XAI架构探索与实践

夏正勋[1]，唐剑飞[1]，杨一帆[1]，罗圣美[2]，张燕[1]，谭锋镭[1]，谭圣儒[1]
1. 星环信息科技（上海）股份有限公司，上海 200233；
2. 中孚信息股份有限公司，江苏 南京 211899

摘要
可解释AI（explainable AI，XAI）是可信AI技术的重要组成。当前，业界对XAI的技术点展开了深入的研究，但在工程化实施方面尚缺少系统性研究。提出了一种通用的XAI技术架构，从原子解释生成、核心能力增强、业务组件嵌入、可信解释应用4个方面入手，设计了XAI基础能力层、XAI核心能力层、XAI业务组件层、XAI应用层4个层级，通过各技术层之间的分工协作，XAI工程化的落地实施得到了全流程保障。基于该XAI架构，可以灵活地引入新的技术模块，支撑XAI的产业化应用，为XAI在行业中的推广提供了一定的参考。

关键词
可解释AI；可信AI；XAI架构

中图分类号：F08，F062.5　　文献标志码：A　　doi: 10.11959/j.issn.2096-0271.2024013

Exploration and practice of XAI architecture

XIA Zhengxun[1], TANG Jianfei[1], YANG Yifan[1], LUO Shengmei[2], ZHANG Yan[1], TAN Fenglei[1], TAN Shengru[1]
1. Transwarp Information Technology (Shanghai) Co., Ltd., Shanghai 200233, China
2. Zhongfu Information Inc., Nanjing 211899, China

Abstract
XAI(explainable AI) is an important component of trusted AI. In-depth research on the technology points of XAI has been carried out in the current industry, but systematic research on engineering implementation is lacking. This paper proposed a general XAI technical architecture, which started from the follow four aspects: atomic interpretation generation, core competence enhancement, business component embedding and trusted interpretation application. We designed four levels: XAI foundation layer, XAI core competence layer, XAI business component layer and XAI application layer. Through the division of labor and cooperation among the technical layers, the implementation of XAI engineering was guaranteed throughout the whole process. Based on the XAI architecture presented in this paper, new technical modules can be introduced flexibly to support the industrialization application of XAI, providing certain reference for the promotion of XAI in the industry.

Key words
explainable AI, trusted AI, XAI architecture

0 引言

近年来，人工智能（artificial intelligence，AI）在量化金融、机器人、智慧交通、生命科学、智能医疗等领域得到了广泛应用[1-4]。AI模型广泛使用的机器学习模型涵盖了统计、深度学习、神经网络等多类型模型，但这些模型大多存在类似“射手假说”的谬误[5]，有被误用、滥用的风险。从数据分析假定（如以回归为代表的统计模型），到数据辅助模型驱动（如支撑向量机的传统机器学习模型），再到数据驱动（如卷积神经网络、蒙特卡洛数等深度学习和强化学习模型），AI模型的参数数量、复杂程度增长迅速，比如近几年大型语言模型的参数规模以年均10倍的速度增长。李国杰院士指出，可解释性是高级人工智能的重要组成部分。然而，在AI新“摩尔定律”的影响下，AI逐渐超出了人类的理解范畴。AI对使用者而言就是一个黑盒，无法保证AI判断的逻辑依据与人类相同，无法对AI的行为结果进行解释。XAI是未来人工智能研究的主攻方向[6]，是可信AI技术体系的重要组成部分[7]，在一定程度上可以解决上述问题。

当前学术界对XAI的研究主要围绕3个问题展开。一是探究AI系统的行为依据和决策原因，该方向致力于解释AI模型的工作原理，对影响AI结果的因子重要性进行度量，协助发现AI应用过程中的新知识。例如，Ribeiro等人[8]提出了LIME方法来解释模型对单个示例的预测结果；Zhou等人[9]提出了CAM方法，以类别热力图可视化方式展示原图对预测输出的贡献分布。二是XAI如何为AI应用提供安全性保障，该方向致力于分析AI系统的工作过程，寻找影响AI结果的因素，通过反事实推理等技术手段预防AI应用过程中的各种意外。例如，Akula等人[10]基于反事实原理提出了CoCoX模型，对CNN模型进行解释；Hsieh等人[11]提出了DiCE4EL算法，基于事件日志生成反事实的解释，主动管理潜在的异常结果[12]。三是XAI如何为AI应用提供社会性保障，该方向致力于使AI满足公平、合规等要求，增加AI应用的社会认可度。例如，Jafta等人[13]使用XAI技术对ML系统的公平性进行了评估；Pradhan等人[14]提出了Gopher系统，通过识别训练数据的一致性子集，解释偏差或意外的模型行为，从而定位偏差的来源并减少偏差。围绕XAI的3个问题，业界进行了深入探索，并取得了多方面的技术成果。这些技术成果有多种划分方式。2020年Du等人[15]根据获得可解释性的时间不同，将其分为内在可解释性方法和事后可解释性方法，根据解释范围的不同将事后可解释性方法分为全局可解释性方法、局部可解释性方法。2020年KDD大会从模型类型角度，将其分为深度学习模型XAI方法、贝叶斯模型XAI方法、强化学习模型XAI方法。2021年曾春艳等人[16]从解释的属性出发，将XAI模型分为自解释模型、特定模型解释、不可知模型解释、因果可解释性4个类型。2022年KDD大会将XAI的应用场景分为多智能体系统（multi-agent system，MAS）、机器学习（machine learning）、计算机视觉（computer vision）、规划（planning）、知识表示与推理（knowledge representation and reasoning，KRR）、人工智能不确定性（uncertaintyin AI，UAI）、搜索（search）、博弈论（game theory）、自然语言处理（natural language processing，NLP）和机器人技术（robotics），分别在这些场景中对XAI技术进行了阐述。

XAI虽然取得了很多技术成果，但在技

术架构、落地实施方案及未来发展的技术路线等方面尚缺乏系统性研究。为此，“如何实现可信可靠可解释人工智能技术路线和方案?”入选中国科协2022年重大前沿科学问题。当前XAI存在的问题包含以下几个方面。

一是XAI研究起步晚、成果少。从应用角度出发，XAI是一个跨学科的复杂问题[17]。Miller[18]提出XAI需要引入哲学、心理学和认知科学中用于定义解释、产生解释、选择解释、评价解释的方法，以此提供更好的解释。Kulesza等人[19]提出了一种以人为本的解释方法，通过用户与解释系统的交互，提升解释的准确性。Hsiao等人[20]从认知状态和认知过程的角度提出了XAI系统的评估指标。上述研究尚处于摸索阶段，未形成系统性的技术解决方案。

二是XAI缺乏对落地实施方案的研究。当前XAI研究偏重于基础理论研究，鲜有对XAI技术落地的研究。例如，Szczepański等人[21]指出同一个事件可能有多种解释，如何解决多个解释中的“罗生门”问题，是XAI技术落地的一个难点。系统运营者关心如何向用户提供一个更易接受的解释结果，为解决这一问题还需要建立一套适用于XAI的评价指标。此外，从AI系统演进的角度，XAI使传统的AI系统转变为交互式AI系统。2020年，Spinner等人[22]将XAI与8个全局监控指标相结合，构建了一个交互式的机器学习系统，但该系统没有对XAI交互能力进行多层次、多维度的分析，也没有对可交互性XAI业务能力进行抽象。因此，学术界与产业界需要一个完善、可支撑工程落地的架构设计来填补上述的技术空白，帮助XAI跨越技术与真实应用场景之间的鸿沟，全方位满足技术落地的需求。

三是缺乏通用XAI业务方法论的研究。当前XAI通常只局限于某一领域的应用，缺少将现有AI系统快速升级为XAI系统的工具链及平台。2019年，Kwon等人[23]实现了一个名为RetainVis的医疗领域可解释RNN模型；2021年，Ohana等人[24]实现了一种对股市危机进行特征归因的局部解释模型；2022年，Park等人[25]实现了一种能够预测经济增长率和经济危机的可解释模型。上述XAI技术方案的实施需要根据项目实际情况进行调整，缺乏通用性。因此，需要归纳XAI的能力并进行业务级的封装，提供通用的XAI业务开发组件，灵活支持各种AI应用场景。此外，基于XAI业务组件可进一步设计完整的XAI业务开发工具链，控制系统改造的影响，在不改变现有AI系统业务流程的前提下，以最小代价将现有系统升级为具备可解释能力的AI系统，有效应对碎片化和多样化的需求，并大幅缩减研发、定制、部署等工程化过程中的人力、时间、费用等成本。

基于对上述问题的思考，结合XAI的工程实践，本文提出了一种通用的XAI架构，其中基础能力层与核心能力层为前两个问题的解决提供了多维度的原子解释模组、合规性归因模组、一致性模组等技术能力模组，业务组件层为第三个问题的解决提供了XAI表征模组、XAI展示模组、Pipeline嵌入模组等业务组件模组。作为一种通用的XAI架构设计，本文提出的XAI架构并不局限于解决上述3类问题，通过各层的分工协作为应用场景提供全流程的XAI技术支持，降低应用门槛，为XAI商用化提供高质、高效的保障。

1 XAI架构实现

从理论研究到工程化落地，XAI需要在理想化的原子解释基础上考虑特定的工程化问题，例如，如何使解释符合AI各相

关方的需求，如何与现有AI系统进行有机融合，如何对XAI应用进行有效管理，如何将解释结果有效地展现给终端用户。基于以上思考，本文提出了一种通用的XAI架构，该架构分为4层，如**图1**所示。

- XAI基础能力层为XAI应用提供基础解释能力，包含原理性可解释、数据可解释、过程可解释、模型可解释与结果可解释等功能，可以满足AI不同应用场景对XAI的不同需求，为XAI的工程化提供了解释基础。
- XAI核心能力层提供XAI工程化落地的核心能力，包含知识库、合规性归因、认知归因、一致性模组、内部评估、反事实归因等模组，保证最终解释结果满足用户的实际解释需求。
- XAI业务组件层为XAI工程化落地提供业务能力支持，包括XAI表征、XAI展示、Pipeline嵌入等模组，降低了XAI业务的开发难度，提升了XAI业务的开发效率。
- XAI应用层包含XAI技术在行业落地的典型应用案例。业务型应用包括可解释推荐系统、可解释溯因系统、异常事件追溯系统、AI决策风险控制系统、AI合规监管系统；平台型应用包括可解释图计算系统、可信AI算法管理平台、可信可解释AI学习平台。

