷。上述分析表明,大模型不具备目标驱动能力的一个具体原因在于不具备检验多重约束的能力。

四、大模型的主要特性

本节进一步梳理、总结大模型的主要特性。通过这些特性,可以更明确地认识到大模型逻辑增强的必要性。

第一个特性:大模型缺乏逻辑可靠性。测试表明,大模型有时不能正确地完成逻辑运算,但在其他情况下却可以。综合两方面情况,为什么说大模型缺乏逻辑可靠性?为了回答这个问题,首先必须明确,大模型属于"机器"的范畴,不属于人的范畴;对于不同范畴的对象,需要应用不同的评判标准。大模型是一种软件,适用于计算机科学的标准,这种标准规定:一个程序在某项任务上是逻辑可靠的,当且仅当该程序在该任务上的运算总是正确的,否则就认为该程序在该任务上不是逻辑可靠的。如果允许一个程序的运算有时对有时错,那就没有逻辑可靠性可言。

逻辑可靠性是由一系列逻辑功能加以保证的,包括计数、等量代换,逻辑否定、约束满足、传递性推理等数学和逻辑的基本能力。现已发现,根据计算机科学的标准,大模型不具备这些数学、逻辑的基本能力[2-3]。因此,大模型不具备逻辑可靠性。

在聊天式对话中,逻辑可靠性似乎不重要。但很多细分行业要求符合计算机科学标准的逻辑可靠性,或至少具备高可信度。

第二个特性:大模型能回答任何问题,但不保证回答总是正确的。根据 L_C 理论的三条公理,可以证明关联度预测能够回答任何问题。根据第一个特性,大模型无法保证回答总是正确的。

在聊天应用中,有必要保证机器总有回复或大多数情况下有回复,这种回复谈不上对错,但如果频繁出现不回复,就聊不下去了。可是在细分行业的应用中,重要的不是有回复,而是给出正确的回答、避免错误的回答。

第三个特性:大模型与人之间只有弱共识。由于大模型从训练语料中提取了字词之间的统计关联,所以在这方面与多数人是一致的,也就是有共识的;此外,大模型没有其他语义,所以在其他语义上与人没有共识。例如,大模型不知道一个词本身是什么意思,但它知道这个词跟别的词是如何关联的。于是,当人和大模型对话的时候,不同的人对大模型输出的同一句话可以有不同的理解,即这些人之间没有形成共识,但都被大模型接受。所以在关联度预测机制下,大模型不跟用户吵架,用户说什么它都说对。其深层原因在于,很多意思是大模型不掌握的,所以也就不可能与用户的意思发生矛盾,这样就不可能与用户吵架。

但是,过去的 AI 和软件都要求强共识,所以软件人员要学习编程,通过学习达成强共识,否则就用不好软件。可是大模型没有强共识,只有弱共识,所以通过自然语言对话人人都可以使用大模型。这就解释了为什么大模型好用,根本原因在于弱共识性。

总结上述特性可知,大模型的工作原理跟人的智能的原理是不同的,但表现却往往是类似的。这符合图灵的机器智能观——机器智能的工作原理与人的智能的工作原理可以相同,也可以不同[7-8]。所以,包括大模型在内的人工智能并不能像人那样,随意地参与一切经济活动。有必要针对不同应用的实际特点,判断大模型应用的适当性,以及可行的应用方式。

五、细分行业需求分析

一个细分行业通常拥有高质量的、数量较大的专业数据,这些数据往往是通用大模型的研发难以获得

的，所以通用大模型很少使用这些数据进行训练。另外，一个细分行业通常拥有行业内公认的领域知识，而通用大模型无法完全掌握各个细分行业的领域知识。所以，通用大模型无法准确地回答细分行业的很多问题。本文将能够胜任细分行业应用的人工智能系统称为行业人工智能系统，这种系统需要满足以下基本要求：

第一，具备专业性，即要求行业人工智能系统提供符合相关细分行业标准的专业性回答，而不是常识性回答。细分行业往往是专业领域，拥有特定专业标准，不同的细分行业有不同的专业标准，这些标准与常识有很大的不同。基础大模型基于统计关联度回答提问，往往较好地符合常识，或者常识性与专业性相混合，却未必符合各个细分行业的专业标准。

