人工智能开源与标准化研究报告

二零一九年四月

专家咨询组

潘云鹤　高　文　戴　红　谭铁牛　吴朝晖　李伯虎
林　宁　于海斌　吴　飞　周志华　董景辰　黄河燕
朱小燕　张德政　朱恺真　曲道奎　左　毅　钱　恒

国家人工智能标准化总体组

组　长：赵　波
副组长：闵万里　胡国平　徐　波
　　黄铁军　吴文峻　欧阳劲松
秘书长：孙文龙

编写单位（排名不分先后）

中国电子技术标准化研究院
中国科学院自动化研究所
华为技术有限公司
北京深睿博联科技有限责任公司
腾讯计算机系统有限公司
成都四方伟业股份有限公司
京东数字科技控股有限公司
英特尔（中国）有限公司
威麟信息技术开发（上海）有限公司
国际商业机器（中国）投资有限公司
深圳前海微众银行股份有限公司
机械工业第六设计研究院有限公司
浪潮软件集团有限公司
深圳市商汤科技有限公司
重庆邮电大学
华夏芯(北京)通用处理器技术有限公司
南京云问网络技术有限公司
西门子（中国）有限公司
中国电力科学研究院有限公司
金税信息技术服务股份有限公司
深圳云天励飞技术有限公司
上海智能制造系统创新中心有限公司
重庆中科云从科技有限公司
天津天大康博科技有限公司
苏州苏相机器人智能装备有限公司

编写人员（排名不分先后）

侯培新　堵俊平　戴东东　代　红　董　建　张　群　马珊珊　汪小娟
王燕妮　关　贺　符海芳　孟繁亮　张文杰　杨晓光　付会文　安耀祖
翁家良　朱兆颖　李　斌　卢丽珊　王功明　庞　宇　杜振东　杨　萌
沈盛宇　刘　鹏　刘　斌　张大伟　章谦一　霍　欣　王伟才　易　明
田　忠　李海杰　颜深根　叶安华　刘　军　陈江宁　张英丽　秦湘军
王　彤　郑文先　陈　斌　代　翔　梅　军　王　飞　李　军　郑晨光
瞿卫新　杨　品

目录

第一章 概述

1.1 背景及目的

自 2013 年以来，随着深度学习技术的不断发展，引发了新一轮人工智能热潮，诸如：AlphaGo、刷脸支付、无人驾驶、AR、无人超市等应用层出不穷。大量资本和并购的涌入，加速了人工智能和产业的结合，人工智能甚至有可能成为是继蒸汽机、电力和计算机之后，人类社会的第四次革命。人工智能（Artificial intelligence，AI）是研究、开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的一门新的技术科学。在历史上，人工智能有过很多个定义，但是迄今为止没有一个官方的、统一的、正式的定义。人工智能最早由麻省理工学院的 John McCarthy 在 1956 年的达特矛斯会议上提出的：人工智能就是要让机器的行为看起来就像是人所表现出的智能行为一样。

世界各国纷纷将发展人工智能作为抢抓下一轮科技革命先机的重要举措。随着人工智能领域国际竞争的日益激烈，2017 年国务院印发《新一代人工智能发展规划》，提出我国新一代人工智能发展的指导思想、战略目标、重点任务和保障措施，为部署构筑我国人工智能发展的先发优势，加快建设创新型国家和世界科技强国构建了基础。

本报告中的开源指源码公开、源数据公开及其他成果形式（如软件、系统或平台架构等）的公开。近年来开源技术蓬勃发展，诸如计算机视觉开源社区 OpenCV、开源数据集 ImageNet、开源智能终端操作系统 Android 和其他大量开源工具及平台，无不表明开源创新与协同有力推动了产业进程。同理，人工智能尤其是深度学习相关的开源蓬勃发展，也将对我国人工智能相关产业产生积极影响。

第一，人工智能开源有助于支撑人工智能领域形成高端产业集群优势，逐步引领世界前沿技术的发展。

第二，人工智能开源有助于吸引更多人才进入人工智能产业，建设多层次人才培养体系。

第三，人工智能开源有助于推动人工智能广泛应用，加快推动人工智能与各

行业的融合创新和赋能。

1.2 本报告的价值

本报告旨在为政府及行业的政策制定者、企业业务决策者、技术决策者提供参考，促进经济社会各领域智能化转型，加速人工智能技术在全行业应用落地。

（1）促进人工智能产业的发展提升

报告集成了各行各业在人工智能领域的经典案例，提供了丰富的知识积累和发展经验，可以帮助决策者快速形成发展思路（包括实现方法和风险评估），促进行业的发展提升。

（2）加速人工智能技术的应用落地

报告描述了机器学习、深度学习开源技术全栈，通过介绍开源工具平台及基准的方法论，降低行业人员学习和应用人工智能的技术门槛，提升研发速度，降低研发和运维管理成本，使前沿技术和新兴算法能快速运用到具体领域业务中并创造价值。

（3）推动人工智能生态圈建设

报告所描述的人工智能领域的经验和需求能够促进人工智能生态圈的良性发展，促进企业的技术创新。标准与开源的联动能使产业发展更加健康。

（4）推动产业以更开放的心态进行协同创新

报告所倡导的开源开放的业态有助于推动中国人工智能开源走向更深层次，例如数据开放协同的文化及平台建设、开源分享思维和隐私保密需求的平衡等。报告会给出开放数据平台的构建思路及四种可供参考的方案。

国家人工智能标准化及开源研究报告对政府或行业的政策制定者,企业和科技决策者提供的4大价值

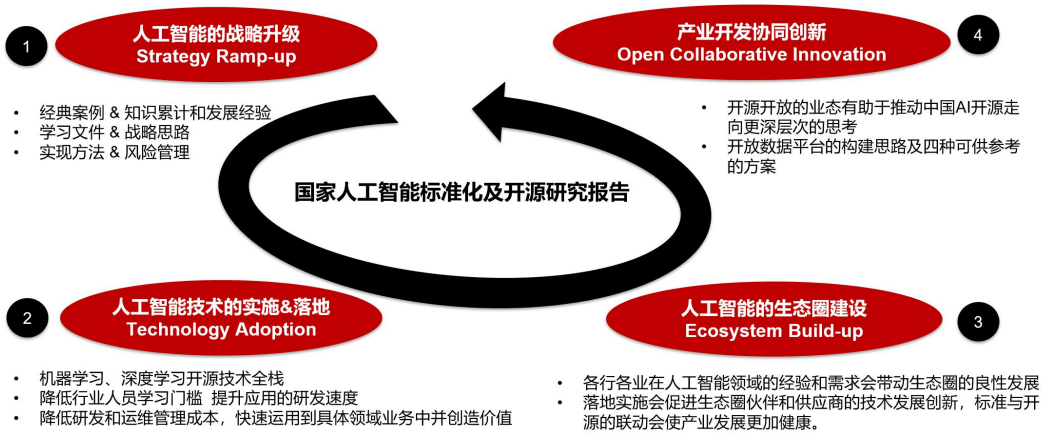

图 1 本报告的四大价值

1.3 本报告的脉络梳理与导读

本次报告分为如下几个部分:

第二章对 AI 的产业现状及人工智能开源落地行业的宏观问题进行描述。首先分析了国际国内当前 AI 产业集聚情况，从基础层、技术层和行业应用层等三个层次深入描述了 AI 产业链现状，提及了工业、医疗、电商等典型领域的应用场景，然后对 AI 开源面对的法律道德、垄断风险、标准统一等系列问题进行了深入探讨。

第三章主要对人工智能开源现状及相关生态做一个全面的分析，包括人工智能开源项目（聚焦机器学习、深度学习）的全栈图，并给出全栈各层的定义。其次从各层挑选一些典型的项目来分析其技术及生态特点、主要的发起及参与的公司及个人。此外也介绍相关的开源组织，包括国内的组织如开源中国、国际组织如 Linux 基金会及 Apache 软件基金会，然后分析各组织、公司在重要开源项目中的角色及目的。

第四章首先介绍了 AI 在应用时的总体工作流，然后系统化地分析当前基于开源的人工智能技术在解决行业具体问题时还有哪些不足、开源起了什么作用、还有哪些短板、在全栈中还有哪些缺失等，试图从技术生产者及技术消费者两个纬度来阐述人工智能开源技术是否可以解决所有问题。

第五章主要阐述新一代数据驱动的人工智能将给传统以代码为核心的开源

理念带来哪些挑战，从政府角度以及一些典型行业出发分析 AI 数据开放和协同中存在的问题，从顶层设计、法律规范、数据治理、开源数据平台建设说明 AI 数据开放和协同的可行性，最后给出四种可行性技术架构推动新一代开源运动 (Open Source Movement) 的升级，实现“开放生态圈平台”（Open Ecosystem Platform）的愿景。

第六章将着重阐述人工智能领域开源与标准的关系和相互促进。首先介绍在云计算、大数据、电信网络等几个成功的开源与标准联动的案例，随后阐述人工智能领域开源与标准的相互关系和联动建议，并针对人工智能落地过程中的问题梳理出标准的机会，最后阐述在标准制定中可能遇到的问题以及相关思考。

第二章 AI 产业现状及开源面临的宏观问题

2.1 AI 产业现状及产业链

现有 AI 开源产品在行业中的应用越来越多，一些企业利用自身的技术优势，重点打造 AI 应用开放平台，提供语音引擎、视觉引擎、自然语言处理引擎等众多 AI 基础技术；围绕开放平台，构建人才生态和行业生态，全面覆盖教育、金融、家电、医疗、手机、汽车、安防等领域，在内业已产生巨大的经济价值和社会价值。伴随着应用场景的快速发展，数据开源会成为新的趋势，数据收集和标注的标准化需求也会越来越迫切，业内也产生了一批从事数据收集和标注的初创公司和平台。

目前，全球涉及人工智能的企业集中分布在美国、中国、加拿大、德国等少数国家或地区，且在美国和中国的企业数量已占全球的半数以上。美国和中国依靠其卓越的技术研发机构及融合丰富应用场景的各类实验室，协同领衔全球人工智能的发展，奠定了雄厚的技术基础。中国当前具有多个人工智能聚集中心和地方特色人工智能发展产业，其中以北京与天津、上海与杭州、深圳与广州为重点城市群抱团发展的产业格局逐步显现，形成三大人工智能聚集中心。

图 2 是我们制定的人工智能参考框架图，图 3 是人工智能领域目前在产业界应用的全景图。在产业全景图中的“基础设施”层对应了参考框架中的“数据”与“算力”，产业全景图中的“关键技术”层对应了参考框架中的“算法”，产

业全景图中的“智能系统”及“行业应用”对应了参考框架中的“产品与服务”。

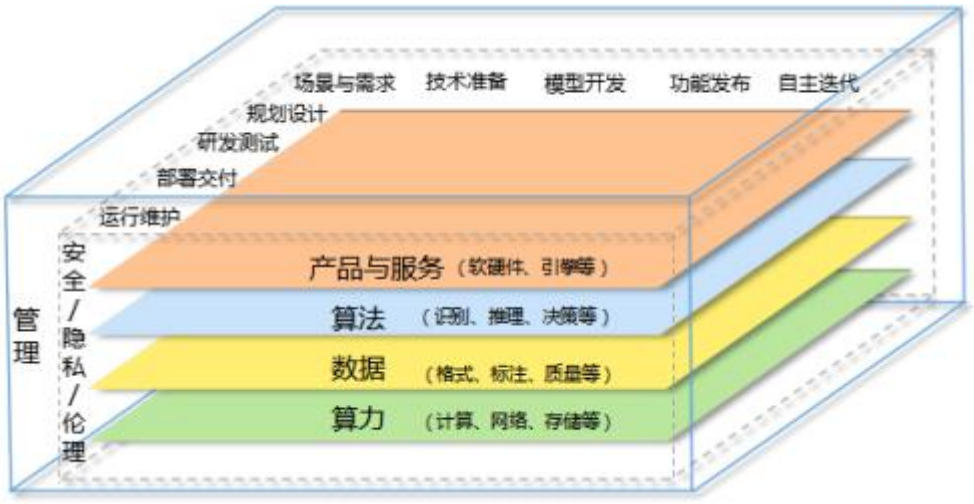

图 2 人工智能参考框架图

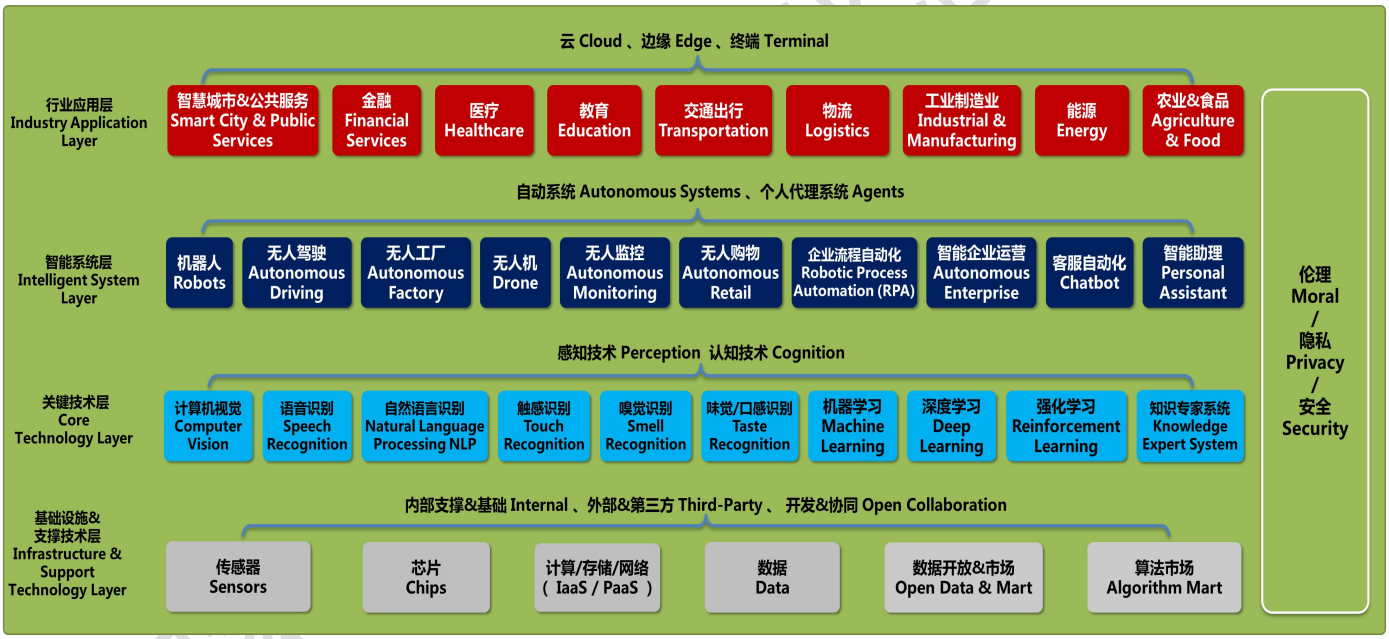

图 3 人工智能产业生态圈全景图

人工智能产业链宏观上由基础层、技术层和应用层等三个层次组成，其中基础与核心技术的研究主要分布在大企业及科研机构，而应用层的研究测试在大中小企业均有涉及，形成了全面开花、全行业覆盖的局面。

2.1.1 基础层

芯片研发作为基础层的核心，已成为人工智能发展的关键因素。芯片在技术架构方面可分为通用类芯片（如 CPU、GPU 等）、半定制化芯片（如 FPGA 等）、

全定制化芯片（如 ASIC 等）和类脑计算芯片。目前 GPU 是深度学习训练平台的主流配置，而 FPGA 的灵活可编程特点可以使得在算法未完全成熟时切入市场，同时其低功耗特性也被大型数据中心所青睐。在专用人工智能芯片领域，自 2016 年 Google 发布了 TPU 芯片后，这一市场热潮不断。国内如寒武纪、地平线、华为海思等公司也纷纷研发出可规模商用的人工智能专用计算芯片。

随着物联网技术的不断发展，传感、计算、通讯、AI 等功能的集成变得尤为重要，若每个功能均依靠单一芯片，不但效率低下，而且能耗和成本都很高，因此将不同的功能整合在一起，构建异构芯片，会极大缓解上述问题。由 AMD、ARM、华为、HXGPT、高通、IMAGINATION 和三星等公司组成的全球异构系统架构(HSA) 联盟在 2017 年成立了中国区域委员会(CRC)。CRC 的任务是以构建 HSA 生态系统为侧重点，提高对异构计算的意识认知，并促进 HSA 在中国的标准化进程。

新一代人工智能依赖于海量数据的处理、存储、传输，因此离不开云计算。云计算是把大量的计算资源封装抽象为 IT 资源池，用于创建高度虚拟化的资源供用户使用。通过动态整合、共享硬件设备供应来实现 IT 投资的利用率最大化，降低了使用计算的单位成本及 IT 运维成本，促进了人工智能产业的商业化进程。

2.1.2技术层

目前技术层中的核心技术主要由科技巨头企业掌控，如微软、亚马逊、Google、Facebook、百度、阿里、腾讯、京东、小米、商汤等。此外，一大批初创企业和开源组织也陆续加入其中。它们共同探索和推进 AI 技术的发展，催生出了一批在业内有深远影响力的开源项目，如 TensorFlow、PaddlePaddle、Caffe、CNTK、Deeplearning4j、PyTorch 、Mahout、MLlib。这些科技企业，通过招募 AI 高端人才及组建实验室等方式加快关键技术研发，并通过开源技术平台构建生态体系。

技术层面，包含机器学习、知识图谱、自然语言处理、虚拟现实或增强现实、计算机视觉、生物特征识别、人机交互等技术与应用场景相结合，从而衍生出大量的智能化产品与服务，包括智能家居、智能机器人、智能搜索引擎、智能问答系统、一体机 VR、无人驾驶汽车、人脸识别系统、智能客服等。

2.1.3 行业应用层

人工智能是制造业数字化、网络化、智能化转型发展的关键引擎，是促进实体经济发展的重点方向。近年来各国政府和产业界纷纷采取行动推进基础性研究及产业实践部署，人工智能的各种应用如机器人、无人驾驶、智能客服等百花齐放，大中小企业均有涉及，形成了全行业全覆盖的局面。本次报告的写作单位涉及了如下行业：工业制造、医疗、电商、公安、金融、消费电子、交通、物流、航空、能源、政务等，因此下面报告中不论是行业案例还是痛点分析等主要以它们为主。同时这些行业也是在 AI 应用中诉求比较明确的领域。

2.1.3.1 工业

人工智能在工业领域深度融合新一代信息通信技术与先进制造技术，贯穿于设计、生产、管理、服务等制造活动的各个环节，引导具有自感知、自学习、自决策、自执行、自适应等功能的新型生产方式。人工智能在工业领域进一步融合拓展的应用方向还有机器视觉检测分拣、人机交互、可视化及 AR/VR、行业知识图谱及知识自动化等，支持工业设备能耗预测与优化，增强工业设备预测性维护和智能故障诊断，为企业生产个性化需求、企业运行优化及产品生命周期控制提供辅助决策，进而提升制造质量水平和企业经济效益。

2.1.3.2 医疗

基于图像分析技术的影像辅助诊断和医学病理分析相结合，提供了更准确的临床诊断，同时提升了医疗服务的效率。在健康趋势分析、疾病预测、影像辅助诊断等领域引入人工智能技术，可以有效预测疫情并防止其进一步扩散和发展，提供患者预前和预后诊断和治疗的评估方法和精准诊疗决策，有效提高医护人员工作效率和诊断水平，从而在整体上为医疗健康领域向更高的智能化方向发展提供了非常有利的技术条件。

2.1.3.3 电商

在电商领域，无人店、无人货架纷纷引入人脸识别、货物识别等先进技术，实现无人值守，融合人工智能的仓储机器人，实现了货物的识别、拣选和自动搬运等功能，极大解放了生产力。通过对消费者历史购买行为的深入分析，提供了更精准的目标客户营销和商品推荐。

2.1.3.4 公共安全

在安防领域，通过支持前端提取信息，如采用在复杂场景下的人车混合多特征结构化信息技术，提取人脸属性、人脸轨迹、车牌车型等特征属性，利用人工智能对视频、图像进行存储和图像比对分析，建立危险人数图像库，从而识别危险隐患并进行安全处理，是构建未来智慧城市安防体系的基础，在反恐维稳、犯罪预警、案件侦破和网络音视频监管等领域具有重要应用价值和广泛的应用前景。

2.1.3.5 金融

在金融领域可以借助大数据，以人工智能为内核支持金融行业的用户画像识别、资产信息标签化、智能获客、身份认证、智能化运维、智能投顾、智能理赔、反欺诈与智能风控、大数据征信、网点机器人服务等应用场景，对于提高金融系统管理效率、拓展金融新业务、防范金融风险等方面意义重大。

2.1.3.6 智能终端/个人助理

以住宅为平台，基于物联网技术，由硬件、软件系统、云计算平台构成的家居生态圈，实现远程控制设备、设备间互联互通、设备自我学习等功能，并通过收集、分析用户行为数据为用户提供个性化生活服务。通过人机交互应用在多种服务行业的咨询、指引、查询、讲解和业务办理等应用场景；与 APP 连接，实现硬件控制、日程管理、信息查询、生活服务、情感陪伴等。

2.1.3.7 交通

借助移动通信、宽带网、射频识别、传感器、云计算等新一代信息技术作支撑，利用摄像头监测交通路况和车辆信息，联通各个核心交通元素，广泛应用人工智能技术、统计分析技术、数据融合技术、并行计算技术等处理海量交通信息数据，实现信息互通与共享，以及交通元素间彼此协调、优化配置和高效使用，形成人、车和交通的一个高效协同环境。

2.1.3.8 物流

利用智能搜索、推理规划、计算机视觉、智能机器人、大数据分析以及射频识别、自动感知、全球定位系统等先进的物联网技术，应用于物流业运输、仓储、配送、包装、装卸等基本活动环节，实现智能物流系统的线路规划、人车资源调配、自动化运作和高效率优化管理，提高物流效率，提升物流行业的服务水平，降低成本，减少自然资源和社会资源消耗。

2.1.3.9 航空

利用机器学习、知识图谱、自然语言处理、人机交互、计算机视觉、生物特征识别、AR/VR 为基础的感知与认知、决策执行与控制、交互与协同、检测与维护等内容，应用于人脸识别安检、智慧航显、航空发动机预测健康管理及航空大数据分析、路径规划、任务规划、集群管理、目标识别、战术决策、毁伤评估、质量评估和可靠性实验检测等方面。不论民用还是军用航空领域，AI 可以实现人与机器智能的结合，全面提升观察-调整-决策-行动（OODA）环的运行速度和运行质量。

