可信AI治理框架探索与实践

夏正勋，唐剑飞，罗圣美，张燕
星环信息科技（上海）股份有限公司，上海 200233

摘要
人工智能进一步提升了信息系统的自动化程度，但在其规模应用过程中出现了一些新问题，如数据安全、隐私保护、公平伦理等。为了解决这些问题，推动AI由可用系统向可信系统转变，提出了可信AI治理框架——T-DACM，从数据、算法、计算、管理4个层级入手提升AI的可信性，设计了不同组件针对性地解决数据安全、模型安全、隐私保护、模型黑盒、公平无偏、追溯定责等具体问题。T-DACM实践案例为业界提供了一个可信AI治理示范，为后续基于可信AI治理框架的产品研发提供了一定的参考。

关键词
可信AI；治理框架；AI公平伦理；AI可解释；AI监管

中图分类号：TP311.13 文献标志码：A doi: 10.11959/j.issn.2096-0271.2022036

Exploration and practice of trusted AI governance framework

XIA Zhengxun, TANG Jianfei, LUO Shengmei, ZHANG Yan
Transwarp Information Technology (Shanghai) Co., Ltd., Shanghai 200233, China

Abstract
Artificial intelligence (AI) has further improved the automation of information systems, however, some issues have been exposed during its large-scale application, such as data security, privacy protection, and fair ethics. To solve these issues and promote the transition of AI from available systems to trusted systems, the T-DACM trusted AI governance framework was proposed to improve the credibility of AI from the four levels of data, algorithm, calculation, and management. Different components were designed to solve specific issues such as data security, model security, privacy protection, model black box, fairness, accountability, and traceability. T-DACM practice case provides a demonstration of the trusted AI governance framework for the industry and provides a certain reference for subsequent product research and development based on the trusted AI governance framework.

Key words
trusted AI, governance framework, AI ethics and fairness, AI interpretability, AI supervision

0 引言

人工智能（AI）的蓬勃发展是人类社会的重大历史事件，其开创了一个新的时代[1]。随着AI技术的快速发展，AI应用的规模愈加庞大，IDC预测AI应用市场规模有望在2024年突破5 000亿美元大关[2]。但随着AI应用的深入，一些深层次问题逐渐暴露出来，比如使用脸部照片、人脸视频、3D头套等方式恶意欺骗人脸识别系统[3]；在交通标志上添加一个不显眼的对抗扰动，误导并改变自动驾驶汽车的行车路线[4]；通过毒化数据的方式，在模型中添加后门，从而控制模型的行为[5]；“大数据杀熟”[6]和“数字剥削”[7]等恶意使用AI的行为。上述AI安全伦理问题引发了人们对AI系统可信性的担忧，阻碍了AI的进一步发展，已经成为亟待解决的问题。

如何打造可信的AI系统已经成为政府、企业的关注重点。本文分析了AI系统在当下遇到的可信问题与挑战，从技术和管理两个维度出发，提出了一种通用的可信AI治理框架——T-DACM（trusted data & algorithm & computation & management）。T-DACM具体包括可信数据、可信算法、可信计算、可信管理4个层次，覆盖了数据安全、模型安全、隐私保护、风险控制、过程管理、可解释性、公平伦理、追溯追责等AI热点问题的解决方法，为企业及监管机构提供了一种可行的可信AI解决方案。

1 国内外研究进展

可信是AI技术的自身要求，也是AI产业发展的要求。统计模型存在类似“射手假说”[8]的问题，AI不可避免地有被误用、滥用、恶意使用的风险。为了降低这些风险，国外对可信AI的研究主要从安全及数据隐私方面入手，兼顾AI使用过程中的伦理道德，制定了相关的法律法规。

2014年Goodfellow I J等人[9]提出用模型对抗训练的方法来防御对抗样本攻击，增强模型的安全性；2016年Ribeiro M T等人[10]提出一种生成解释方法来解释模型的决策过程，提升模型决策的透明性；2018年Gidaris S等人[11]提出一种元学习方法来提高模型在未知领域数据上的泛化能力，降低模型的误用风险；2019年Nguyen H H等人[12]提出使用多任务学习的方法来提高模型的领域泛化能力，降低模型误判的风险；2019年Iosifidis V等人[13]提出一种基于重采样的方法来消除数据中的偏见，提升模型的公平性；2021年Sauer A等人[14]提出一种反事实图片的生成方法来协助分类模型找到更稳定的因果特征，提升模型的泛化能力。

与此同时，在可信AI治理指导原则方面，2018年欧盟提出了《人工智能道德准则》，提出可信AI的5个指导原则：福祉原则、不作恶原则、自治原则、公正原则、可解释性原则[15]；2020年Shin J等人[16]提出构建一个可信AI系统需要重点考虑AI应用的再现性、可解释性和公平性；2020年英国信息专员办公室提出了《解释AI做出的决定》，提出算法透明度可以从原理解释、责任解释、数据解释、公平性解释、安全与可靠性解释及影响解释6个方面实现[17]。2021年，Winter P M等人[18]提出，可信的AI应用需要重点关注机器学习（machine learning，ML）关键理论理解、质量评估手段、领域泛化、置信度量手段、道德伦理、用户可接受度、抗攻击能力等方面。

国内对可信AI的技术研究基本与国外同步，在可信AI治理配套的法规方面，从

网络安全、数据安全、个人信息多维度入手，更加全面，职责明确，易于实施。在可信AI技术研究方面，2017年Meng D Y等人[19]提出基于数据增强和样本检测的MagNet方法来防御对抗样本攻击；2018年Kuang K等人[20]提出一种基于样本加权的因果预测算法DGBR来提高模型对未知环境的适应能力；2019年Zhang Q S等人[21]提出一种基于决策树的替代模型方法来解释卷积神经网络（convolutional neural network，CNN）模型的决策逻辑；2019年FENG R等人[22]提出一种基于对抗框架的数据表示方法来消除数据集中对特定群体的偏见；2021年Zhao Y Y等人[23]提出一种多源元学习框架来增强模型的领域泛化能力。

在可信AI治理指导原则方面，我国于2016年颁布了《中华人民共和国网络安全法》，要求网络运营者保障网络免受干扰、破坏或者未经授权的访问，防止网络数据泄露或被窃取、篡改。2019年我国提出了《新一代人工智能治理原则——发展负责任的人工智能》，强调了和谐友好、公平公正、包容共享、尊重隐私、安全可控等8条原则，从而逐步实现可审核、可追溯、可信赖的AI[24]；2020年中国信息通信研究院提出了《人工智能安全框架（2020年）》，其从AI安全技术、AI安全管理两个方面设计AI安全框架，对AI安全能力进行分级，提出了合法合规、功能可靠可控、数据安全可信、决策公平公正、行为可解释、事件可追溯的AI安全目标[25]；2021年6月通过的《中华人民共和国数据安全法》要求机构开展数据处理活动时，应当遵守法律、法规，尊重社会公德和伦理，遵守商业道德和职业道德，诚实守信，履行安全保护义务、承担社会责任，不得危害国家安全、公共利益，不得损害个人、组织的合法权益；2021年8月通过的《中华人民共和国个人信息保护法》要求机构在利用个人信息进行自动化决策时，要建立算法影响评估制度进行事前风险评估，建立算法审计制度审计活动遵守法律、法规的情况，保证决策的透明度和结果公平合理。

综上所述，围绕可信AI的问题，学术界、产业界及政府均进行了不同方向的探索，取得了一些成果，但可信AI的落地尚存在一些问题：首先，大多可信方面的研究还停留在理论分析阶段，在产业界尚未进行大规模应用；其次，可信AI涉及的技术领域较广，目前尚缺少一个通用模式将诸多可信AI技术有机结合起来；最后，针对AI系统的行为，缺乏完善的监管系统。因此，需要一个统一的可信AI治理框架，对可信AI涉及的管理及技术因素进行统筹考虑，以满足学术界、产业界及政府三方对可信AI的迫切需求，完成AI系统由可用系统向可信系统的转变。

2 可信AI治理框架简介

本文在现有理论及技术研究的基础之上，结合应用场景，对可信AI治理进行了探索与实践，提出了T-DACM ，如**图1**所示。

T-DACM由下而上分为4层，分别是可信数据（trusted data）层、可信算法（trusted algorithm）层、可信计算（trusted computation）层、可信管理（trusted management）层，具体如下。

- 可信数据层为可信AI提供了数据基础。其具体包含异常数据检测、偏倚消除、偏见消除、数据增强等组件，可以检测异常样本，保证模型正常工作，对异常数据、异质性数据进行处理，引入公平性算法消除数据中的歧视与偏见，通过样本变换或者样本生成来扩增数据集，增强模型的鲁棒性。