第二，具备逻辑可靠性或高可信度。一个系统具有逻辑可靠性，指的是它的输出相对于输入具有保真性——只要输入在某种意义上是真的，则输出在相同意义上也一定是真的。这样的系统不会由于自身的问题引入错误。对于行业人工智能系统，往往要求回答的准确性达到 99.9%甚至更高，不具备逻辑可靠性的人工智能技术由于自身会引入错误，无法达到这么高的准确性，如目前的各种大模型。

第三，具备知识能力，即符合如下 3 条知识公理，其中 K 是一个模态词，代表“知道”；p 是一个命题变元，代表一个命题；$\supset$ 是逻辑连接词“蕴含”，代表“如果，则”；$\neg$ 是一个逻辑连接词“否定”，代表“并非”。

(1)知真知公理：$Kp \supset p$。直观解释为：凡是行业人工智能系统知道的，在对应细分行业中都是真的；

(2)知所知公理：$Kp \supset KKp$。直观解释为：行业人工智能系统知道自己知道什么，所以需要时它能够用自己的知识给出正确回答，不会明明知道正确答案却给出错误回答或者不回答。

(3)知不知公理：$\neg Kp \supset K\neg Kp$。直观解释为：行业人工智能系统知道自己不知道什么，所以遇到自己不知道的，不会做出回答。

前文的 3 个测试例分别表明，大模型不满足 3 条知识公理。所以，大模型不具备知识能力。

然而部分研究者不同意上述分析结论，他们认为大模型已经具备了逻辑推理能力，有时表现不佳只是因为训练不足，只需增加大模型的参数和训练，就可以消除大模型的一切缺陷，包括逻辑能力不足的问题，从而满足行业人工智能系统的三项基本要求。

表 2 给出了详细的对比，表明大模型的“推理”与逻辑推理是根本不同的。为了加以区别，本文将大模型的“推理”称为“类 L_C 推衍”。根据表 2，类 L_C 推衍与逻辑推理在决定推理特性的 7 个方面都是不同的。因此，大模型推衍与逻辑推理之间的差别不是程度上的差异，而是根本性质上的差别，这种差别无法通过更多数据的训练而加以消除。

另外，最近在 *Nature* 上发表的一项实验测试[9]表明，参数较多的模型相对于参数较少的模型可能更不可靠，更容易提供错误回答，所以增加参数未必能够消除大模型的一切缺陷。事实上，由于自然语言是非封闭的，而非封闭场景不可能获得充分的训练数据[1, 8]，所以单纯依靠更多数据的训练不可能让大模型具备逻辑能力。

根据以上分析可以确认，大模型不满足行业人工智能系统的三项基本要求，不能胜任这一职能。因此，针对细分行业的应用，有必要进行大模型的逻辑增强，通过技术手段弥补大模型的不足，使得行业人工智能系统满足全部三项基本要求。与三项基本要求密切相关的是可解释性和可控性，下文将结合解决方案进一步讨论。

另外，行业人工智能系统还需要满足一些附加要求，如可操作性，即能够进行细分行业所需的专门运算。例如，在数据处理方面[10]，需要以下运算：检索(在细分行业检索特定的结果，而大模型的回答经常不是用户想要的)；分类(把对象分成不同的类)；比较(对

表 2　类 L_C 推衍与逻辑推理的对比

	类 L_C 推衍	逻辑推理
表示对象	语元和语元序列	命题(关于世界的判断)
规则来源	用数据训练出大量规则	手工编写少量规则
表达粒度	描述关联细节的实例性规则	描述抽象模式的概括性规则
可推标准	基于语境的统计相关性	逻辑可靠性
抽象能力	不遵守逻辑公理或数学公理	遵守逻辑公理和数学公理
语义学	?(非概念化)	模型论语义学(概念化)
性能	所有提问都有回应，所有回应符合统计相关性	所有结论是逻辑可靠的，有时没有结论