XAI技术架构为XAI工程化落地提供全流程支持，基于不同层级能力组件的协作，可以保证XAI解决方案实施的效率与质量，输出有理有据、通俗易懂的解释结果，满足不同角色的可解释性需求。

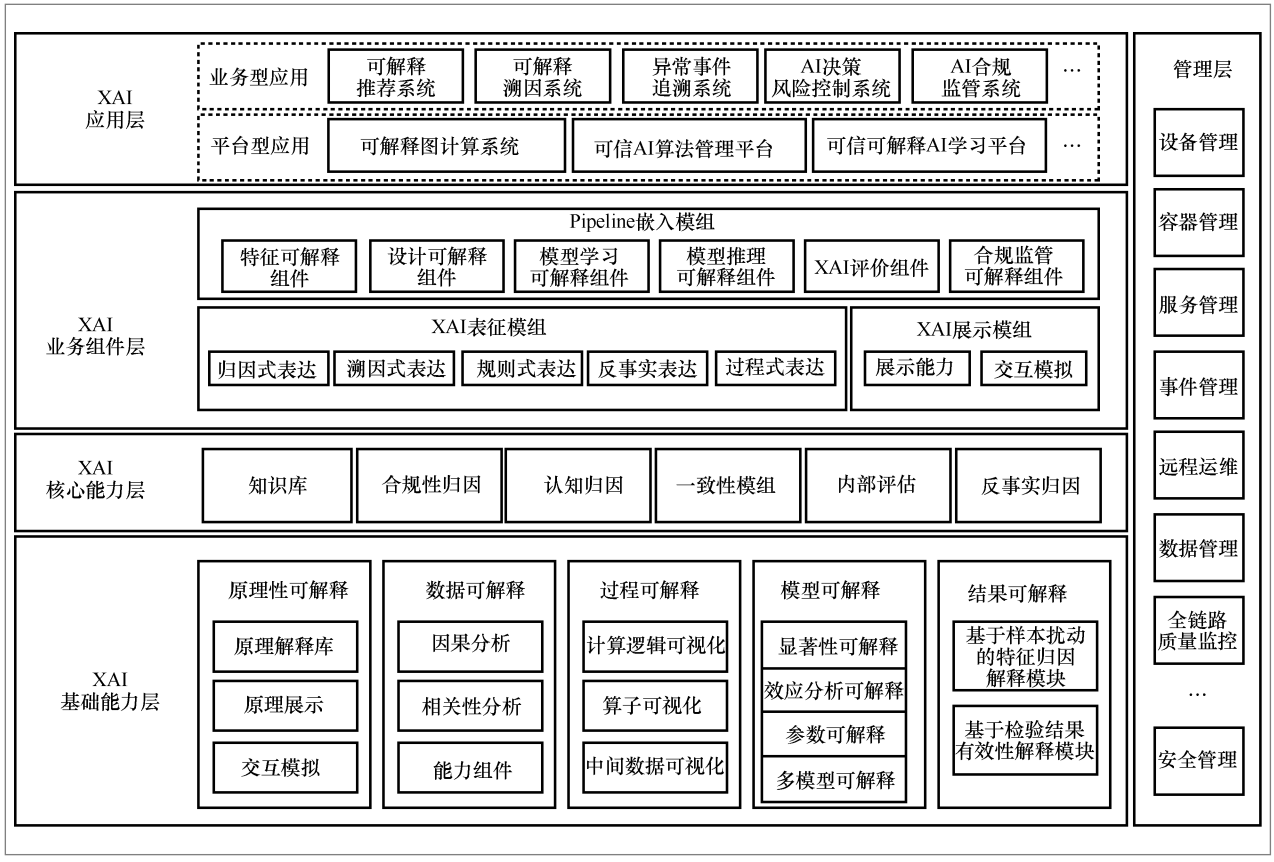


图1 通用 XAI 框架

1.1 XAI基础能力层

XAI基础能力层对原子解释能力进行集中管理，为上层提供原子解释。XAI基础能力层输出的原子解释是最终解释的“原料”，是XAI工程化落地的基础。XAI基础能力层如**图2**所示，包含原理性可解释模组、数据可解释模组、过程可解释模组、模型可解释模组与结果可解释模组，每个模组由多个模块组成。

1.1.1 原理性可解释模组

原理性可解释模组从人工智能的定义、过程及影响的角度对AI的结果或行为进行解释，帮助相关方从内部视角理解AI的核心实现方式及过程，属于概念性解释，通常以图文、公式等方式显性表示。原理性可解释模组主要由原理解释库模块组成，具体介绍如下。

原理解释库模块要使AI具有可解释性，不仅需要保证算法透明，更需要保证AI全流程是可解释的。原理解释库模块包含AI全流程涉及的原理性知识，可以为AI全流程的可解释性提供知识基础。原理解释库模块由样本管理原理、样本增强原理、特征工程原理、模型设计原理、算法选择原理与校验原理组件组成。样本管理原理组件提供与样本管理相关的知识，例如，如何对数据集进行版本管理，如何合理划分训练集、测试集与验证集，如何进行K-fold划分等。样本增强原理组件提供样本增强相关的知识，如何选择合适的样

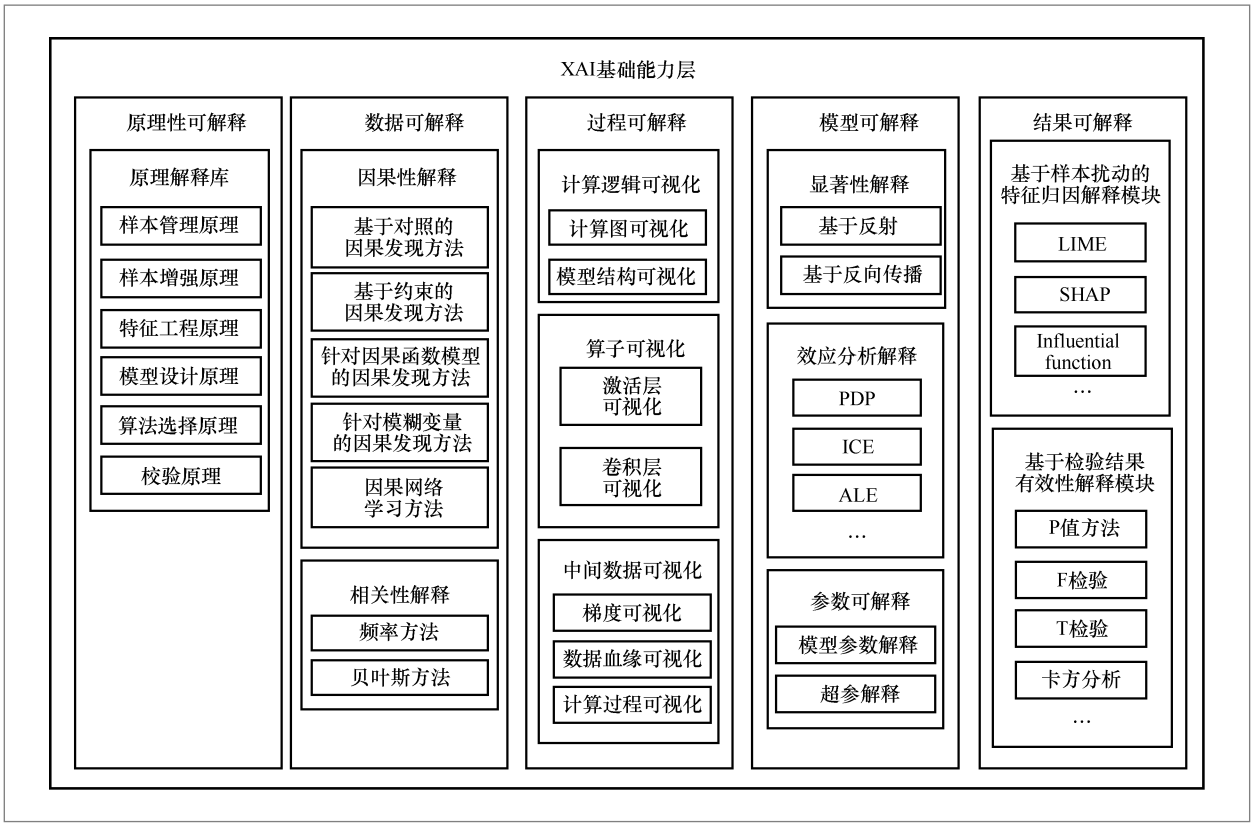


图 2 XAI 基础能力层

本增强方法，样本增强方法的原理等。特征工程原理组件提供与特征工程相关的知识，例如，如何选择合适的特征工程算法、特征工程算法原理等。模型设计原理组件提供机理模型[26]或数据模型关于“如何解决问题”的设计细节，包含理论假设、设计模式、工作机制、流程设计、预期输出等。算法选择原理组件提供与算法相关的知识，例如，对每个算法的假设条件、适用范围和建模思路，算法间的择优方法，算法的调优方法等。校验原理组件提供算法的各种检验指标的说明，如准确率、精度、召回率、PR曲线、F1值、ROC曲线、AUC值等。

1.1.2 数据可解释模组

从数据的视角来看，数据可解释模组基于对数据集的观测与干预对AI推理或决策进行归因，产生对AI结果或行为的解释。从数据集视角来看，解释分为因果性解释和相关性解释，通常用因果图或特征重要性的形式来展现，具体组成如下。

（1）因果性解释模块

该模块采用因果分析算法对观察数据进行分析，可以找出数据中隐含的因果关系。该模块由基于对照的因果发现方法、基于约束的因果发现方法、针对因果函数模型的因果发现方法、针对模糊变量的因果发现方法、因果网络学习方法等组件组成。基于对照的因果发现方法组件通过不同组之间的对照，比较得出平均干预效应来完成因果学习。该类方法包括实验性方法及观测性方法两大类，其中实验性方法包括随机对照试验方法、A/B测试方法等，观测性方法包括分层、匹配、重赋权等。基于约束的因果发现方法组件采用先确定因果关系结构再确定结构中的方向的方法来完成因果学习，这一类方法包括IC算法、FCI算法等。针对因果函数模型的因果发现方法组件利用因果数据生成机制引起数据分布不对称从而分析变量之间的因果关系，这一类方法包括ANM算法、LiNGAM算法等。针对模糊变量的因果发现方法组件可以对包含隐变量影响的数据进行因果分析，通常采用Cornfield不等式、工具变量法、阴性对照法等。因果网络学习模块利用构建因果网络来表示数据特征及标签之间的完整因果关系，因果网络学习方法包括局部学习、主动学习与分解学习3种：局部学习方法通常会给定一个目标变量来寻找原因与结果，得到局部的因果网络，并以此为基础构建出完整的因果网络；主动学习方法通过干预某变量的方法，将变量之间的相关关系转变为因果关系，从而得到因果网络；分解学习方法通过将大网络的学习分解为小网络的学习来构建整体网络。