2.2 AI 开源所存在的问题

2.2.1 法律道德问题

随着人工智能的发展，其已经逐渐涉及到违法犯罪的黑色领域，被大肆用于

诈骗、色情、犯罪、甚至未来的战争中。例如，无人车、无人机等设备可以在不依赖人的情况下自主做出决策，做出危及人身安全的动作。这些新的情况将会带来伦理、道德和法律上的一系列危机，亟待相关专家给出合适的解决方案。

高质量、大规模的数据一般会认为是企业的重要资产，开源或开放后可能导致丧失竞争优势，缺少让数据开源贡献者获得合理回报的机制。另外，数据和模型很难保护自身知识产品不被竞争对手抄袭，甚至直接使用，这对企业进行数据开源形成了很大阻力，因此需要建立合适的政策保护机制。

2.2.2 潜在锁定风险

目前虽然有大量的开源技术和软件可以使用，但是背后的厂商如谷歌、Facebook、亚马逊、苹果对这些开源技术也掌握着绝对话语权。一旦使用开源软件的某些厂商利益跟上述公司相冲突，不排除被取消软件授权或者相关软件不再更新的可能性。企业基于自身的相关考虑，将相关项目进行开源，然而由于企业自身存在大量业务开展，因此导致其开源的相关项目的维护不及时，一旦项目停止维护，项目的使用者则面临进退两难的困境，平台迁移成本太高，但若不迁移平台，业务也无法得到平台新的支持。

2.2.3 安全问题

AI 开源工具虽有开放、共享、自由等特性，企业在享受开源技术带来的便利的同时，也存在巨大的安全风险。由于源代码公开，所有发现的漏洞都会被第一时间公布，因此也容易被攻击者利用；由 AI 开源技术形成的软件，其最终使用用户往往得不到最及时的更新，并且在软件开发和验收过程中，不易准确判断软件里包含哪些开源组件，容易造成安全隐患。另一方面，AI 开源代码在社区中一般由相应的团队或个人开展维护工作，缺乏对应的激励机制保障代码查找漏洞或及时更新，也会导致用户疑虑，降低 AI 开源技术及产品的推广使用。

2.2.4 标准统一问题

不同于其他开源软件，当前 AI 开源模式不够充分，仅限于开源 AI 框架，数

据开源力度不足，对 AI 技术的应用形成了壁垒。深度学习方面，AI 已开源框架、工具缺乏基本的统一标准，造成不同框架下的模型算法兼容困难；硬件优化方面，AI 开源软件大多在 X86 和 GPU 上进行优化，很少有在其他体系结构上进行优化的项目；数据格式方面，AI 开源目前多是针对深度学习的开源项目，而深度学习需要大量的训练数据，数据问题将许多公司卡在门外；模型算法方面，从数据和模型研究到形成产品方案之间存在明显差距。很多开源的 AI 算法，仅在所限定的理想条件下有效，难以适应复杂的实际应用环境，且在大规模分布式计算与存储环境下效果不佳。

2.2.5 版本兼容性问题

不同开源工具的兼容性问题导致整合困难，同一开源工具的不同版本之间也存在兼容性问题。开源社区涌现了一批以 Caffe、MxNet、TensorFlow、Torch 等为代表的热门 AI 开源开发框架，这些框架简化了 AI 技术的工程实现难度，但是每个框架之间接口不统一，模型格式不一致，在一定程度上造成了在各个框架之间迁移成本较高的问题，使得模型的复用较为困难，同时也增加了用户的学习成本，为在不同场景下使用不同开发工具造成了一定的障碍。即使对同一 AI 开发框架，接口调整较频繁，每次升级都会导致不少额外工作量；变化内容较为激进，项目自身向上兼容能力较差，导致企业/个人在更新开源软件时带来极大风险，同时也增加了用户的学习成本。

2.2.6 行业问题

虽然当前 AI 已开始逐步应用，但各行业因为自身的属性，均面临一些棘手的难题，制约着 AI 朝更深入、更广泛的方向应用。由于前期研发周期较长，相关领域技术人才缺乏，且雇佣成本较高，实际经济回报难以预估，许多企业不敢冒险尝试。另外，智能制造领域中的人工智能标准及开源代码仍然相对较少，无法满足当前人工智能技术的标准化需求，并制约着我国人工智能应用的有序、规范、健康发展。

传统金融机构历史包袱重，多数核心系统难以迅速采用开源 AI 技术；金融行业注重客户的数据隐私保护，数据的使用制约限制了人工智能相关模型的有效

性；既懂金融业务、又懂开源 AI 技术的人才也极度稀缺。

航空航天行业背景特殊，需要有针对性地进行开源。目前完全出于航空航天领域考虑的 AI 框架少，技术架构不明晰，且军事领域由于出于安全问题考虑，公共技术移植也较少；此外，在民用领域，从飞控系统的开源开始，就不断打开了无人机的进入门槛，但是对于开源的安全性和稳定性还有待考虑，对于开源的质量评定等还处在探索阶段。

第三章 AI 开源生态现状

3.1 AI 开源全栈（聚焦机器学习及深度学习）

《人工智能标准化白皮书（2018 版）》中对 AI 技术栈分为：智能芯片、智能传感器、机器学习、知识图谱、自然语言处理、人机交互、计算机视觉、生物特征识别及虚拟现实/增强现实等。考虑到目前 AI 领域开源主要是在机器学习，尤其是深度学习领域，而且这也是目前 AI 技术创新与落地的主要领域，本报告总体上聚焦分析深度学习中的开源技术、所面临的问题以及相关的技术创新、标准化的机遇等。基于如上考虑，后续除非特别指出，在本报告的语境中，AI 主要对应的是机器学习/深度学习这一领域。

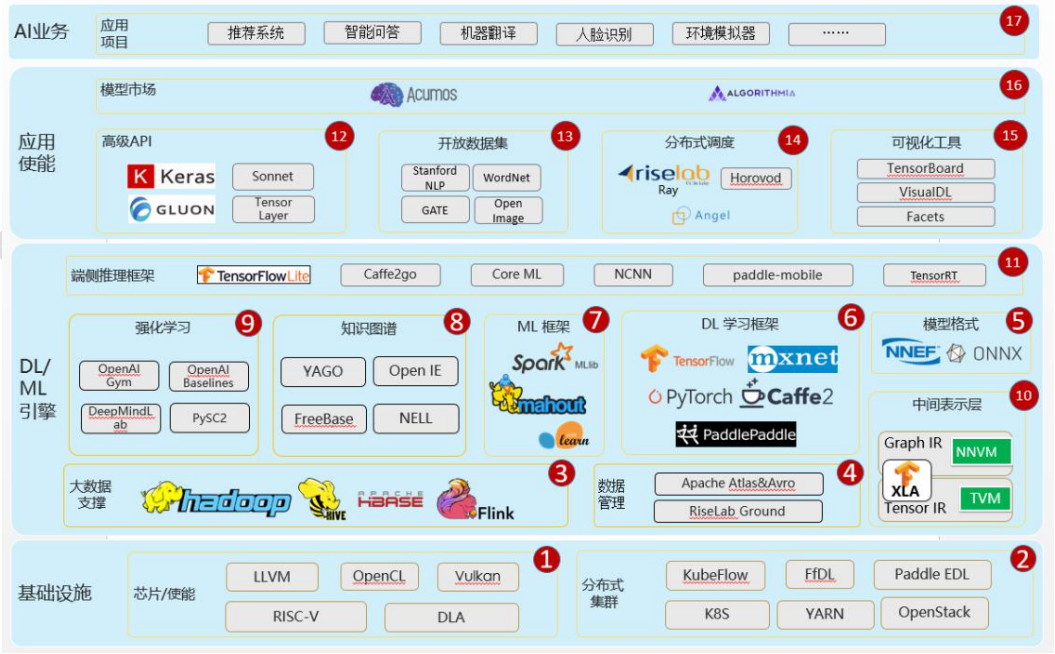

图 4 AI 开源全栈示意图

上图是目前比较活跃的机器学习及深度学习的开源社区及项目所形成的开源全栈示意图，整个开源全栈被分为四个层次：

（1）基础设施

AI 芯片是为 AI 计算设计的高性能芯片，业界有 AI 芯片设计、指令集、编程框架等开源芯片使能项目；同时 AI 工作负载（训练、推理、数据准备与治理、生命周期管理等）计算量大，对存储、网络传输等要求高，需要依赖分布式部署和承载的软件平台。这两部分构成了 AI 框架的基础设施层。

（2）深度学习/机器学习（DL/ML）引擎

深度学习/机器学习（DL/ML）引擎主要指深度学习/机器学习框架，包括训练与推理，以及与二者相关的模型格式、框架内数据格式等，同时也包括训练及推理之前的数据准备与数据管理。

（3）应用使能

应用使能包括数据科学家、AI 算法工程师在准备数据与使用训练框架之外的所有周边工作，主要包括任务视角的集群管理、数据可视化、框架易用性，以及模型/数据市场及端到端的生命周期管理。

（4）AI业务

AI 业务主要指基于 AI 基础及通用能力之上构建的领域 AI 服务，涵盖视频、语音、图像文本等。

以下是各组成部分的主要分类、描述和相关项目（按图中数字标识顺序）：

3.1.1 芯片使能

深度学习需要大量重复执行矩阵乘法、激活函数（如 sigmoid、tanh）等计算过程。通用 CPU 执行上述计算的性价比较低，需要采用专用的计算芯片。适合 AI 计算的芯片包括 GPU、FPGA 或者 ASIC 方案，它们通过把 AI 中常用函数计算硬件化来提升硬件计算速度、降低功耗。其中热门的开源技术包括开源的指令集、开源硬件实现方案、异构计算框架和编译器等。

DLA 是 NVIDIA 基于 Xavier SoC 的一个深度学习开源加速平台，适合于端侧推理场景的芯片 SoC。DLA 是一个卷积神经网络加速器（只能推理，并不能进行训练），它还需要外部的 CPU 和内存单元才能完整驱动整个加速器，CPU 通过

中断和 CSB 总线控制 NVDLA 加速器。链接：http://nvdla.org/

RISC-V 是基于精简指令集计算(RISC)原理建立的开放指令集架构(ISA)。RISC-V 是开源的指令集，可以免费地用于所希望的设备中，允许任何人设计、制造和销售 RISC-V 芯片和软件。基于 RISC-V 指令集架构可以设计服务器计算芯片、家用电器计算芯片、工控计算芯片和比指头小的传感器计算芯片。链接：https://riscv.org/

LLVM 是构架编译器(compiler)的框架系统，用 C++编写而成。项目启动于 2000 年，最初由美国 UIUC 大学主持开展发起的一个开源项目，目前 LLVM 已经被苹果 IOS 开发工具、Xilinx Vivado、Facebook、Google 等各大公司采用。链接：https://llvm.org/

OpenCL 当前由 Khronos 集团管理，是一个为异构平台编写程序的框架。此异构平台可由 CPU、GPU 或其它类型的处理器组成。OpenCL 由一种用于编写 kernels（在 OpenCL 设备上运行的函数）的语言（基于 C99）和一组用于定义并控制平台的 API 组成。链接：https://www.khronos.org/opencl/

Vulkan™ 也由 Khronos 集团开发，使软件开发人员能够全面获取 Radeon™ GPU 与多核 CPU 的性能、效率和功能，大幅降低了 CPU 在提供重要特性、性能和影像质量时的“API 开销”，而且可以使用通过 OpenGL 无法访问的 GPU 硬件特性。链接：https://www.khronos.org/vulkan/

Cyborg 是 OpenStack 社区中的一个官方项目，能够提供异构加速硬件通用管理框架。Cyborg 提供面向异构加速硬件的基础生命周期管理能力(CRUD 操作)，通过抽象通用的数据模型与统一的管理操作 API，为用户提供统一的异构计算资源使用体验，而无需针对每一种异构加速硬件特别构建管理模块。同时，Cyborg 提供异构加速硬件管理元数据的标准化，使得资源描述和业务需求之间的映射关系更加准确，在调度上可以更加通用与便捷。作为通用的异构计算资源管理框架，Cyborg 提供了如下标准化接口：

- 挂载与卸载异构计算设备
- 创建异构计算设备
- 删除异构计算设备
- 更改异构计算设备

- 查询异构计算设备
- 可编程异构计算设备的烧写
- 异构计算设备的租户配额控制

基于以上的标准化接口，用户可以通过云计算基础设施平台为其 AI 业务分配合适的异构计算资源，从而避免 GPU 资源调度不均、新的异构计算设备(如 ASIC 等)接入困难等平台相关的难题。链接：https://wiki.openstack.org/wiki/Cyborg

3.1.2 分布式集群

云计算平台对人工智能的基础支撑包括异构计算的部署和开发工具两个方面。在异构计算方面，GPU 已成为深度学习训练平台的主流配置；FPGA 的灵活可编程特点可以使得在算法未完全成熟时切入市场并方便地进行迭代，同时其低功耗特性也被大型数据中心所青睐；2016 年 Google 发布了 TPU 芯片，为机器学习提供定制化加速。在开发工具方面，现阶段业界有很多机器学习开发框架，云平台将这些框架服务化并进行针对性的优化，极大降低了人工智能开发的准入门槛。当前，AI 训练平台的部署已经形成以容器技术为基础的自动化部署趋势。

各种 AI 公司或者互联网公司的 AI 部门都在尝试如何在 Kubernetes (K8s) 上运行 TensorFlow、Caffe/2+PyTorch、MXNet 等分布式学习任务。不同的 AI 平台对集群分布式部署有不同的需求，也带来了配置难题。针对这些问题，各种部署工具也应运而生。

Google KubeFlow 正在不断完善 TensorFlow 在 K8s 集群上的部署、运维、参数调优等功能支持，并且已经有 TNN 等公司向 KubeFlow 贡献特性。链接：https://github.com/kubeflow/kubeflow

由 IBM 发起的 FfDL 是针对 TensorFlow、Caffe、PyTorch 等多个 AI 平台在 K8S 进行部署的工具，目前只实现了基本的部署功能。链接：https://github.com/IBM/FfDL

百度 Paddle 非常推崇在 K8S 上进行 AI 平台的部署，Paddle EDL 项目在 K8S 上提供了资源利用率、弹性调度、容错等部署能力。链接：http://www.paddlepaddle.org/

OpenStack 社区也加强对 GPU 等硬件的支持，同时也出现了一些支持 AI 平

台部署的项目。链接：https://www.openstack.org/

3.1.3大数据支撑

Apache Hadoop 已经成为大数据处理领域事实上的标准。Hadoop 作为一个完整的大数据处理生态圈，包括多个组件：分布式文件系统 HDFS、并行化计算框架 MapReduce、非关系型数据库 HBase、分布式协调系统 Zookeeper 等。国内外知名的 IT 公司，例如 Yahoo、亚马逊、百度、阿里巴巴等，都利用 Hadoop 集群批量处理上 PB 级别的数据。链接：http://hadoop.apache.org/

Hadoop 技术为大数据技术的应用提供了很好的支撑环境，优势主要体现在以下几处：

- Hadoop 本身是开源社区，方便定制；
- 扩展性好，安全性高；
- 社区活跃，得到多个大公司的支持；
- 成本低，开发周期短，技术成熟。

Apache Spark 是一个快速、通用、开源的集群计算系统，是适用于大规模数据处理的统一引擎。它可提供 Java、Scala、Python、R 语言的高级 API，能够高性能执行图计算，支持诸如 Spark SQL、结构数据处理、机器学习、图计算、流计算等多种高级工具。Spark 既可以单机方式运行，也可通过 YARN 在 Hadoop 集群上运行，它可兼容 Hadoop 数据，能够处理任何 HDFS、HBase、Cassandra、Hive 以及其它 Hadoop Input Format 的数据。Spark 既可运行批处理作业，又可处理流计算、交互查询、机器学习等新型作业。

Apache Hadoop YARN（Yet Another Resource Negotiator，又一种资源协调者），作为一个开源的通用资源管理系统，是一种新的 Hadoop 资源管理器，为上层应用提供统一的资源管理和调度，为集群在利用率、资源统一管理和数据共享等方面带来巨大好处。YARN 的基本思想是将资源管理、任务调度与监测分散到不同的进程中。资源管理器（RM）掌控全局，每个应用有自己的应用主控器（AM），应用要么是单独作业，要么是有向无环图作业。YARN 是 Hadoop 的扩展，它不仅可以支持 MapReduce 计算，还能管理诸如 Hive、Hbase、Pig、Spark/Shark 等应用，从而使得各种类型的应用互不干扰地运行在同一个 Hadoop 上面，并通过

YARN 从系统层面进行统一管理，共享整个集群资源。链接：https://hadoop.apache.org/

3.1.4数据管理

作为 AI 训练的前置需求，数据发现一直是企业在实施 AI 过程中消耗资源最为巨大的部分。DataCatalog 作为解决数据发现难题的手段，被主流厂商和社区所重视。DataCatalog 旨在通过对复杂场景下的多数据中心、多种数据源进行元数据（MetaData）统一管理来解决数据逻辑统一、物理分布的数据共享问题。

在开源社区领域，作为大数据端到端的数据治理方案包括：Apache Atlas（元数据治理）、Avro（统一数据交换格式）。伯克利大学的 RISELab Ground 项目也是专注于 Big Meta Data 管理，通过管理数据上下文，解决数据使用效率低下、治理困难等问题。

Apache Atlas（元数据治理）是一个可扩展核心数据治理服务集,支持数据分类、集中策略引擎、数据血缘、安全和生命周期管理，该项目支持管理共享元数据、数据分级&分类、审计、安全性以及数据保护。使企业能够有效地和高效地满足数据的合规性要求。链接：https://atlas.apache.org/

Apache Avro 可以将数据结构或对象转化成便于存储或传输的格式。Avro 设计之初就用来支持数据密集型应用，适合于远程或本地大规模数据的存储和交换。avro 支持跨编程语言实现（C, C++, C#，Java, Python, Ruby, PHP）。avro 依赖于一套可定义的 Schema，通过动态加载相关数据的 Schema，可以有效减少写入数据的开销，使得序列化快速轻巧。链接：https://avro.apache.org/

RISELab Ground 是一个数据湖（data lake）context 管理系统。它提供了一个 RESTful 服务的机制，让用户去推论他们拥有什么数据，数据从哪里来向哪里去，谁在使用数据，数据何时变化，为什么会有这种变化等。通过管理数据上下文，解决数据使用效率低下、治理困难等问题。Ground 提供了一个通用 API 和追踪信息的元模型，可以和很多数据储存库一起工作。链接：https://rise.cs.berkeley.edu/projects/ground/

3.1.5模型格式

ONNX 是微软、Facebook 为联手打造 AI 生态系统，推出的 Open Neural Network Exchange（ONNX，开放神经网络交换）格式。这是一个用于表示深度学习模型的标准，可使模型在不同框架之间进行转移。ONNX 是迈向开放生态系统的第一步，AI 开发人员可以轻松地在工具之间转换，选择最适合他们的组合。现在支持 ONNX 的框架有 Caffe2、PyTorch、Cognitive Toolkit、MXNet。谷歌的 TensorFlow 尚不支持 ONNX。链接：https://github.com/onnx/onnx

NNEF 是由 Khronos 集团主导的跨厂商神经网络文件格式，定义了压缩网络正式语义、结构、数据格式、通用操作（例如卷积、池化、正则化等），解决神经网络分裂化等问题。NNEF 计划支持包括 Torch、Caffe、TensorFlow、Theano、Chainer、Caffe2、PyTorch、MXNet 等几乎所有 AI 框架的模型格式转换。目前，已经有 30 多家芯片企业参与其中。链接：https://www.khronos.org/nnef

3.1.6 深度学习框架

TensorFlow 在 Google Brain 团队支持下，已经成为最大的活跃社区。它支持在多 GPU 上运行深度学习模型，为高效的数据流水线提供使用程序，并具有用于模型检查、可视化、序列化的配套模块。TensorFlow 今年对生态系统进行了大量的扩充，将 TensorFlow 的触角延伸到更多领域：支持 Keras 高级 API 封装，提高了开发效率；构建模型集，构建完善的常用模型库，方便数据科学家使用；发布 TensorFlow Hub，为再训练和迁移学习提供常用模型算法，共享多种精度预先训练好的模型；通过 TensorFlow.js 占据浏览器端深度学习生态，成为 TensorFlow 当前一个重要的发展方向。链接 https://www.tensorflow.org/

MXNet 是亚马逊（Amazon）主导的深度学习平台，性能优良，目前是 Apache 孵化器项目。MXNet 可以在任何硬件上运行（包括手机），支持多种编程语言：Python、R、Julia、C++、Scala、Matlab、Javascript 等。为了减低学习和使用的难度，MXNet 推出了 Gluon 高级 API 封装。链接：https://mxnet.apache.org/

Caffe/2+PyTorch 是 Facebook 主导的深度学习平台，目前已合并到 PyTorch 进行统一维护。在图像处理领域，Caffe 有着深厚的生态积累，同时 PyTorch 作为一

个易用性很强的框架，受到越来越多数据科学家的喜爱。在国内，很多 AI 图像处理团队在试用 PyTorch、TensorFlow、MXNet 后，往往选择 PyTorch 作为其主要工作平台。链接：https://pytorch.org/

PaddlePaddle 是百度旗下深度学习开源平台，Paddle（Parallel Distributed Deep Learning）表示并行分布式深度学习。其前身是百度于 2013 年自主研发的深度学习平台，且一直供百度内部工程师研发使用。PaddlePaddle 是一个功能相对全面、易于使用的深度学习框架，一些算法封装良好，如果仅仅只需要使用现成的算法（VGG、ResNet、LSTM、GRU 等），那么直接执行命令，替换数据进行训练。PaddlePaddle 的设计和 Caffe 类似，按照功能来构造整个框架，二次开发要从 C++ 底层写起，因此适用于使用成熟稳定模型处理新数据的情况。它的分布式部署做得很好，支持 Kubernetes 的部署。链接：http://www.paddlepaddle.org/

BigDL 是一种面向 Apache Spark 的分布式深度学习库。用户可以通过 BigDL 将深度学习应用编写为标准的 Spark 程序，这些程序可以直接在 Sparkshedu Hadoop 集群上运行。BigDL 提供了丰富的深度学习支持，结合英特尔 MKL 和多线程应用，因此有极高的性能，可以实现高效的横向扩展，执行大规模数据分析。链接：https://bigdl-project.github.io/0.6.0/