注：目前尚未建立类 L_C 推衍的语义，故表中用问号加以标记，但已确定这种语义是非概念化的[2]。

不同对象的属性进行比较);反向搜索(找到符合条件的对象)。与基本要求相比,附加要求比较容易满足,故本文余下部分不再讨论附加要求。

六、行业人工智能系统设计:主要难点与封闭化方案

构建行业人工智能系统主要有两条途径。第一条是"机制兼容"途径,即通过人工智能基础研究,改造大模型的底层机制——关联度预测,使改造后的新机制能够满足三项基本要求。第二条是进行系统集成,开发出的集成系统包含两类部件,一类是大模型部件(可以有多个),另一类是人工智能强力法部件,如形式化推理、知识图谱、人工智能规划、检索增强的生成(RAG)等部件,通过两类部件各自功能的集成,以满足行业人工智能系统的三项基本要求。

目前几乎所有相关研究都采取系统集成途径。例如,2024 年 6 月报告的一项研究[11]试图通过验证大模型的回答是否正确,来提高回答的逻辑可靠性。作者设计了一些算法,让大模型调用外部推理部件进行验证,取得了一定效果。其中一个数据集上,测试结果的准确率达到了 99.60%,不过这个数据集比较简单,在其他几个数据集上,准确率为 80%上下。同类其他工作的效果是类似的。

系统集成途径的基本想法是:让大模型部件根据用户提问和提示,生成当前任务的求解目标和求解条件,输入强力法部件,由强力法部件完成任务求解。由于强力法部件是可解释的,并具备逻辑可靠性,从而保证了行业人工智能系统的任务求解是可解释的,并具备逻辑可靠性。

然而上述想法隐含着一个深层的核心难点。一方面,自然语言表达通常不是逻辑良形的(logically well-formed),大模型生成的求解目标和求解条件通常也不是逻辑良形的。另一方面,强力法部件不能有效处理非逻辑良形的表示[1]。于是,被集成的两类部件无法有效地协调工作,以实现预期的功能。

逻辑公式和数学公式都是逻辑良形表示的例子。非逻辑良形的表达很容易引起推理的错误,大模型也不能幸免。例如,2024 年 6 月菲尔兹奖获得者 Timothy Gowers(下文称为 Gowers)对大模型解答"过河问题"进行了测试,测试中的一个问题是:一个农夫带两只鸡过河,如果一条船只能容纳一个人和两个动物,那么农夫带着两只鸡渡河所需的最少渡河次数是多少?结果大模型回答:至少需要渡河 5 次。这个问题本身并不存在逻辑错误,但其表达不是逻辑良形的,因为没有显式表达出"农夫"和"人"以及"鸡"和"动物"的逻辑关系。Gowers 对过河问题的测试结果表明,当提问存在隐式逻辑关系时,大模型的出错率至少为 80%,对某些测试例甚至高达 99%以上。

非逻辑良形表示的一些典型样例包括:未定义的概念、条件缺失、直接冲突、间接冲突[12]。复旦大学的研究团队构造了包含这些样例的"逻辑陷阱",并通过测试发现,对于带逻辑陷阱的提问,大模型的准确率仅为 24.3%。这与 Gowers 的测试结果是一致的。

处于人工智能当前发展阶段,在封闭性或封闭化场景中,上述核心难点是可以被攻克的;而在非封闭场景中,无法保证逻辑可靠性和知识公理的成立[1,8]。作为人工智能研究的一个里程碑,AlphaGo Zero 是封闭化的第一个成功范例[8],具有不可忽视的参考价值。在决策论规划(一种人工智能强力法技术)的经典理论中,行动值函数 Q 被定义为

$$Q(s,\ a) = R(s,\ a)\ + \gamma \sum\nolimits_{s' \in S} T(s,\ a,\ s')V\ (s') \tag{4}$$

其中:$T(s,\ a,\ s')$为状态 s 下执行行动 a 到达状态 s'的转移概率。在围棋问题中,该转移概率是与对手有关的,而人工智能系统无法完全把握对手的情况,从而产生了非封闭性。在 AlphaGo Zero 中,行动值函数被重新定义为

$$Q(s,\ a) = \sum_{s'|a,s\to s'} \frac{V(s')}{N(s,\ a)} \tag{5}$$

其中:平均胜率估计 $V(s')$取代了转移概率 $T(s,\ a,\ s')$。由于平均胜率是与对手无关的,所以通过重新定义行动值函数,实现了围棋问题的封闭化。AlphaGo Zero 以重新定义的行动值函数为基础,通过自博获取平均胜率估计,并以平均胜率估计作为决策依据,最终达到了远超人类的性能,并彻底颠覆了人类的围棋知识。这意味着,AlphaGo Zero 不仅在围棋领域应用了人工智能技术,更重要的作用是实现了围棋领域的知识