（2）相关性解释模块

该模块采用相关性算法，分析观察数据，找到数据中隐含的相关关系，由频率方法与贝叶斯方法两个模块组成。频率方法模块基于频率学派的思想来计算特征之间的相关程度，模块中有PCC、KCCA、CMI等算法，PCC算法（皮尔逊相关系数）可以检测两个变量之间的线性相关关系；KCCA算法组件可以把低维的数据映射到高维的核函数特征空间，然后在核函数空间分析变量间的关联关系；CMI算法组件利用信息论中的条件互信息来衡量随机变量之间的相关性；贝叶斯方法组件基于贝叶斯学派思想计算特征之间的相关度，例如概率图模型生成特征之间的贝叶斯网络来表示特征与标签之间的因果关系。

1.1.3 过程可解释模组

过程可解释模组可以将AI运行的黑

盒打开，提供AI系统的计算逻辑、算子处理及中间结果的可视化解释，帮助用户理解模型的运行机理，从而达到解释计算逻辑的目的。过程可解释模组的具体组成如下。

（1）计算逻辑可视化模块

该模块通过可视化模型的计算逻辑为相关方提供模型工作过程的直观解释。计算逻辑可视化模块由计算图可视化组件与模型结构可视化组件组成，计算图可视化组件采用有向无环图（directedacyclic graph，DAG）或控制流图（control flow graph，CFG）来表示模型计算过程中数据的流转与计算方以及模型内各种计算之间的相互依赖关系；模型结构可视化组件以图文的形式展示模型的静态结构，直观解释了模型的组成。

（2）算子可视化模块

该模块通过对模型计算过程中卷积层、激活层等组成算子的效果进行可视化展示，帮助相关方理解模型中各算子的效果。算子可视化模块由激活层可视化与卷积层可视化组件组成，可以对激活层与卷积层在计算过程中的输出进行可视化。

（3）中间数据可视化模块

该模块通过对模型计算过程中的中间数据进行可视化展示，帮助相关方理解模型训练或推理过程中的数据变化。中间数据可视化模块由梯度可视化、数据血缘展示和计算过程可视化组件组成。梯度可视化组件可以对模型产生的梯度进行可视化，在一定程度上解释了模型对样本的决策依据；数据血缘组件可以对AI应用中使用的数据进行血缘展示，帮助AI应用相关方了解数据的来龙去脉；计算过程可视化组件对AI应用过程中数据的变化进行可视化展示，帮助AI应用相关方更好地了解数据的变化过程。

1.1.4 模型可解释模组

模型可解释模组可以对模型的决策逻辑进行解释，通过生成输入的显著图、效应分析及提供参数的意义说明等方式，帮助相关方理解模型做出最终决策的逻辑。另外，通过对多模型的编排逻辑进行说明，帮助相关方理解当前多模型编排的理由。该模组由显著性解释、效应分析可解释、参数可解释与多模型可解释模块组成，具体组成如下。

（1）显著性解释模块

该模块通过可视化影响预测的重要输入像素来解释模型做出决策的原因，侧重于输入，但忽略了对模型如何做出决策的解释。该模块由基于映射的显著性方法组件与基于反向传播的显著性方法组件组成，基于映射的显著性方法有CAM与GradCAM等。CAM组件通过对特征图进行线性加权来获得类别热力图，与原图叠加来可视化对预测输出的贡献分布。GradCAM[27]组件利用模型输出对激活映射（特征层）的梯度信息计算权重，生成显著性映射；基于反向传播的显著性方法组件由LRP、DeepLIFT、Guided BP等方法组成。LRP[28]和DeepLIFT[29]组件利用分层相关传播思想，利用自上而下的相关传播规则，生成视觉解释；Guided BP[30]组件通过ReLU非线性反向传播时操纵梯度的视觉解释，反向传播时保留了梯度与激活值均为正的部分，生成显著图。

（2）效应分析解释模块

该模块由PDP、ICE与ALE等组件组成。PDP算法组件可以显示特征对机器学习模型的预测结果的边际效应，显示结果与特征之间的复杂关系；ICE算法组件可以显示特征更改时实例的预测如何改变；ALE算法组件可以显示特征平均对机器学

习模型预测的影响。

(3)参数可解释模块

该模块通过对模型中参数的解读来帮助用户更好地理解模型，由模型参数解释组件与超参解释组件组成。模型参数解释组件会对参数意义进行说明，并对模型调参依据的原因进行解释，例如，在决策树模型中，对决策树中每一个节点的划分逻辑进行说明，解释决策路径的形成逻辑；超参解释组件会说明超参数的意义和超参调整的逻辑要求。比如，再随机森林算法需要对n_estimators超参进行解释，该参数代表决策树数量，在一定范围内，提高树的数量有助于提升模型性能，但达到一定数量后，性能会保持稳定，提升的空间程度较小，同时增加该参数会增加模型的计算量。

1.1.5 结果可解释模组

结果可解释模组可以对模型的决策结果进行解释。一方面，可以对模型做的具体决策进行解释；另一方面，可以对模型决策的有效性进行说明，帮助相关方了解模型具体决策的可信程度。需要注意的是，在使用结果可解释模组时要注意数据分布外(out of distribution，OOD)问题，即当分析的目标样本与模型使用的训练数据存在偏倚时，结果可解释模组的结果不再是可信任的。因此，在使用结果可解释模组之前，需要对模型样本及待分析样本做分布分层分析。结果可解释模组由基于样本扰动的特征归因解释模块与基于检验结果有效性解释模块组成，具体组成如下。

(1)基于样本扰动的特征归因解释模块

该模块通过观察对输入进行扰动后模型预测的变化，对模型对该样本的决策进行特征归因。该模组由LIME[8]、SHAP[31]等算法组件组成。LIME组件基于想要解释的预测值及其附近的样本构建局部的线性模型或其他代理模型。SHAP是基于博弈论思想的一种黑盒模型事后归因解释方法，通过将模型的预测解释分解为每个特征的贡献，计算每个特征的夏普利值(Shapely value)，从而了解每个特征对最终决策的贡献度。

(2)基于检验结果有效性解释模块

该模块评估估计值与总体参数的合理误差，以概率的形式量化结果的有效性。常用的衡量方法包括P值方法、F检验、T检验以及卡方分析等。

通过原理性解释模组、结果可解释模组、数据可解释模组、过程可解释模组、模型可解释模组的协作，框架具备了原子解释能力，这些原子解释将作为后续环节的输入，为XAI的工程化落地提供解释基础。

1.2 XAI核心能力层

XAI核心能力层在基础能力层之上对原子解释进行进一步的处理，提供XAI工程化落地的核心能力。基础能力层的原子解释通常存在主客观评价不一致问题、解释缺乏客观性问题[32]、解释的合规程度问题、对同一事件产生多个“真实”解释的罗生门问题[21]等。这些问题导致了原子解释的“有效性”与“可用性”存在着先天不足，这也是XAI难以商用落地的主因之一。

XAI核心能力层的目标是解决上述问题，提供“有效”且“可用”的解释。“有效”指的是对同一问题的不同视角的解释能够保持一致，“可用”指的是解释能够适应AI应用场景需要，满足场景中不同角色的解释需求。如**图3**所示，XAI核心能力层由知识图谱(knowledge graph，KG)库、

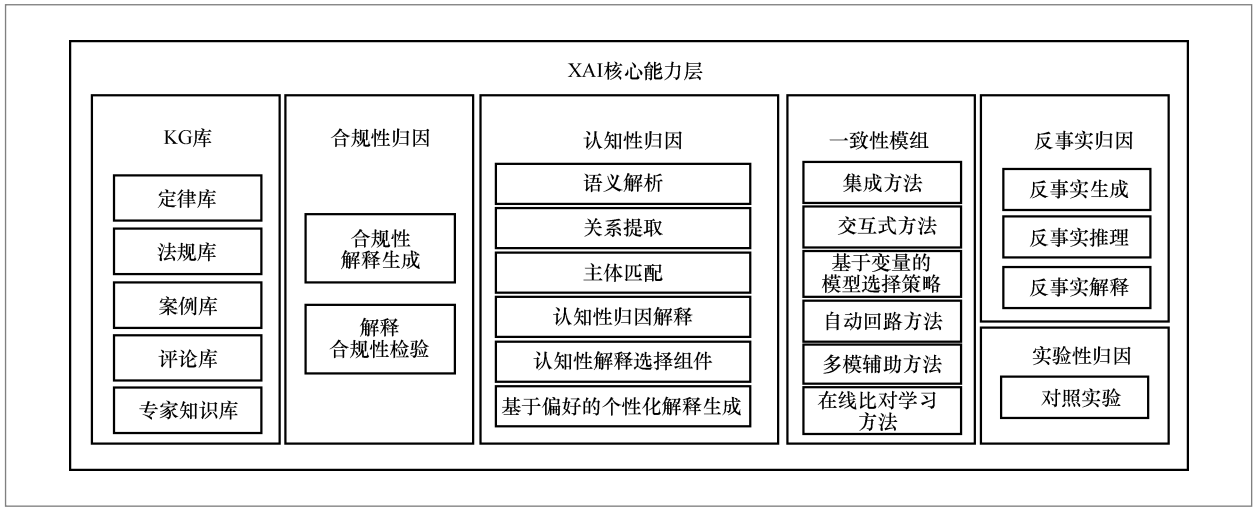


图 3 XAI 核心能力层

合规性归因模组、认知性归因模组、一致性模组、反事实归因模组与实验性归因模组组成。KG库为原子解释提供了世界知识的支撑，世界知识指的是世界上发生的一些真实事件，包含主客观知识，通常分为事实型知识及常识性知识。合规性归因保证归因的结果合乎法律规定，认知性归因确保归因的结果与用户的客观评价保持一致。一致性模组针对罗生门问题，可以保持对同一事件产生的多个解释的一致性。反事实归因保证归因的结果经得起假设场景的验证。