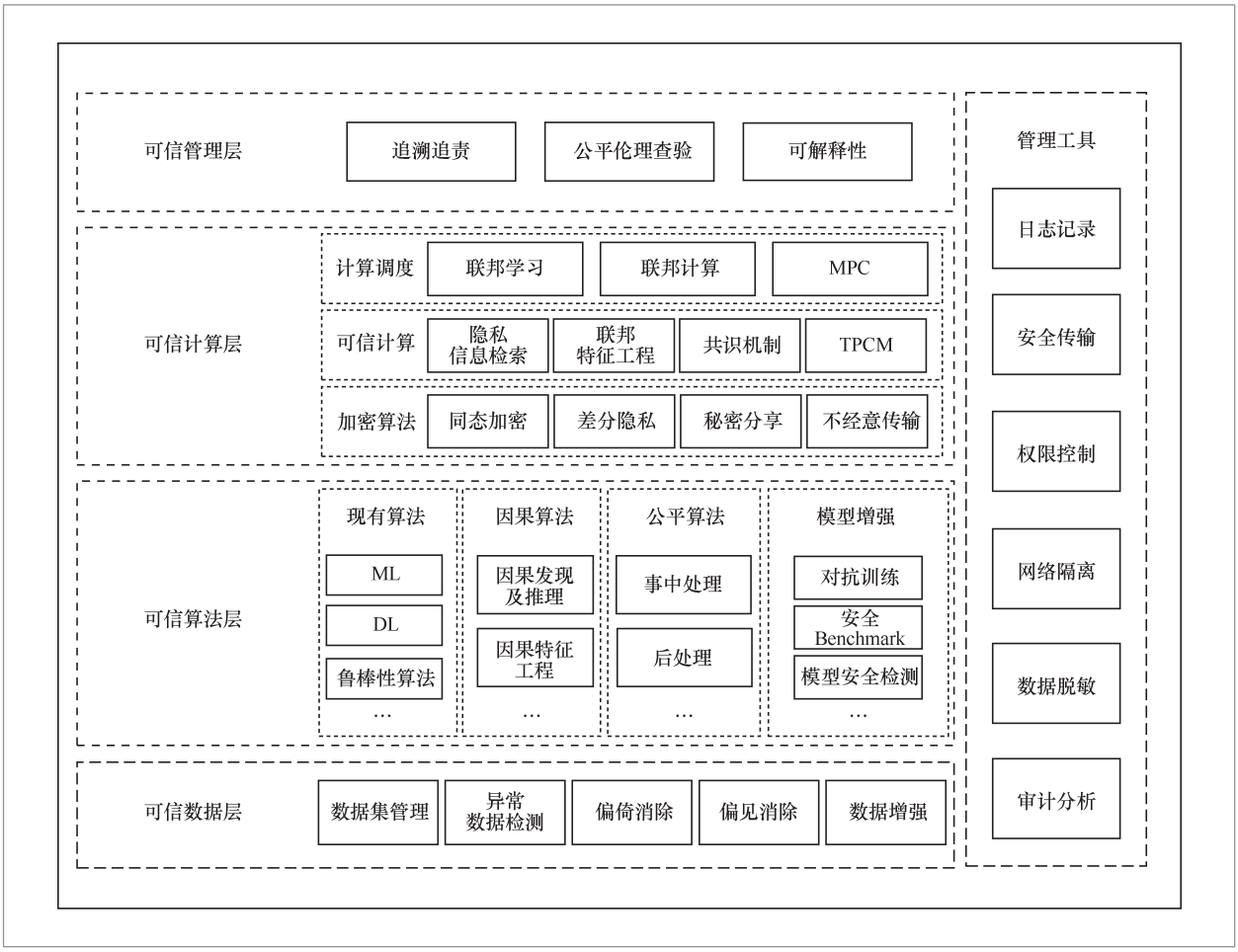

图1 T-DACM

● 可信算法层为可信AI提供安全与强鲁棒性的核心能力。除了现有算法，其还包含因果算法、公平算法、模型增强等算法组件，可以从数据中发现稳定特征，提高模型对不同环境的适应能力，去除模型的决策逻辑对弱势群体的歧视，提升模型对常见攻击方式的防御能力。

● 可信计算层为可信AI提供可信计算的能力。其具体包括加密算法、可信计算、计算调度等模组，可以使用加密算法及可信平台控制模块（trusted platform control module，TPCM）保障单方计算场景下的数据安全，使用联邦学习、联邦计算、安全多方计算（secure multi-party computation，MPC）等组件为多方参与的模型训练、模型推理、计算等保障数据安全。

● 可信管理层为可信AI提供可追溯、可监管、可理解的管理能力。其具体包括追溯追责、公平伦理查验、可解释性3个组件，可以通过事故发生之后的回溯对责任进行认定，通过分析系统的实时行为来监控违反法规伦理的行为，对模型的决策逻辑进行解释，为模型的优化提供参考。

T-DACM基本覆盖了AI从模型学习、模型应用到系统管理的全流程。以个性化保险定价为例，可信数据层使用Relabelling算法[26]消除数据集中对弱势群体的偏见；可信算法层使用公平优化（fair

optimization）方法[27]增强定价算法的公平性；可信计算层使用联邦学习计算框架[28]在保护隐私的前提下联合第三方银行数据共同建模，定价结果更加准确可信；在可信管理层，既能回溯某一定价的决策过程并做出解释，也能实时检测在运行过程中整个定价系统是否存在违反法律、伦理规则的行为。T-DACM内层与层之间的无缝协作可以保障AI系统全流程的可信性。

3 可信AI治理框架实现

3.1 可信数据层

数据是人工智能的基础，模型学习依赖于大量训练样本，模型推理也依赖于输入数据的质量，但当数据异常时，模型的输出结果往往脱离预期，为人工智能系统带来潜在的风险。2016年，微软聊天机器人Tay受到一些偏激言论的语料影响，行为异常，微软被迫临时关闭了Tay的在线学习能力；2019年，腾讯科恩实验室发现存在扰动信息的路面误导了自动驾驶系统，致使车辆驶入反向车道。上述例子均是由异常数据引起的模型行为异常。除此以外，数据还与模型的鲁棒性、公平性密切相关，具体而言，当推理数据与训练数据分布不同时，模型的精度往往会下降，严重时甚至无法工作。当训练数据中弱势群体的数据远少于强势群体的数据时，模型的决策往往会更倾向强势群体，从而造成对弱势群体的偏见或不公平对待。因此，构建一个可信的数据层是可信AI系统的首要任务。

可信数据层是可信AI治理的首要环节。可信数据层解决由数据导致的不可靠、不可信问题，方法主要有两大类：一类通过检测方法找到异常数据并删除，另一类通过数据增强方法对数据进行可信性增强。在训练阶段，其输入为训练数据集，输出为剔除了恶意数据或增强了公平性的可信数据集；在推理阶段，其根据具体任务，对输入数据进行检测，拒绝对抗样本、伪造数据等恶意输入，或通过技术手段过滤掉对抗样本、毒化样本中的恶意信息，保证服务的延续性。

可信数据层具体包括数据检测模组与数据增强模组，如图2所示。数据检测模组由样本集检测模块与单样本检测模块组成：样本集检测模块可以对样本集进行整体检测，单样本检测模块用来对样本进行独立检测。数据增强模组由样本增强、偏见消除与偏倚消除3个模块组成：数据增强模块可以降低对抗样本等恶意数据的干扰，也可以扩展数据集；偏见消除模块可以平衡数据集中弱势群体与优势群体的比例，从而消除对弱势群体的偏见；偏倚消除模块可以消除样本集中的异质性影响。

图2中数据检测模组具体如下。

- 样本集检测模块：由分布检测、偏见检测、毒化样本检测与偏倚检测等组件组成。分布检测可以将样本集中不同分布的数据检测出来，通过控制模型输入数据的边界来降低鲁棒性风险，该类方法包括孤立森林、DBScan等；偏见检测可以检测数据集中强势群体与弱势群体在特定公平规则下的比例是否失衡，用来判断使用的样本集中是否存在偏见，该类方法包括Equal Opportunity[29]、Equalized Odds[30]等；毒化样本检测可以检测样本集中是否包含被毒化的数据，可以将毒化数据从数据集中剔除，典型的毒化数据检测方法有AUROR[31]等；偏倚检测可以用来检查是否采集过程的瑕疵导致了样本的分布不均衡，进而导致了数据的异质性，偏倚检测的经典方法为卡方检验[32]。