Analytics Zoo 是一个基于 Spark 和 BigDL 的端到端智能分析流水线，通过提供高水平的流水线 API、内置深度学习模型、参考用例等，可以轻松地在 Spark 和 BigDL 上构建和生成深度学习应用程序，特别适合大数据集群分析和深度学习。链接：https://github.com/intel-analytics/analytics-zoo

3.1.7 机器学习框架

Scikit-learn 是 BSD 证书下开源基于 Python，构建于现有的 NumPy(基础 n 维数组包)、SciPy(科学计算基础包)、matplotlib(全面的 2D/3D 画图)、IPython(加强的交互解释器)、Sympy(Symbolic mathematics)、Pandas(数据结构和分析)之上，做了易用性的封装。Scikit-learn 提供一系列特征工程能力：降维(Dimensionality Reduction)、特征提取(Feature extraction)、特征筛选(Feature selection)能力等，同时对分类、回归、聚类、交叉验证、流型计算等机器学习算法和模型提供了标准实现。作为简单且高效的数据挖掘、数据分析的工具，被广泛应用在 ML 领域。

链接：http://scikit-learn.org/

Mahout 是 Apache Software Foundation（ASF）旗下的一个开源项目，主要关注协同过滤 (Collaborative Filtering, 简称 CF)领域的推荐引擎（协同过滤）、聚类和分类支持。Mahout 初期使用 Hadoop 的 MapReduce 作为计算框架实现，目前已经迁移到 Spark 和 Flink 为主的平台实现。链接：https://mahout.apache.org/

3.1.8知识图谱（知识库）

Freebase,Yago2 作为 Curated KBs（Curated KBs 知识库是由结构化的三元组（entity, relation/property, entity）所表达的知识组成的知识库，如果把这个三元组用起点-边-终点来表示，这个知识库就可以被表示为指示图）的代表，从维基百科和 WordNet 等知识库中抽取大量的实体及实体关系，可以把它们理解为是一种结构化的维基百科，被 google 收购的 Freebase 中包含了上千万个实体，共计 19 亿条 triple。像维基百科这样的知识库，与整个互联网相比，仍然数据量太小。

Open Information Extraction (Open IE), Never-Ending Language Learning (NELL) 作为 Extracted KBs 的代表，直接从上亿个网页中抽取实体关系三元组，涉及到 entity linking 和 relation extraction 两大关键技术。与 Freebase 相比，这样得到的知识更加具有多样性，而它们的实体关系和实体更多的则是自然语言的形式，当然，直接从网页中抽取出来的知识，其精确度要低于 Curated KBs。

3.1.9 强化学习

其他许多机器学习算法中学习器都是学怎样做，而强化学习(RL，Reinforcement Learning)是在尝试的过程中学习到在特定的情境下选择哪种行动可以得到最大的回报。RL 最重要的 3 个特点在于：

- 基本是以一种闭环的形式；
- 不会直接指示选择哪种行动；
- 一系列的行动和奖励信号都会影响之后较长的时间。

其核心就是运用马尔可夫决策过程（Markov Decision Processes,MDPs）。MDPs 就是一个智能体采取行动从而改变自己的状态获得奖励与环境发生交互的循环

过程。

OpenAI Gym 是 OpenAI 的一个开源项目。OpenAI 成立于 2015 年底，是一个非营利组织，它的目的是“建立安全的人工通用智能(AGI)，并确保 AGI 的福利被尽可能广泛和均匀地分布”。除了探索关于 AGI 的诸多问题之外，OpenAI 对机器学习世界的一个主要贡献是开发了 Gym 和 Universe 软件平台。Gym 是为测试和开发 RL 算法而设计的环境/任务的集合。Gym 让用户不必再创建复杂的环境，Gym 用 Python 编写，它有很多的环境，比如机器人模拟或 Atari 游戏。Gym 还提供了一个在线排行榜，供人们比较结果和代码。同时 OpenAI 还开源了与 Gym 配套的高质量强化学习算法实现项目 Baseline。链接：https://gym.openai.com/

DeepMind Lab 类似于 3D 游戏的平台，它的研发工作都基于智能体，透过仿真智能体的眼睛以第一人称的视角观察周围环境。智能体可以采集果实、走出迷宫、穿越危险的悬崖峭壁、玩激光游戏、快速学习和记忆随机变化的环境。同时 DeepMind 联合暴雪推出的星际争霸 2 开发环境 PySC2，封装了暴雪提供的机器学习 API，为通过 DeepMind Lab 实验强化学习提供了很大的方便。同时 DeepMind 还开源了 Control Suite，基于 MoJoCo 物理引擎设计了一组有着标准化结构、可解释奖励的连续控制任务，还为强化学习 Agent 提供一组性能测试指标。链接：https://deepmind.com/

3.1.10 模型中间表示层 IR

在深度神经网络中，中间层 IR 的覆盖范围比较广泛，其核心思想借鉴了 LLVM。IR 是为解决在不同硬件平台编译运行而产生的中间层表示形式，它是解决模型推理侧运行在不同硬件平台的重要描述方法，主要包括 NNVM/TVM 和 TensorFlow XLA 两大阵营。但实际上类似 ONNX、NNEF，模型交换格式的核心是对各种中间层表示的定义。中间表示层 IR 是打通在深度学习中多种不同前端训练框架与多种不同后端的表达桥梁，从而更有效实现它们之间的优化和映射。目前业界的中间表示层都一致地采用了 Graph IR + Tensor IR 两层优化结构，Intel nGraph、Apache SystemML 等都是如此。按照目前业界的共识，“IR”的竞争将是未来 Framework 之争的重要一环。

3.1.11 端侧推理框架

Caffe2go 是最早出现的移动端推理框架，让深层神经网络在手机上高效运行。由于端侧的 CPU 配置差异、设备性能有限，因此优化是非常重要的，Caffe2go 是基于 CPU 的优化进行设计。链接：https://github.com/caffe2

TensorFlow Lite 是运行在 Android 和 iOS 平台的计算框架，结合 Android 生态的 NN Runtime，能够实现较为高效的 AI 移动端应用速度，支持根据硬件情况自动切换 CPU 或 GPU。链接：https://www.tensorflow.org/lite/

NCNN 是腾讯开源的移动端 AI 执行框架，支持多种训练框架的模型转换，包括 Caffe、PyTorch、MXNet、ONNX。NCNN 主要面向 CPU 的 AI 模型应用，无第三方依赖，因此具有较高的通用性。在 CPU 领域，其 AI 模型运行速度明显强于 TensorFlow Lite，是国内目前较为广泛使用的移动端 AI 框架。链接：https://github.com/Tencent/ncnn

Core ML 是苹果公司的 iOS AI 组件，能够对接 Caffe、PyTorch、MXNet、TensorFlow 等绝大部分 AI 模型，并且自身提供了常用的各种手机端 AI 模型组件。链接：https://developer.apple.com/machine-learning/

paddle-mobile 是百度自研的移动端深度学习框架，将 paddle 模型部署在手机端。目前，在 iOS 系统中支持 GPU 计算，在 Android 系统中仅支持 CPU 计算。从社区的整体评价来看，功能比较单一，支持比较有限。链接：https://github.com/PaddlePaddle/paddle-mobile

对于计算量需求庞大的 CNN，需要通过压缩神经网络来提高效率，神经网络压缩最关键的方法是剪枝和量化。TensorRT 使用量化的方法，将 FP32 位权值数据优化为 FP16 或者 INT8，而推理精度没有明显的降低。关于 TensorRT，需要明确以下几点：

- TensorRT 是 NVIDIA 开发的深度学习推理工具，只支持推理，不支持训练。目前 TensorRT3 已经支持 Caffe、Caffe2、TensorFlow、MXNet、PyTorch 等主流深度学习库。

- TensorRT 底层针对 NVIDIA 显卡做了多方面的优化，不仅支持量化，还可以和 CUDA CODEC SDK 结合使用。

- TensorRT 独立于深度学习框架，通过解析框架文件来实现，不需要额外

安装 DL 库。

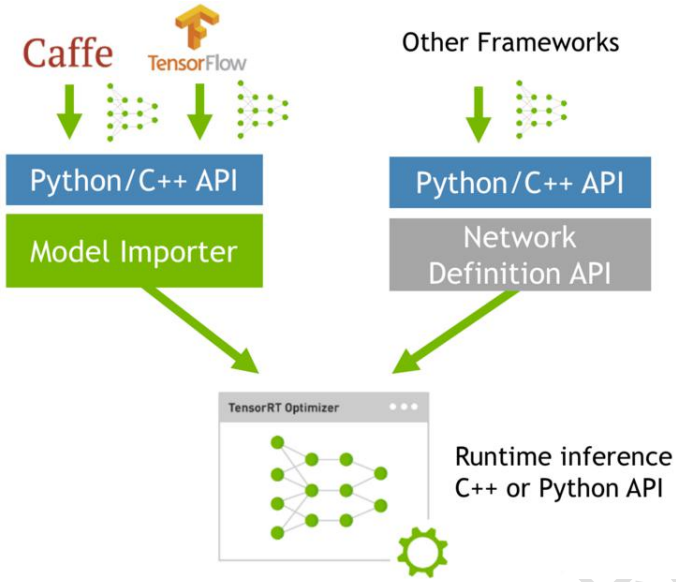

图 5　C++或 Python API 运行界面

如果模型在 ONNX 格式或其它流行框架上训练，如 TensorFlow 和 Matlab，则很容易导入模型到 TensorRT 中进行推理。链接：https://developer.nvidia.com/tensorrt

3.1.12　高级 API

为降低 AI 技术的使用难度、吸引更多的开发者，需要将 TensorFlow、Caffe/2、MXNet 等主流框架封装成高级 API，也称为 AI 前端框架。在设计上，此类高级 API 的实现方式、风格都很类似，支持不同领域的差异性。

Keras 是一个极简的、高度模块化的神经网络库，采用 Python 开发，能够运行在 TensorFlow 和 Theano 任一平台，可以在此平台上完成深度学习的快速开发。链接：https://keras.io/

Gluon 是 Amazon 开源以易用性为主的可以同时支持静态图和动态图 AI 平台，在灵活性和速度上都有优势，弥补了 MXNet 难于使用的短板。链接：https://github.com/gluon-api/gluon-api

PyTorch 由 Facebook 提出，是接近 AI 的高级 API 定义，并与 Caffe/2 进行合并。同时由于其强大的可调试性和易于开发性，以及对 caffe/2 的良好支持，PyTorch 在 AI 图像视频领域一直有很高的占有率。链接：https://pytorch.org/

Sonnet 是被 google 收购的 deepmind 团队开源、支持数据科学家基于

TensorFlow 搭建复杂的神经网络。链接：https://github.com/deepmind/sonnet

TensorLayer 来自英国帝国理工大学以华裔为主要核心人员的开源项目。从实用角度出发，TensorLayer 封装了基于 TensorFlow 的常规神经网络各部分实际功能需求（神经网络层、损失函数、数据预处理、迭代函数、实用函数、自然语言处理、强化学习、文件、可视化、激活函数、预训练模型、分布式），获得了 2017 ACM Multimedia 年度最佳开源软件奖。链接：https://github.com/tensorlayer/tensorlayer

3.1.13 开放数据集

PTB（Penn Treebank，宾州树库）是由美国宾夕法尼亚大学在上世纪 90 年代发起的 NLP 语料标注项目，主要对华尔街日报中的英文文章进行分词、词性、句法树 3 个层次的标注，是进行 NLP 研究的基础语料。其最早的对外发布版本是 1995 年的 PTB 2.0，后来不断推出新的版本，目前最新版本是 English Web Treebank Propbank（2017 年），它在原先基于网络语料 PTB 基础上添加了论元标记，便于进行语义角色标注。

CTB（Penn Chinese Treebank，宾州中文树库）是由美国宾夕法尼亚大学在 1999 年发起的中文 NLP 语料标注项目，最初是对新华社的中文新闻稿进行分词、词性、句法树 3 个层次的标注，后来又扩大到杂志、广播、博客、论坛等多种类型的语料。其标记符号和 PTB 类似，但根据中文语法做了一定程度的修改。其最早的对外发布版本是 2001 年的 CTB 2.0，后来不断推出新的版本，目前最新版本是 CTB 9.0（2017 年）。

UD（Universal Dependencies，统一依存句法）源于 NLP 领域杰出人物（Joakim Nivre、Marie de Marneffe、Filip Ginter 等）在 2013 年发起的项目 Universal Dependency Treebank (UDT)，其目的是构建跨语言的依存句法标记体系。2013 年建立 6 种语言的标记体系，2014 年初扩大到 11 种，2014 年底扩大到 30 种，目前最新版（UD2.2）已经包含 71 种语言的 122 套标记体系，以中文为例，包括 4 种标注体系：GSD、PUD、HK、CFL。在 UD 官方网站上，可以下载所有已经标记好的依存句法语料，是进行 NLP 语义分析的重要基础材料。UD 开放语料的重要性已经得到学术界认可，UD2.0 和 UD2.2 分别被作为 NLP 领域盛会 CONLL 2017

和 CONLL 2018 的指定语料，评测全球学术和产业界在 NLP 领域最新产品的性能。

WordNet 是由普林斯顿大学的心理学家、语言学家和计算机科学家联合设计的基于认知语言学的英语词典，得到美国国家科学基金和 Tim Gill 基金的赞助。它将名词、动词、形容词和副词各自按照含义构成同义词网络，每个同义词网络包含多个同义词集合，每个同义词集合代表一个基本概念，这些集合之间通过各种关系相连。常用关系包括：上下级关系、整体部分关系等。当一个词含有多个含义时，该词会出现在不同的同义词集合中。WordNet 的四个子网（名词网、动词网、形容词网、副词网）通过 cross-POS（part of Speech）指针连接语义形态相似的词汇，从而形成全网。链接：https://wordnet.princeton.edu/

ImageNet 是谷歌公司 2016 年发布的数据集，包含 900 多万张图像链接，包括训练集（约 900 万张）、验证集（约 4 万张）、测试集（约 12 万张）三部分。所有图像通过标签注释被划分为 6000 多类，每张图像都标注了图像级标签和边界框。每张图像被赋予唯一的 64 位标识码，保存在 CSV 文件中。在 CSV 文件中，每行对应一张图像，包含图像的 URL 地址、标识码、标题、作者和 License 信息。链接：http://www.image-net.org

中文数据比较流行的是清华大学提供的 30 小时语音数据库 THCHS-30。英文数据比较主流的数据库包括两个：LibriSpeech 和 AMI。LibriSpeech 是霍普斯金大学提供的包含 960 小时的安静和带噪语音。AMI 是远场数据库，AMI 大约 100 小时左右。另外还有：

- 演讲、语言公共数据集http://www.openslr.org/resources.php
- SQuAD 斯坦福问答数据集
- AWS 公用数据集 https://aws.amazon.com/cn/public-datasets/
- UC Irvine Machine Learning Repository
  http://archive.ics.uci.edu/ml/index.php
- Kaggle 竞赛数据集 https://www.kaggle.com/competitions
- KDnuggets 数据集 https://www.kdnuggets.com/datasets/index.html
- 持续更新的数据集清单
  https://github.com/awesomedata/awesome-public-datasets

语音合成方面目前公开的数据集还比较少，一个典型的公开数据集是卡内基

梅隆大学 2003 年发布的 Arctic 英文训练集，包括 3 个男性和 1 个女性发音人，每人 1150 句。数据集网址见 http://www.festvox.org/cmu_arctic。

3.1.14 分布式调度

随着复杂 AI 模型的规模不断扩大，与之相伴的是模型越来越复杂，参数量越来越大，例如: Inception v3 参数量约 25 million，ResNet 152 参数量约 60 million，VGG16 参数量约 140 million，Deep Speech 2 参数量超过 300 million，一些语言模型参数量甚至超过 1 billion。数据并行训练方式要求每个 GPU 节点拥有一份完整的模型参数副本，并在融合梯度时发送和接收完整的梯度数据，巨大的通信数据量给多机多卡并行训练带来了极大的网络通信压力。

Ray 是伯克利 RISELab 开源的高性能分布式执行框架，用于解决类似增强学习（Reinforcement Learning）领域大规模仿真需要的 Billions 级别数据资源调度，着眼于 AI 领域特定算法。针对 AI 程序员的核心诉求，它提供了灵活的高性能框架支持。Apache 顶级项目 Arrow（内存数据交换格式）是 Ray 的一个副产品，目前已经在大数据领域广泛使用。链接：https://rise.cs.berkeley.edu/projects/ray/

Angel 是由腾讯与香港科技大学、北京大学联合研发的第三代计算平台，使用 Java 和 Scala 语言开发，是一个面向机器学习的高性能分布式开源计算框架。它采用参数服务器架构，解决了上一代框架的扩展性问题，支持数据并行及模型并行的计算模式，支持十亿级别维度的模型训练。Angel 还采用了多种业界最新技术和腾讯自主研发技术，如陈旧同步并行 SSP（Stale synchronous Parallel）、异步分布式 SGD、多线程参数共享模式 HogWild、网络带宽流量调度算法、计算和网络请求流水化、参数更新索引和训练数据预处理方案等，这些技术使 Angel 性能大幅度提高，达到常见开源系统 Spark 的数倍到数十倍，能在千万到十亿级的特征维度条件下运行。在系统易用性上，Angel 提供丰富的机器学习算法库及高度抽象的编程接口、数据计算和模型划分的自动方案、参数自适应配置等。同时，用户能够像使用 MR、Spark 一样在 Angel 上编程，建设了拖拽式的一体化的开发运营门户，屏蔽底层系统细节，降低用户使用门槛。另外，Angel 还支持深度学习，它支持 Caffe、TensorFlow 和 Torch 等业界主流的深度学习框架，为其提供计算加速。链接：https://github.com/Angel-ML/angel

Horovod 是 Uber 开源的一个深度学习工具，它吸取了 Facebook 与百度 Ring Allreduce 的优点，可以帮助用户实现分布式训练。当处理数据较多时，分布式 TensorFlow 虽然具备可扩展性，但当 GPU 超过一定数量时，硬件利用率明显下降，计算处理能力与硬件规模不再呈线性关系。而 Horovod 基于 MPI 实现了 Ring Allreduce，从而解决这个问题。此外，相对于标准分布式 TensorFlow，Horovod 在模型代码实现上也提升了用户体验。在 GPU 较多时，Horovod 可以使算力的发挥提升了近一倍。链接：https://github.com/uber/horovod

3.1.15 可视化工具

可视化工具是 AI 训练中的重要工具，可以提高 AI 开发效率，主要用于展示训练过程中的统计数据（最值、均值等）变化情况和数据的分布图等。目前，TensorBoard 提供最强的可视化工具支持，其它 AI 平台通过社区贡献对接 TensorBoard 功能。百度的 VisualDL 也基本实现了 Tensorborad 的类似功能，而且兼容 PaddlePaddle、PyTorch、MXNet、Caffe2 在内的大部分主流 DNN 平台。在图像数据集管理方面，Google 开源了数据集可视化工具 Facets，可以帮助开发者洞察数据的分布情况。

3.1.16 模型市场

Acumos 项目是一个开发和共享 AI 模型和搭建 AI 工作流的平台，为 AI 领域工程化、产品化的应用场景服务。例如，如何将 AI 能力封装成接口，服务更多实时性、大并发的应用场景。其目标是在掌握大量业务数据的企业和 AI 技术公司之间搭建起一套标准交付流程。链接：https://www.acumos.org/

Algorithmia 用 App Store 的模式为“算法”量身打造了一个类似的应用商店，让开发者可以到这个商店里发布自己的算法，或者寻找并购买自己需要实现的算法。Algorithmia 作为服务协调者提供部署服务，解决数据+模型应用的“最后一公里”的难题。2017 年 Algorithmia 获得了 Google AI 基金的投资。链接：https://algorithmia.com/

MAX（Model Asset Exchange）是 IBM 为数据科学家和 AI 开发者建立的一个发布、寻找和使用免费或开源模型的一站式市场，能够支持包括 TensorFlow、

PyTorch 和 Caffee2 在内的机器学习引擎，并提供标准化的模式为这些模型进行分类、标注、部署。链接：https://developer.ibm.com/code/exchanges/models/

3.1.17 应用类项目

应用类项目种类繁多，主要包括推荐系统、智能问答、机器翻译、人脸识别、环境模拟器等。

3.1.17.1 推荐系统

SVD Feature 是由上海交通大学 Apex 实验室开发的 Feature-based 协同过滤和排序工具，采用 C++语言编写，代码质量很高。在 KDD Cup 2012 中获得第一名，在 KDD Cup 2011 中获得第三名，相关论文发表在 2012 的 JMLR 中。它包含一个很灵活的 Matrix Factorization（矩阵分解）推荐框架，能方便地实现 SVD、SVD++等方法, 是单模型推荐算法中精度最高的一种。SVDFeature 代码精炼，可以用相对较少的内存实现较大规模的单机版矩阵分解运算。此外，它还含有 Logistic regression 的 model，可以很方便的用来进行 ensemble learning（集成学习）。链接：http://apex.sjtu.edu.cn/projects/33

EasyRec 是由奥地利国家研究中心（Research Studios Austria Forschungsgesellschaft mbH）开源的推荐系统，具有易集成、易扩展、功能强大、可视化等诸多特性。架设 EasyRec 服务器后，通过申请 tenant 就能方便集成。EasyRec 使用不同的数据收集 API 收集网站的用户行为，通过离线分析就能产生推荐信息。EasyRec 自下而上包括三层：持久层、业务层和展现层。横向采用模块化集成方案，包括数据录入模块、管理模块、推荐模块、离线分析模块、特定领域模块等。链接：http://easyrec.org/

3.1.17.2 智能问答

AIML 的全称是 Artificial Intelligence Markup Language（人工智能标记语言），是一种创建自然语言软件代理的 XML 语言，由 Richard Wallace 和 Alicebot 开源软件组织在 1995-2002 年间发明创造。AIML 是利用 XML 标准定义的一种服务于

人工智能领域需要的特定语言，它描述了被称为 AIML 对象的一组数据对象，并且描述了处理这些数据对象的程序的行为。AIML 的雏形是名称为“ALICE”的聊天机器人，它总共赢得 3 次每年度的 Loebner 奖，并且在 2004 年获得 Chatterbox Challenge 的冠军。设计 AIML 的最初意图就是为了能够用最简单的方式来创建人工智能聊天机器人，而且在语法上接近常用的 HTML 语法。AIML 定义了一套具有特殊含义的标记，使得以 AIML 为核心的聊天系统具有强大的功能，此外，用户也可以根据需要定义各种新标记来扩展系统的功能，因此 AIML 具有很好的扩展性。