1.2.1 KG库

KG库是结构化的语义知识库，在XAI领域用来存储并提供定律、法规、案例、评价、专家知识等方面的现有知识。其中定律知识泛指当前公认或受到广泛认同的客观定律，如物理定律、化学公式等；法规知识库指从现行的法律法规中抽取出来的知识，不同应用领域的法规知识库各不相同，且具有很强的时效性；案例知识是从具体案例中抽取出来的知识，是对法规库的有力补充，也需要跟随法规库的变动进行更新；评论库包含从各种交互渠道中提取出来的知识，比如用户评价、评分表等，这些包含人们对客观事物的感受、理解以及思考的评价结果，有一定的认知合理性，可以作为重要依据辅助优化XAI的解释结果。这些知识以三元组等形式存储在KG库中，便于XAI核心能力层的其他组件使用。KG库模组由定律库、法规库、案例库与评论库等模块组成，具体介绍如下。

（1）定律库

定律库提供各领域的客观知识，而客观知识的融合可以使XAI解释结果更加可靠。定律库中的知识来源于不同的领域，可根据应用场景、用户行为等上下文信息选择合适的领域知识，从而提供多维度、有逻辑的解释。以基于神经网络的嵌入向量（embedding）构建的个性化推荐系统的可解释性为例，引入商品知识来寻找推荐的可解释路径。例如，引入类似WordNet[33]中的物品层级来解释给客户推荐萨摩耶犬的原因，客户浏览过哈士奇与阿拉斯加的相关信息，因此给客户推荐了

同为工作犬的萨摩耶犬。引入定律库使解释更加符合客观定律及通用常识。

(2)法规库

该模块可以从各领域内的法律规范条文中提取相关知识，例如，从法律文本中提取出类似于(人员甲，提供辩护，人员乙)、(检察院，起诉，人员丙)的三元组，供合规性解释生成、合规性检验等模块使用。合规性的要求在强监管领域更重要。法规库的难点在于实时更新，需要投入人力构建自动化流程，监控领域内的法规变化，自动抽取最新的法规知识并结合人工监督形成确实可用的法规库。

(3)案例库

该模块是对法规库的补充，法规库提供的法规知识通常缺乏细节，在生成细粒度的解释时，需要使用案例库中的知识进行细节补充。案例库的难点有两个：一是时效性，需要与法规库保持一致；二是解读方式，不同的解读方式会提取出不同的领域知识，需要借助领域专家的经验提取出正确的案例知识。

(4)评论库

该模块在提供者与接收者之间形成交流闭环，保证最终生成的解释与用户的需求一致。解释接收者来源广泛，可以为解释提供多视角、多模态的评价。解释接收者的满意度也是XAI的重要目标，可以从客户的评价中抽取知识作为解释的一个优化标准。

1.2.2 合规性归因模组

合规性归因模组的作用是保证解释符合伦理法规要求。具体来说，该模组可以将KG库中的合规性知识(如法规库知识、案例库知识)作用在原子解释上，有两种作用方式：一是在解释的生成阶段引入合规性知识，生成原生性的合规解释；二是在解释生成后，基于合规性知识对解释进行筛选，挑选出最优的合规解释。合规性归因模组由合规性解释生成模块与解释合规性检验模块组成，具体介绍如下。

(1)合规性解释生成模块

该模块在解释的生成阶段引入合规性知识，通过在解释生成之前或过程中注入合规性知识，保证生成的解释能够符合法律法规的要求。例如，在信贷审批业务中，存在“年龄小于18，则不能贷款”这一合规性规则，在遇到未成年申请贷款时，优先基于该规则给出拒绝放款的解释。

(2)解释合规性检验模块

该模块在解释结果已经生成的情况下，使用KG库中的合规性知识对解释结果进行进一步的筛选，去除不合规范的解释，挑选符合场景需要的合规解释。例如，在解释推荐贷款产品的原因时，“因为用户为女性，所以推荐了该信用贷产品”这一原子解释与“男女平等”的合规知识相冲突，因此该原子解释被去除。

1.2.3 认知性归因模组

认知性归因模组可以让用户参与解释的生成过程[34]，使解释尽可能满足人们的可解释性需求[35]。认知性归因模组使用的知识来源于KG库中的评论库，可以通过语义解析、关系提取、主体匹配等技术，从评论库提取出认知性知识，进一步构建认知性的归因解释。同时，认知归因模组可以基于认知性知识对原子解释进行评估，预测解释接收者对该解释的认可程度。此外，认知归因模组还可以基于用户画像生成该用户最易接受的解释内容。认知归因模组由认知性知识解析模块、认知性归因解释生成模块、认知性解释选择模块、个性化解释生成模块组成，具体介绍如下。

（1）认知性知识解析模块

该模块由语义解析、关系提取、实体匹配3个组件组成。语义解析组件对评论库中的评论进行分析，得到评论的逻辑表示；关系提取组件从文本中抽取两个或多个实体之间的语义关系；实体匹配组件从文本提取的所有实体中找出所有相同实体。通过3个组件的协同工作，该模块提取出评论中包含的知识，构建并更新评论知识图谱，是该模组中的其他模块的认知性知识基础。

（2）认知性归因解释生成模块

该模块基于与用户的交互信息生成解释，保证生成的解释符合用户的解释要求。用户对历史解释的评价保存在评价库中，通过认知性知识解析模块抽取用户评价，形成当前领域的评价知识图谱。由于知识图谱本身具有一定的可解释性，可以结合评价知识图谱中的路径来解释模型的决策，这样生成的认知性归因解释不会偏离用户的解释要求。

（3）认知性解释选择模块

该模块可以过滤不符合用户解释要求的解释结果，可以应用于认知性归因解释生成模块生成的解释，也可以应用于其他模块生成的普通解释。通过分析评论库中的历史交互信息，该模块可以统计出用户解释要求中的重要因素，然后过滤掉缺乏这些因素的解释，得到符合用户解释要求的解释。

（4）个性化解释生成模块

该模块对认知性解释选择组件进行了增强，由于最终用户的背景、偏好不同，最终用户对同样的解释的满意度也不尽相同。该模块通过对最终用户历史评价信息的分析，统计该用户对解释的偏好，在筛选生成后的解释时，优先选择符合该用户偏好的解释，这样能有效提升用户对解释的满意度。

1.2.4 一致性模组

一致性模组主要是解决XAI解释结果的罗生门问题的模组。该模组除了包含集成方法、交互式解释和基于变量的模型选择策略等传统罗生门问题的解决方法，还加入了自动回路、多模辅助、在线对比学习等新方法，这些新方法弥补了传统方法适用范围不广泛、解决问题不彻底等缺陷，可以更好地解决罗生门问题。具体介绍如下。

（1）集成方法模块

该模块使用Bagging算法让罗生门集合中的所有模型共同表决测试样例的输出[36]。以Bagging方式聚合罗生门集合中的大量竞争模型，或者说对罗生门集合中的差异性进行平均，可以在一定程度上提高稳定性和准确性，减少非唯一性。虽然集成方法可以在一定程度上缓解罗生门问题，但并没有从根本上解决罗生门问题。

（2）交互式方法模块

现有的交互式模型选择方法包含全自动、半自动和人工选择等方式。交互式方法结合了降维、线性加权与用户的连续反馈，从罗生门集中选择出备择的最优解释。通过用户的连续反馈，可以在不断变化的环境中持续改进模型预测[37]，在复杂多变的罗生门问题中不断精进最佳选择。交互式方法通过引入用户反馈提高了效率以及用户对XAI决策的信任度，但该方法也存在一定的局限性，比如使用范围较为局限，大部分现有的交互式模型选择方法的搜索空间为整个假设空间，而非罗生门集合。

（3）基于变量的模型选择策略模块

该模块可以得到不同变量对不同方案的重要程度以及不同变量之间的依赖程度和影响，因此，可以根据实际情况选择最

合适的解决方案。基于变量的模型选择策略可以进一步提高模型的可解释性，在一定程度上揭示了变量与模型的关系以及变量之间的关系，帮助用户做出决策。但是该方法的使用范围也具有一定的局限性，对某些对变量要求不高的模型分辨能力较差。

（4）自动回路方法模块

该模块将XAI目标的求解过程转化为基于假设及检验的探索实证处理过程，使得解释结果的准确性、适用性更高，同时增加了用户对解释结果的信任程度。具体来说，该模块将人类的理性认知作为原始解释结果的一种类型，自动化地从数据、经验、模型和认知评价4个方向对原始解释进行扩增处理，得到更全面的解空间。然后，通过图结构校验方法对不确定解释结果进行结构化，使其符合正常的推理逻辑。接着，对结构图中的不可直接观测的变量进行自动化指代替换，从而对解释中的因果关系进行代理验证。最后，对解释结果进行进一步的解析，并通过自动化的方式对待检验的“因果对”进行实证检验，引入了人类对解释结果的理性评判，输出具有置信度、推理结构、实证支撑的解释性结果，使结果更准确、可靠、可信。

（5）多模辅助方法模块

该模块通过数据、经验常识、人类认知等多维度辅助信息来解决XAI的罗生门问题。在XAI模型学习的过程中，能够同时支持数据面、经验知识面、人类认知面的多维度数据输入，并根据辅助学习的优化需求，在模型学习过程中提取数据统计面、经验知识面、人类认知面的解释结果的共因，使XAI解释结果既符合人类决策优化的常理，又兼备逻辑完备性，实现内在的一致性。

（6）在线对比学习方法模块

该模块针对XAI技术落地实施过程中出现的不一致性问题，根据实际情况选择不同类型的在线学习方法，然后以孪生学习的形式对多个比对结果的一致性进行考察，经过多轮在线学习对XAI模型进行优化，实现所有比对结果的最终一致性。

1.2.5 反事实归因模组

反事实归因模组通过反事实方法提升解释的泛化能力。反事实就是对实验组中的研究对象实施不同的干预，从而得到不同干预对应的潜在反事实结果，在潜在结果比较的基础之上，可对因子的影响进行分析，对已有的结果进行归因性解释。反事实解释通常具有很强的因果性，因此基于解释结果还可以进一步执行反事实推理。该模组由反事实生成、反事实推理与反事实解释组件组成，具体组成如下。