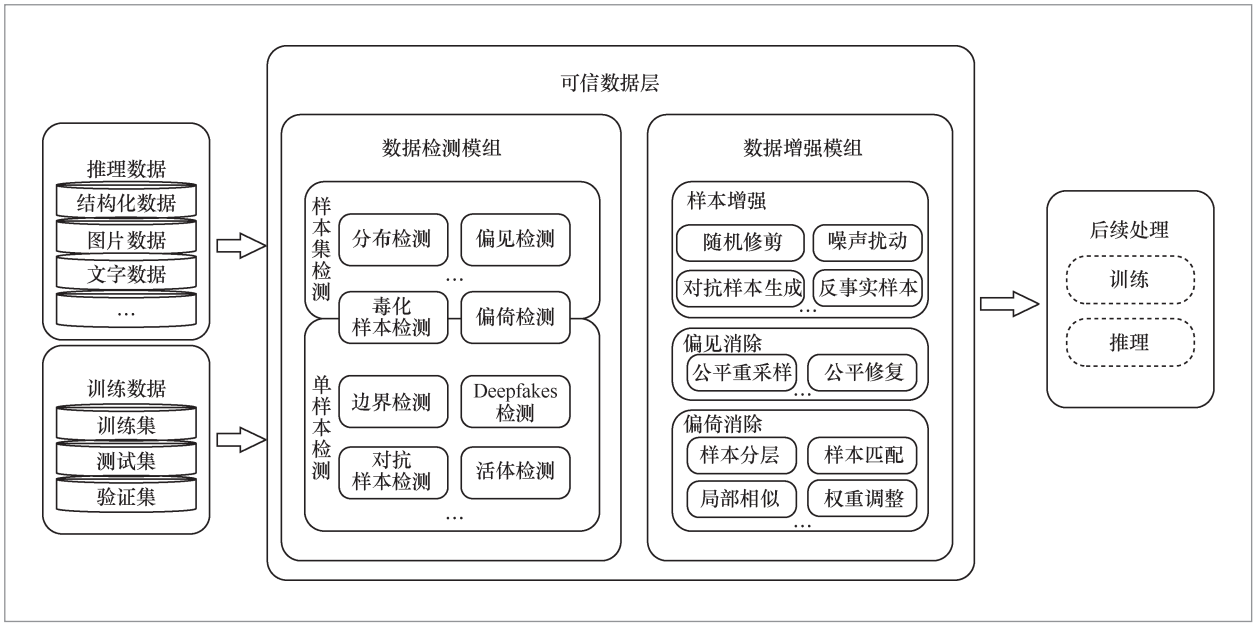

图 2 可信数据层架构

● 单样本检测模块：由毒化样本检测、偏倚检测、边界检测、Deepfakes检测、对抗样本检测、活体检测等组件组成。毒化样本检测和偏倚检测组件的功能与样本集检测模块类似；边界检测指的是检测数据的常规指标，判断其是否超出阈值，可以判断样本中是否存在明显异常；Deepfakes检测、对抗样本检测、活体检测指的是对伪造视频、对抗样本、假脸数据等恶意数据进行针对性检测，这些恶意数据检测模块可以在一定程度上解决AI应用被恶意误导的问题，目前，Deepfakes、对抗样本、假脸数据都有许多成熟的检测方法，如使用MesoNet[33]来检测Deepfakes视频，使用MagNet[19]来检测对抗样本，使用De-Spoofing[34]来进行活体检测。

图2中数据增强模组具体如下。

● 样本增强模块：由随机修剪、噪声扰动、对抗样本生成、反事实样本等组件组成。随机修剪、噪声扰动等是机器学习中常用的数据增强方法，这些组件通过对原数据进行随机裁剪、添加随机噪声以及进行翻转、尺度变换来生成新的样本数据，生成的样本可以对原样本集进行扩充，从而丰富样本集，也可以对原始图片进行替换，从而在一定程度上防御对抗样本、毒化样本等的恶意攻击。对抗样本生成组件可以用来生成对抗样本，生成的对抗样本可以用于对抗训练，提升模型对对抗攻击的防御能力。反事实样本组件可以基于结构因果模型，采用类似于反事实生成网络（counterfactual generative network）[14]的方法，使用事实中不存在的特征组合，生成反事实样本，从而提高模型的泛化能力。

● 偏见消除模块：由公平重采样、公平修复等组件组成。对于样本集中的不公平，公平重采样组件通过对样本集进行重新采样，重新生成弱势群体与强势群体均衡的样本集，如按照指定的公平原则进行重采样的方法[35-36]。公平修复组件通过修改标签或直接改变训练数据中一个或多个变量的分布来修复数据集中的不公平，如训练专

用模型来修正标签的方法[26, 37]。

● 偏倚消除模块：由样本分层、样本匹配、局部相似、权重调整等组件组成。样本分层方法通过对数据集进行分层，保证每一层中的数据是同质的、均衡的，从而达到消除偏倚的目的，如基于倾向得分进行分层的方法[38-39]；样本匹配方法用匹配的方法将数据集中同质的数据筛选出来，保证筛选出来的数据是均衡的，这一类方法以PSM[38]为代表；局部相似组件基于流行假设/局部近似理论，找到一个好的样本集划分方式，使得每个样本组在特征空间上足够近且样本量上足够，并认为细小到分支（局部）的数据是近似的、同质的、均衡的，这一类方法以SITE[40-41]为代表；权重调整方法给样本分配不同的权重，使得实验组与对照组的分布类似，以实现均衡，其代表方法有逆概率加权（inverse propensity weighting，IPW）[38]、数据驱动变量分解（data-driven variable decomposition，D²VD）[42]等。

得益于异常数据检测模组与数据增强模组的协作，经可信数据层处理后的数据大概率是可信的，将这些正常、无偏、公平的数据作为后续环节的输入，为整个AI应用的可信打下了坚实基础。

3.2 可信算法层

算法与模型是AI系统的核心，但当前基于统计理论的人工智能算法只完成了输入数据与输出数据的“曲线拟合”[43]，因此模型输出的结果缺乏内在逻辑，难以解释，也缺乏公平性及安全性方面的考量，由此引发了很多问题。比如，一个模型可以很好地识别草地上的狗，但难以识别水中的狗，这是因为训练过程中更倾向于将草地特征与狗的特征直接“拟合”，由此可见，这种Shortcut Learning[44]的学习方式使模型预测过程缺乏严谨逻辑，预测结果只适用于特定场景，特别容易产生鲁棒性问题。此外，由于概率空间边界的模糊性，在理论上，统计模型始终无法防御对抗性攻击[45]，只能持续提升模型抗攻击的能力。在AI大规模应用的过程中，除了鲁棒性问题、安全性问题，还存在更隐蔽的歧视偏见、公平伦理问题，比如美国的一款案件管理和决策支持工具COMPAS倾向于给少数族裔较高的“累犯”得分。针对上述的问题，需要在AI研发过程中，对算法训练、模型决策过程进行约束、监管及核查，使模型不违背人类社会规则及伦理道德。

现有算法为了提升模型鲁棒性，通常采用元学习[46]、多任务学习[47]、网络架构设计[48]等方法，T-DACM在现有算法之外，扩展了因果算法、公平算法及安全增强3个模组，如**图3**所示。其中，因果算法模组用因果关系代替相关关系提升模型的稳定性；公平算法模组通过消除推理过程中的偏见来提升模型的公平性；安全增强模组持续提升模型的安全性，从而为AI构建一个可信的算法层。

● 因果算法模组由因果发现及推理模块、因果启发稳定学习模块组成，消除弱相关特征或错误特征对决策结果的干扰，确立输入数据与输出结果之间的因果逻辑，基于稳定的因果逻辑进行AI决策。

● 公平算法模组由事中处理（in-processing）模块与后处理（post-processing）模块组成，分别在模型学习过程中、模型学习完成后两个阶段对模型的公平性进行检查、纠正，保证模型决策过程及结果与群体的肤色、人种、性别等受保护属性[49]无关，从而保证AI的公平性。

● 安全增强模组由对抗训练、安全基准测试及模型安全检测模块组成。对抗训练组件通过使用迭代更新的对抗样本不断提升模型的对抗攻击防御能力；安全基准

测试模块可以对模型的安全性能进行衡量；模型安全检测模块通过不同的参数对安全提升方法的效果进行衡量，从而持续提升模型的安全性。

图3中因果算法模组具体如下。

● 因果发现及推理模块由基于对照的因果模型、基于约束的因果模型、针对因果函数的因果模型、针对隐变量的因果模型等组件组成。基于对照的因果模型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，其中实验性方法有随机对照试验方法、A/B测试方法[50]，观测性方法有Stratification、Matching、Re-weighting、Tree-based等[51]；基于约束的因果模型组件通过先确定因果关系结构、再确定结构中方向的方法来完成因果学习，这一类方法包括IC算法[52]、FCI算法[53]等；针对因果函数的因果模型组件利用因果数据生成机制引起数据分布的不对称性来分析变量之间的因果关系，这一类方法包括ANM算法[54]、LiNGAM算法[55]等；针对隐变量的因果模型组件可以对包含未观测到因素（隐变量）影响的数据进行因果分析，通常采用的方法有Cornfield不等式[56]、工具变量法[57]、阴性对照法等[58]。

● 因果启发稳定学习模块由因果特征工程、反事实辅助学习等组件组成。因果特征工程组件可以利用因果发现及推理算法从特征中找到强相关的因果特征，剔除无关及弱相关特征，提高模型的稳定性，这一类方法包括IAMB[59]、MMMB[60]、MMPC[61]等；反事实辅助学习组件生成反事实图片，并将其加入训练，以提高模型的泛化能力，这一类方法有反事实生成网络[14]等。