创新。

有些场景本身是封闭的，这种情况相对较为简单。还有一些场景不是封闭的，但可以被封闭化，使之变成封闭的，本文集中考虑这种情况，提出一个攻克行业人工智能系统核心难点的封闭化方案，包括以下主要步骤。

步骤 1：封闭化。针对给定的细分行业，识别它的一个相对独立的局部，称为领域 D，对 D 进行封闭化。假设 D 原来是用某个专业语言描述的，对 D 进行封闭化将产生一个描述 D 的新语言 Σ，称为领域 D 的标准语言。Σ 是逻辑良形的，并且用 Σ 描述 D 可以消除一切不确定性。正如在 AlphaGo Zero 中，当描述围棋决策的语言从基于转移概率更换为基于平均胜率估计，消除了围棋决策中的不确定性。但制造业中的封闭化通常不限于描述语言的改变，而是同时涉及领域本身的改变，包括制造装备、工艺流程、生产组织等方方面面的调整。

步骤 2：领域知识重构。以 Σ 为基础，重新构建 D 的知识库 $K_\Sigma(D)$，使得 $K_\Sigma(D)$ 能够充分表达 D，不存在不确定性。具体要求为：对 Σ 语言描述的任何求解目标 g 和求解条件 c，在 $K_\Sigma(D)$ 中可以判定 $c \supset g$ 是否成立，其中 $c \supset g$ 的直观解释为：如果 c 成立，则 g 成立。所谓对 D 进行领域知识重构，就是找到满足上述条件的 $K_\Sigma(D)$。

步骤 3：行业大模型的开发。基于 $K_\Sigma(D)$，开发符合细分行业专业标准的行业大模型 $M_\Sigma(D)$，保证 $M_\Sigma(D)$ 生成的任何求解目标 g 都符合标准语言 Σ 对求解目标的要求、$M_\Sigma(D)$ 生成的任何求解条件 c 都符合 Σ 对求解条件的要求，从而保证 g 和 c 都是逻辑良形的并符合领域 D 的专业标准。为了开发 $M_\Sigma(D)$，可能需要根据 Σ 调整 D 的数据结构，也可能需要收集更多相关数据，特别是领域调整后的数据，生成一个新的数据集并用于训练 $\mathrm{M}_\Sigma(D)$。

步骤 4：推理机的开发。开发由一组强力法部件组成的、可判定的推理机 $B_\Sigma(D)$，使之能够完成如下推理功能：对行业大模型 $M_\Sigma(D)$ 产生的任何求解目标 g 和求解条件 c，判定 g、c 与 $K_\Sigma(D)$ 是否相容，若相容则判定 $c \supset g$ 是否成立。注意，行业人工智能系统中的推理是由强力法部件构成的推理机 $B_\Sigma(D)$ 负责的，行业大模型 $M_\Sigma(D)$ 只负责“转述”用户提出的求解目标和求解条件。

步骤 5：逻辑验证。逻辑验证的目的是确认 $B_\Sigma(D)$ 的逻辑可靠性，$B_\Sigma(D)$ 的逻辑可靠性定义为：对行业大模型 $M_\Sigma(D)$ 产生的任何求解目标 g 和求解条件 c，如果 $B_\Sigma(D)$ 判定 $c \supset g$ 成立，则对 Σ 的任何抽象解释 I_Σ，如果 $I_\Sigma(c)$ 是抽象真的，那么 $I_\Sigma(q)$ 也是抽象真的。逻辑验证可借助于现有自动证明工具完成。

步骤 6：领域验证。领域验证的目的是确认，由 $B_\Sigma(D)$ 推理得出的结论在领域 D 中都是真的，定义为：给定一个重言式 t，对任何求解目标 g，如果 $B_\Sigma(D)$ 判定 $t \supset g$ 成立，那么 g 在领域 D 中是真的。领域验证涉及领域 D 中代表性数据的收集和确认。