（1）反事实生成模块

该模块可以基于干预方式生成反事实结果（潜在结果），还可以通过反事实样本生成消除数据集的偏差，帮助提升模型泛化能力，消除模型中的偏见[38]。

（2）反事实推理模块

反事实推理是一种特殊的因果推理方法[39]，通过未发生的条件来推理可能的结果。可以替换不真实的条件或可能性并进行因果推理，反事实推理组件可以辅助打开模型黑盒，对AI模型进行更深入的探索[40]。

（3）反事实解释模块

该模块可以回答“如果输入的特征发生了某种特定变化之后，输出的结果将如何变化”这一问题，并在特征和预测结果之间建立因果关系，帮助确定对模型的决策具有因果关系的特征，解释模型的决策过程。

1.2.6 实验性归因模组

实验性归因模组通过对照实验给出解

释结果，可以在客观状态下以及在排除其他自变量干扰的条件下，操控一个或多个假定有关的自变量，并观测其对某些因变量的效应差。具体介绍如下。

对照试验模块进行随机对照实验的增强型测试，可以同步进行多组堆叠性测试，实施形式可以为A/B测试[41]、蓝绿部署[42]、冠军/挑战者试验[43]等。测试在不同干预情况下某一个变量产生的效应差异，基于效应差异分析造成不同结果的原因。例如，A/B测试通过对比只有一个变量不同的两个版本的表现来研究该变量的作用以及影响，其中版本A可能是当前正在使用的版本，而版本B是改进版，通过比较版本A、B的使用效果差异来确定测试变量的改进方向。

XAI核心能力层是XAI工程化落地核心能力的提供者，是XAI工程化落地的最重要环节。XAI核心能力层使解释主客观一致、合规合理，与XAI业务中各角色的核心要求保持一致，符合用户解释要求等。

1.3 XAI业务组件层

XAI业务组件层是连接XAI核心技术能力与XAI应用实施落地的纽带，提供了XAI应用开发必需的业务能力组件。XAI业务组件层主要解决了XAI应用落地的表达、展示和与AI Pipeline结合3个问题。

如**图4**所示，XAI业务组件层由XAI表征模组、XAI展示模组、Pipeline嵌入模组组成。其中，XAI表征模组负责解决XAI解

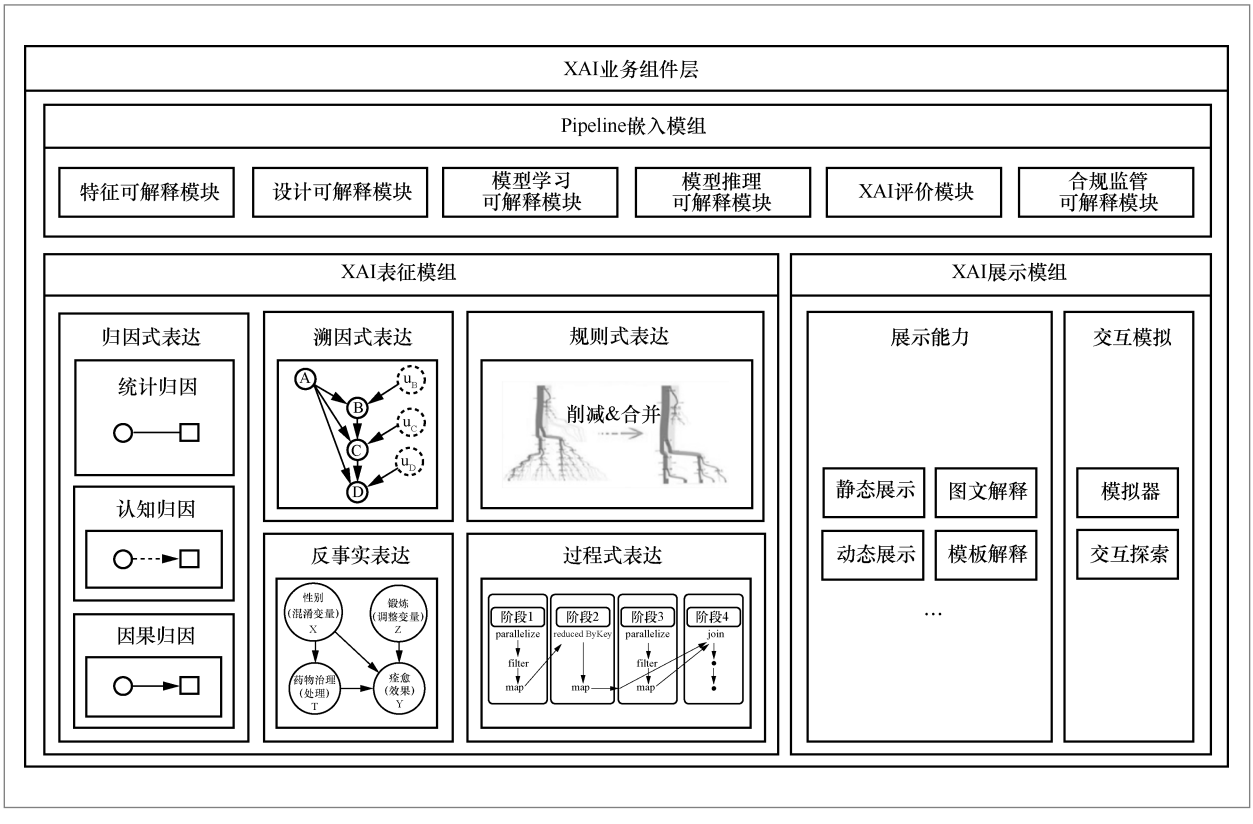


图 4 XAI 业务组件层

释结果表达的问题；XAI展示模组负责解决XAI解释结果如何在具体应用中展示的问题；Pipeline嵌入模组负责将XAI能力嵌入AI应用的Pipeline中，为传统AI应用赋予可解释能力，增强AI应用的可信可理解程度。3个组件相互协作，完成XAI技术的业务流程。

虽然AI应用开发者设计了算法模型，但通常不了解某个参数的权重以及产生某个结果的原因。AI应用开发者、XAI系统使用者、XAI最终用户、监管者等不同角色的知识背景、对可解释的要求并不相同，因此，需要使用不同的表征方式来呈现XAI解释结果。开发者需要借助可解释性技术来充分理解AI系统，及时发现、分析、修正缺陷；XAI的最终用户需要理解AI系统的内在决策过程，从而提高对AI系统的信任程度；监管者更看重XAI技术对AI应用监管效率的提升。由于这些不同角色的需求各不相同，知识背景也参差不齐，需要使用不同的表征方式来呈现XAI解释结果。

1.3.1 XAI表征模组

XAI表征模组由归因式表达、溯因式表达、规则式表达、反事实表达与过程式表达模块组成，可以为各角色提供最合适的解释表征形式，具体介绍如下。

（1）归因式表达模块

该模块负责将AI模型的输出结果归因到模型输入的具体特征或特征组合，同时采用特征权重来解释特征或特征组合对模型决策结果的影响。根据归因方式的不同，归因式表达模块可分为统计归因子模块、认知归因子模块与因果归因子模块，其中统计归因通过统计变量与推理结果之间的相关关系来完成相关性归因；认知归因将模型的推理结果与人类的认知习惯相结合，计算来源于主观评价或知识库的认知因素与推理结果之间的相关关系，从而完成认知性归因；因果归因采用因果分析方法，分析特征与推理结果之间的因果关系，从而完成因果性归因。

（2）溯因式表达模块

该模块用来对溯因式解释进行表达，溯因式解释从结果出发，推测出事件发生的原因，通常是基于特征之间的因果关系，生成特征的影响链，可以通过对故障或决策的溯因，找到故障或决策的根本原因。通常来说，可以使用DAG图来完成溯因式表达。

（3）规则式表达模块

该模块可以将解释结果表达为类似于“if … then …”形式的规则。规则类知识符合人类认知规律，是人类容易理解的知识形式之一。规则类知识主要来源于专家知识积累，也可以使用DeepRED等算法从模型中抽取，最终得到的规则类知识存储在KG库中。

（4）反事实表达模块

该模块通过反事实的形式来表达解释，通过潜在结果模型，研究特定干预下模型的潜在结果，在特征和预测结果之间建立因果关系，从而帮助XAI用户更好地理解模型。

（5）过程式表达模块

该模块通过展示全流程的过程状态来解释AI应用，帮助最终用户理解模型的决策原因，这种解释方法的理解具有一定的知识门槛，更适合具有一定专业背景的专业人士。

1.3.2 XAI展示模组

XAI展示模组由展示能力与交互模拟组成，负责展示XAI解释结果，具体介绍如下。

（1）展示能力模块

该模块提供各种解释的展示能力，可

以选择静态展示、动态展示等展示模式，也可以选择图文或模板式等展示方式。静态展示组件可以对原理展示模块中的知识进行静态展示，也可以借助Matplotlib等绘图模块对数据集、模型、模型结果进行图表展示，如数据特征与标签的因果图展示等；动态展示组件使用预录制好的视频资料、动态页面等方式，为用户提供算法原理的动态演示，例如逻辑回归算法原理演示视频、图形化的GRS检验物理意义展示等。在展示方式上，可以使用图像或文字的方式对解释结果进行展示，更进一步地，还可以使用模板式展示方式，将可解释输出嵌入事先准备好的可解释模板。使用模版生成的解释可以复用过去的经验，减少了生成难度，同时，模版化的解释可以与业务相结合，便于最终用户的理解。

（2）交互模拟模块

该模块由原理模拟器与算法探索组件组成，为用户提供一种模拟验证、交互探索的方式。交互可以让解释易于理解，让最终用户理解原理时感觉更加方便、容易，更高效地完成原理学习，达成预期目标。原理模拟器子模块向用户提供算法的模拟数据、模拟逻辑和交互接口等，在交互过程中向用户提供算法的原理性解释。例如，使用者通过主动操作、知识问答等方式与XGBoost等算法运行原理的示例进行互动，从而加深对算法原理的理解；算法探索组件借助架构提供的交互式算法模板，组件中的各类数据探索方法可以在交互过程中帮助用户完成数据分布、异常值等探索，多模块组合探索功能可以帮助用户完成对算法中各模块组合效率的探索，帮助用户设计出最合适的算法流程。例如，通过模块中的引导步骤，引导用户探索不同的特征工程对模型结果的影响，辅助用户完成算法构建过程中的各种探索。