DrQA 是 FaceBook 在 2017 年 7 月开源的开放域问答系统，对应文章发表在 ACL 2017。DrQA 是基于阅读理解的 Open QA 系统，执行过程包括先后两部分：Retriver 和 Reader；Retriever 根据问题在维基百科语料库中检索出最相关的 5 篇候选文章，核心算法是二元语法哈希（Bigram Hashing）和 TF-IDF；Reader 从候选的 5 篇文章中提取出问题答案，核心算法是 RNN 编码预测。DrQA 专注于回答事实性问题，其数据来源仅有维基百科，属于单一信源的问答系统，相关证据可能仅有一处，不能像 IBM DeepQA 使用知识库、词典、新闻、书籍等多种来源的信息冗余获得正确答案，所以其搜索算法的精度很高。

3.1.17.3 语音识别/机器翻译

Kaldi 是一个非常强大的语音识别工具库，主要由 Daniel Povey 开发和维护。目前支持 GMM-HMM、SGMM-HMM、DNN-HMM 等多种语音识别的模型的训练和预测。其中 DNN-HMM 中的神经网络还可以由配置文件自定义，DNN、CNN、TDNN、LSTM 以及 Bidirectional-LSTM 等神经网络结构均可支持。链接：http://kaldi-asr.org/

Sockeye 是亚马逊在 2017 年 7 月开源的基于 Apache MXNet 的机器翻译框架，使用 Python 实现，其核心是神经序列编码器-解码器模型。它自身包含三大主流神经翻译架构：注意力循环神经网络（Attention RNN）、自注意力变换器（Self-attentional Transformer）、全卷积网络（Fully CNN），可以进行训练和扩展。此外，Sockeye 还支持多种优化器以及归一化、正则化技术，并利用当前的 NMT 文献提升了推断能力。用户可以很轻松地运行标准的训练流程，探索不同

的模型设置，验证新的想法。链接：https://github.com/awslabs/sockeye

FairSeq 是 FaceBook 在 2017 年 5 月开源的机器翻译项目，它通过多跳注意和精准门控等策略改进 CNN，实现了卷积序列到序列学习中的完全卷积模型，支持在单个机器上使用多 GPU 进行训练，其运行速度是基于循环神经网络(RNN)系统的 9 倍（谷歌的机器翻译系统使用的就是这一技术）。链接：https://github.com/facebookresearch/fairseq

3.1.17.4 人脸识别

DeepFace 是由 FaceBook 在 2014 年开源的人脸识别项目，使用 Python 研发，对应论文发表在 CVPR 2014，它是采用深度学习技术进行人脸识别的奠基之作，其技术在后继的人脸识别项目中均有体现。DeepFace 的人脸识别过程是“检测->对齐->表示->分类”，2014 年在 LFW（人脸比对数据集）上的正确率达到 97.25%，逼近人类自身的人脸识别正确率 97.5%。

此外，Facebook 还开源了三款人工智能图像分割软件，将能够识别一种图片中的人物及物体等，并判断出它们在图像中的具体位置。链接：https://github.com/pytorch/fairseq

FaceNet 是由谷歌在 2015 年开源的基于 TensorFlow 的人脸识别项目，由 Python 编写，对应论文发表在 CVPR 2015。FaceNet 的原理是直接将人脸图像映射到欧几里德空间，图像在欧氏空间的距离代表其相似性。在算法上，其特色有两点：舍弃了传统的 SoftMax 学习算法，通过三元组（样本、最远正例、最近反例）构建目标函数进行优化，从而提高了训练收敛速度。此外，FaceNet 采用两种 CNN 网络（Zeiler&Fergus 架构、GoogleNet 架构）提取图像特征向量。2015 年在 LFW 上的正确率达到 99.63%，在 YouTube Face 上的正确率达到 95.12%。链接：https://arxiv.org/abs/1503.03832

SeetaFace 是由中科院计算所山世光研究组在 2016 年开源的人脸识别项目，代码基于 C++实现，不依赖第三方库。该引擎包括搭建一套全自动人脸识别系统所需的三个核心模块：人脸检测模块 SeetaFace Detection（结合传统人造特征和多层感知机检测人脸）、面部特征点定位模块 SeetaFace Alignment（级联多个深度模型来回归人脸 5 个关键特征点的位置）以及人脸特征提取与比对模块

SeetaFace Identification（采用 9 层 CNN 提取人脸特征）。2016 年在 LFW 上的正确率达到了 97.1%，处理速度是每图 120 毫秒。和 FaceBook、谷歌等公司开源的人脸识别项目相比，其性能未必最好，但在速度和精度方面基本达到可用程度，更重要的是，该软件是国内科研领域为数不多的优秀开源软件，在开源方面具有示范意义。链接：https://github.com/seetaface/SeetaFaceEngine

3.1.17.5 其它应用类开源项目

Detectron 是 Facebook AI 研究院（简称 FAIR）开源的物体图像识别平台，该项目自 2016 年 7 月启动，遵循 Apache 2.0 开源许可证。

Pattern 由安特卫普大学计算语言学与心理语言学（简称 CLiPS）研究中心所打造，提供多种人工智能功能，具体包括：数据挖掘、自然语言处理、机器学习、网络分析以及可视化。它基于 Python，且提供 50 多个示例与 350 多项单元测试。GitHub 用户为其给出了超过 6000 个星评，fork 次数亦超过 1100 次。链接：https://github.com/facebookresearch/Detectron

Aerosolve 是由爱彼迎创建的一款人工智能工具，特别擅长处理地理数据。因为它拥有“人性化”的设计思路，所以声名大噪。其主要功能包括：基于资源节约理念的特征表达、特征转换语言、可调试模型、Java 与 Scala 支持以及图像内容分析代码。在 GitHub 页面中，该项目拥有超过 4200 个星评、550 次 fork。链接：http://airbnb.io/aerosolve/

DSSTNE 是由 Amazon 公司打造的推荐引擎原型，代表着“深层可扩展稀疏张量网络引擎”。网络购物巨头利用它来创建自己的推荐引擎，但其出色的能力也引起了其它零售商与在线企业的关注。根据开发人员所言，其在机器学习训练数据稀缺类用例当中发挥着巨大的作用。在 GitHub 上，该项目拥有 4000 多个星评、超过 660 次 fork。链接：https://github.com/amzn/amazon-dsstne

CaffeOnSpark 最初由雅虎开发而成，可以将 Caffe 深度学习框架引入到 Hadoop 和 Spark 集群，主要用于图像搜索和内容分类以及其它使用场合。链接：https://github.com/yahoo/CaffeOnSpark

DeepDetect 是基于 Caffe、TensorFlow 和 XGBoost 的开源深度学习服务器系统，被空中客车和微软之类的企业组织所使用，它为图像分类、对象检测、文本

及数字数据分析提供了一套易于使用的 API。链接：https://deepdetect.com/

DeepMind Lab 是由谷歌 DeepMind 部门开发的 3D 游戏环境，适用于深度强化学习研究。链接：https://deepmind.com/research/publications/deepmind-lab/

《星际争霸 II》API 库是由谷歌的 DeepMind 和暴雪娱乐公司在共同开展一个项目，可使用《星际争霸 II》视频游戏作为 AI 研究平台。它是一种跨平台的 C++库，可用于构建脚本化的机器人程序。链接：https://us.battle.net/forums/en/sc2/topic/20758616786

Numenta 组织提供了与层级实时记忆（HTM）有关的众多开源项目。实际上，这些项目试图基于生物界对于人类新大脑皮层的了解来生成机器智能。链接：http://numenta.org/

Open Cog 是旨在生成有益的人工通用智能（AGI）的项目，而不是专注于狭义的 AI（如深度学习或神经网络）。该项目正致力于打造能够拥有类人智能的系统和机器人。链接：http://opencog.org/。

3.2 开源组织

开源基金会组织也进入 AI 领域，参与 AI 开源社区的治理，为开源项目的长期、健康、持久运行提供技术支持和机制保障，比如 Linux 基金会、OpenStack 基金会和 Apache 基金会。本节就一些比较著名的国内国际开源组织进行介绍。

3.2.1 开源中国

开源中国(OSCHINA.NET)成立于 2008 年 8 月，目前已建立了一个全球内容最完善的开源软件分类数据库，收录全球知名开源项目近 5 万款，涉及几百个不同的分类。同时围绕这些开源项目为中国开发者提供最新开源资讯、软件更新资讯、技术分享和交流的技术平台。经过在开源领域超过十年的深耕，以及与中国本土开源环境的结合，推动了中国开源领域快速发展。开源中国已发展成为目前国内最大的开源技术社区，长期致力于推动国内开源软件的应用和发展，提升本土开源能力，以及为开源生态环境的优化提供支持，目前开源中国社区已有近 500 万的开发者，网站全球排名 700。

2011 年，开源中国举办了首场名为源创会的线下开源技术交流活动，秉承着

“开源和创新”的理念，鼓励开放和自由。到目前为止，开源中国联手国内 IT 公司、开源作者、开源组织和其他开源社区，在全国各地举办了将近 100 场源创会，总参会人数超过 5 万人。源创会不止于分享技术，还努力传播开源精神和理念。此外还会举办年度开源技术盛会【源创会年终盛典】，评选年度开源项目等活动，极大的活跃了国内开源交流的氛围。源创会仍然是目前国内历史最悠久、规模最大、覆盖全国最多城市的技术交流会议。

2013 年，开源中国推出码云代码托管平台(https://gitee.com/)，成为国内首个自己的开源项目托管平台。码云为中国广大开发者提供代码托管、项目管理和文档管理的平台。目前已托管国内项目超过 500 万个，其中开源项目超过 150 万个。码云已发展成为全球第二大代码托管平台。经过多年的不断打磨，码云平台不仅在功能、性能、易用程度方面都得到了全方面的发展，并且成为了国内超过 6 万家企业的选择，在码云上开展项目管理和开发工作。在推动开源的发展上，码云推出了 GVP 精选开源项目栏目，为广大的开发者推荐值得信赖的开源项目。同时，码云还专门为高校计算机老师打造了“高校版”， 目前已经成为国内高校老师进行软件工程教学的首选平台，有近 1000 个高校的老师通过码云高校版提升软件工程教学的过程。真正为助力计算机专业教学改革与「新工科」实践落地贡献自己的力量。

2018 年，码云推出了开源项目抄袭检测工具 copycat.gitee.com ，开源软件作者可以通过使用该工具来检查目前国内存在比较严重的项目抄袭现象；推出了“项目指数” 工具，对开源项目各项指标进行量化，对开源项目的长期、全方位发展提供了很好的视角和完善建议。

在未来发展方向上，开源中国将不遗余力的继续通过社区、工具的强强结合，不断的推动国内开源的发展，推动社会化协作开发，从而提升整个中国的软件产业的水平。

3.2.2 开源社

开源社（http://www.kaiyuanshe.cn/）于 2014 年成立，是由国内外支持开源的个人、社区及企业，依“贡献、共识、共治”原则所组成的非营利开源联盟，旨在共创健康可持续发展的开源生态体系，并推动中国开源社区与项目成为全球

开源软件的积极参与及贡献者。

开源社是国内第一个专注于开源治理、社区发展、国际接轨，以及开源项目的开源组织，完全由志愿贡献于开源事业的个人会员组成。开源社与支持开源的社区、企业以及政府相关单位紧密合作，但始终不忘初心，维持着厂商中立，公益非营利，推广并贡献开源的使命与愿景。

2017 年，开源社转型成为由纯粹个人成员（类似 GNOME Foundation，Apache Software Foundation）的治理模式组成，并于 2017 年底由全体正式个人成员选举出 7 名理事，组成理事会及执行委员会（下设 9 个工作组：成员发展、基础设施、财务、法律事务、媒体、文案/翻译/设计、线下活动、社区合作、高校合作等），由企业及社区开源专家组成的顾问委员会，以及法律咨询委员会。

在开源治理方面，开源社与欧洲最知名的开源治理社区 - OSS Watch 合作，在开源社官网提供了国内最完整的开源治理文档与知识库、开源许可证介绍、开源流程指导、以及开源许可证选择器自动化工具等。同时也成立了法律顾问委员会，为社区免费提供了开源治理方面的咨询服务。2015、2016 及 2018 年，开源社陆续发布了中国开源生态系统年度报告，以及开源社区参与调查报告等，为描绘中国国内开源的发展图像尽一份心力。2016 年 1 月，开源社作为中国首家开源组织加入 Open Source Initiative (OSI) 成为联盟成员，为推动开源治理与合规的开源软件许可证与 OSI 携手合作。

在社区发展方面，开源社举办了几十场的高校系列巡回宣讲 -【开源者行】，用自己搭建的【开放黑客松云平台】举办了 20 多场线上与线下的编程马拉松。开源社于 2015 年 10 月在 Apache Software Foundation 的支持下主办了【2015 阿帕奇中国路演】，2016 年 10 月联合国内多家开源社区/企业/联盟共同筹办首届【2016 中国开源年会-COSCon'16】，2017 年 11 月举办了国内首次以开源社区运营以及项目贡献为主题的盛会 -【2017 中国开源年会-COSCon’17】，2018 年 10 月举办了包括开源硬件与开源教育论坛的【2018 中国开源年会-COSCon’18】。

在国际接轨方面，开源社秉持着【带进来，走出去】的理念，与众多国际顶级社区，如 Apache Software Foundation (ASF), AllSeen Alliance, FreeBSD Foundation, FOSS.Asia, GNOME Foundation, Linux Foundation, Node.JS Foundation, Open Source Initiative (OSI), Open Innovation Networks 等，持续地在国内外密切交

流与合作，同时为国内诸多高质量的开源项目进入国际顶级基金会如 Apache Software Foundation 孵化器搭桥铺路。

在开源项目方面，目前开源社结合社区贡献者开发了两个拳头项目，如 KCoin 项目（开源贡献激励平台）与开放黑客松云平台（原微软公司开源项目，于 2018 年正式捐赠给开源社，是第一个由国际顶级企业捐赠给中国开源社区的项目），并期盼未来有更多更好的开源项目以及贡献者加入开源社。

3.2.3 OpenI 启智开源开放平台

OpenI 启智平台是新一代人工智能产业技术创新战略联盟（AITISA）组织产学研用通力协作共建共享的开源软件开源硬件开放数据超级社区，肩负“新一代人工智能开源开放平台”的使命与梦想，英文名称 OpenIntelligence，简称 OpenI。平台旨在促进人工智能领域的开源开放协同创新，构建 OpenI 的技术链、创新链和生态链、推动人工智能产业健康快速发展及其在社会经济各领域的广泛应用。

OpenI 启智平台初始阶段就发布了代码开源许可证 OIL1.0，前期主要参与发起单位有：鹏城实验室、北京智源人工智能研究院，华为、国防科技大学、百度、北京大学、阿里、北京航空航天大学、腾讯、讯飞、商汤、滴滴、美团、小米、字节跳动、微软、寒武纪、Intel、NVIDIA、中科院微电子所、中国科技大学、清华大学等。近期有突出贡献能力的核心成员鹏城实验室、华为、百度、国防科大率先联合共建与贡献，发布 OpenI 启智平台及 AI 框架，包括：“OpenI 章鱼”智能资源管理系统，“OpenI 珊瑚”异构资源集群调度系统，“启智 Trustie”群体化协同创新环境，“启智 VisualDL”深度学习框架核心项目；相关代码已经通过内部知识产权审核签署贡献者许可协议 CLA，进入 OpenI 立项管道与社区步道，对外正式开放。OpenI 启智平台基础设施及支撑环境共建工程正在鹏城实验室 AI 中心大楼及港辖河套“人工智能国际研发中心”进行，包括世界级的 AI 超算“云脑”系统，OpenI 启智深圳基地，OpenI 北京智源社区，新一代人工智能产业创新联盟“启智空间”等。未来分布全国的新一代人工智能重大基础设施正在规划建设中。

OpenI 核心成员北京智源研究院旷视智能模型设计与图像感知联合实验室(孙剑博士团队)，新发布并开放世界最大物体检测数据集 Object365，积极贡献自

动训练算法及 Brain++框架部分功能；成员微众银行(杨强教授团队)贡献“OpenI 纵横”联邦数据学习系统；数据堂计划贡献大量语音开放数据集；相关项目签署 CLA 后即将进入立项管道，加入社区步道。同时，OpenI 组建了国内及国际顶级开源与互联网法律及知识产权专家、道德伦理学者团队，开始共同制定面向开放数据集的下一代启智许可证 OIL2.0、以及新一代人工智能科技产权与治理策略；OpenI 将在此基础上联合广大数据科技业者共同构建 OpenI 生态环境开放数据社区。

被誉为 AI 黄埔军校的微软研究院作为 OpenI 最早的发起参与成员之一，正与中国最大开发者社区 CSDN 合作，基于鹏城云脑发起 AI 教育项目，精心设计 AI 教育体系与课程体系。为此 OpenI 发起“启智学院”计划及建立人工智能人才培养基地的倡议，在国防科大贡献的“启智 Trustie”群体化协同创新环境以及 AI 人才实训平台基础上，联合产教学研各界力量共同打造 OpenI 学习社区和开发者社区。

3.2.4 Linux 基金会

Linux 基金会是非盈利组织，成立于 2000 年。基金会源于围绕 Linux Kernel 的开源推动组织 Open Source Development Lab （OSDL）和 Free Software Foundation （FSF）。该组织的核心目标就是推动 Linux 系统的发展、保护其成员和社区资源。为了确保 Linux 的开放性和技术领先，基金会邀请了 Linus Torvalds 等重要角色参与项目的开发管理，和来自世界各地的开发人员开展合作。截止 2017 年底，1400 多家公司约 15,600 名开发者以及大量个人开发者贡献了内核代码。同时，基金会还推出了 Linux 使用、系统管理、虚拟化技术等在线培训课程和认证计划来普及 Linux 技术。此外，Linux 基金会还有成熟的社区合作方式、会员制度和技术会议组织经验。

随着开源项目的不断发展，Linux 基金会开始致力于围绕更多开源项目构建可持续的生态系统，加速技术开发和商业落地。先后托管了众多知名的合作项目，比如 OpenDayLight、OPNFV、CNCF、openSDS、Meego、Tizen、Hyperledger、ACRN 等。2018 年 3 月，Linux 基金会宣布成立 LFDL（Linux Foundation Deep Learning）及其第一个人工智能相关的项目 Acumos AI。目前比较重量级的项目包括百度贡

献的 EDL 项目，腾讯贡献的 Angel 项目、Uber 贡献的 Horovod 等。目前，已有 10 多家企业参与，包括华为、百度、腾讯、中兴等中国公司。Linux 基金会和这些合作项目的关系可以总结为：

（1）Linux 基金会为项目提供运作平台，确保一些通用的社区基础治理工作可以供多个项目共享，如法律、秘书处、章程、会议主办。

（2）Linux 基金会帮助合作项目成立，但是对合作项目没有管辖权。基金会不干涉项目的日常运行，每个项目由项目维护者自行管理。

（3）每个合作项目都可以根据自己的情况邀请各自的会员。合作项目的会员仅在合作项目里有效，一个公司能否成为基金会的会员，同合作项目的会员关系或会员等级没有直接联系。

（4）Linux 基金会和这些项目的合作规则确保了各个项目能够高效发展，同时借助 Linux 平台可以得到更好推广。现在，一个开源项目能否成为 Linux 基金会的合作项目已经成为了它“是否代表行业方向、是否真正开放治理、是否高质量可商用落地”的衡量标准。

Linux 基金会采用的是企业会员+个人会员制。

企业会员分为：银级、金级、白金级三个等级，各级会员需承担的责任有所差别。白金级是最高等级，对成员的要求最高，白金会员同时拥有董事会席位，每年需缴纳会费 50 万美元。目前成员包括 AT&T、思科、富士通、谷歌、日立、华为、IBM、英特尔、NEC、甲骨文、高通、三星、腾讯、VMWare 和微软。

个人会员没有太多限制，任何开源软件专业人员、开发人员、系统管理员和学生都可以以个人名义加入。基金会会员有权利在基金会下提交议案创建并管理新项目，也可以参与基金会重要会议开源策略的讨论，另外可以享受基金会提供的各种优惠资源和服务。

3.2.5 OpenStack 基金会

OpenStack 基金会是一家于 2012 年成立的非盈利组织，旨在推动 OpenStack 云操作系统在全球的发展、传播和使用。OpenStack 基金会的目标是在全球范围内服务开发者、用户及整个生态系统，为其提供共享资源，以扩大 OpenStack 公有云和私有云的成长，从而帮助技术厂商选择平台，助力开发者开发出行业最佳

的云软件。

OpenStack 目前已经成为仅次于 Linux 的第二大活跃开源社区，也是全球成长最快的开源组织之一。OpenStack 拥有来自 176 个国家的 31894 名成员，得到了 555 家公司的支持，已经拥有 94 项产品及服务。

基金会分为企业会员和个人会员两大类。企业会员根据各公司赞助会费的情况，分成白金会员、黄金会员、企业赞助会员以及支持组织者，其中白金和黄金会员的话语权最大。目前，中国已经成为 OpenStack 市场增长最快的区域。截止 2017 年，华为已成为唯一一家来自中国的白金会员公司，且在 24 个黄金会员席位中中国公司占据了半壁江山（包括两家台湾公司），中国正逐步成为 OpenStack 的主角。个人会员是免费无门槛的，他们可凭借技术贡献或社区建设工作等参与到 OpenStack 社区中。

根据 stackalytics.com 网站提供的社区贡献统计，截止 2018 年 3 月 6 日，这也是 OpenStack 自诞生以来公布的第 17 个版本-Queens 代码贡献中，共有 200 多家企业和组织上榜，这其中包括 Redhat、IBM、Intel、Rackspace、SUSE 等，这些老牌企业依旧在全球贡献处于领先地位。中国以华为、九州云、中兴、烽火、麒麟云、海云捷迅、易捷思达为代表的几十家企业也积极参与 Queens 代码贡献，成为全球的 OpenStack 技术的中坚力量。

3.2.6 Apache 基金会

Apache 软件基金会（Apache Software Foundation，ASF）正式创建于 1999 年 7 月，是专门为支持开源软件项目而办的一个非盈利组织。