图3中公平算法模组具体如下。

● 事中处理模块提供在模型训练过

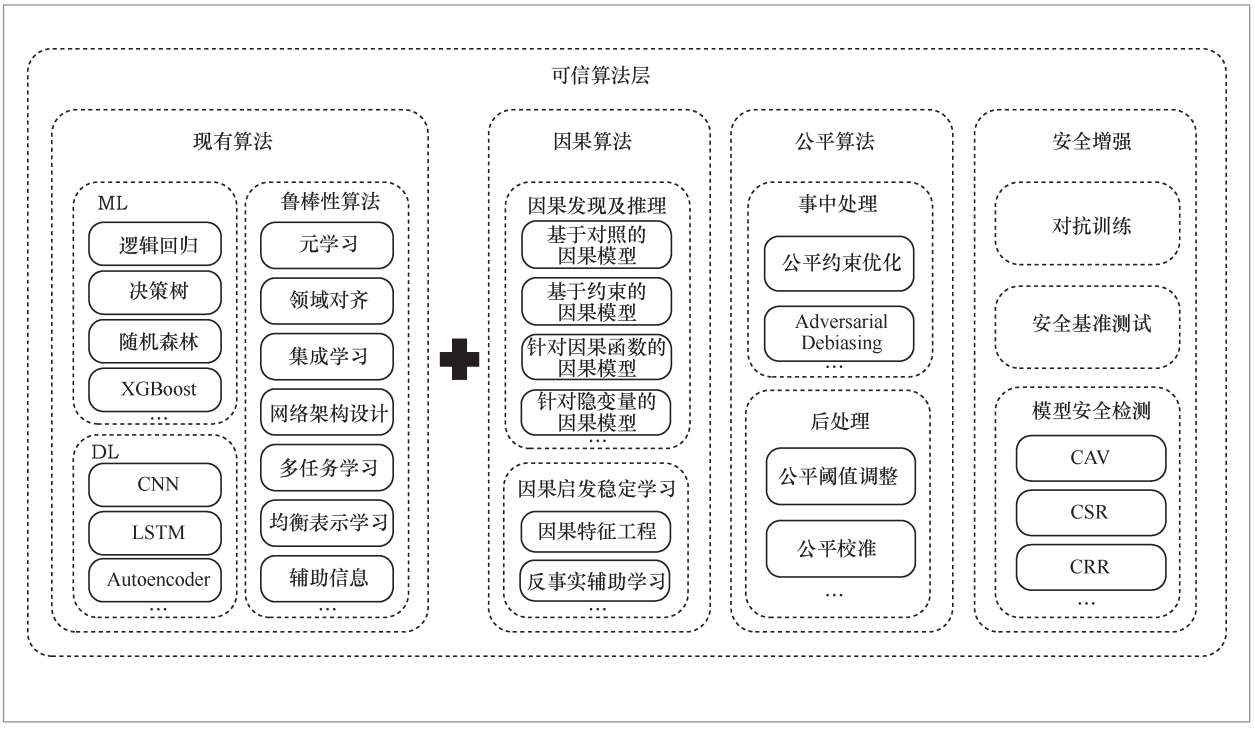

图 3 可信算法层架构

程中解决公平问题的工具，由公平约束优化、Adversarial Debiasing等组件组成。其中，公平约束优化组件可以在模型的目标函数中增加以公平为目标的正则项来保证模型的公平性，该类方法有Prejudice Remover Regularizer[62]等；Adversarial Debiasing指的是在训练的过程中尝试训练一个对抗组件[63]，对模型的公平性进行衡量，同时根据衡量结果持续优化模型的公平性，该类方法有One-Network[64]等。

● 后处理模块通过对模型的推理结果进行后处理来解决公平问题，由公平阈值调整（fair thresholding）与公平校准（fair calibration）等组件组成。其中，公平阈值调整指的是通过调整模型的决策阈值来保证模型的公平性，这一类方法有Equalized Odds Post-Processing[29]等；公平校准组件可以通过优化模型的得分输出来修改模型的决策结果，保证模型的公平性，这一类方法有Calibrated Equalized Odds Post-Processing[65]等。

图3中安全增强模组具体如下。

● 对抗训练模块通过对抗训练的方法，使用可信数据层生成的对抗样本，对模型持续进行微调（fine-tuning），可以持续不断地提升模型对对抗攻击的防御能力。

● 安全基准测试模块可以设置不同的基准测试对不同场景下的模型进行安全性能评估。通过为每个测试场景设置测试数据、攻击方案、基准配置等，测试模型在各攻击方案下的安全性能，并与基准配置进行比较排序。

● 模型安全检测模块由分类精度差异（classification accuracy variance，CAV）、分类牺牲率（classification sacrifice ratio，CSR）与分类矫正率（classification rectify ratio，CRR）等检测指标组成。CAV指标衡量的是应用防御方法前后模型对同一批对抗样本的准确率的变化情况，该指标越大，表示抗攻击的效果越好；CSR指标表示模型增强前防御成功的样本在模型增强后防御失败的比率，该指标越高，表示抗攻击的效果越差；CRR指标表示模型增强前防御失败的样本在模型增强后防御成功的样本中的比例，该指标越高，表示抗攻击的效果越好。

可信算法层从因果、公平、增强3个角度提供了多样的可信AI能力，为其他层的可信能力提供了支持与补充，是可信AI治理框架可信能力的核心。

3.3 可信计算层

大规模分布式训练可充分利用跨组织的数据资产及计算资源，极大地提升人工智能模型的精度，但此过程中往往存在数据安全及隐私泄露的风险。比如，2019年研究人员发现，分布式训练过程中参与方之间交换的梯度信息能够造成训练样本泄露[66]。跨地域的分布式服务系统同样存在类似的问题，早在2016年研究人员就发现，AI模型仅在几千次使用后就被窃取数据[67]。分布式计算框架最初被设计用来解决单机性能不足的问题，节点间的数据交互主要发生在内部网络，受到网络防火墙的保护，不易被外界攻击，因此数据泄露的风险较低。但在跨域计算的场景下，节点间的数据交换暴露在公共网络中，因此计算过程中的数据安全问题不能被忽视。本文设计的T-DCAM的可信计算层通过引入联邦学习、联邦计算、安全多方计算等隐私计算方法，减少交互过程中敏感数据的传输，并对数据通道进行加密，保证计算全流程和跨组织计算的安全性。其具体如**图4**所示。

可信计算层包括加密算法、可信计算、

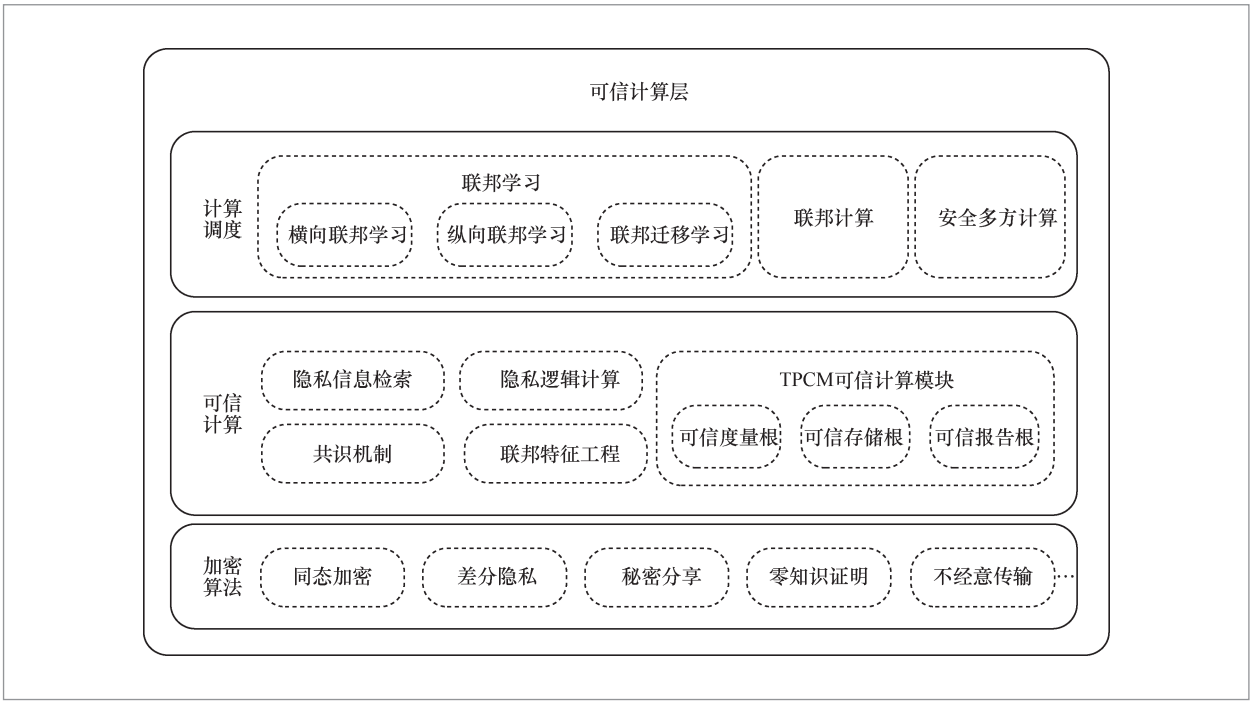

图 4 可信计算层结构

计算调度3个模组。加密算法模组由同态加密、差分隐私、秘密分享、零知识证明、不经意传输等组件组成，保证数据传输的安全；可信计算模组由隐私信息搜索、隐私逻辑计算、共识机制、联邦特征工程、TPCM可信计算模块等组件组成，提供跨节点的可信计算能力；计算调度模组由联邦学习、联邦计算及安全多方计算模块组成，提供了一个多方参与的可信计算调度框架，从架构流程、计算规范上保证跨域计算的可信性。