1.3.3 Pipeline嵌入模组

Pipeline嵌入模组由特征可解释模块、设计可解释模块、模型学习可解释模块、模型推理可解释模块、XAI评价模块、合规监管可解释模块组成，覆盖了AI应用的整个生命周期，其目标是将XAI与现有AI Pipeline进行融合。传统AI应用已经形成了比较成熟完善的AI应用流程，XAI的工程化落地应该在不影响原有AI应用的前提下，尽可能提供更高的附加价值，将传统AI应用升级为可信可解释AI应用。具体介绍如下。

（1）特征可解释模块

该模块将特征相关的XAI基础能力层与核心层中的模块嵌入AI流程中。以研发过程中数据探索过程为例，该模块将XAI基础能力层中的数据可解释模组嵌入当前的流程中。当研发人员在平台上进行特征工程时，数据可解释模组不但不会影响原有的特征层工程流程，还可以通过因果性解释模块帮助研发人员发现数据中的因果关系，更好地完成特征工程工作。

（2）设计可解释模块

该模块应用于AI Pipeline的算法设计、模型工程设计阶段，一方面，XAI可以应用于单个算法的设计，例如基于XAI基础能力层中的模型可解释模组，辅助开发人员更科学地设计出有理论依据的模型；另一方面，在模型编排设计过程中，同步提供模型编排设计的原理性解释，帮助设计者完成更合理的多模型系统性设计。此模块可以帮助开发者、用户以及管理者从设计原理层面更科学地开展研发、使用、监管等相关工作。

（3）模型学习可解释模块

该模块将XAI应用嵌入模型训练阶段以提升模型的质量与效率。借助该模块，可以在模型训练过程中嵌入过程可解释模

块，通过计算图可视化、卷积层可视化、梯度可视化等组件了解模型训练的状态；也可以在模型训练中嵌入模型可解释模组，借助参数可解释模块解释参数、完成参数的选择，借助显著性解释与效应分析解释模块对模型进行可解释性分析，判断得到的模型是否可靠、可信；也可以在模型训练过程中加入反事实归因模组，提高模型的鲁棒性。

(4)模型推理可解释模块

该模块在模型推理阶段对XAI技术进行编排。在模型推理阶段，除了借助LIME、SHAP等样本扰动方式对样本推理结果进行解释之外，还需要注意引入XAI核心能力层中的一致性模组、KG库等模组，解决解释的罗生门问题、主客观不一致问题等。以推荐系统的推荐结果解释为例，该模块可以将结果可解释模块嵌入推荐结果生成的流程中，生成解释结果，与推荐结果一起提供给最终用户，还需要在流程中嵌入认知性归因模组与KG库，提取用户对推荐解释的评价中的知识，并存储在KG库中。

(5)XAI评价模块

该模块负责对生成的解释进行评价，AI应用中各角色的需求并不相同，对XAI评价也各有侧重，因此，需要多种评价方法对解释的好坏进行评估。该模块由技术级评估、用户级评估与系统级评估3个组件组成。

技术级评估组件由效率指标、质量指标、交互性指标、性能指标与效能指标子组件组成。效率指标评估系统给出解释的速度。质量指标关注评价的质量，考量不同的解释形式表达的充分性，如可视化程度、文字描述的可读性、复杂性等，在解释的内容部分会考量解释的透明度和完备性等。交互性指标考察解释是否满足之前的问题解决建议、是否具有可修正性等。性能指标考量解释的准确性、稳定性、可比较性等。效能指标考察解释对问题解决的帮助程度、人工干预程度和异常事件发现能力等。

用户级评估组件由解释满意度、解释优良性与信任评估子组件组成。解释满意度指的是解释的接收方对解释的满意程度，需要在生成解释后，收集解释的接收者对解释的满意程度，并根据结果对解释结果进行调整，提高最终生成的解释的质量。解释优良性是指框架内部对生成的解释的评价，需要在内部形成解释优良性的评价标准，该标准需要考虑到客户满意度、资源消耗、生成效率、专业度等多种客观指标，并通过权重调整来完成对不同场景的适配。信任评估指的是给出解释后，解释接收者对该解释结果的信任程度，需要在收集的最终用户对解释结果的信任程度的基础上形成一个评估预测系统，对解释结果进行过滤，只保留信任评估得分较高的解释。

系统级评估由整体公平性与整体合规性两个子组件组成。整体公平性子组件负责判断系统行为是否违反公平性标准，针对具体的AI应用，设计出专用的公平算法，例如使用机会均等的公平标准，判断系统最近有无针对老年人的歧视行为发生等。整体合规性子组件负责提供判断系统行为是否违反相关法律法规的标准，需要对AI应用可能涉及的法律法规及伦理问题进行事前、事中及事后结果的监管。

(6)合规监管可解释模块

该模块从模型、事件、系统3个层面对AI系统进行合规监管，引入XAI技术，实现AI决策可解释、事件可追溯、责任可定位，行为符合法律法规的监管要求。模型级合规监管追踪模型内部的决策过程，并对决策结果进行解释；事件级合规监管在异常事件发生后对该事件的全流程进行回溯，通过将问题的原因定位到具体的子流

程来对异常事件的追溯进行细化；系统级的合规监管需要对AI系统的整体行为进行实时监控及周期性复盘，通过XAI等技术对违规行为进行预警与处理。该模块的目标是通过这3个层面的合规监管的协作，建立合法合规、公平公正、行为可解释、结果可追溯的可靠、可控、可信的AI系统。

XAI业务组件层为XAI业务应用开发提供技术支撑，关注XAI的表征、展示以及与传统AI业务的融合，以Pipeline无缝嵌入的方式为XAI的实施落地提供便利，提高了XAI工程化落地的效率。

1.4 XAI应用层

基于XAI能力组件及业务组件，可以实现丰富新颖的可信AI应用。如**图5**所示，XAI应用层包括可解释图计算系统、可信可解释AI学习平台、可信AI算法管理平台、可解释推荐系统、可解释溯因系统、异常事件追溯系统、AI合规监管系统、AI决策风险控制系统等。

XAI应用层中的不同组件可以满足各相关方的不同需求，如可信可解释AI学习平台可以帮助平台开发者对研发过程中的模型进行解释，提升模型研发的效率；异常事件追溯系统可以帮助业务运营者追溯业务运行过程中的异常，快速定位业务问题，异常事件追溯系统和AI合规监管系统可以帮助监管者判断AI平台在可解释过程中的行为是否合规，并对违规的行为进行溯源定责；可解释推荐系统可以帮助最终用户获取AI系统生成推荐的具体原因，提升用户满意度。

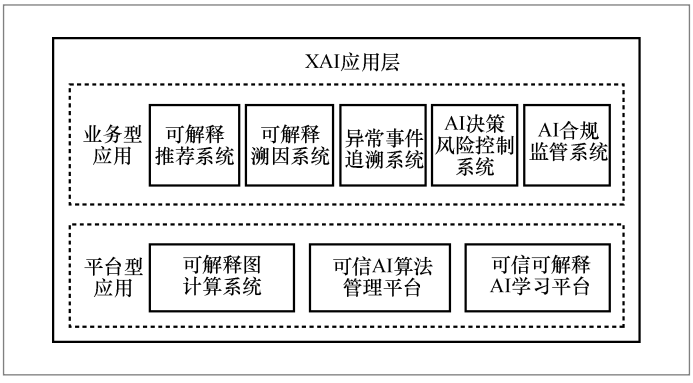


图 5 XAI 应用层

可解释图计算系统引入KG库等外部知识来辅助解释图计算系统的决策行为。可信AI算法管理平台通过XAI实现算法透明性、公平性、安全性的评估及管理，进一步打造双向透明可信的算法或模型交易市场。可信可解释AI学习平台是对传统AI训练平台的增强，实现了AI模型学习全流程的透明可信，做到数据集管理、算法设计、模型学习、模型优化、模型测试全流程的可解释。

可解释推荐系统在提供推荐结果的同时，提供推荐的依据及原因，可以提升推荐系统的透明度，提升用户对推荐系统的信任度、接受度及满意度。可解释溯因系统的作用是提供模型、系统、事件级的溯因分析，实现AI行为可解释、事件可追溯。异常事件追溯系统可以提供针对异常事件的检测、发现、诊断、溯因等应用能力。类似地，基于XAI对AI系统过程及决策结果的解释能力，AI决策风险控制系统可以对OOD、对抗样本等问题进行主动管理，AI合规监管系统可以判断AI系统是否存在违反伦理道德、法律法规的行为。这些场景仅仅是XAI应用的一小部分，相信随着XAI技术的发展及深入应用，XAI会在更多的领域开花结果。