ASF 采用会员制，以确保 Apache 的项目可以在没有个人志愿者参与的情况下依然能够持续存在。独立个体若要加入 Apache，需要证明自己能够在开源软件的开发中通力合作，并通过在基金会的项目中持续地参与和贡献。目前共有 8 位白金赞助商、9 位金牌赞助商、8 位银牌赞助商和 14 位铜牌赞助商，成员总数达为 731 名，拥有超过 6700 位代码提交者。基金会的白金会员包括 Google、Microsoft、Facebook 等。2018 年 9 月，腾讯成为基金会白金会员，这也是中国首家成为 AFS 白金会员的公司。

AFS 目前为超过 350 个开源项目提供支持，涵盖人工智能和深度学习、大

数据、构建管理、云计算、内容管理、DevOps、物联网和边缘计算、移动、服务器和 Web 框架等众多领域。

在 Apache 支持的各种项目中，Mahout 是一个开源机器学习框架，具有构建可扩展算法的编程环境。机器学习服务器 PredictionIO 可以帮助开发人员和数据科学家为机器学习任务创建预测引擎和服务。阿里巴巴集团向 Apache 捐赠开源实时计算系统 JStorm，并在 Apache Storm 里孵化；华为发起并捐献的 Apache® CarbonData™项目已正式成为 Apache 顶级项目，用于大数据高效存储格式解决方案；网易研究院和新加坡国立大学发布的开源分布式深度学习平台 Apache SINGA 是 ASF 资助的第一个深度学习项目，可通过不同的运算符（神经网络层）构建深度学习模型，广泛应用于科研、医疗、金融等领域；Apache 顶级项目 System ML 由 IBM 发起和捐赠，该工具可帮助电脑从海量数据中找到相同的形态，用于预测搜索关键词、辨认人脸、探测股价异常波动等任务。同时亚马逊联合华盛顿大学构建的深度学习框架 MXNET 也是 Apache 基金会中的顶级项目。

3.3 组织/机构参与开源的角色及目的

开源的价值是多方面的，不同组织/机构投入到开源中来的价值取向和价值获得也是多样的，主要包括六个方面：

（1）开放选择：开源减少了厂商和特定实现技术的锁定，组织/机构今天做出的决定，不会限制其未来的选择。

（2）灵活便捷：组织/机构无论对内对外，不论其技术选择如何，都能容易地实现互联互通。

（3）敏捷开发：开源技术依托来源广泛的开源社区，通常采用更灵活的开发、测试、集成方法。

（4）快速适应变化：外部环境、可用资源、适用场景的变化，容易反馈至多社区驱动的项目开发，使得其具备新能力，快速迭代以适应新形势。

（5）技能普及：由于开源社区的普遍性、广泛性和跨地域特性，所需技能不局限于一地，容易获得。

（6）公平公开：开源使得生态系统的利益相关方能够处在一个更加公开公正、平等互利的环境中。

对于行业引领者，他们往往要从战略的角度，更多地考虑生态系统的建设，使其更具粘性，更能吸引技术、应用和市场的开发者、维护者、宣传者以及最终使用者。这样的引领者，往往能在技术萌芽早期发现其价值（无论这个技术来自组织内部还是组织外部），并战略性地将其推向开源（尤其是基金会制度下的开源社区）。这确保了相关战略的长期性和持久性，并快速提升影响力，成为社会主流，从而战略性保障相关领域不被其他利益相关方以技术、市场或其它手段垄断。比如，在电子商务刚刚兴起的时候，IBM 和若干企业、非赢利组织将一款轻便的 HTTP 服务器系统推向开源，并成立 Apache 基金会进行管理。这一举措使得 HTTP 服务器这种电子商务基础技术不被垄断势力把持、分割，从而对电子商务的繁荣提供了重要保障。针对稳定市场，开源往往也能起到不同凡响的鲶鱼效应，例如 Mozilla 浏览器和 Eclipse 集成开发环境。

对于行业的跟随者，他们往往需要从自身业务的角度，选择一个更有前途的技术确保不被淘汰，开源项目的存在就给了他们一个更好的选择。他们可以根据业务特点选择一个合适的项目，并且通过社区投入的方式获取一定的话语权。随着跟随者数量的增加，生态就变得更为繁荣，跟随者的投入就更有价值。比如，当 Linux 出现后，越来越多公司选择基于 Linux 开发中间件和应用，随着生态的爆发式增长，Linux 上的应用被得到广泛认可。尤其在云领域，Linux 以及上面的应用已经成为主流。

对于行业的用户，除了质量和成本，还需要考虑会不会被供应商绑定。开源确保了产品的代码公开透明，大大降低了被绑定的风险。

综上，组织/机构依据自身战略需求，可以以不同形式参与到开源社区中，包括开源社区的资金赞助商，技术开发、维护的代码贡献者，技术环境（CI、CD）的运维者，技术应用的增值开发者，以及社区的布道者、培训者、营销者、出版者等。

第四章 AI 开源技术目前在落地中存在的问题与差距

AI 开源技术打破了建立专利技术公司的壁垒，大大提高了 AI 技术的迭代速度，同时也由于 AI 开源技术产出者对应用环境背景不了解，导致开源代码的鲁

棒性、稳定性和可拓展性往往不能达到 AI 开源技术消费者使用需求。AI 开源技术可以用于早期的算法验证，真正能够用于大规模推广产品仍然需要进行大数据技术拓展以及大规模测试和基于应用场景的先验知识。

4.1 AI 在应用时的总体工作流

4.1.1概述

人工智能主要是通过机器来模拟人类认知能力的技术，最核心的能力就是根据给定的输入作出判断或预测。

在二十世纪八十年代一度兴起的专家系统就是基于人工定义的规则来回答特定问题，基于专家知识数据库来做出判断。但专家系统的人工定义规则的方式有着很多的局限性：一方面，在复杂的应用场景下建立完备的规则系统往往是一件非常昂贵而耗时的过程；另一方面，很多基于自然输入的应用，比如语音和图像的识别，很难以人工的方式定义具体的规则。当实际应用涉及现有知识数据库不存在的新知识，需要人工获取或者定义新的知识。

近年来，随着新型算法、大数据以及高性能计算硬件等技术的发展，基于深度学习的人工智能取得了重大突破，以多层神经网络为基础的深度学习被推广到多个应用领域。

下图是 80 年代传统专家系统和近年基于深度学习的人工智能系统工作流对比示意图。

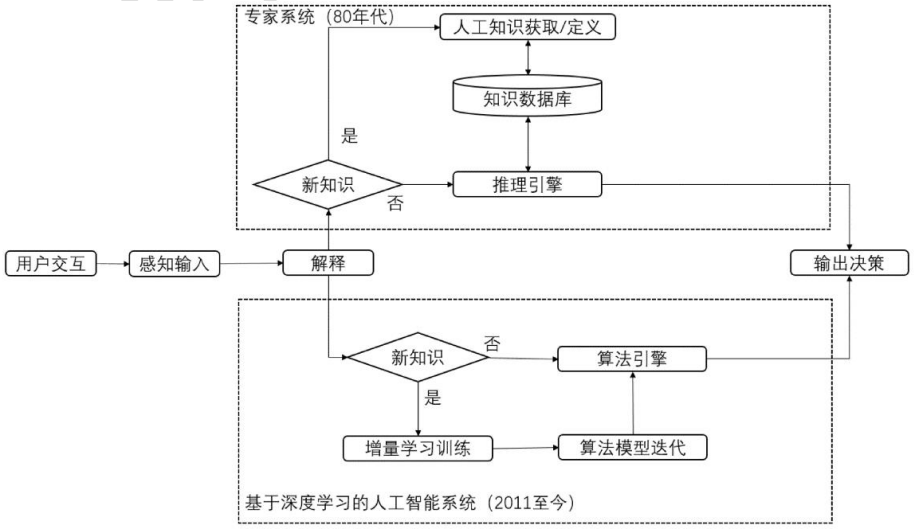

图 6 基于深度学习的人工智能系统工作流对比示意图

当前的人工智能普遍通过基于深度学习的机器学习来获得进行预测和判断的能力。机器学习分为根据数据学习的监督学习/半监督学习/无监督学习，以及根据行动结果形成激励反馈的强化学习。这样如果在得到足够的算法、算力和数据支撑下，系统可以在无人工或者少量人工干预的情况下完成迭代，智能地处理新知识所对应的业务。

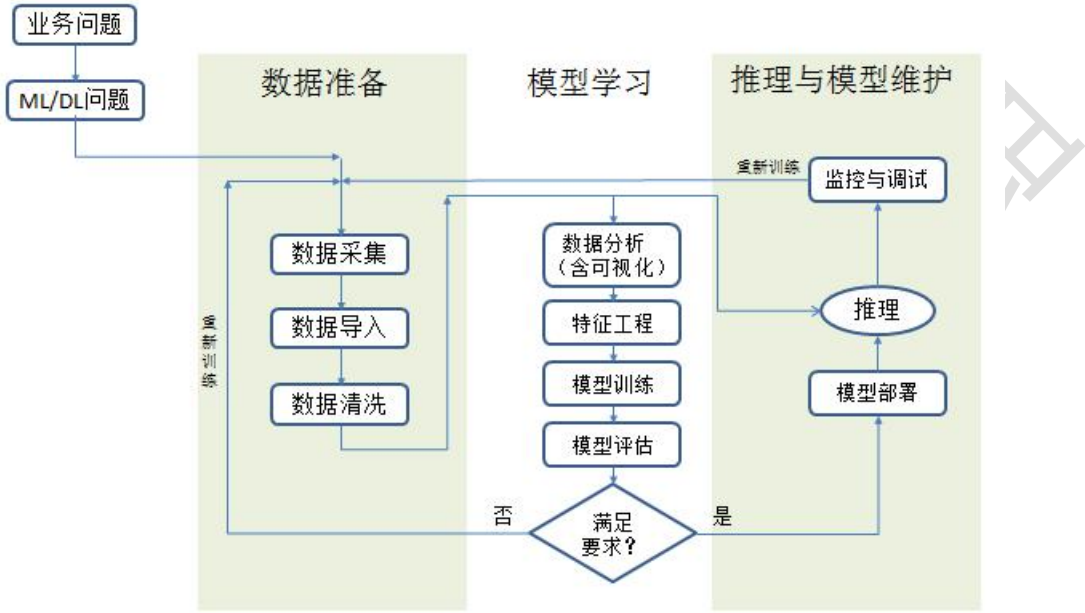

图 7 基于机器学习及深度学习的 AI 工作流

基于机器学习及深度学习的 AI 工作流主要包含数据准备、数据特征、模型训练、模型评估和优化以及模型应用和部署。

（1）数据准备

数据准备阶段包含数据管理、数据导入、数据加工、数据选择等步骤，这是 AI 工作流的初始阶段，该阶段侧重理解项目目标和需求，理解数据的分布与产生，将原始数据进行各种变换与格式转换，归一化后存储到不同类型的数据库或者表中。输出或处理后的数据可提供以图像、图形、表格、矢量文件、音频、图表或任何其他所需格式。

（2）数据特征

数据特征包含特征工程、数据探索、概率统计等步骤。特征工程是从现有数据中提取有用信息或特征的过程，是机器学习的重要组成部分；数据探索是通过统计和可视化技术来描述数据，以便从收集的信息中形成真实的分析；概率统计是提取特征的一种方法，通过概率统计的方式找出数据隐藏的规律，由这些规律

建立新的特征，根据特征对已知数据进行归类，对未知类别的数据进行预测。

（3）模型训练

模型训练指将准备好的数据按准备好的数据特征建立模型。数据建模是一种将结构化数据进行定义和组织的过程，并对结构内的数据施加限制和约束。数据模型是基于业务约定的，按照数据模型的业务约定可将数据转换成计算机最终能够理解并能由此建立规律的数据。

（4）模型评估和优化

AI 的主要目标是精确预测，可以通过修改访问数据和模型来改进和优化模型，要求模型使用的训练数据规则，可以良好地推广到新的数据。模型评估主要分为离线评估和在线评估两个阶段，不同的算法评价指标也不同。我们可以使用各个优化方法优化模型，在数据没有改变和新增的情况下可以通过改变模型参数，更换模型底层算法，模型融合等方式来优化模型。

（5）模型应用和部署

模型的创建不是 AI 的最终目的，建模是为了增加更多有关数据的规律，但这些规律需要以一种客户能够使用的方式组织和呈现。在许多情况下，往往是客户而不是数据分析师来执行部署运用阶段，尽管数据分析师不需要处理部署运用阶段的细节工作，但是预先了解需要执行的任务从而正确使用已构建的模型是非常重要的。

算法、算力、数据是支撑基于机器学习、深度学习的人工智能系统的基础，所以对应人工智能系统工作流需要三大模块：业务系统、数据系统、深度学习训练系统。业务系统（包括算法引擎）是人工智能业务的用户接口，是整个总体工作流的起点和终点，应该包括数据采集入口、部署的算法模型以及最终的人工智能生产业务。数据系统是对数据进行处理，包括导入、清洗、标注等等。深度学习训练系统利用算力，进行模型训练和测试。

三个子系统互相支持形成业务闭环，如下图所示，系统在少量或者无人工干预的情况下完成对新知识的学习，自动匹配未来快速增长和转型的新业务。

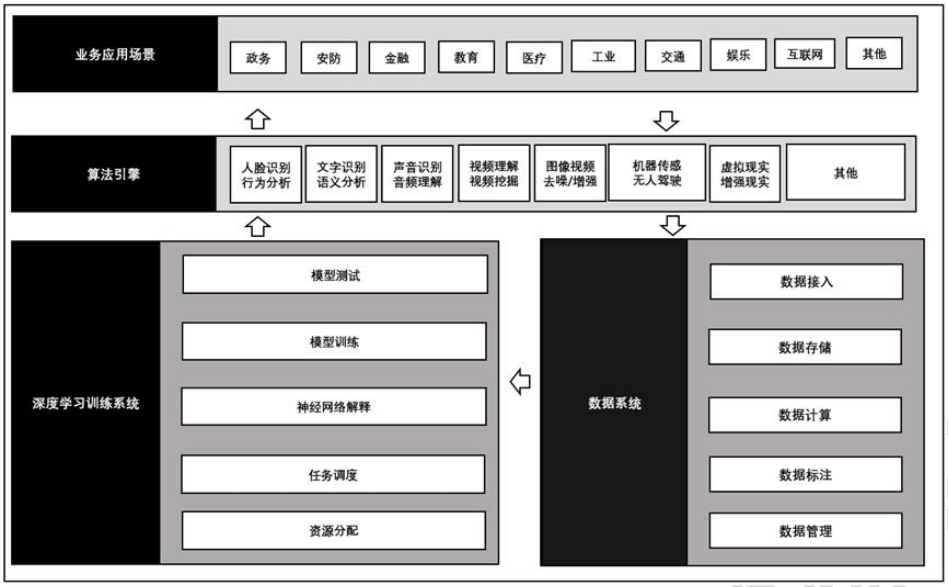

图 8 三个子系统

4.1.2 经过抽象的工作流实现

根据上述概述，人工智能系统的工作流实现需要经过以下子系统：算法引擎、数据系统、深度学习训练系统。主要工作步骤如下：

（1）根据业务应用感知数据等输入，并将其输入到算法引擎进行判断和推理；

（2）如果是新建系统，或者新类型数据，则优先对接数据系统和深度学习系统，进行新模型训练和部署；

（3）如果是已有类型数据，则优先根据现有算法引擎里面的模型进行推测；

（4）根据算法模型和输入数据推理出的机器预测或决策返回给上层业务。

4.1.2.1 业务系统及算法引擎工作流

算法引擎属于人工智能业务系统的一部分，部署核心算法模型，主要功能是对感知的数据进行判断和推理。根据不同业务，还需要对接具体不同的硬件设备系统，比如图像/视频识别需要 GPU 进行视觉数据处理。该算法模型类别可包括但不限于以下范围：人脸识别及行为分析、文字识别及语义分析、声音识别及音频理解、视频理解及视频挖掘、图像视频去噪及增强、机器传感及无人驾驶、虚拟现实/增强现实等。

业务系统及算法引擎主要工作流如下：

（1）作为整个生态系统的初始入口，感知原始数据，并接入到数据系统进行存储、计算和管理；

（2）判断数据是否符合已知知识体系的结构特征数据；

（3）如果是，根据现有部署的算法模型进行推理和分析，并返回结果；

（4）如果非已知知识特征数据并判断需要进行重新训练，则经数据系统和训练系统进行机器学习，部署新模型。

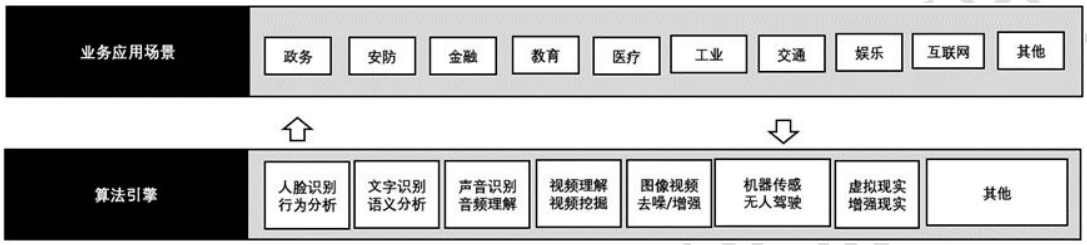

图 9 业务系统及算法引擎主要工作流

4.1.2.2 数据系统工作流

数据系统可基于 Hadoop 等通用大数据架构实现。

数据系统主要工作流如下：

（1）数据接入：从业务应用等对接系统接入数据，兼容多种形式与多种格式，数据接入层的主要功能是为了对数据进行统一标准化的处理，整合各个数据入口；

（2）数据存储：接入数据保存到核心数据存储系统，进行数据分布式存储，包括对上层业务系统存入的原始数据进行持久存储，以及后期经过标注等处理的数据所实现的高速访问与备份；

（3）数据标注：对数据进行标注，给数据增添丰富的结构化信息，形成供研究人员使用的训练集或测试集；

（4）数据计算：给数据增添丰富的结构化信息，包括机器学习所需的特征标注，形成可用的训练集或测试集，经过标注处理的结构化数据可以通过数据接入流程返回进行下一步数据存储和数据计算；

（5）数据管理：对数据进行访问授权与分类管理，根据租户和业务进行数据隔离，保障数据安全；

（6）经过上述工作流处理的数据，分类为业务数据和机器学习数据：

如果是业务数据，则通过可定制的 API 应用接口返回给业务应用。应用接口应是可定制开发的开放接口，为开发者提供简单、丰富的数据处理与算法的 API，轻松实现分布式的数据计算。

如果是机器学习数据，则通过训练系统接口与训练系统无缝对接，进行高速数据传输为深度学习提供训练数据和测试数据。

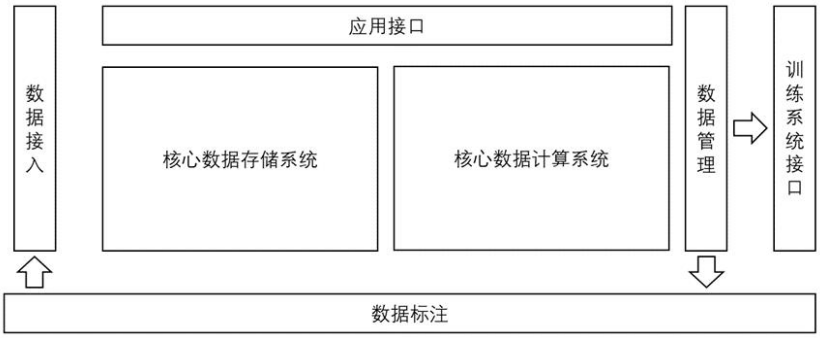

图 10 数据系统主要工作流

4.1.2.3 深度学习训练系统工作流

深度学习训练系统是当代人工智能的核心，通过高性能算力赋予机器以学习的方式迭代组建出模仿人类感知、判断和推测的算法模型。

深度学习训练系统主要工作流如下：

（1）资源分配：深度学习训练过程需要大规模的高性能计算硬件支撑，而且复杂模型的训练周期可能会长达数天甚至数星期，因此需要训练前期通过云平台进行合理的资源分配，并制定完善策略在训练结束后及时回收；

（2）任务调度：同一个训练系统可能存在多个不同的训练任务，成熟的大规模训练可以兼容分布式同步训练和异步训练，因此训练系统完成资源分配之后需要进行调度任务，合理安排队列；

（3）神经网络解释：目前深度学习训练主要针对的是人工神经网络模型，需要对复杂的算法模型进行降维；

（4）模型训练：机器学习可以分为从数据进行的监督学习、半监督学习和无监督学习，以及从行为中进行的强化学习。无论哪一种机器学习，都需要对模型不断进行成千上万次的迭代，才能得到接近可用于实际业务应用的模型。该迭

代训练过程中，需要持续调用高性能计算平台 CPU、GPU、TPU 或 NPU 等异构高性能硬件的计算能力。

（5）模型测试：经过训练的模型，在生产业务系统上进行部署之前，需要先经过测试校验模型的精度是否已经达到生产要求。没有通过测试的模型需要返回上一步重新进行迭代训练直至精度满足要求。