图4中加密算法模组为可信计算层提供了各种加密算法，具体如下。

● 同态加密组件为可信计算层提供同态加密[68]算法。该算法可以对数据进行加密，且加密后的数据可以直接计算，并可以通过解密获得正确的计算结果。

● 差分隐私组件为可信计算层提供差分隐私[69]算法。该算法可以对数据添加干扰噪声，以此来保护数据中的用户隐私信息，且不会显著改变数据的计算结果。

● 秘密分享组件为可信计算层提供秘密分享[70]的能力。该组件可以将需要加密的数据信息以适当的方式拆分，拆分后的每一份信息由不同的参与者管理，单个参与者无法恢复秘密信息，只有若干个参与者协作才能恢复信息。

● 零知识证明组件为可信计算层提供零知识证明[71]算法。通过该算法，证明者能够在不向验证者提供任何有用信息的情况下，使验证者相信某个论断是正确的。

● 不经意传输组件为可信计算层提供不经意传输[72]能力。该组件可以从数据集合中发送部分数据给接收者，但事后不清楚具体发送了哪些数据，以保护数据接收者的隐私。

图4中可信计算模组具体如下。

● 隐私信息检索组件可以在查询的过程中保护查询方的隐私信息，数据提供方无法获知查询方具体查询了哪个对象。隐私信息检索组件基于n选1的不经意传输[72]，

查询方将查询需求加密后发给数据提供方，数据提供方基于加密信息返回*n*条查询结果给查询方，查询了方从*n*条信息中计算出自己需要的查询结果，在此过程中，隐私信息检索组件保证查询方得到了匹配的查询结果却不留查询痕迹。

● 隐私逻辑计算组件基于混淆电路[73]或不经意传输[74]技术来实现多方安全比较，可以在不提供自己的数据给其他参与方的情况下，判断各方数据的大小关系，判断各方数据是否相等。

● 共识机制组件可以在分布式系统中各节点状态不一致时，提供特定机制奖励提供资源维护区块链的使用者，惩罚恶意的危害者，使得分布式系统中的各节点达成共识，保证状态的一致。共识机制组件中的工作量证明（proof of work，PoW）机制基于节点的工作量对节点进行奖惩，权益证明（proof of stake，PoS）机制基于节点持有区块的比例及持有时间来进行奖惩。

● 联邦特征工程组件可以对即将进行联邦学习的各参与方数据进行联合的特征工程。由于联邦学习对数据隐私保护的要求，特征工程不能直接使用各参与方的数据，需要借助特定的算法来完成。联邦特征工程组件包含联邦采样、联邦特征分箱、联邦特征选择、联邦特征归一化、联邦独热编码等算法。

● TPCM可信计算模块可以为软件的执行提供一个可信的执行环境，该模块包含可信度量根（root of trust for measurement，RTM）、可信存储根（root of trust for storage，RTS）及可信报告根（root of trust for reporting，RTR）等组件。当实体请求访问可信执行环境时，根据请求的资源类型，使用RTM、RTS或RTR对实体进行度量，完成实体的身份认证，即判断得到的度量值是否在可信环境中有记录，若有，则认为该实体可信，否则，认为该实体不可信。

图4中计算调度模组具体如下。

● 联邦学习模块在保护数据参与方数据隐私的前提下，利用多方数据完成模型训练、推理等功能。该模块由横向联邦学习、纵向联邦学习、联邦迁移学习3个组件组成：横向联邦学习模块针对参与方拥有相同特征但样本分布不同的情况，让各参与方利用私有数据在本地进行训练，再通过模型聚合方式不断更新模型；纵向联邦学习模块适用于参与方数据特征不同，只有一方有标签数据的情况，此时模型需要多方的数据才能训练，推理时也需要多方数据才能完成，进行纵向联邦学习首先需要对各参与方的加密实体进行对齐，然后使用特定的纵向联邦学习算法进行加密模型训练；联邦迁移学习模块适用于参与者间特征和样本重叠都很少的场景，不同的数据方首先训练各自的模型，然后在保证隐私的前提下，多方对这些模型进行联合训练，最后得出最优的模型，并将其返回给各个数据所有方。

● 联邦计算模块可以在不直接进行数据交换的前提下，集成来自不同数据库、数据平台产品的异构数据源，按协议统筹调度，各参与方先计算各自的中间结果，再汇总所有数据源的计算结果，计算出全体数据的计算结果。

● 安全多方计算模块可以在无可信第三方参与的情况下，让多个计算参与方利用各方的秘密数据计算一个预先达成共识的函数，计算结束后任意一方可以得到己方的结果，但无法获得其他信息。

借助加密算法、可信计算、计算调度模组提供的技术支撑，在多方参与的跨域计算过程中，可信计算层在满足《中华人民共和国数据安全法》要求的前提下，实现数据资产的跨企业协同，保障数据安全的合规使用。

3.4 可信管理层

多国政府从顶层设计对AI系统提出了可信要求[15,17,24-25]，目标是建立合法合规、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。可信管理层是T-DACM的最顶层，如**图5**所示，从模型、事件、系统3个层级对AI系统进行管理，实现AI行为可解释、事件可追溯、责任可定位，并符合法律法规的监管要求。

模型级监管可跟踪模型内部的决策过程，对决策结果进行解释，其功能具体由可解释性模组实现；事件级监管可在事后对某一事件全流程进行回溯，准确定位到问题发生的子流程，进行进一步的问题诊断及责任细分，其功能具体由事件追溯模组实现；系统级监管可对AI系统的整体行为进行实时监控或周期性复盘，对违反法律法规与伦理道德的行为进行预警及处理，其功能具体由公平伦理核查模组实现。

可解释性模组由自解释方法、生成解释方法、代理模型可解释方法、可视化的解释方法组件组成，解决黑盒模型不可解释的问题，具体如下。

- 自解释方法指的是线性模型、树模型等本身可解释性较好的模型，可以通过模型自身来解释其决策逻辑。
- 生成解释方法使用分类和语言生成模型生成解释性文本，相关方法有Generating Visual Explanations[74]等。
- 代理模型可解释方法通过训练一个局部近似的自解释性模型来解释原模型的行为，LIME[10]是这一类方法的代表。
- 可视化的解释方法指的是利用热图、特征图等方法对模型决策过程进行可视化的展示，针对模型行为提供直观、可理解的视觉解释。

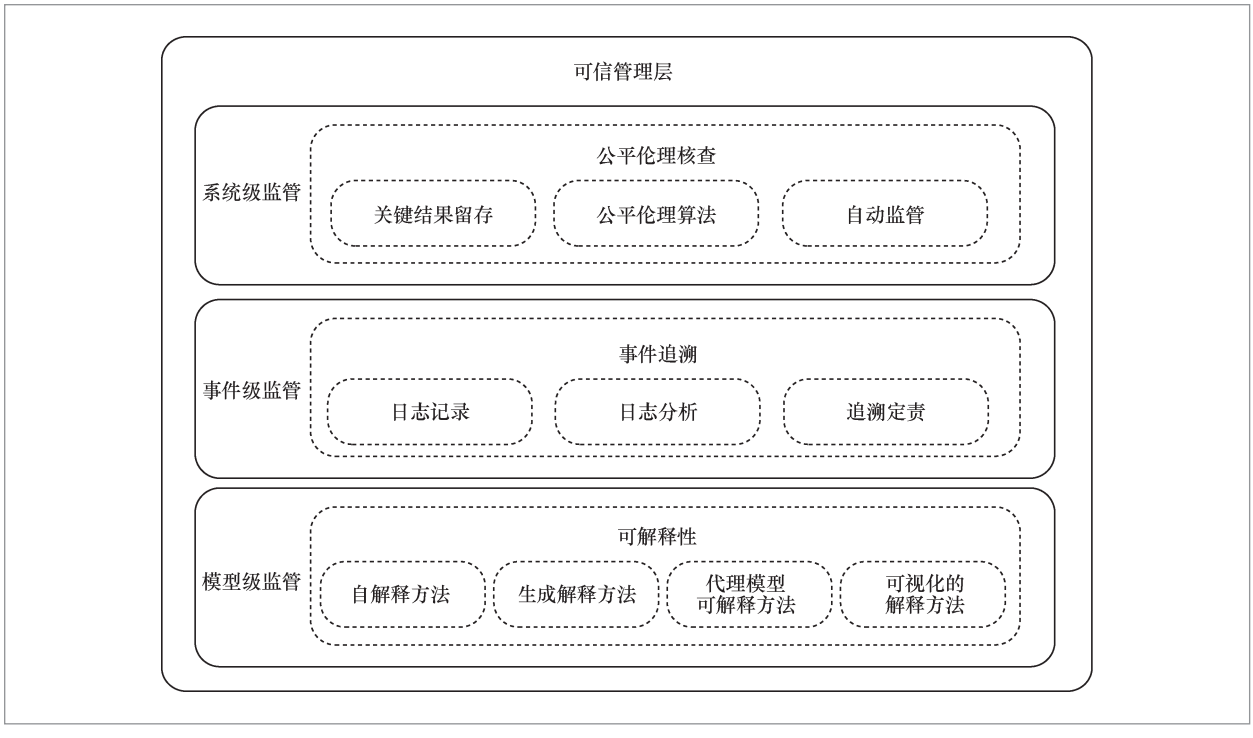

图 5 可信管理层结构

事件追溯模组由日志记录、日志分析、追溯定责组件组成，记录AI系统决策过程，解决事后问题定位及定责难题，具体如下。

● 日志记录组件根据事先约定，在AI应用特定行为被触发、特定时间点或特定时间间隔后，进行日志记录。记录内容包括时间、行为主体、行为内容、执行上下文环境等。

● 日志分析组件在事件完成后被触发，可以根据事件的日志记录生成事件的报告，按事件中子流程发生的先后顺序或子流程之间的依赖关系对子流程进行展示，方便后续的分析。

● 追溯定责组件在事件完成后对事件报告进行检查，判断有无事故发生。若有事故发生，则按照子流程的时间顺序或依赖关系定位到问题发生的源头，按照责任划分规则，进行故障的自动认定。