2 应用案例

Transwarp Sophon-XAI是星环信息科技（上海）股份有限公司打造的可信可解释AI套件，在Sophon人工智能平台中的位置如**图6**所示。

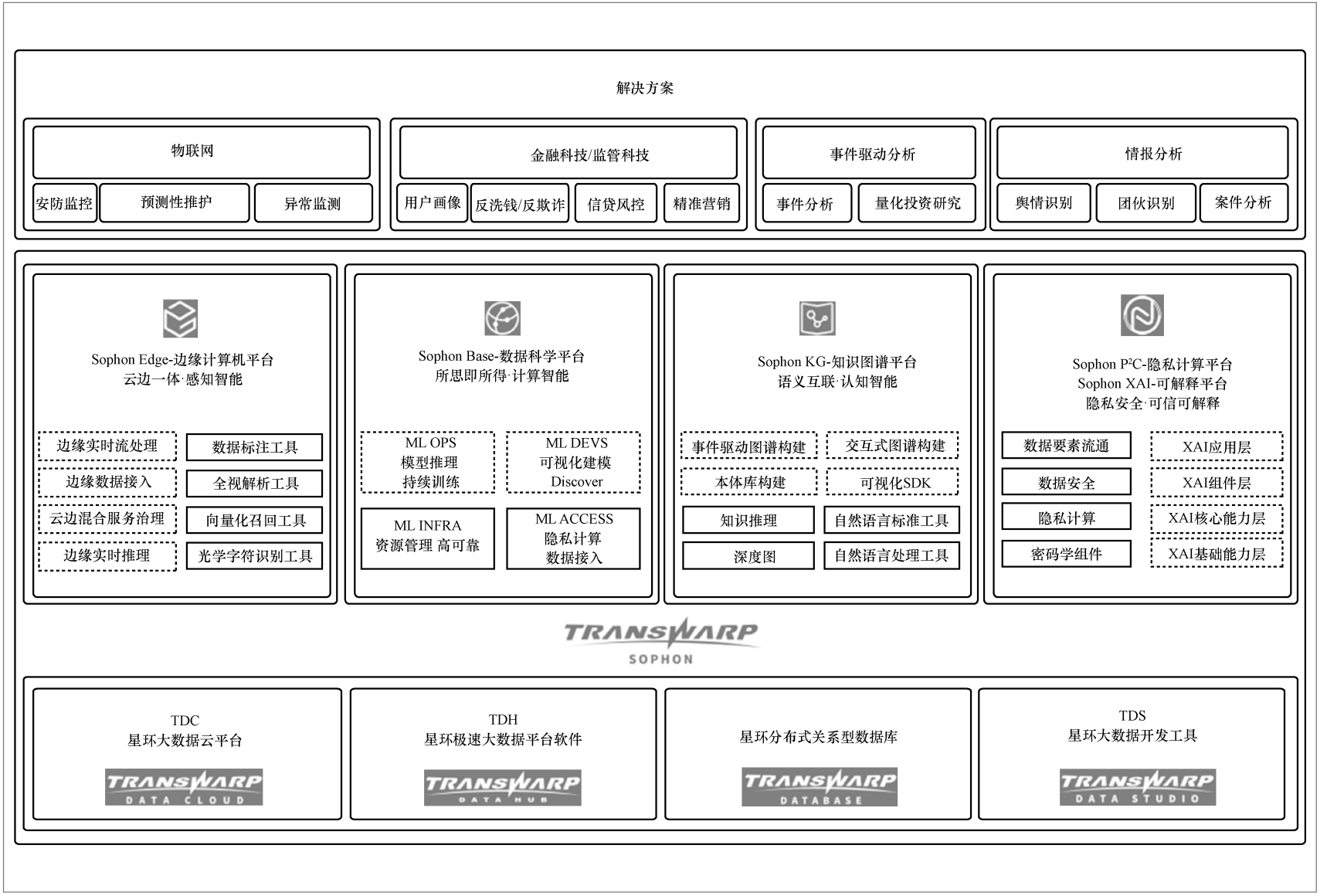

解决方案
物联网
安防监控
预测性推护
异常监测
金融科技/监管科技
用户画像
反洗钱/反欺诈
信贷风控
精准营销
事件驱动分析
事件分析
量化投资研究
情报分析
舆情识别
团伙识别
案件分析
Sophon Edge-边缘计算机平台
云边一体·感知智能
边缘实时流处理
数据标注工具
边缘数据接入
全视解析工具
云边混合服务治理
向量化召回工具
边缘实时推理
光学字符识别工具
Sophon Base-数据科学平台
所思即所得·计算智能
ML OPS
模型推理
持续训练
ML DEVS
可视化建模
Discover
ML INFRA
资源管理 高可靠
ML ACCESS
隐私计算
数据接入
Sophon KG-知识图谱平台
语义互联·认知智能
事件驱动图谱构建
交互式图谱构建
本体库构建
可视化SDK
知识推理
自然语言标准工具
深度图
自然语言处理工具
Sophon P²C-隐私计算平台
Sophon XAI-可解释平台
隐私安全·可信可解释
数据要素流通
XAI应用层
数据安全
XAI组件层
隐私计算
XAI核心能力层
密码学组件
XAI基础能力层
TRANSWARP
SOPHON
TDC
星环大数据云平台
TRANSWARP DATA CLOUD
TDH
星环极速大数据平台软件
TRANSWARP DATA HUB
星环分布式关系型数据库
TRANSWARP DATABASE
TDS
星环大数据开发工具
TRANSWARP DATA STUDIO

图 6　Sophon 架构

Sophon是一款集数据处理、模型加工、线上监控以及数据分析为一体的人工智能基础平台，可以帮助用户快速构建最适应场景的解决方案，完成从特征工程、模型训练到模型上线和维护的机器学习全生命周期。SophonXAI遵循本文提出的通用XAI架构，完全满足XAI工程落地的实际需要，已率先在业界进行实施推广。

以Sophon XAI在某银行信贷风控项目的实施为例。在引入SophonXAI前，该银行经常收到贷款用户投诉，客户满意度在82%左右，投诉主要针对系统给出的信用等级评分解释。调研后发现，由于现有的黑盒客户信用评分模型缺乏透明性，通常采用原始的相关性分析方法给出一个笼统的解释，当用户需要进一步的详细解释时，需要人工审查才能给出用户较为满意的解答，整个过程大约需要1~2小时。整体来说，现有系统解释的质量较低，给出解释的效率较低，需要引入新方法解决此问题。

为了解决上述问题，在原项目中引入SophonXAI组件，最终形成了如**图7**所示的系统架构。首先，在算法库中引入结果可解释模块，并采用KernelShap方法进行客户授信评分分析，从而得到初始的原子解释。其次，引入XAI核心能力层对原子解释进行增强，引入合规性归因模块保证解释结果的合规，并通过存储经验知识的KG库、认知归因模块对解释进行迭代增强，保证解释结果与最终用户的要求相一致。然后，借助XAI表征模块中的丰富解释表征形式，帮助客户更好地接受解释结果，最终有效提升了解释的质量，大大减少了用户的投诉。最后，在管理层中引入Pipeline嵌入组件，将XAI嵌入原AI业务中，形成自动化流程，降低了人工干预的必要性。通过这些尝试，客户满意度提高到94%，

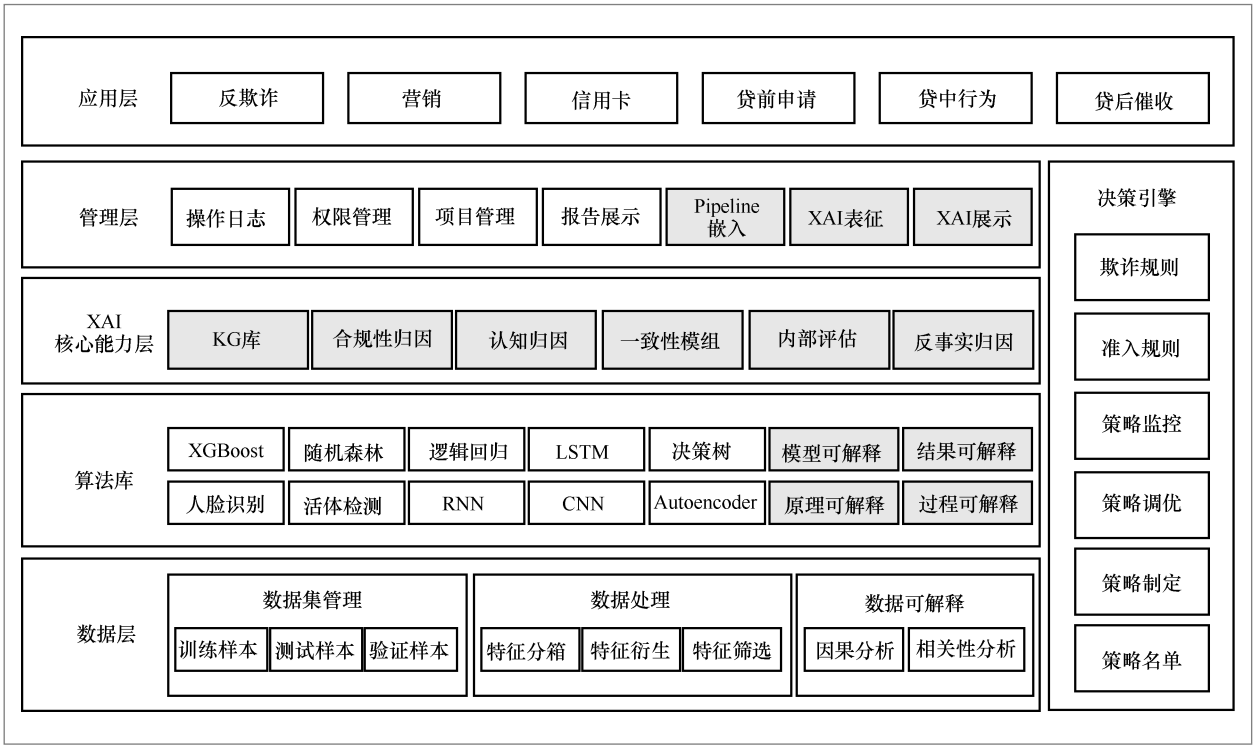


图 7 改造后系统架构

解释生成的时间也缩短到2 min之内。SophonXAI的引入解决了原有项目中解释质量差、生成效率低的问题，该案例可以为XAI在行业应用提供一定的参考。

3 结束语

人工智能已进入社会生活的各个领域，但其不透明性阻碍了AI技术在敏感领域的推广及应用，XAI技术满足了AI产业生态不同角色打开模型黑盒的期望，为建设可信AI生态提供了有力的实现路径。本文率先提出了一种通用的XAI架构，并在产业界进行了实践与落地，这是笔者对如何打造一个端到端的、可实施的XAI架构的思考、探索与实践。XAI的发展任重而道远，由于知识水平的局限，本文尚存在一些考虑不周的地方，比如从复杂系统理论[44]审视XAI、数据分布或上下文背景差异对XAI的影响等，但随着XAI研究的不断深入、新方法新技术的提出，XAI架构也会越来越完善，为打造下一代AI系统的建设贡献力量。

参考文献:

[1] 叶津汶. 大数据、人工智能在金融领域应用研究[J]. 理财, 2022(8): 13–15.
YE J W. Research on the application of big data and artificial intelligence in the financial field[J]. Finance, 2022(8): 13–15.

[2] 谢芬. 人工智能技术及其在智能机器人领域的应用[C]//2018年智慧教育与人工智能发展学术会议论文集. 香港: 香港新世纪文化出版社, 2018: 21–23.
XIE F. Artificial intelligence technology and its application in intelligent robot field [C]//Proceedings of the 2018 Intelligent Education and Artificial Intelligence Development Conference. Hong Kong: Hong Kong New Century Culture Publishing House, 2018: 21–23.