最终，完成上述步骤之后，输出符合生产的算法模型，将会部署到算法引擎。

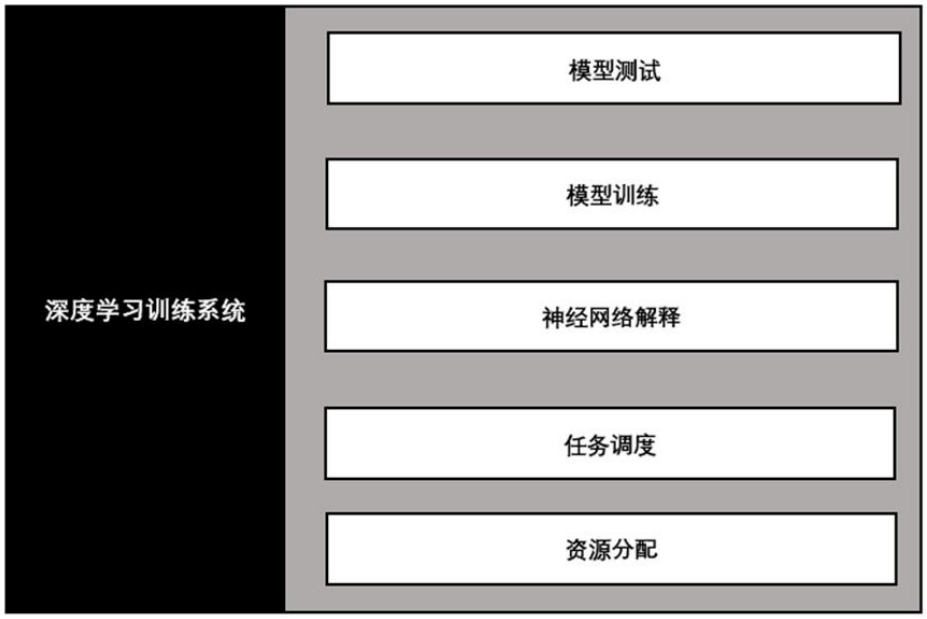

图 11 深度学习训练系统主要工作流

4.1.3 实际应用的 AI 工作流应具备的特点

AI 应用工作流应该支持但不仅限于以下特点：

（1）生态开放

支持行业内开源平台库和开源算法库，支持行业研究员开发者对深度学习模型与算法的定制扩展，支持各种主流的操作系统平台和主流厂商硬件。

（2）API 接口

支持 API 接口，易于进行二次开发，并且可以通过第三方管理平台进行统一管理。

（3）可视化管理

支持所有子系统的工作流程，兼容开源可视化框架，易于部署和运维管理，并且具备完善的日志系统进行记录用户操作行为和系统告警。

（4）强扩展性

能进行大规模多节点的集群运行，支持弹性伸缩，可进行节点扩容减容。

（5）安全可靠

支持权限控制和资源访问安全控制，支持用户数据隔离等安全机制。

4.2 当前 AI 技术在行业应用中的现状及问题

借助数据存储、处理和分析技术，算法与分布式基础设施，AI 开源技术在各行业中的落地场景越来越多，然而在实际发展中也出现了不少问题，如 AI 开源技术对数据的有效利用程度，算法实现模块功能与应用存在差距，以及目前的基础设施能否起有效支撑。本报告从交通、油气、公共安全、工业、电力、金融和医疗七大领域进行分析。

4.2.1 交通领域

随着城市化的发展和交通设施的快速建设和升级，交通行业信息化建设取得质的飞跃：传感器、摄像头、感应线圈等在交通领域被广泛应用。交通疏堵、应急指挥系统、辅助决策系统以及一大批服务于各个专业的信息管理系统的逐步建成与应用，带来了规模巨大、类型多样的交通数据。这些数据相对于传统数据而言具有量大、分布广、结构复杂和数据维度高等特点。

图 12 AI 开源技术在交通行业的应用

近年来在大数据、云计算、算法理论的推动下，AI 开源技术在交通行业的应用越来越广泛。AI 开源大数据技术被应用于交通信息监测、出行者服务、道路规划等业务场景。人工智能技术也被应用于紧急救援与安全、交通管理与规划、车辆安全与辅助驾驶等交通领域，给人们的生产和生活带来诸多便利。但是，AI 开源技术在交通行业的应用仍然存在一些问题。

4.2.1.1 AI 开源和数据技术的差距

因为交通领域的数据具有量大、分布广、结构复杂和数据维度高等特点，现有的数据处理分析技术无法满足实际需求。另外，因缺乏统一标准和技术规范，智能交通系统项目的建设先于行业统一标准的推出，在缺乏标准的条件下，许多地区的智能交通系统自成体系，缺乏应有的衔接和配合，标准互不统一。

（1）数据系统可靠性与稳定性有待提高

交通系统复杂度和整合程度越来越高，涉及底层计算资源调度、分布式数据存储、流式计算和批处理计算融合、二/三维一体化数据管理等复杂系统组件的整合，而系统的健壮性没有得到同步提高。

（2）数据存储与交换同步不成熟

交通领域数据底层存储主要以 Oracle、DB2 等关系型数据库为主，但随着智能交通的发展，产生了较多音视频、图像、地理信息等非结构化数据，数据存储使用到图形数据库 Neo4j、文档存储型数据库 MongoDB、GIS 空间数据库等 NoSQL 数据组件，这些组件之间的数据采集、实时同步等数据传输与交互难以实现。

（3）异构数据协同整合挖掘复杂度高

综合城市路网、路况监测、道路摄像头、GPS 等信息，对多种异构数据进行管理和协同计算，通过轨迹数据的挖掘分析，以目前的开源 AI 技术实现难度较高。将不同组件的现有开源框架（比如 Hadoop、Spark、TensorFlow ）直接整合适用于多种异构数据协同计算的 AI 平台难度较高。现有的人工智能开源技术、数据挖掘技术对分析空间、轨迹、视频图像组成的混合数据尚未有成熟的算法。

4.2.1.2 AI 开源的算法实现和应用的差距

AI 开源的算法实现和应用的差距主要表现在交通信息系统中的信号控制系统中的优化技术以及智能车辆中的智能控制技术。

（1）信号控制系统中的优化技术还需完善

信号控制中的自适应信号控制是智能交通的典型系统，该系统涉及人工智能中的智能体、神经网络、机器学习等技术。目前 AI 开源算法在自适应信号控制系统的应用还存在交通流预测中的不确定性、旅行时间估计的困难以及缺少自调节机制的问题。

（2）智能车辆中的智能控制技术仍待优化

智能车辆涉及人工智能中的模糊控制、神经网络控制和自适应控制等技术，目前 AI 开源算法在智能车辆中的智能控制技术还面临三大核心问题：一是对变幻莫测的道路环境如何做出快速反应；二是在突发技术故障时如何保证无人驾驶的持续稳定性；三是驾驶辅助，及时防范潜在事故，比如监测司机驾驶习惯和状态。

4.2.1.3 AI 开源对分布式基础设施的需求与差距

随着系统规模扩大，前端设备点位和设备故障点增加，存在海量设备管理的潜在问题；当前智能交通的研究偏重于功能的实现，在信息收集、传输、处理各个环节存在严重的信息安全漏洞风险；汇集的大量数据需要在高安全管理标准的数据中心进行存储，而交通系统中原有的数据存储中心的不规范会威胁现有的存储和安防。

4.2.2油气领域

AI 开源技术在油气领域主要用于智能勘探、油井监控、生成控制等。其中涉及的地质数据早已达到 PB 级以上，人工智能、大数据分析已经率先成功应用在勘探开发领域，随后在管道运输、炼油化工及成品油销售领域开始发挥作用。由于不确定的模型变量无法建立关系数据模型，可借助人工智能、大数据分析，从海量数据中发掘规律性关联，找出解决方案。

图 13　AI 开源技术在油气领域应用

在勘探开发领域，针对海量地质数据的存储和处理，已有案例尝试在 Hadoop 环境中进行地质数据预处理；在钻井方面，建立基于决策树、神经网络、逻辑回归等算法的概率模型，全面分析历史作业数据，识别过去反生和卡管相关的数百个特征，使钻井卡管的实时预测模型精度达到较高水平。虽然 AI 开源技术在能源行业的应用逐渐增多，不少问题也逐渐涌现。

4.2.2.1　AI 开源和数据技术的差距

由于油气行业相对封闭，支持 AI 分析的数据技术基础和环境还不够成熟，现有数据采集系统的分布杂乱无章、重新布线投资过大、PLC 较慢且不稳定。在油气行业，传统的关系型数据库、嵌入式数据库被广泛地应用于油气管网调控、汽油消费分析等领域；各种开源软件之间在数据采集、指标口径、分类目录、交换接口、访问接口、数据质量、安全保密等方面没有关键共性标准；数据结构和标准的不同，数据读取时需要适配不同的数据库，将耗费大量的人力。

实时数据整合困难、行业特定数据规范和业务规范复杂且不统一，油气行业信息化长期处于自行发展，存在大量的各类型实时数据库，后续数据整合存在标准难统一、瞬时数据量巨大、与业务数据整合困难的问题。现有的开源框架仅被用于单个业务领域的分析挖掘，尚未出现统一的大数据分析平台支持安全生产统

一研判、生产安全平稳运行、下游精准销售与客服等各业务环节及全产业链的优化。

4.2.2.2 AI 开源的算法实现和应用的差距

AI 开源算法在油气行业使用分为两大模块：能源生成的优化和安全、能源配送的优化和安全。

（1）油气生成的优化和安全。

在生产领域，通过在各个设备上加装传感器，检测设备的运行状态，并利用神经网络和异常检测建立设备故障的模型。若异常数据和正常数据的分布重合，异常检测算法便很难做出准确的判断；且神经网络是一个黑箱算法，不能解释推理过程和依据，且在数据不充分时无法正常运行，因此理论和学习算法都有待进一步完善和提高。

（2）油气配送的优化和安全

国内外已经开发的多种人工智能工具，如专家系统、人工神经网、模糊理论、启发式搜索、遗传算法等，都在油气行业展开了研发和应用。专家系统在实际应用中还存在一些不足，如知识获取瓶颈问题、知识难以维护、不能有效处理不确定因素等。遗传算法虽能在复杂且庞大的搜索空间自适应搜索寻找最优解，但是遗传算法编程实现比较复杂、参数一般靠经验选择、训练时间比较长、对初始化有一定的依赖。

4.2.2.3 AI 开源对分布式基础设施的需求与差距

随着分布式能源和微网系统的增加，能源所有权的转移，将推动新的能源商业模式兴起；物联网带来了巨量的微型电源的需求，随着物联网技术的成熟，物联网的连接节点数量高出一个或多个数量级，近期可达几百亿，每个传感器及连接节点几乎都是全天候能量消耗器，其能耗和能量续航成为痛点。

4.2.3 公共安全领域

随着公共安全立体化和信息化的社会治安防控体系建设，金盾工程、天网工

程、雪亮工程等构建起合理网络架构、共享基础资源、统一公共平台、以及可行可控的信息应用安全。当前公安系统已由 IT 时代逐步转向 DT 时代，业务数据类型结构化转变为多样化，包括音视频、图片、文档、矢量等；数据分析对象包括视频、社交、网络、行为等，对计算底层和计算方式提出更高的要求。传统的数据挖掘分析可靠性低，很难达到要求，利用强大的计算能力及 AI 服务能力从海量复杂的数据中锁定轨迹，实现目标精确定位、线索智能检索和事件预测预警，在一定程度上具有重大意义。

图 14　AI 开源技术在公共安全领域应用

近年来，公共安全行业基于 AI 构建情报研判预警体系、智能警务标准体系、扁平化一级指挥体系等业务场景开展了深化应用，从而掌握潜在治安隐患、增强预防能力，形成有效防控违法犯罪发生和重大恶性案件爆发的一套运行机制。针对公安机关办案所产生的电子卷宗材料，利用深度学习、知识图谱和自然语言处理技术，自动智能研判文书报告、梳理办案过程、证据链条追溯、识别卷宗文书瑕疵，从而规范办案程序、确保办案质量和提高办案效率。基于 AI 开源的安全防范领域主要集中生物特征识别、计算机视觉等应用方面，包括人脸识别、车辆识别、违禁物品识别、可疑行为识别、轨迹分析、公共场所人流分析。同时，AI 开源技术在公共安全行业的应用仍然存在大量问题亟待解决。

4.2.3.1　AI 开源和数据技术的差距

（1）AI 开源技术的数据安全问题

随着 AI 开源技术的推广，AI 开源技术的使用效率提高，学习成本也大大降低，但不乏 AI 开源技术的安全隐患问题，而当前开源技术的漏洞是无法预知的。公安行业的信息数据大多为敏感数据，基于 AI 开源框架构建的智慧公安体系一旦爆发高危漏洞就会引发系列社会问题，因此构建一套完善的安全防控机制对于智慧公安体系是亟待解决的事情。

（2）AI 开源框架的不兼容性

基于 AI 开源框架构建行业业务分析系统，采用单一的开源框架往往不能满足需求。如重点人员犯罪预警通常会利用深度学习、机器学习交替使用完成特征工程，以及利用深度学习、强化学习和迁移学习完成模型的训练，因此实现不同 AI 开源框架如 Caffe、Spark MLlib、Tensorflow、PyTorch 等算法框架的交互和模型共享显得尤为重要。但是目前大多数开源框架之间相互不兼容，需要构建一种能够在不同开源框架之间进行通讯的成熟的工具或标准方法。

（3）数据形态多样化

基于 AI 开源技术挖掘公安行业数据中存在的价值，需要对警务综合、PGIS、治安防控、情报研判、出入境、音视频数据库以及外围业务系统数据进行治理和分析。由于业务需求和业务理解不同，数据存储方式和结构不统一，难以将 GIS、图像、视频等多样化数据转化为结构化数据，难以实现将 Orale、MySQL 等关系型数据库与 Hive、Hbase 、HDFS 等分布式数据库进行关联、同步传输、交互与整合。

4.2.3.2 AI 开源的算法实现和应用的差距

AI 开源技术提供了大量的通用算法，主要包括深度学习、计算机视觉、机器学习、自然语言处理等，但应用于公共安全行业的算法理论体系还不成熟，且标准不统一，没有一套专用于公共安全行业的通用算法理论。同时，开源算法深度不够，一般无法直接满足业务需求，需要耗费大量的人力对算法进行优化。此外，随着侦查技术的不断发展，视频人脸识别技术在公安犯罪识别中拥有巨大的应用前景。不同于传统静态图像，视频图像具有同时多个主题多帧融合、实时信息等特点，更多受到采集条件和运动的影响，视频序列中采集的人脸图像分辨率低且人脸模糊，人脸图像超分辨率技术和图像复原技术是目前解决的一种思路。

单一的 AI 开源软件往往无法满足复杂的业务需求，需整合各 AI 开源软件（如 Standford、Hanlp、IRSTLM 等）的功能，但是对语料、词性的标注目前还没有统一的标准，不同的开源框架趋向于运用不同的知识库、分词词典、情感词典等，缺乏公安行业紧密结合的语料资料，因此在实际业务应用中难以对开源软件算法进行扩充、深度优化。

4.2.3.3 AI 开源对分布式基础设施的需求与差距

AI 开源技术在公共安全领域的分布式计算主要体现在图像处理和高并发的电信网络数据处理，利用机器学习和深度学习框架、大数据分布式计算进行分析处理，有助于有效预警防控违法犯罪发生和重大恶性案件的爆发，极大提升公共安全监管水平。

以智能手机为依托的移动互联网时代，尤其是新一代通信技术的蓬勃发展，网络违法犯罪日益突出，违法犯罪人员的反侦查能力不断提升，利用 AI 开源技术完成对这些监管对象的实时分析，将有助于提高公共安全管理效率。而目前的流式计算相关的开源技术还无法与 AI 开源技术相互融合，当前的 AI 开源技术仍然专注于对离线图像和轨迹数据进行分析，而对时效性有较高标准的公共安全领域，从数据的预处理、数据的标注，再到利用机器学习和深度学习框架进行模型训练都需要漫长的过程，同时也对开源计算平台和硬件提出了较高的要求。

4.2.4 工业领域

智能制造的快速发展亟需 AI 开源在工业领域的推动，而基于 CPS(Cyber-Physical Systems)的智能制造技术仍在探索研发阶段，且相关代码及数据掌握在极少部分巨头企业手中。目前由于与智能制造相关的 AI 开源项目较少，因此 AI 开源主要以离散点的形式在智能制造中展开，其所涉及的共性技术主要包括：机器学习、生物特征识别、计算机视觉、自然语言处理和知识图谱等。同时，上述应用主要围绕产品质量检测、工艺分析与优化等特定重复性问题展开，并为企业管理者或车间运维人员提供辅助优化、辅助决策以提升企业的生产效率，减小人员的工作强度。

4.2.4.1 AI 开源和数据技术的差距

工业大数据是指在工业领域中，围绕典型智能制造模式，从客户需求到销售、订单、计划、设计、研发、工艺、采购、供应、制造、库存、发货和交付、售后、运维、报废或回收再制造等整个产品全生命周期中各个环节所产生的各类数据及相关技术和应用的总称。由此可以看出，工业大数据多种多样，在不同的信息采集软件、不同的数据采集设备以及产品生产周期的不同环节，获取的数据格式及种类均有一定的差别。而海量数据是 AI 开源算法得以充分发挥作用的关键场景，因此工业数据技术对工业中 AI 开源应用具有重要作用，其两者之间的差距主要表现为：

（1）工业数据采集及通信方面

我国许多工厂内部的 ERP 还没有做好，数据采集还不全面，许多数据还需人工采集，且数据采集系统方式杂乱无章，导致数据短缺、后期数据处理困难。另外，工业现场的数据通信标准之间通常不能兼容，总线网关协议繁多，这些协议之间不能直接互联互通，信息孤岛的情况在工业界广泛存在，无法满足人工智能技术对优化建模数据量的数据完整性需求。

（2）工业数据挖掘

面对多种多样的采集数据，能够自我感知、自我记忆的数据采集感应系统尚未建立，处理复杂数据结构的数据处理技术仍需优化，高效的数据库维护和管理机制还需完善。虽然已有部分 AI 开源框架用于大数据挖掘，但用于工业大数据的 AI 开源框架较少，且均是针对某一方面的工业数据挖掘。

（3）工业数据信息转化

在从工业大数据中获取了有价值数据之后，还需通过筛选、存储、关联、融合、索引、调用等形式将数据进行模型训练，模型上线后还需保证系统的稳定性、时效性和吞吐等等指标，之后才可变为对工业生产有用的信息，因此如何训练出好的模型、如何去选择好的参数、如何进行特征组合，都需要很专业人员去完成，这都大大增加了 AI 在工业界落地的难度。

4.2.4.2 AI 开源的算法实现和应用的差距

AI 算法应用于智能制造的各个环节，其中主要包括：

（1）基于增强现实（AR）的人员培训

AR 设备能够为学员提供实时可见、现场分步骤的指导，尤其是在产品组装等领域，通过将图纸转换为可视三维模型，指导操作人员完成所需的步骤。

（2）预测性维护

预测性维护依据实时采集的设备运行数据，通过机器学习算法辨识故障信号，从而实现对故障设备的提前感知与维护，最终减少设备所需的维护时间与费用，提高设备利用率，更大程度地减少因设备故障所引起的损失。

（3）动态智能排产

智能排产系统通过机器学习算法等帮助企业进行资源和系统的整合、集成与优化，动态实现排程优化，进而帮助企业实现按需生产，提高运行效率，缩短产品周期，提升企业的产能。

（4）智能在线检测

智能在线检测技术依据传感器采集的产品照片，通过计算机视觉算法检测残次品，从而提高产品检测速度及质量，减少因漏检、错检所引起的损失。

然而面对有监督学习在工业中的应用，数据采集、数据格式以及采用哪种 AI 算法来，实现稳定的生产是有待验证的问题。另外，面对无监督学习在工业中的应用如 AR 的实现，对图像识别算法的精度及处理速度要求很高，虽然关于图像识别的开源算法很多，但用于 AR 眼镜的图像识别算法，其处理现实物体与虚拟物体遮挡问题的能力还有待改善，另外当虚拟物体与现实物体叠加时因延迟而产生的抖动也是一个挑战。

4.2.4.3 AI 开源对分布式基础设施的需求与差距

智能制造采用各类标识技术自动识别零部件、在制品、工序、产品等对象，在仓储、生产过程中实现自动信息采集与处理，通过与国家工业互联网标识解析系统对接，实现对产品全生命周期管理。同时要求工厂采用工业以太网、工业 PON（(Passive Optical Network)、工业无线、IPv6 等技术，实现生产装备、传感

器、控制系统与管理系统等互联，以推动数据的采集、流转和处理。进一步利用IPv6、工业物联网等技术，实现与工厂内、外网的互联互通，支持内、外网业务协同。智能制造的实现是多生产线、多工厂间的设备互联，每个设备、零部件、产品都用相应的信息数据，数据量十分庞大。处理工业大数据，并作用于工业生产的各个环节，且保证数据处理速度及在线处理能力，分布式基础设施十分关键，但当前各 AI 框架与平台的处理能力参差不齐，面对不同的工业大数据，还没有形成统一的公共能力或事实标准。

4.2.5 电力领域

随着电力信息化的推进，智能变电站、智能电表、在线监测系统、现场移动检修系统、测控一体化系统以及一大批服务于各个专业的信息管理系统的逐步建成与应用，智能电网数据的规模和种类快速增长。电力系统已从以往类型较为单一、增长较为缓慢的数据时代逐渐步入海量、多源、异构、分布控制产生、复杂、动态内联的大数据时代。电力数据具有体量大、类型多、实时性、价值大、复杂性等大数据典型特征，现有的分析处理方法已不能满足要求，需要开发利用新的技术，进行跨领域、多维度的复杂关联分析。

图 15　AI 开源技术在电力领域应用

随着云计算、大数据及人工智能等前沿技术在互联网、安防等行业的不断演进，其在电力领域应用也越来越广泛。大数据技术在电网负荷预测、电动汽车需求分析、电网可靠性影响因素分析、用户能效评估及客户缴费行为分析等业务场景开展了深化应用。人工智能也广泛应用于电力系统仿真分析，基于环境识别、复杂内外部条件认知，以数据为基础，通过深度学习自动提取电网稳定特征，实

现对电网稳定运行方式和有效措施的快速判断。然而在利用前沿的开源技术过程中也遇到了各种各样的问题。

4.2.5.1 AI 开源和数据技术的差距

（1）存储多样性，缺乏统一标准

电力业务体系非常复杂，产生的数据复杂多样，包括结构化数据、非结构化数据、实时量测数据等，分别存储在关系型数据库、分布式文件系统、MPP 数据库、NoSQL 数据库等多种异构存储介质中，由于这些存储介质在数据存储方式和结构上存在差异，无法实现跨异构数据库数据同步、关联查询与整合，当前开源技术在这些方面还存在不足，无法满足电力领域对海量异构数据的交互式探索分析。

（2）软件跨版本升级难度大

在电网调度实际应用中，将 Hadoop、HBase、Hive 等开源组件同步升级到高版本的时候，复杂度高而且周期很长，无法实现快速跨版本的快速平滑升级。

（3）缺少开源组件评测标准

缺少对存储组件 NoSQL 数据库、MPP 数据库、分布式文件系统等，计算组件 Storm、Spark、Mapreduce、Impala 等功能与性能的基准测试及评测标准。