步骤 7：行业服务。由领域知识库 $K_\Sigma(D)$、行业大模型 $M_\Sigma(D)$ 和推理机 $B_\Sigma(D)$ 组成行业人工智能系统 $S_\Sigma(D)$。在领域 D 中，由既掌握行业知识、又懂 $S_\Sigma(D)$ 的领域工程师，运用经过验证的 $S_\Sigma(D)$ 为本行业的专业用户服务。每一次服务中，要求专业用户描述当前任务的求解目标 $g*$ 和求解条件 $c*$，由行业大模型 $M_\Sigma(D)$ 将 $g*$ 和 $c*$ “转述”为 Σ 语言表达的 g 和 c，由领域工程师确认 g 与 $g*$ 语义相符、c 与 $c*$ 语义相符。然后启动推理机 $B_\Sigma(D)$ 执行用户任务，并向用户报告运行结果。

上述封闭性方案在一定条件下满足行业人工智能系统的三项基本要求。第一，如果在步骤 7，领域工程师对 g 与 $g*$ 语义相符、c 与 $c*$ 语义相符的确认是真实有效的，即行业大模型 $M_\Sigma(D)$ 的“转述”没有发生专业性失真，那么之后 $B_\Sigma(D)$ 的推理过程也不可能出现专业性失真，所以 $S_\Sigma(D)$ 的完整运行满足专业性要求。

第二，推理机 $B_\Sigma(D)$ 具备逻辑可靠性，行业人工智能系统 $S_\Sigma(D)$ 具备领域可靠性。逻辑验证保证了 $B_\Sigma(D)$ 的逻辑可靠性。结合逻辑验证与领域验证，可以进一步证明 $S_\Sigma(D)$ 在 D 中的领域可靠性：对行业大模型 $M_\Sigma(D)$ 产生的任何求解目标 g 和求解条件 c，如果 $B_\Sigma(D)$ 判定 $c \supset g$ 成立，那么在领域 D 中，若 c 是真的则 g 一定也是真的。

第三，行业人工智能系统 $S_\Sigma(D)$ 具备知识能力。由已证明的 $S_\Sigma(D)$ 的领域可靠性可知，$S_\Sigma(D)$ 具备知真知能力。由 $B_\Sigma(D)$ 的可判定性可以证明，$S_\Sigma(D)$ 具备知所知和知不知能力。

总之,依照封闭性方案开发的 $S_{\hat{Y}}(D)$ 基本满足行业人工智能系统的三项基本要求。之所以不是完全满足,是因为存在着专业性要求方面的一个可能漏洞——由于人和大模型都难免犯错,所以无法彻底排除发生专业性失真的可能性。如果不出现这种失真,三项基本要求将得到完全满足。

进一步,由于行业人工智能系统 $S_{\hat{Y}}(D)$ 的推理是由强力法推理机 $B_{\hat{Y}}(D)$ 负责的,而强力法具有可解释性和可控性,所以 $S_{\hat{Y}}(D)$ 的推理是可解释的和可控的。由于 $S_{\hat{Y}}(D)$ 的任务即求解目标和求解条件是由用户提出、由行业大模型 $M_{\hat{Y}}(D)$ "转述"的,所以不具备严格意义上的可解释性,但领域工程师的介入保证了任务的可控性。

七、结论与展望

制造业的高质量发展对我国具有重大战略意义。针对这一目标,本文梳理了制造业细分行业对人工智能技术的三项基本要求,提出了开发行业人工智能系统的封闭性方案,并论证依该方案开发的行业人工智能系统满足三项基本要求——具备专业性、逻辑可靠性和知识能力,以及推理过程的可解释性和任务求解的可控性。这表明,人工智能技术能够促进制造业的高质量发展,从而成为新质生产力。

在封闭性方案中,行业大模型的作用是"转述"用户的任务描述,由强力法推理机负责任务求解。这一方案的设计考虑来源于对大模型及其奇异表现的科学解释。科学解释表明,大模型不具备逻辑可靠性和知识能力,无法满足行业人工智能系统的三项基本要求,也不具备可解释性和行业应用所需的可控性。

封闭性方案的最大特色在于细分行业或其局部的封闭化。经过封闭化,行业人工智能系统中大模型部件与强力法部件之间的"鸿沟"得以消弭,设计目标得以达成。然而,封闭化并非现有技术成果的简单应用,而是一种应用驱动的技术创新,包含着生产过程的改变,从而孕育着行业的重大变革。