[3] 杨晓光, 马成元, 王一喆, 等. 交通人工智能及其发展综述研究[J]. 人工智能, 2022(4): 18–29.
YANG X G, MA C Y, WANG Y Z, et al. Research on artificial intelligence in transportation and its development[J]. Artificial Intelligence, 2022(4): 18–29.

[4] 高奇琦, 吕俊延. 智能医疗: 人工智能时代对公共卫生的机遇与挑战[J]. 电子政务, 2017(11): 11–19.
GAO Q Q, LYU J Y. Intelligent healthcare: opportunities and challenges to public health in the era of artificial Intelligence[J]. E-government, 2017(11): 11–19.

[5] 刘慈欣. 三体[J]. 意林, 2019(12): 67.
LIU C X. The three-body problem[J]. Yilin, 2019(12): 67.

[6] 张钹. 人工智能进入后深度学习时代[J]. 智能科学与技术学报, 2019, 1(1): 4–6.
ZHANG B. Artificial intelligence entering the post deep learning era[J]. Journal of Intelligent Science and Technology, 2019, 1(1): 4–6.

[7] 夏正勋, 唐剑飞, 罗圣美, 等. 可信AI治理框架探索与实践[J]. 大数据, 2022, 8(4): 145–164.
XIA Z X, TANG J F, LUO S M, et al. Exploration and practice of trusted AI governance framework[J]. Big Data Research, 2022, 8(4): 145–164.

[8] RIBEIRO M T, SINGH S, GUESTRIN C. "Why should I trust you?": explaining the predictions of any classifier[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016: 1135–1144.

[9] ZHOU B L, KHOSLA A, LAPEDRIZA A, et al. Learning deep features

for discriminative localization[C]// Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 2921–2929.

[10] AKULA A, WANG S, ZHU S C. CoCoX: generating conceptual and counterfactual explanations via fault-lines[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(3): 2594–2601.

[11] HSIEH C, MOREIRA C, CHUN O Y. DiCE4EL: interpreting process predictions using a milestone-aware counterfactual approach[C]//Proceedings of 2021 3rd International Conference on Process Mining (ICPM). Piscataway: IEEE Press, 2021: 88–95.

[12] RUBIN D B. Causal inference using potential outcomes[J]. Journal of the American Statistical Association, 2005, 100(469): 322–331.

[13] JAFTA G,DE WAAL A, DERKS I, et al. Evaluation of XAI as an enabler for fairness, accountability and transparency[C]//Proceedings of the Second Southern African Conference for Artificial Intelligence Research. Cham: Springer, 2022.

[14] PRADHAN R, ZHU J, GLAVIC B, et al. Interpretable data-based explanations for fairness debugging[EB]. arXiv preprint, 2021, arXiv: 2112.09745.

[15] DU M N, LIU N H, HU X. Techniques for interpretable machine learning[J]. Communications of the ACM, 2020, 63(1): 68–77.

[16] 曾春艳, 严康, 王志锋, 等. 深度学习模型可解释性研究综述[J]. 计算机工程与应用, 2021, 57(8): 1–9.
ZENG C Y, YAN K, WANG Z F, et al. Survey of interpretability research on deep learning models[J]. Computer Engineering and Applications, 2021, 57(8): 1–9.

[17] PÁEZ A. The pragmatic turn in explainable artificial intelligence (XAI)[J]. Minds and Machines, 2019, 29(3): 441–459.

[18] MILLER T. Explanation in artificial intelligence: insights from the social sciences[J]. Artificial Intelligence, 2019, 267: 1–38.

[19] KULESZA T, STUMPF S, WONG W K, et al. Why-oriented end-user debugging of naive Bayes text classification[J]. ACM Transactions on Interactive Intelligent Systems. 2011, 1(1): 1–31.

[20] HSIAO J H W, NGAI H H T, QIU L, et al. Roadmap of designing cognitive metrics for explainable artificial intelligence (XAI)[EB]. arXiv preprint, 2021, arXiv: 2108.01737.

[21] SZCZEPAŃSKI M, CHORAŚ M, PAWLICKI M, et al. The methods and approaches of explainable artificial intelligence[C]//Proceedings of the International Conference on Computational Science. Cham: Springer, 2021: 3–17.

[22] SPINNER T, SCHLEGEL U, SCHÄFER H, et al. ExplAIner: a visual analytics framework for interactive and explainable machine learning[J]. IEEE Transactions on Visualization and Computer Graphics, 2020, 26(1): 1064–1074.

[23] KWON B C, CHOI M J, KIM J T, et al. RetainVis: visual analytics with interpretable and interactive recurrent neural networks on electronic medical records[J]. IEEE Transactions on Visualization and Computer Graphics, 2019, 25(1): 299–309.

[24] OHANA J J, OHANA S, BENHAMOU E, et al. Explainable AI (XAI) models applied to the multi-agent environment of financial markets[C]//Proceedings of the Explainable and Transparent AI and Multi-Agent Systems. Cham: Springer, 2021: 189–207.

[25] PARK S, YANG J S. Interpretable deep learning LSTM model for intelligent economic decision-making[J]. Knowledge-Based Systems, 2022, 248: 108907.

[26] 周健民, 沈仁芳. 土壤学大辞典[M]. 北京: 科学出版社, 2013.
ZHOU J M, SHEN R F. Dictionary of soil science[M]. Beijing: Science Press, 2013.

[27] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization[J]. International Journal of Computer Vision, 2020, 128(2): 336-359.

[28] BACH S, BINDER A, MONTAVON G, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation[J]. PLoS One, 2015, 10(7): e0130140.

[29] SHRIKUMAR A, GREENSIDE P, KUNDAJE A. Learning important features through propagating activation differences[C]//Proceedings of the 34th International Conference on Machine Learning. New York: ACM, 2017: 3145-3153.

[30] SPRINGENBERG J T, DOSOVITSKIY A, BROX T, et al. Striving for simplicity: the all convolutional net[EB]. arXiv preprint, 2014, arXiv:1412.6806.

[31] LUNDBERG S, LEE S I. A unified approach to interpreting model predictions[EB]. arXiv preprint, 2017, arXiv: 1705.07874.

[32] 孔祥维, 唐鑫泽, 王子明. 人工智能决策可解释性的研究综述[J]. 系统工程理论与实践, 2021, 41(2): 524-536.
KONG X W, TANG X Z, WANG Z M. A survey of explainable artificial intelligence decision[J]. Systems Engineering-Theory & Practice, 2021, 41(2): 524-536.

[33] MILLER G. WordNet: a lexical database for English[J]. Commun ACM, 1995, 38: 39-41.

[34] SCHOONDERWOERD T A J, JORRITSMA W, NEERINCX M A, et al. Human-centered XAI: developing design patterns for explanations of clinical decision support systems[J]. International Journal of Human-Computer Studies, 2021, 154: 102684.

[35] LIAO Q V, VARSHNEY K R. Human-centered explainable AI (XAI): from algorithms to user experiences[EB]. arXiv preprint, 2022, arXiv: 2110.10790.

[36] BREIMAN L. Statistical modeling: the two cultures (with comments and a rejoinder by the author)[J]. Statistical Science, 2001, 16(3): 199-231.

[37] OUYANG L, WU J, XU J, et al. Training language models to follow instructions with human feedback[EB]. arXiv preprint, 2022, arXiv: 2203.02155.

[38] BHAT S, JIANG J Q, POOLADZANDI O B, et al. De-biasing generative models using counterfactual methods[EB]. arXiv preprint, 2022, arXiv:2207.01575.

[39] JOHANSSON F, SHALIT U, SONTAG D. Learning representations for counterfactual inference[C]//Proceedings of the International Conference on Machine Learning.[S.l.:s.n.], 2016: 3020-3029.

[40] PEARL J, MACKENZIE D. The book of why: the new science of cause and effect[M]. New York: AAAS, 2018: 855-855.

[41] 王萍. A/B测试方法的教育应用研究[J]. 电化教育研究, 2015, 36(8): 58-66.
WANG P. Research on the educational application of A/B testing method[J]. e-Education research, 2015, 36(8): 58-66.

[42] YANG B, SAILER A, JAIN S, et al. Service discovery based blue-green deployment technique in cloud native environments[C]//Proceedings of 2018 IEEE International Conference on Services Computing (SCC). Piscataway: IEEE Press, 2018: 185-192.

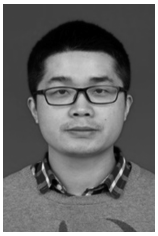
[43] KIM E, LEE J, SHIN H, et al. Champion-challenger analysis for credit card

fraud detection: hybrid ensemble and deep learning[J]. Expert Systems With Applications, 2019, 128: 214-224.

[44] 刘曾荣, 李挺. 复杂系统理论剖析[J]. 自然杂志, 2004, 26(3): 149-151.
LIU Z R, LI T. Analysis of complex system[J]. Ziran Zazhi, 2004, 26(3): 149-151.

作者简介

夏正勋(1979-),男,星环信息科技(上海)股份有限公司高级研究员,主要研究方向为人工智能、大数据、数据库、流媒体处理技术。

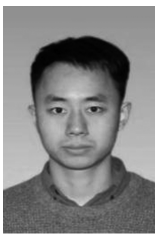


唐剑飞(1986-),男,星环信息科技(上海)股份有限公司大数据技术标准研究员,主要研究方向为大数据、数据库、图计算。

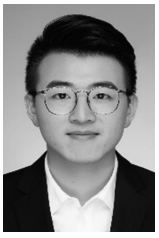
杨一帆(1986-),男,博士,星环信息科技(上海)有限公司产品总监、首席科学家,主要研究方向为统计(统计计算、生存分析、时间序列和生物信息)、图计算、强化学习。

罗圣美(1971-),男,博士,中孚信息股份有限公司副总裁,主要研究方向为大数据、数据安全、人工智能。

张燕(1985-),女,星环信息科技(上海)股份有限公司大数据技术研究员,主要研究方向为大数据、人工智能等。



谭锋镭(1990-),男,就职于星环信息科技(上海)股份有限公司,主要研究方向为AI平台、分布式系统研发。



谭圣儒(1995-),男,星环信息科技(上海)股份有限公司人工智能产品经理,主要研究方向为机器学习、模型管理、模型运维监控、模型可解释性等。

收稿日期: 2023-06-08