4.2.5.2 AI 开源的算法实现和应用的差距

电力行业是资产密集型行业，资产总量庞大、分布广泛，如果传统方法无法有效管理，将难以支撑电网的可持续性运行和发展。电网网络庞杂、设备点多面广、运行特性各异，传统的运维检修办法难以对设备状态进行精准评价和针对性地投资、改造和运维，从而利用 AI 开源技术来进一步提升电力资产管理和智能化运维水平。但在 AI 开源技术框架的实际应用中，存在算法开发调用复杂度高、不同数据集和模型文件兼容性差、可视化交互不足等诸多问题，无法满足电网的业务需求，另外，在电力巡检和监控影像的目标识别与缺陷检测方面，目前 AI 开源算法尚未达到满足生产需求的精度与效率。最后，电力系统领域的故障预测、

极端天气预报、设备健康状态评价等场景存在着可用数据量小的困境，在小样本学习开源技术方面还存在匮乏，无法为应用提供有力的技术支撑。

4.2.5.3 AI 开源对分布式基础设施的需求与差距

电网的电力调度、保护测控、安全运维、在线监测等领域涵盖大量的传感器，并需要实现物理设备与信息感知终端耦合，以及网络基础设施的大范围空间部署，然而传统业务系统在采集终端、存储系统、计算资源等方面上越来越存在瓶颈，因而对开源的分布式基础设施提出了更高的要求来满足实时采集、边缘计算、深度挖掘、线性扩展等业务需求。

图 16 AI 开源对分布式基础设施的需求与差距

4.2.6 金融领域

AI 开源技术在金融行业的应用较为广泛，目前已经在金融子行业（银行、证券、保险等）的智能客服、智能营销、远程身份认证、智能化运维、反欺诈与智能风控、大数据征信等环节中有所体现。同时，AI 开源技术在金融行业的应用仍然存在大量问题亟待解决。

4.2.6.1 AI 开源和数据技术的差距

金融行业具有牵涉面广、高度信息化和高频交易的特点，沉淀了海量数据，相对于其他行业在数据质与量上有一定优势。但是由于金融行业受严格的法律法规与监管政策制约，牵涉用户敏感信息，数据的使用制约在一定程度上也限制了人工智能开源相关模型的有效性。

4.2.6.2 开源的算法实现和应用的差距

智能客服是以亿级海量聊天信息为数据基础，并通过深度学习训练而成的智能产品。相对于人工客服，智能客服在服务流程优化、成本节省、客户效果体验等方面都表现出了一定的优势。目前在实现 AI 开源算法与应用的主要差距在于需要持续地根据实践数据，进行知识库的补充与标注。通观行业的类似产品，在上下文多轮复杂对话、情绪识别、场景切换上与用户的要求仍然相差甚远。

金融行业的反欺诈工作是基于之前的业务数据进行数据挖掘、建立模型，并应用到当前的生产数据，挖掘新的异常模式。利用机器学习可以将现有的专家规则方法的灵活性提高，降低误报率，建立观察对象的关联关系网络图等。目前挑战主要在于对海量数据的高效利用，如有效数据的筛选与抓取、算力支持等。

4.2.6.3 AI 开源对分布式基础设施的需求与差距

金融的核心业务对安全性、可靠性有很高的要求，尤其是一些发展时间较长的大型金融机构，历史包袱比较重，因此传统金融业务和多数核心系统还难以迅速采用分布式基础设施。目前少部分互联网银行或创新的金融产品或辅助性的金融业务已逐步采用分布式基础设施，在低成本、标准化的开放硬件和开源软件的基础上，采用数据复制、多副本、读写分离等技术弥补基础软硬件的不足。另外在满足系统高性能、高可用和容灾备份等方面要求的前提下，采取了自主可控的分布式架构，加快产品创新的周期，节约相关的系统建设和运维成本，实现系统的快速迭代与升级等。

4.2.7医疗领域

随着中国现代化高速发展，城镇化建设稳步提高，越来越多的人口涌入一、二线城市。而医院的建立与扩建需要长时间的规划，培养一名专业医生更是需要多年的学习及实践经历，导致部分医院，尤其是三甲医院的就诊压力逐年增高，且医生工作强度大。与此同时，随着近年来深度学习的兴起，AI 水平日趋成熟，工作效率与准确性有着显著提升，已经在一些识别、检测任务中达到甚至可以超越人类的水平。因此，结合 AI，把医生从低级重复性的工作中解放出来，把精力放在更重要、更关键的诊断决策上，是未来医疗领域发展的必然趋势。

由于医疗诊断本身的敏感性，其最终决策往往需要结合患者的完整信息（包括且不限于：病症的准确描述，病人的基本信息，病史及家族史，影像检查数据，相关的血液、尿液、甚至基因的检查报告），依赖多方信息的相互支持，不限于单个影像分析或是检测报告。即便如此，同时获得患者的完整数据需依靠医院内多科室间甚至多医院间的沟通与协调，往往涉及到大范围的信息系统升级，在现阶段显得尤为困难。医疗影像（如 x 光、断层扫描 CT、磁共振 MRI、超声 Ultrasound 等）因为可以在极低副作用甚至无副作用的前提保证下，快速、准确、直观地展现出患者体内的状况，往往在常规检查甚至疑难杂症的诊断过程中被主要采用。因此，当下人工智能大多以医疗影像分析为切入点，为医生提供辅助诊断。同时，AI 开源技术在医疗行业的应用仍然存在大量问题亟待解决。

4.2.7.1 AI 开源和数据技术的差距

与其他领域所面临的数据问题类似，医疗领域数据也存在着诸如标准不统一，质量、格式参差不齐，开放性极低等问题。具体关于医疗数据收集的要求与挑战，请参考第 5 章节。本节主要探讨医疗数据自身几点值得思考的问题：

（1）数据的分布偏差

数据分布的不同主要来自两个方面：不同医疗器材医疗厂商间的设备在成像过程中使用的方法及材料不尽相同，同时各大医院影像科室的操作方法也是因人而异，导致最终产出的影像上会有细微差异。其次，不同地区的人种及病症也有着不小的差异，此前 IBM 在海外设计训练的诊断模型 Dr.Watson 在国内的测试结

果就不尽人意。这些因素都会对 AI 预测带来极大挑战，尤其在与医疗机构的首次合作时。

（2）数据的采样偏差

医疗影像数据本身信息量十分庞大，通常单个三维影像会有上百 Mb 甚至 Gb 级的数据。为了提高病例检出的敏感性，在训练过程中通常会调低健康患者的比例，调高罕见病例的出现频率，致使 AI 的训练数据分布与真实场景严重失衡。如何平衡病症的检出与误报也是一个值得深刻思考的问题。

（3）精标注的定义与缺失

与获取数据同样重要的问题是获取数据相对应的标注。可惜的是，医院本身的诊断报告往往只对少数最主要的病症有简单描述，往往还不是结构化数据，不能直接转化为训练目标。虽然半监督、弱监督训练已经是热门的研究方向，但其效果相较于更为成熟的监督学习还有一定差距。又因为标注的质量即代表着算法本身所能达到的上界，所以现阶段医疗领域的人工智能应用需依靠影像学专家对收集来的数据进行详细标注。如何以更高效的交互方式获取标注，如何保证数据标注的准确性，如何规范不同医生之间的标注差异，这些都是 AI 从一个开源项目延伸到一个实际解决方案过程中亟待解决的问题。

4.2.7.2 开源的算法实现和应用的差距

随着深度学习在自然图像识别、定位、分割等领域的高速发展，AI 在医疗领域也迎来的新的突破。同为影像数据，医疗影像分析可拆分为许多子任务，并借鉴许多自然图像中的经典方法与深度学习方法。甚至早期深度学习算法直接利用在大规模自然图像识别任务里学习得的模型进行迁移学习，与只在有限的医疗影像数据上训练的模型相比，取得了更加优异的成果。

当然，医疗图像本身有着许多特有的挑战，诸如：数据数量少，单个数据信息量大、沉余多；病灶组织多变，相较健康组织占比极少；标注难获取，数据质量及标注质量难统一，甚至其真实结果无法确定等。这些因素都会在算法设计的时候有着多种考量，而且不同病种也会有不同的需求及预期。一个完备的解决方案通常是多种算法多个模型的融合，其中在一些数据稀缺的疑难问题上仍会采用传统机器学习方法以取得最泛化的效果，在一些不可犯错的边界条件上仍会加入

人工定义的边界条件。这些考虑往往不是一个单纯以准确性为目标函数的开源算法率可以涵盖的，但一些新的算法仍可替换、改进解决方案中的某一步，带来最终效果上的提升。

关于模型的可解释性也是AI在医疗领域推行过程中很值得探讨的一个问题。一方面，传统机器学习十分注重医生在影像数据上的推理过程，常会提取一些可模拟人类理解图像的特征，可被理解的分类器。但很不幸的是，其预测结果往往不如深度学习在同任务中的表现。另一方面，深度学习在一些特定任务里有着可比医生的效果，但是会在意想不到的地方出现人类不会犯的低级错误；而且每当这种状况出现的时候，因为无法解释深度学习模型的逻辑，AI 结果的置信度就会降低。怎样去融合传统方法与深度学习的优点，减少甚至杜绝意外状况的发生，增强算法模型的鲁棒性，也是从开源到实际应用中需要解决的问题。

4.2.7.3 AI 开源对分布式基础设施的需求与差距

由于医疗数据的敏感性，其最终归属与使用的合法性目前仍是一个有争议的问题。为了保证数据的安全，目前的解决方案多倾向于在医院本地部署运算设备，或是基于加密的传输渠道，在合作企业的私有云进行远程运算。但是医疗数据的体量巨大，随着 AI 在医疗领域的运用逐渐成熟，安全、可靠的分布式运算必将是其最终的发展方向。

4.3 问题总结及应对思路

从上面各行业、各场景中应用 AI 开源技术的实际情况看，在数据、算法、算力（基础设施平台或硬件平台）这三个纬度存在一些共性问题。基于这些共性问题，本报告试图给出一些解决思路，供产业及技术组织参考，其中一部分在本小节提出，另一些会在第五章及第六章进行进一步描述。

4.3.1 AI 开源软件的数据支持

4.3.1.1 数据支持问题

AI 开源软件的数据支持匮乏问题，主要表现在以下几个方面：

（1）某些开源软件只提供算法，不提供数据，从而很难模拟训练过程。

（2）数据来自于不同的组织，数据需要满足格式要求才能被统一使用。

（3）数据在不同组织间共享和交换缺乏权威的数据共享许可协议。

（4）开源软件使用的数据或者语料的标注标准多样化，不便对语料进行扩展和完善。

（5）AI 开源软件对接的数据来源无法访问，或部署后无法使用。

（6）某些 AI 开源软件缺乏数据收集和选择标准，不利于构建实际环境下的语料。

（7）更多的数据就意味着更精准的模型，并且数据是有产权的，当前还缺乏不同组织间的数据汇聚、应用和保护机制。

（8）AI 数据量大，数据的有效处理需要建立从数据产生、保存和一次处理入手。

4.3.1.2 AI 开放数据治理及数据格式标准化

当前 AI 一个重要特点就是数据驱动，数据的来源、数量及覆盖度直接影响了 AI 的落地效果。数据的开放可获得性是 AI 大规模应用所不可回避的问题，本文第五章将对这一问题给出详细的分析及建议思考。

同时 AI 的模型训练是一个不断重复、迭代的过程，在开发过程中快速有效的读取训练数据是 AI 训练的必然需求。统一的数据格式有利于各种存储平台有针对的进行分布式、高并发的训练数据读取和分发。而各种 AI 训练平台的训练数据格式不统一，更加重了从各种数据存储基础设施中快速分布式读取训练数据的对接难度和工作量。

在数据仓库、大数据等技术普遍应用的今天，如何将分散在不同地域、不同业务系统中的数据有效进行描述，同时建立起逻辑统一的数据交换标准成为了普

遍需求。数据目录服务是解决这一现状的有效手段，通过统一的规则标准来描述数据的血缘关系、访问授权等信息(数据自身属性 metadata)，实现数据源可搜索、可控授权访问、有效同步等能力。

不稳定的数据依赖是影响 AI 数据处理的重要因素，今年来为解决大数据实施后的数据稳定性问题提出了冻结副本(frozen copy)概念，数据湖(data lake) 平台可以有效聚合结构化、半结构化、非结构化数据为 AI 提供稳定的数据支持。

针对 AI 训练特点的数据格式能够有效加速 AI 模型的训练工作效率。在本文第六章将给出 AI 数据格式标准化的建议。

4.3.2 AI 开源软件的算法

4.3.2.1 算法问题

AI 开源软件的算法难以满足实际应用问题，主要表现在以下几个方面：

（1）大多数 AI 开源软件缺少足够的技术支持，有些没有技术支持，和作者沟通十分困难，不利于解决实际中的问题，影响实用效果。

（2）AI 开源软件数量多，且对模型和算法的标识方式不一，相互间的成果不能互相借鉴，不利于推理侧 AI 模型的部署；不同训练框架软件的模型进行迁移，可能还会造成模型精度损失等潜在风险；同时造成 AI 创新分散，不利于 AI 产业发展。

（3）某些 AI 开源软件自身存在 Bug，没有在测试中发现，如果用于实际项目的核心模块，可能会造成比较严重的后果。

（4）AI 开源软件的开发语言和环境以及规范存在多样性，没有统一的标准，代码兼容性差，移植到实际项目中的成本很高。

（5）AI 开源软件给后端芯片提供不同的接入方式，接入成本高，也难以发挥出芯片的优势。

4.3.2.2 算法&模型格式标准化

算法的持续创新必然是一个永不停止的过程，上述各行业内关于算法质量、精度及适用性的问题还会靠学术界及产业界共同探索解决。但同时也应看到还有

一些算法的问题是由于各种AI平台的分散造成基础算法实现不够统一所引起的，这严重影响了不同行业的AI的应用与落地以及持续迭代。建议从整个人工智能领域的常用、通用算法层面定义基础算法的统一标准和接口，形成能够被不同行业快速利用的基础算法库。

定义统一的模型格式是推动AI发展的重要基石，当今AI领域各行业的经典模型算法都被广泛使用，并且利用已有模型进行迁移学习并组合多种AI模型形成复杂业务系统已经成为常态。但是由于缺乏统一完善的模型格式定义在各种平台间共享模型存在极大困难，（ONNX模型格式也未能完成对TF模型格式的支持）。同时模型格式的统一也是AI产品从开发到部署过程中实现可升级、可管理、可测试的重要基础。在本文第六章将给出相关的标准化建议。

4.3.3 AI开源软件的分布式基础设施

4.3.3.1 分布式基础设施支持问题

AI开源软件对分布式基础设施支持不足，主要表现在以下几个方面：

（1）某些AI开源软件来自高校、研究所等科研机构，其实验环境是单机、小规模物理或模拟集群，对于实际的分布式云计算环境支持不足。

（2）由于AI开源软件和传统软件存在差异，目前仅有的大型公共分布式集群环境无法作为其运行和测试的载体，所以导致目前的AI开源软件在发布前缺少在大型分布式环境中的测试，因此对于分布式基础设施的支持不佳。

（3）AI开源软件框架对分布式支持自身存在不足，不能发挥分布式集群环境在训练上的优势。

（4）AI开源软件需要更多的分布式计算资源的支撑，然而成本以及基础建设方面的困难导致很多算法模型无法得到及时有效的验证和演进。

4.3.3.2 分布式 AI 框架的标准化及构建分布式数据管理平台的思考

目前 AI 训练和推理平台（框架）的各种实现分散不统一，对 AI 落地实施过程中各层次组件的扩展、优化造成了极大障碍。此外，在大规模机器学习、增强学习等领域中都严重依赖分布式能力，如：参数服务、网络通信、梯度融合等，但当前各 AI 框架与平台在这些方面能力参差不齐，没有形成统一的公共能力或事实标准。由于 AI 模型的快速迭代、不断升级对生产环境持续更新部署 AI 模型提出了挑战，需要定义统一的 AI 模型版本管理平台，通过支持对 AI 模型输入、输出数据的元数据描述管理来实现统一的 AI 模型部署 Pipeline 流程。当前 AI 训练平台的部署已经形成以容器技术为基础的自动化部署趋势，各种 AI 公司或者互联网公司的 AI 部门都会尝试在 Kubernetes 上运行分布式训练平台。但不同的 AI 平台对集群分布式部署有不同的需求也带来了配置难题。以上这些 AI 框架的问题很多还是会在开源社区中，通过不断创新、迭代来解决，新的开源框架还会出现，但同时需要考虑各个层次之间的接口，尽快在基础能力、常见参数、实现方式等方面进行规范，形成统一接口，并推动开源社区参考并遵循。同时还应推动重量级开源框架进行基金会化的开放治理，来共同形成事实标准。在这个领域内的相关标准化的思路请参考第六章。

当前除了缺乏 AI 框架的相关标准化之外，高效的数据管理同样是痛点。之前提到的开放数据治理还是解决数据来源的问题，但 AI 数据管理则是如何在分布式场景下通过集群技术高效的增、删、改、查、同（步）用于 AI 的海量数据，是 AI 基础能力的重要组成部分。目前在开源界还没有系统化、综合性的此类项目。如前所述，人工智能的发展离不开海量的高质量数据支持，随着近几年大数据平台在各行业的落地部署，企业积累了大量的业务数据。但由于在大数据平台建设过程中不同业务系统独立实施，造成数据孤岛情况普遍存在，使得数据治理问题也成为困扰企业使用数据为人工智能服务的重要问题。目前人工智能的数据使用问题主要面临着数据存储分散，数据源不稳定，数据难以发现、共享、布控困难等障碍。只有将异构数据源（结构化、半结构化、非结构化数据并存）的数据高质量、高效率地整合到一起，进而进行加工、分析、挖掘和展现，才能更精

准的通过 AI/ML 来驱动商业和业务决策。在 AI 领域数据特征质量是训练模型质量的最重要直接影响因素，自动化特征标注、知识图谱等工具的缺乏对支撑 AI 快速实施的限制也越来越明显。

整体来看目前虽然人工智能领域发展迅速，但是纯粹的人工智能开发只占整体工作的小部分，大量准备性、支撑性、平台性工作的开展依然缺乏端到端的支撑，现阶段各行业已经形成了自己基于结构化、半结构化、非结构化数据混合存在的大数据形态。对于 AI 发展而言，有效的整合各种支撑平台为 AI 所用是当前面临的重要问题。

第五章 AI 数据开放及协同

5.1 AI 数据的关系和需求

5.1.1 面对的挑战

AI 在深度学习和机器学习领域的突破高度依赖于大量数据支撑，数据特性是决定 AI 技术有效性的重要环节，当前在数据层面的关键挑战有如下多个方面：

（1）可用性（Availability）：保证数据的可用性才能走出人工智能训练的第一步，没有数据，算法再好也无法训练监督学习模型。

（2）可访问性(Accessibility) ：即使数据可用，如果它们存在加密或权限受限，无权便捷获取，则等同于没有数据。

（3）规范性(Standardability) ：如果数据可用且可访问，但格式不标准或不规范，那么将会导致数据不可导入人工智能的训练，或者影响数据训练的实用效果。因此，需要制定相关标准，提高数据格式的规范性。规范性主要包括编码规范、标注规范、标签规范等。

（4）兼容性(Compatibility) ：在相同应用中，不同厂商设备所产生的数据标准存在差异，通用性差，那么将使数据在成本、规模和性能方面存在一定问题。此外，各种 AI 训练平台的训练数据格式不统一，加重了从不同基础设施中快速读取训练数据的对接难度，同时加大了相关开发的工作量。

（5）质量(Quality) ：准确标注有代表性的数据是高质量数据集的特征，数

据集的质量是生成准确推理模型的一个重要环节，高质量数据训练集往往比更大的数据训练集更好。

（6）关联性(Relevancy) ：训练过程的数据相关性是高度主观的，并且可以在不同的行业部门和使用的上下文中有不同的解释和定义。

（7）完备性(Completeness) ：没有足够大小和覆盖范围的样本可能会导致产生不同偏见的训练结果，因此通过标准化来定义数据输入的相关规范，使其保持完整性和充分性。

（8）机密性(Confidentiality) ：需要有法律框架和其他机制来保证数据不会被非法使用，以保护数据机密及其隐私。

（9）安全性(Security) ：数据消费者使用数据时，需要数据提供者提供对数据本身进行一定程度访问的方法，并确保在消费前中后期的数据安全性，目前产业界还在这一领域进行探索。

（10）所有权(Ownership) ：一旦数据由数据提供者提供，数据的所有权归属会引发争议。目前开放数据平台和数据消费者对数据本身的存储、消耗和分配等模块缺乏管理和控制的定义。

（11）可复制性（Reproducibility）：数据和模型都是软件产品，通过复制操作，很容易被竞争对手抄袭，甚至直接使用。这对企业进行数据开源形成了很大阻力，因此需要建立合适的政策机制保护企业自身知识产权，不被竞争对手抄袭和直接使用。

（12）可转换性（Convertibility）：对于已经清洗和标注的数据，依然面临着需要洗牌、批处理等数据转换方面的问题，所以各主流 AI 框架都实现了自己的数据集格式，由于不同格式的数据无法互操作，因此需要实现彼此之间的互相转换。

（13）可融合性（Integration）：在实际应用中，各行业都已经形成了结构化、半结构化、非结构化数据混合存在的大数据形态，因此需要数据支撑平台能够融合不同形态的数据，将数据高质量、高效率地整合起来，进而实施加工、分析、挖掘和展现的有效支持。

（14）可发现性（Discoverability）：在数据仓库、大数据等技术普遍应用的今天，如何将分散在不同地域、不同业务系统中的数据有效进行描述，使原本分

散在各业务系统的“生数据”转变为可为 AI 所用的“熟数据”，需要建立逻辑统一的数据交换标准。

（15）标注细分（Tag Refinement）：数据标注是进行人工智能训练的基础，公开数据集标注相对粗糙，仅能够满足通用型人工智能算法验证的需要，无法有效验证蓬勃发展的人工智能细分领域应用型算法，特别是针对专用场景相关的行业特殊算法。所以，需要根据行业细分的实际需要，采用分类、画框、注释等方式细致、准确地标注数据，但是对数据进行细分标注是一项极其消耗精力和财力的事情。

（16）成本效益（Cost-Effectiveness）：为了保证数据可用，使用者需要进行数据准备、数据治理、数据消岐等操作，在某些实际应用中上述工作甚至占用大约 90%以上的工作量，是企业在实施 AI 过程中消耗资源最为巨大的部分。

（17）激励机制 (Rewarding Mechanism) ：高质量的大规模数据是企业的重要资产，开源或开放后可能导致企业丧失竞争优势，目前缺少让数据开源贡献者获得合理回报的机制。此外，如何将所提供的数据价值与所产生数据的使用结果相链接，以及如何让数据提供者对数据的贡献认可和奖励，目前并未达成共识。