在传统行业的以往实践中,新技术的应用主要表现为大量微创新,不涉及重大的新发现和新发明,所以也没有依据创新的需要投入相应的科技力量。在人工智能时代,这种传统理念已经过时,制造业的发展驱动力正从"新技术的行业应用"向"新技术驱动的行业创新"转移,资源配置、产业生态、行业管理和产业政策乃至社会治理都将随之发生根本性变革,带来全新的发展机遇。

参考文献

[1] 陈小平. 人工智能中的封闭性和强封闭性——现有成果的能力边界、应用条件和伦理风险[J]. 智能系统学报, 2020, 15(1): 114-120.

[2] 陈小平. 大模型关联度预测的形式化和语义解释研究[J]. 智能系统学报, 2023, 18(4): 894-900.

[3] 陈小平. 大模型: 从基础研究到治理挑战[J]. 中国人工智能学会通讯, 2024, 14(1): 2-9.

[4] ARKOUDAS K. GPT-4 can't reason[EB/OL]. Preprints 2023, 2023080148. https://www.preprints.org/manuscript/202308.0148/v1.

[5] YEHUDAI G, KAPLAN H, GHANDEHARIOUN A, et al. When can transformers count to n?[EB/OL]. arXiv: 2407.15160, 21 Jul 2024.

[6] QIN C W, ZHANG A, HANG Z S, et al. Is ChatGPT a general-purpose natural language processing task solver?[EB/OL]. arXiv: 2302.06476, 15 Feb 2023.

[7] TURING A M, Intelligence machinery[M]. The Turing Digital Archive, 1948. DOI:10.1093/oso/9780198250791.003.0016.

[8] 陈小平. 人工智能伦理导引[M]. 合肥: 中国科学技术大学出版社, 2021.

[9] ZHOU L X, SCHELLAERT W, MARTÍNEZ-PLUMED F, et al. Larger and more instructable language models become less reliable[J]. Nature, 2024, 634: 61-68.

[10] ZHU Z A, LI Y Z. Physics of language models: Part 3.2, knowledge manipulation[EN/OL]. arXiv: 2309.14402v2, 16 Jul 2024.

[11] XU J D, FEI H, PAN L M, et al. Faithful logical reasoning via symbolic chain-of-thought[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok: Association for Computational Linguistics, 2024: 13326-13365.

[12] ZHAO J, TONG J Q, MOU Y R, et al. Exploring the compositional deficiency of large language models in mathematical reasoning through trap problems[EB/OL]. arXiv: 2405.06680v2, 2024.

[13] SIlVER D,SCHRITTWIESER J, SIMONYAN K , et. al. Mastering the game of go without human knowledge[J]. Nature, 2017, 550: 354-359.

Logical Enhancement of Large Language Models and Industrial Innovation Driven by Artificial Intelligence Technologies

Chen Xiaoping[1,2]

(1. School of Computer Science, University of Science and Technology of China, Hefei 230026, China;
2. Intelligent Manufacturing Institute of Guangdong Academy of Sciences, Guangzhou 510070, China)

Abstract: The significance, challenges and opportunities of driving innovative development in manufacturing industry with artificial intelligence technologies including large language models were explored. The technological system of large language models was analyzed, its basic engineering concepts were clarified, and the pan-L_C theory—a scientific explanation for next token prediction—was presented. Based on the theory, the causes and consequences of some weird behaviors of large language models were explained, giving a more comprehensive and in-depth understanding of large language models. On the basis, three main requirements for artificial intelligence technologies in manufacturing industry were sorted out, and the core difficulties in the integration of large language models and the artificial intelligence brute-force technology were revealed. A closedness-based solution is proposed for the construction of artificial intelligence systems in manufacturing sectors, such that these systems satisfy the main requirements of specialization, logical validity and knowledge ability, as well as explainability and controllability. Finally, the trend of shifting from "industrial application of new technologies" to "sector innovation driven by new technologies" in the high-quality development of manufacturing industry is discussed briefly.

Keywords: large language models; artificial intelligence; explainability; manufacturing industry; sector-oriented artificial intelligence; sector innovation