公平伦理核查模组由关键结果留存、公平伦理算法、自动监管组件组成，对系统行为进行合规性监管，保证AI系统符合法律法规及公平伦理要求，具体如下。

● 关键结果留存组件负责留存系统运行过程中的关键数据。对于涉及公平伦理、法律法规的关键数据，如员工的调度数据等，在系统做出决策的同时，将决策结果与决策的上下文数据存入数据库等持久存储介质中，方便后续的核查。

● 公平伦理算法组件负责提供判断系统行为是否违反公平伦理及法律法规的标准，针对具体的AI应用，需要根据其可能涉及的公平伦理及法律法规问题，设计出专用的公平伦理算法。如可以使用机会均等（equal opportunity）[29]方法，对系统一段时间内的优惠券发放行为进行审查，判断有无对老年人的歧视行为发生，也可以使用广义熵指数（generalized entropy index，GEI）[75]方法对针对某位员工的调度指令进行检查，判断该调度是否公平。

● 自动监管组件针对AI系统中类似“大数据杀熟”“数字劳工”等违反伦理或法规的系统敏感行为，进行事前数据、事中处理、事后结果的监管，实时或定期地从公平伦理算法组件中选择合适的算法对系统行为进行监管。

可信管理层是可信技术与管理要求的结合，T-DACM通过模型、事件、系统3个层级的协作满足了AI可解释、可追溯、可监管的要求。可信管理层可以帮助人们更好地掌控AI的行为，是可信AI不可或缺的组成部分。

4 应用案例

T-DACM通过可信数据层、可信算法层、可信计算层、可信管理层的联合协作，满足了可信AI应用及管理的要求，并可通过组件扩展的方式满足未来需求。其应用全景图如**图6**所示，在研发阶段，可通过可信数据治理、可信算法设计、模型学习等相关过程实现可信AI模型的开发工作；在应用阶段，可通过可信数据治理、模型推理、安全多方计算等相关过程实现可信AI解决方案的落地工作；在管理阶段，可通过事件

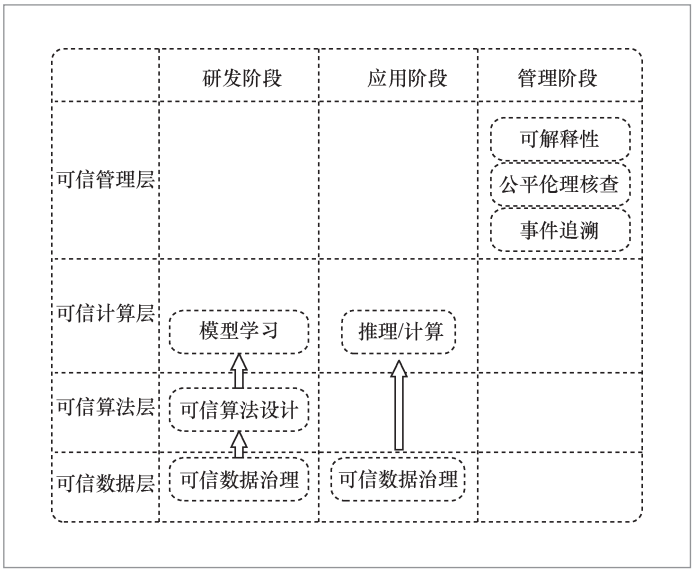

图 6 可信 AI 治理框架应用全景图

追溯、公平伦理核查、可解释性等能力接口对AI系统进行可信监管。

以星环信息科技（上海）股份有限公司Sophon可信AI治理套件（Sophon trusted-AI toolkit）在某银行风控项目的实施为例。该项目目标是对原授信系统进行可信化升级改造。原授信系统架构如**图7**所示，由数据层、算法库、管理层、应用层与决策引擎构成。该架构存在以下问题及需求。

业务推广过程中高风险客户的识别率仅达到87%，较低的风险识别率提高了项目的财务风险；此外，还存在大约3%的优质客户联系方式失效的问题，导致客户资源流失。

现有黑盒客户信用评分模型无法对客户评分差异进行解释，授信结果缺乏透明性及可比较性，给授信工作的顺利开展增加了困难，需要引入新方法解决此问题。

对于高风险客户识别率较低的问题，尽管尝试了多模型融合、参数优化等多种模型优化方法，模型精确率仍只有91.7%。针对模型优化效果不佳的情况，引入T-DACM的可信计算层，通过纵向联邦学习组件利用第三方风控数据进行联合建模，从数据优化的维度提升模型精度。联邦学习的方式解决了模型学习过程中数据风险特征不足的问题，在不直接引入运营商风险特征数据的前提下，将风控模型精度提升到99.2%，达到了商用要求。对于客户资源流失的问题，通过T-DACM可信计算层的安全多方计算组件，使用隐匿集合求交功能，实现与第三方社交媒体用户资源的匹配，从而找到失联客户，并进行接触推广，最终将失联用户率降低至0.4%，有效地挽回了客户资源。在本案例中，T-DACM可信计算层在不共享数据的前提下，完成了模型学习和联合计算，既保护了双方的数据隐私安全，又充分利用了数据的价值。

对于模型黑盒问题，项目初期尝试了线性回归、决策树等自解释性较好的模型，但未能达到项目使用要求。通过引入T-DACM可信管理层的可解释性模块，具体使用LIME[77]来解释信用评分模型的决策行为，LIME方法适用于多种模型及数据的解释，其解释结果容易理解，解决了

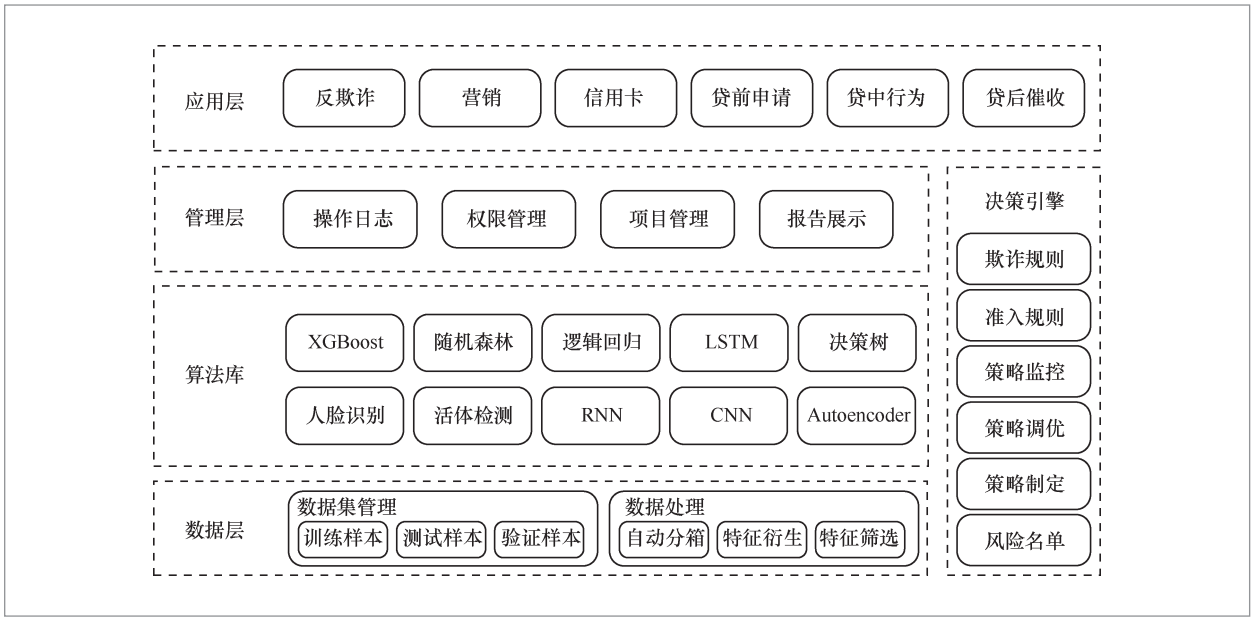

图7 原授信系统架构

对客户的评分差异无法解释的问题。

综上所述，引入T-DACM对原风控项目进行了如图8所示的改造。

T-DACM的引入解决了原有项目开发、运营过程中模型精度不足、模型黑盒难以解释等问题，为AI应用提供了可信化改造的指导。该案例对类似项目的可信AI问题的解决具有借鉴意义。

5 结束语

近10年来，人工智能应用呈现爆发式增长，在人脸识别、自动驾驶、对话系统、金融风控等领域得到了广泛应用。但是，外部恶意攻击与内部机理引发的可信事故给AI的深入发展带来了新的挑战。本文提出的T-DACM结合数据、算法、计算、管理4个维度的可信方法，提供了一个端到端的可信AI解决方案，并在产业界进行了实践与落地。可信AI是一件任重而道远的事情，当前模型的黑盒仍未完全解开，彻底的可信尚未达到。相信随着人们对AI研究的愈加深入、新方法新技术的不断提出，可信AI治理框架必将越来越完善。

参考文献:

[1] 杨庆峰. 从人工智能难题反思AI伦理原则[J]. 哲学分析, 2020, 11(2): 137-150, 199.
YANG Q F. An analysis of ethical principle of AI: based on the difficult problem of AI[J]. Philosophical Analysis, 2020, 11(2): 137-150, 199.

[2] IDC. IDC forecasts companies to spend almost $342 billion on AI solutions in 2021[Z]. 2021.

[3] ZHANG M G, ZENG K H, WANG J W.

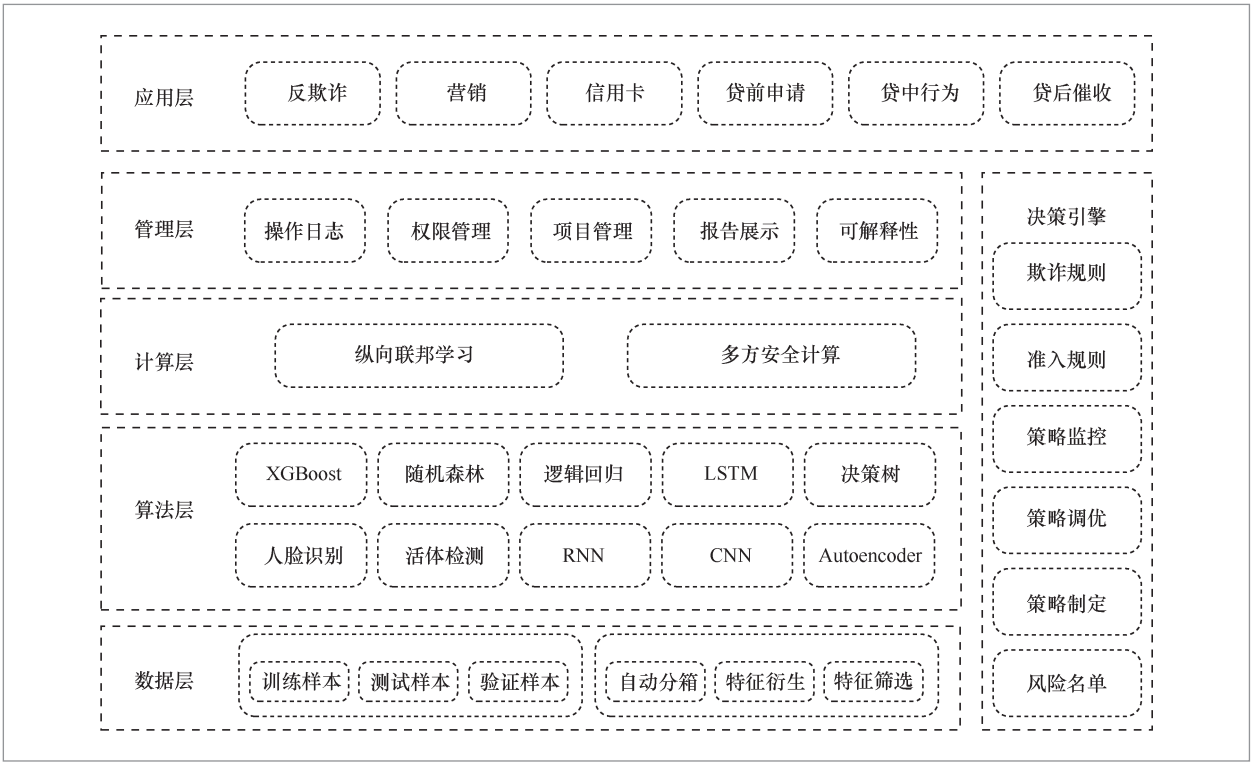

图 8 改造后系统架构

A survey on face anti-spoofing algorithms[J]. Journal of Information Hiding and Privacy Protection, 2020, 2(1): 21-34.

[4] LU J J, SIBAI H, FABRY E. Adversarial examples that fool detectors[J]. arXiv preprint, 2017, arXiv:1712.02494.

[5] GU T Y, DOLAN-GAVITT B, GARG S. BadNets: identifying vulnerabilities in the machine learning model supply chain[J]. arXiv preprint, 2017, arXiv:1708.06733.

[6] 邢根上, 鲁芳, 罗定提. 政府监管下的电商大数据“杀熟”演化仿真分析[J]. 湖南工业大学学报, 2021, 35(2): 65-72.
XING G S, LU F, LUO D T. An evolution simulation analysis of E-commerce big data-based price discrimination under government supervision[J]. Journal of Hunan University of Technology, 2021, 35(2): 65-72.

[7] 朱悦蘅, 王凯军. 数字劳工过度劳动的逻辑生成与治理机制[J]. 社会科学, 2021(7): 59-69.
ZHU Y H, WANG K J. Logical generation and governance mechanism of digital labor overwork[J]. Journal of Social Sciences, 2021(7): 59-69.

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

[9] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv preprint, 2014, arXiv:1412.6572.

[10] RIBEIRO M T, SINGH S, GUESTRIN C. “Why should I trust you? ”: explaining the predictions of any classifier[C]// Proceedings of 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. Stroudsburg: Association for Computational Linguistics, 2016.

[11] GIDARIS S, KOMODAKIS N. Dynamic few-shot visual learning without forgetting[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018.

[12] NGUYEN H H, FANG F M, YAMAGISHI J, et al. Multi-task learning for detecting and segmenting manipulated facial images and videos[J]. arXiv preprint, 2019, arXiv:1906.06876.

[13] IOSIFIDIS V, FETAHU B, NTOUTSI E. FAE: a fairness-aware ensemble framework[C]//Proceedings of 2019 IEEE International Conference on Big Data. Piscataway: IEEE Press, 2019: 1375-1380.

[14] SAUER A, GEIGER A. Counterfactual generative networks[J]. arXiv preprint, 2021, arXiv:2101.06046.

[15] 梁春丽. 欧盟AI道德准则草案出炉[J]. 金融科技时代, 2019, 27(2): 91.
LIANG C L. Draft EU AI ethics code released[J]. Financial Technology Time, 2019, 27(2): 91.

[16] SHIN J, BULUT O, GIERL M J. Development practices of trusted AI systems among canadian data scientists[J]. International Review of Information Ethics, 2020, 28: 1-10.

[17] ICO. Explaining decisions made with AI[Z]. 2021.

[18] WINTER P M, EDER S, WEISSENBÖCK J, et al. Trusted artificial intelligence: towards certification of machine learning applications[J]. arXiv preprint, 2021, arXiv:2103.16910.

[19] MENG D Y, CHEN H. MagNet: a two-pronged defense against adversarial examples[C]//Proceedings of 2017 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2017.

[20] KUANG K, XIONG R X, CUI P, et al. Stable prediction across unknown environments[J]. arXiv preprint, 2018, arXiv:1806.06270.

[21] ZHANG Q S, YANG Y, MA H T, et al. Interpreting CNNs via decision trees[C]// Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE

Press, 2019: 6254-6263.
[22] FENG R, YANG Y, LYU Y H, et al. Learning fair representations via an adversarial framework[J]. arXiv preprint, 2019, arXiv:1904.13341.
[23] ZHAO Y Y, ZHONG Z, YANG F X, et al. Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 6277-6286.
[24] 国家新一代人工智能治理专业委员会. 发展负责任的人工智能: 新一代人工智能治理原则发布[J]. 科技与金融, 2019(7): 2-3.
The National New Generation Artificial Intelligence Governance Specialist Committee. Developing responsible AI: a new generation of AI governance principles released[J]. Sci-Tech Finance Monthly, 2019(7): 2-3.
[25] 景慧昀, 魏薇, 周川, 等. 人工智能安全框架[J]. 计算机科学, 2021, 48(7): 1-8.
JING H Y, WEI W, ZHOU C, et al. Artificial intelligence security framework[J]. Computer Science, 2021, 48(7): 1-8.
[26] JIANG H, NACHUM O. Identifying and correcting label bias in machine learning[J]. arXiv preprint, 2019, arXiv: 1901.04966.
[27] HUANG L, VISHNOI N. Stable and fair classification[C]//Proceedings of the 36th International Conference on Machine Learning. [S.l.:s.n.], 2019: 2879-2890.
[28] YANG Q, LIU Y, CHEN T J, et al. Federated machine learning: concept and applications[J]. arXiv preprint, 2019: arXiv:1902.04885.
[29] HARDT M, PRICE E, SREBRO N. Equality of opportunity in supervised learning[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. New York: ACM Press, 2016: 3323-3331.
[30] BERK R, HEIDARI H, JABBARI S, et al. Fairness in criminal justice risk assessments: the state of the art[J]. arXiv preprint, 2017, arXiv:1703.09207.
[31] SHEN S, TOPLE S, SAXENA P. AUROR: defending against poisoning attacks in collaborative deep learning systems[C]//Proceedings of the 32nd Annual Conference on Computer Security Applications. New York: ACM Press, 2016: 508-519.
[32] 李立明, 詹思延, 叶冬青, 等. 流行病学[M]. 北京: 人民卫生出版社, 2020.
LI L M, ZHAN S Y, YE D Q, et al. Epidemiology[M]. Beijing: People's Medical Publishing House, 2020.
[33] AFCHAR D, NOZICK V, YAMAGISHI J, et al. MesoNet: a compact facial video forgery detection network[C]//Proceedings of 2018 IEEE International Workshop on Information Forensics and Security. Piscataway: IEEE Press, 2018: 1-7.
[34] JOURABLOO A, LIU Y J, LIU X M. Face De-Spoofing: anti-spoofing via noise modeling[C]//Proceedings of the European Conference on Computer Vision, [S.l.:s.n.], 2018: 290-306.
[35] KAMIRAN F, CALDERS T. Data preprocessing techniques for classification without discrimination[J]. Knowledge and Information Systems, 2012, 33(1): 1-33.
[36] ZHANG Z, NEILL D B. Identifying significant predictive bias in classifiers[J]. arXiv preprint, 2016, arXiv:1611.08292.
[37] CALDERS T, VERWER S. Three naive Bayes approaches for discrimination-free classification[J]. Data Mining and Knowledge Discovery, 2010, 21(2): 277-292.
[38] ROSENBAUM P R, RUBIN D B. The central role of the propensity score in observational studies for causal effects[J]. Biometrika, 1983, 70(1): 41-55.
[39] HULLSIEK K H, LOUIS T A. Propensity score modeling strategies for the causal analysis of observational data[J]. Biostatistics, 2002, 3(2): 179-193.
[40] CHIPMAN H A, GEORGE E I,

MCCULLOCH R E. BART: Bayesian additive regression trees[J]. The Annals of Applied Statistics, 2010, 4(1): 266–298.

[41] YAO L Y, LI S, LI Y L, et al. ACE: adaptively similarity-preserved representation learning for individual treatment effect estimation[C]// Proceedings of 2019 IEEE International Conference on Data Mining. Piscataway: IEEE Press, 2019: 1432–1437.

[42] KUANG K, CUI P, LI B, et al. Treatment effect estimation with data-driven variable decomposition[C]//Proceedings of the 31st AAAI Conference on Artificial Intelligence. [S.l.:s.n.], 2017: 140–146.

[43] BISHOP J M. Artificial intelligence is stupid and causal reasoning won't fix it[J]. arXiv preprint, 2020, arXiv:2008.07371.

[44] GEIRHOS R, JACOBSEN J H, MICHAELIS C, et al. Shortcut Learning in deep neural networks[J]. Nature Machine Intelligence, 2020, 2(11): 665–673.

[45] MOOSAVI-DEZFOOLI S M, FAWZI A, FAWZI O, et al. Universal adversarial perturbations[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 86–94.

[46] HOSPEDALES T M, ANTONIOU A, MICAELLI P, et al. Meta-learning in neural networks: a survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021: 1.

[47] ZHANG Y, YANG Q. A survey on multi-task learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2021: 1.

[48] HE X, ZHAO K Y, CHU X W. AutoML: a survey of the state-of-the-art[J]. Knowledge-Based Systems, 2021, 212: 106622.

[49] CATON S, HAAS C. Fairness in machine learning: a survey[J]. arXiv preprint, 2020, arXiv:2010.04053.

[50] TANG DIANE, AGARWAL A, O'BRIEN D, et al. Overlapping experiment infrastructure: more, better, faster experimentation[C]//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2010.

[51] YAO L Y, CHU Z X, LI S, et al. A survey on causal inference[J]. ACM Transactions on Knowledge Discovery from Data, 2021, 15(5): 1–46.

[52] VERMA T, PEARL J. Equivalence and synthesis of causal models[C]// Proceedings of the 6th Annual Conference on Uncertainty in Artificial Intelligence. [S.l.:s.n.], 1990: 255–270.

[53] SPIRTES P, GLYMOUR C, SCHEINES R. Causation, prediction, and search[M]. [S.l.]: The MIT Press, 2001.

[54] HOYER P O, JANZING D, MOOIJ J, et al. Nonlinear causal discovery with additive noise models[C]//Proceedings of the 21st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2008: 689–696.

[55] SHIMIZU S, HOYER P O, HYVÄRINEN A, et al. A linear non-gaussian acyclic model for causal discovery[J]. The Journal of Machine Learning Research, 2006, 7: 2003–2030.

[56] CORNFIELD J, HAENSZEL W, HAMMOND E C, et al. Smoking and lung cancer: recent evidence and a discussion of some questions[J]. International Journal of Epidemiology, 2009, 38(5): 1175–1191.

[57] ANGRIST J D, IMBENS G W, RUBIN D B. Identification of causal effects using instrumental variables[J]. Journal of the American Statistical Association, 1996, 91(434): 444–455.

[58] MIAO W, GENG Z, TCHETGEN E J. Identifying causal effects with proxy variables of an unmeasured confounder[J]. Biometrika, 2018, 105(4): 987–993.

[59] TSAMARDINOS I, ALIFERIS C F. Towards principled feature selection: relevancy, filters and wrappers[C]// Proceedings of International Workshop on Artificial Intelligence and Statistics. [S.l.:s.n.], 2003: 300–307.

[60] TSAMARDINOS I, ALIFERIS C F, STATNIKOV A. Time and sample efficient discovery of Markov blankets and direct

causal relations[C]//Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2003: 673–678.

[61] TSAMARDINOS I, BROWN L E, ALIFERIS C F. The max-min hill-climbing Bayesian network structure learning algorithm[J]. Machine Learning, 2006, 65(1): 31–78.

[62] KAMISHIMA T, AKAHO S, ASOH H, et al. Fairness-aware classifier with prejudice remover regularizer[C]// Machine Learning and Knowledge Discovery in Databases, 2012: 35–50.

[63] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139–144.

[64] ADEL T, VALERA I, GHAHRAMANI Z, et al. One-Network adversarial fairness[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33: 2412–2420.

[65] PLEISS G, RAGHAVAN M, WU F, et al. On fairness and calibration[J]. arXiv preprint, 2017, arXiv:1709.02012.

[66] ZHU L G, LIU Z J, HAN S. Deep leakage from gradients[J]. arXiv preprint, 2019, arXiv:1906.08935.

[67] TRAMÈR F, ZHANG F, JUELS A, et al. Stealing machine learning models via prediction APIs[J]. arXiv preprint, 2016: arXiv:1609.02943.

[68] GENTRY C. Fully homomorphic encryption using ideal lattices[C]//Proceedings of the 41st annual ACM symposium on Symposium on theory of computing. New York: ACM Press, 2009: 169–178.

[69] DWORK C, MCSHERRY F, NISSIM K, et al. Calibrating noise to sensitivity in private data analysis[C]//Theory of Cryptography. Berlin: Springer, 2006: 265–284.

[70] SHAMIR A. How to share a secret[J]. Communications of the ACM, 1979, 22(11): 612–613.

[71] GOLDWASSER S, MICALI S, RACKOFF C. The knowledge complexity of interactive proof systems[J]. SIAM Journal on Computing, 1989, 18(1): 186–208.

[72] RABIN M. How to exchange secrets with oblivious transfer[J]. IACR Cryptol EPrint Arch, 2005: 187.

[73] YAO A C C. How to generate and exchange secrets[C]//Proceedings of 27th Annual Symposium on Foundations of Computer Science. Piscataway: IEEE Press, 1986: 162–167.

[74] HENDRICKS L A, AKATA Z, ROHRBACH M, et al. Generating visual explanations[C]//Computer Vision – ECCV 2016. [S.l.:s.n.], 2016.

[75] SPEICHER T, HEIDARI H, GRGIC-HLACA N, et al. A unified approach to quantifying algorithmic unfairness: measuring individual & Group unfairness via inequality indices[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM Press, 2018: 2239–2248.

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,主要研究方向为大数据、数据库、人工智能、流媒体处理技术等。

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、图计算等。

罗圣美(1971-),男,博士,星环信息科技(上海)股份有限公司大数据研究院院长,主要研究方向为大数据、并行计算、云存储、人工智能等。

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。

收稿日期: 2021-10-26