由于开放数据访问和协作的上述挑战，数据提供者和数据消费者之间存在错综复杂的关系。传统上，数据的聚集、访问和协作是由政府或有中心化权威的互联网或云平台实现（如华为、腾讯、阿里巴巴、JD、谷歌、脸谱网、苹果、微软、亚马逊等）。数据访问在零售业、公共服务业等应用领域已经得到了很好的解决，但在诸如金融、医疗、教育等更为敏感和零散的仓储产业领域，还没有较为成熟的解决方法。此外，数据提供者的权利监管，需要根据集权机构或平台已发布或未公布的隐秘策略来判断，而这些策略很难被数据提供者理解，因此最终往往以牺牲数据提供者为代价，导致数据提供者失去对数据所有权、保密性和隐私权的控制。即使政府作为一个强大的集权机构，也面临着实现开放数据访问和协作的挑战。

5.1.2 AI 数据开放和协同中的相关方

AI 数据中的相关方是指自然人或法人，可以是一人或多人组成的实体，是 AI 数据开放生态中的利益相关方，一般可分为客户方、提供方和关联方。客户方

主要指使用 AI 开放数据的用户或业务相关方，提供方主要指提供 AI 数据的相关方，关联方是指为客户方和提供方提供相关支持的相关方。

（1）客户方

AI 开放数据的使用方和管理方，包括数据业务功能管理、数据生命周期管理等，可以是个人、开源社区、联盟、企业、非盈利组织等。客户方是重点的应用场景方（如智能金融、智能安防、智能物流、智能医疗、智能教育、智能客服、智能运维、知识图谱等应用场景），通过社区或市场获得 AI 产出与成果，其目的可能是用于自我能力提升、服务体验改善、产品智能化，也可能用于为相关方提供智能增值服务等。

（2）提供方

为 AI 开放数据提供技术支持，提供的数据主要包含有偿数据和无偿数据。主要负责开放数据技术和平台的运营，包括部署、变更、配置、安全和风险管理等，如开源社区和开源联盟等组织，提供了开放的数据测试集或案例等，主要集中的领域有认知构架、深度学习、机器学习、自然语言处理、深度神经网络和虚拟现实等。

（3）关联方

主要为客户方和提供方提供相关支持，如监管机构、投资机构、中介机构（法律、审计、评估）等，对客户方和提供方的发展具有重要的促成作用。关联方在人工智能开源发展的过程中，存在 AI 开放数据的流通和交易等需求，其内容包括领域数据与知识服务、知识资产与成果评估和审计等，可用以满足特定需求的有形或无形成果的重要输入要素。监管机构研究拟定 AI 数据开放方针政策和总体规划，制订 AI 开放数据产业与技术发展政策、技术体制和技术标准等，推动产业和行业发展等；投资机构通过参与 AI 开放数据产业链中的投资，促进上下游协调发展。

5.2 AI 数据开放和协同中相关行业分析

，当前 60%以上企业把内部业务平台数据、客户数据和管理平台数据作为大数据应用的主要来源，只有约 1/3 的企业使用外部互联网数据或其他行业企业数据。现阶段并未形成企业内外融合互动的数据采集与处理模式，外部数据应用水

平有待进一步提高。以下将从政府角度以及一些典型行业出发，分析 AI 数据开放和协同中存在的问题。

5.2.1 政府角度分析

在政府开放数据领域，主要存在下述问题：

（1）地方政府缺乏“开放”授权意识

地方政府在数据开放中很少提及数据授权，有些地方政府仅提倡向企业开放数据，或者以交易形式来开放数据，导致普通大众很难获取开放数据。

（2）地方政府对“开放”的定义不统一

国务院的《促进大数据发展行动纲要》并未明确其所提“公共数据资源开放”的标准，不同机构部门缺乏统一共识，按照各自的理解开放数据。

（3）地方政府开放数据缺乏统一顶层设计

目前只确定中央层面的“国家政府数据统一开放平台”，但缺少统一性的顶层设计方案，并且地方政府层面的发展规范，无法统一指导各地政府具体的数据开放行为。

（4）各地政府的数据开放平台难以互联互通

各地政府缺乏统一、开放、易用的数据，数据开放平台缺乏交互性、人性化设计，开放数据的流通性很差，容易成为“睡眠”数据。

（5）开放数据格式混乱

部分地方政府对于开放数据的管理比较混乱，各部门并未对开放数据的采集、存储、治理、消费、质量等做出统一规范，导致同一平台上不同部门提供的数据格式差异较大，数据之间横向交互和协作整合能力匮乏，无法实现更大规模的数据共享与互联，影响数据应用价值的体现。

（6）缺乏政府数据开放的应用支撑

数据应用是政府开放数据的建设目标，但由于应用的高度专业性，政府作为行政管理部门，在技术和市场方面都无明显优势，难以制定符合需要的数据应用。

（7）地方政府未区分开放数据和开放信息的关系

开放数据强调的是原始数据的开放，是客观的，无意义的符号形式，而开放信息是被加工和赋予意义的数据和事实，二者的管理方式存在显著差异。开放信

息一般用于满足公民的知情权，而开放数据应便于公众的获取、加工以及传播和再利用，方能释放开放数据的价值红利。目前，政府的开放数据大多存在于政府信息公开栏目中，从本质上属于开放信息，而目前地方政府对开放数据仍如开放信息一般管理和限制，阻碍开放数据发挥其真正价值。

鉴于以上问题，政府应为公共部门制定更加具体的且有实质性的开放数据战略，并提供支持采用的工具和指导。我们建议政府在公共部门的开放数据战略中采取如下的措施：

- 给出公共部门数据的一个通用定义并制定出不能公开数据部分的准则；
- 在数据采集和处理的所有步骤中默认数字化信息；
- 支持下游数据使用的信息生命周期；
- 确保不能公开数据的安全性、隐私和机密性。

政府公开数据政策应遵循公共部门数据重用的可用性、可负担性和透明度的一般基本原则，通过这些原则来指导和实施具体的数据开放操作。

（1）可用性：所有公共部门的数据，包括公共部门企业的数据，都应该可以重新使用。但是，出于国家安全原因，涉及执法、商业机密、个人数据保护、法律禁止披露信息、第三方拥有知识产权时，可能会限制访问某些数据。

（2）可负担性：公共部门的数据应该可以重复使用而不收费。如果有必要的话，原则上应该限于边际成本。

（3）透明度：重新使用公共部门数据的条件应该公开，不得歧视。作为规范，不应该允许排他性安排。如果需要的话，应该定期审查安排的原因。

5.2.2 医疗行业分析

健康数据一般是指民众在院外自行测量的数据，其准确性和可追溯性较差，在临床上一般不作为诊断依据使用，可用于健康状况评估、健康趋势预测等。医疗数据一般是指在院内专业人员、标准环境、使用计量可靠并经常检查维护的仪器所采集的医学数据，可用于临床诊断。不管是健康大数据还是医疗大数据，都是具有高附加值的信息资产，通过对海量、来源分散、格式多样的数据进行采集、存储、学习和挖掘，可以从中发现新知识、创造新价值、提升新能力，从而进一步推动健康医疗服务产业。因此，健康医疗大数据的发展关乎国计民生，具有重

大的战略意义。目前，医疗领域开放大数据存在如下的问题：

（1）健康医疗大数据的共享和开放程度不高

在目前的医疗体制下，医疗卫生机构很难有动力去共享健康医疗数据，医疗卫生机构和医疗卫生机构之间、医疗卫生机构和社会公众领域之间，均存在不同程度的数据壁垒。数据孤岛效应一方面造成了患者数据重复采集和医疗资源浪费，另一方面也阻碍了健康医疗大数据的系统性开发和建设。

（2）健康医疗大数据领域的法律体系不完善

根据卫健委和药监局的监管观点，患者数据的所有权在患者方，患者有条件分享给院方获得医疗服务用。在人工智能领域，患者的健康数据是极其有价值的，可用于科研和教学中，帮助医院的信息化系统提升智能。在这种情况下，患者往往不具有知情权或难以使用知情权。如何建立相关法律法规介入人工智能对患者数据的使用，是目前亟需研究和解决的重要问题。

（3）健康医疗大数据来源多样性

健康医疗大数据来源较广泛且质量不易控制，有大概率出现不可及的碎片化数据、可及但错误的数据，可及正确但残缺、无法修补的数据。同时，原有自给自足的数据治理方式，解决不了数据共享问题，亟需新的数据治理方式。

（4）健康医疗数据敏感性高

健康医疗数据涉及患者主体及其隐私问题，在使用前必须要进行数据脱敏。常用脱敏策略是匿名化，无法回溯到具体患者，但是通过关联分析等方法，某些患者信息还有可能被泄露，因此开放健康医疗大数据时有必要采用更完善的预处理方法，避免可能引起的法律纠纷。

（5）医疗数据使用的不平等性

民众更倾向前往三甲等大型医院治病，导致医疗数据集中在少数知名医院或机构手中，这些医院或机构享有行业话语权。由于医疗数据的重要性，通过挖掘和分析医疗数据，提高自身医疗水平，进而保持其行业领先地位。对于这些机构来说，医疗数据还包含诊疗方案、药剂配方、操作流程等敏感信息，其它机构获取后会影响其技术优势。因此，这些充分享有医疗数据红利的机构不愿意公开自己的红利之源，即使按照上级主管部门的行政命令进行公开，也仅是一些价值不高的常规指标（如体检等）或统计数据。这种医疗数据分布的不平衡性，导致中

小型医疗机构无法获取足够有效的数据进行研究，使它们和大型医疗机构的技术差距越来越大。

（6）不同医疗数据格式不统一

医疗开放数据种类繁多，涵盖医学影像数据、结构化表单数据、早期的手写病案以及混有手书的半结构化数据。从文件格式上来说，医疗开放数据主要包括 XML、RDF、CSV、JSON、TXT、DICOM、VTK 等，这些复杂的格式类型，加重了医疗开放数据的处理负担。

（7）数据计算模型在健康医疗上不具有可解释性

健康医疗行业对于人工智能辅助诊疗生成的结果有时需要可解释性，以便从业人员能方便的观察模型产生的判定基于哪些因素，是否与医学常识和经验相吻合。当前人工智能的不可解释性，在健康医疗上意味着危险和不可预测性，对健康医疗这样的特殊领域有时较难容忍。因此，哪些应用场景可以有效使用人工智能，而哪些不建议使用人工智能，需要得到充分关注。

另外在医疗领域，也要关注到患者的需求。一方面，患者隐私容易受到侵犯，医疗数据的最大特征是涉及患者主体及其隐私问题，这些信息一旦泄露，不仅违反患者信息保密性的规定，引起法律诉讼，也破坏患者对医生的信任。在某些情况下，这些数据被相关公司企业获取后，会根据患者情况频繁发送广告，给患者生活带来烦恼。另一方面，患者无法享有数据开放的红利，医疗数据开放给相关企业后，这些企业通过分析处理，建立疾病模型，牟取高额利润。但作为数据源头的患者，并没有获取自己出让数据的报酬。因此，有些民众认为医疗数据是个人财产，相关机构无权“直接征用”，反对企业以此牟利，对于采集数据的相关医院，也存有怨言，从而加重医患矛盾。

5.2.3 金融行业分析

由于金融行业的性质，为了符合监管要求，金融机构在金融相关数据的数字化和获取方面已经非常先进，尤其是对数据的准确性和完整性的要求很高。同时随着大规模高并发的网络金融业务及渠道类应用交易日益普及，金融行业机构基于海量数据的处理、分析需求不断增多，这些海量数据对 AI 在金融行业的不同应用会对金融行业带来革命性的变化，而开源技术能够帮助金融行业机构实现 IT

系统的敏捷高效、精细化管理、可扩展以及可管可控，解决其痛点问题。在金融服务业部署人工智能的应用并不缺数据，但仍要面对以下三个主要挑战:

（1）如何利用数据

利用数据的关键是如何开发良好的数据质量治理，这是人工智能充分发挥其潜力的关键。

（2） 部门内部的数据仓库孤岛

无论金融机构是零售银行、财富管理还是投资银行，部门之间的数据共享和聚合对于有效的人工智能应用程序都是至关重要的。

（3）遵守数据驱动应用的监管要求

数据驱动的人工智能应用常常面临监管机构强加的未知或模糊的限制。欧盟的通用数据保护法规 (General Data Protection Regulation, GDPR) 就是一个例子，该法规要求向客户提供有关其算法决策的“解释权”。当前基于机器学习和深度学习的人工智能系统的本质是：该算法是一个黑盒子范式，因此无法解释。然而，金融机构被卷入其中，不得不遵守 GDPR，但技术上实现不了。一种可以考虑的方法是定下一个“人工智能责任框架”，作为人工智能设计和实施的指导和原则，并制定到内部政策和程序。

5.2.4 交通行业分析

随着车路协同理念的出现，基于车联网、大数据以及云计算的交通信号控制系统，可以对道路系统中的交通状况、交通事故、气象状况和交通环境进行实时的监视，依靠先进的车辆检测技术和 AI 技术，获得有关交通状况的信息，并根据收集到的信息对交通进行有效规划，重新定义交通控制。由传统固定配时信号控制到感应式信号控制，再到车路协同环境下的交通感知与控制，完成了从宏观到微观、从路权粗放式管理到道路资源全时空精细化分配的进阶。智能交通系统在交通控制行业得到越来越广泛的运用，如信号灯控制、发布诱导信息等，乃至根据手机定位、微博留言等数据对于交通系统的性能进行评估和调整。

目前，随着城市化的发展和交通设施的快速建设和升级，交通疏堵、应急指挥系统、辅助决策系统以及一大批服务于各个专业的信息管理系统的落地，形成规模巨大、类型多样的交通数据。这些数据相对于传统数据而言具有量大、分布

广、结构复杂和数据维度高等特点，而现有的数据处理和数据分析技术难以满足实际需求。目前，交通领域开放大数据存在如下的问题：

（1）缺乏统一标准和技术规范

智能交通系统项目的建设先于行业统一标准的推出，在缺乏标准的条件下，许多地区的智能交通系统自成体系，缺乏应有的衔接和配合，标准互不统一。

（2）数据系统可靠性与稳定性有待提高

交通系统复杂度和整合程度较高，涉及底层计算资源调度、分布式数据存储、流式计算和批处理计算融合、二/三维一体化数据管理等复杂系统组件的整合，而系统的健壮性并未同步提高，存在牵一发而动全身的制衡问题。

（3）数据存储与交换同步不成熟

交通领域数据底层存储主要以 Oracle、DB2 等关系型数据库为主，但随着智能交通的发展，产生了较多音视频、图像、地理信息等非结构化数据，数据存储使用到图形数据库 Neo4j、文档存储型数据库 MongoDB、GIS 空间数据库等 NoSQL 数据组件，这些组件之间的数据采集、实时同步等数据传输与交互难以实现。

（4）数据存储中心不规范

汇集的大量数据需要在高安全管理标准的数据中心进行存储，而交通系统中原有的数据存储中心的不规范会威胁现有的存储和安防措施。

5.2.5 物流行业分析

数据化是物流向下一代升级，真正实现智能物流的关键转型，虽然当前物流企业在数据获取方面存在天然劣势，但可尝试与第三方企业合作，及早在物联网、大数据、人工智能等方面积累技术，积极对现有信息系统换代升级。目前物流行业智能化面对两个很大的数据整合和应用的挑战：

（1）物流行业数据碎片化

在物流行业能够有效利用新一代的人工智能科技如机器学习和深度学习之前，其最大的挑战之一是供应链上数据的碎片化。在物流业中有许多参与者，包括供应商、船运公司、仓库和最后一英里交付服务等，并且每个参与者仅在该过程的一部分具有可视性和记录数据。

这意味着，不像京东、阿里巴巴、亚马逊，物流公司在整个供应链上没有一

个单一的可见性/单一版本的真相。为了使数据真正有效，对于物流公司提供最好的端到端的传递服务，现有的数据差距必须解决。如果不是这样的话，物流公司就无法与京东、阿里巴巴、亚马逊这样的公司竞争，而且会遭受巨大的效率损失，影响到客户满意度和底线。

事实上，就在不久前，很多物流企业包括世界最大的集装箱航运公司马士基披露了它对亚马逊和阿里巴巴等电子商务公司侵犯业务的担忧。为了克服这些问题，物流公司应该尝试与供应商谈判，建立合作，并让他们分享他们的数据。在很多情况下，供应商知道他们将要运送什么，但是他们大多数不关心或觉得很麻烦与运输者共享数据。所以物流公司需要提供优惠，例如折扣，来说服供应商分享他们的数据，虽然需要付出成本，但是得到对供应链更加有深度的可见度是值得的。

（2）物流数据不一致性和数据低质量

物流行业从供应商向消费者运输货物的服务商的数量很大，数据常常是不一致的，并且数据的记录方式是无组织的。不同的服务商记录他们的数据在不同的系统，有时甚至用不同的测量单位（重量、尺寸、度量/帝国等）。因此，要把整体供应链的数据收集、集成和应用起来，面对的挑战很大。所以，物流行业有必要考虑和研究如何把整体供应链的数据标注化、如何开发和协同才能有效地做数据分析，洞察和部署到人工智能系统做机器学习和深度学习。但物流行业数据不仅不一致，而且质量低、不准确。以这个例子为例，许多物品上的条形码与计算机系统中的条形码不匹配，导致货物进入仓库的速度远远快于他们外出仓库的时间。

为了解决这个问题，物流公司可以发展和遵守一定的标准来记录数据，以避免使用不同类型的测量，并便于数据分析。2007 年发布的物流互联互通模型（LIM - Logistics Interoperability Model）是一个可以考虑的标准。另外公司可以彼此同意数据共享特定规则，建立平台来共享数据和确保数据的准确性，同时也和学术界、工业界和各国政府共享其数据并共同工作，把数据标准化。很可惜行业内的公司还没有开始认识到标准化模型的网络效益，许多公司没有遵循，只是记录他们认为合适数据。

如果物流行业要能够参与下一轮人工智能的技术革命和推动智能化的产业

升级和转型，就必须解决整个物流产业生态圈和参与者之间的零散数据开发和协同的问题。

5.2.6 制造行业分析

当前，我国还处于促进制造业智能优化升级的探索阶段，对大多数企业而言，能够自我感知、自我记忆的数据采集感应系统尚未建立，分析复杂数据结构的数据处理技术仍需优化，高效的数据库维护和管理机制还需完善。工业大数据具有自己独特的特征，是指在工业领域中，围绕典型智能制造模式，从客户需求到销售、订单、计划、研发、设计、工艺、制造、采购、供应、库存、发货和交付、售后服务、运维、报废或回收再制造等整个产品全生命周期所产生的各类数据及相关技术和应用的总称，其以产品数据为核心，极大延展了传统工业数据范围，同时还包括工业大数据相关技术和应用。

工业大数据不完全等同于企业信息化软件中流淌的数据，从业界的共识看，主要来源有三类。第一类是企业经营相关的业务数据，这类数据来自企业信息化范畴，包括企业资源计划(ERP)、产品生命周期管理(PLM)、供应链管理(SCM)、客户关系管理(CRM)和环境管理系统(EMS)等，此类数据是工业企业传统的数据资产。第二类是机器设备互联数据，主要是指工业生产过程中，装备、物料及产品加工过程的工况状态、环境参数等运营情况数据，通过 MES 系统实时传递，目前在智能装备大量应用的情况下，此类数据量增长最快。第三类是企业外部数据，这包括了工业企业产品售出之后的使用、运营情况的数据，同时还包括了大量客户、供应商、互联网等数据状态。

工业大数据的应用，主要是利用大数据推动信息化和工业化深度融合，研究推动大数据在研发设计、生产制造、经营管理、市场营销、售后服务等产业链各环节的应用，研发面向不同行业、不同环节的大数据分析应用平台，选择典型企业、重点行业、重点地区开展工业企业大数据应用项目试点，积极推动制造业网络化和智能化。

在工业领域实现数据开源，主要需包括以下四方面工作：

（1）工厂需按照智能制造的相应标准，采用先进的传感技术、通信技术、物联网技术等，完善自身的数据采集系统，获取大量原始数据。大部分工业数据

集中在工厂内网中，要让工厂和社会都接受互联互通的网络，学会应用，并把理论上的技术转为现实的运用并且使之透明化。

（2）工业数据通常为结构化数据，其存在低质性、碎片化和隐匿性的特点。工业数据的质量较差，对预测和分析结果的容错率很低，而且通常缺乏完整的正常工作和故障模型，此外工业数据模式可能极为短暂，具有隐匿性的特点；普通商业数据可以通过大量的数据来弥补数据的不良品质，例如通过插值等技术手段补全数据，然而对于工业数据来讲，各个指标均有明确的物理定义，数据的完整性非常重要；不全或者错误的记录都将影响各变量之间的关系，对预测的准确性有致命影响；工业数据常常存在数据不均衡的情况，限制了算法的应用。因此对工业数据进行数据清洗以及预处理十分重要，在结合领域知识的情况下，为数据科学家、研究者等提供相应的信息，才会利于对工业数据的后续分析。

（3）数据匿名化/脱敏的标准化。考虑到工业领域数据安全和保密的要求，例如制造业中工艺流程与配方数据是企业的核心竞争力，一旦开源将会对企业产生不可逆转的损失。针对重要工业领域，需要制定数据匿名化/脱敏的标准，在保证企业数据和商业机密安全的前提下，生成开源数据，助力科学研究和多领域合作以提升生产力。 相应的，数据匿名化标准也可以推动数据开源的市场推广，赢得企业的支持。另一方面，工业数据大多具有明确的物理意义，而这些物理意义在领域知识的指导下，能够对数据分析结果产生极大的影响，如何能够在保证数据安全的情况下尽可能保留在数据分析时的关键信息成为一项重要的任务，而数据匿名化/脱敏的程度则需要在开源需求和客户要求之间取得平衡。

（4）数据采集的接口的标准化。由于开发时间或开发部门的不同，企业中往往有多个异构的信息系统同时运行，这些系统之间的数据源彼此独立，相对封闭，使得数据难以在系统与系统之间、企业与企业之间进行交流、共享和融合。为解决这一信息孤岛的问题，就需要考虑针对历史数据、实时数据和异构数据分别建立采集接口、通信协议的标准。对数据进行有效的集成管理将成为企业增强竞争力必然选择，同时也将有力加强工业数据开源的可操作性。

5.2.7 教育行业分析

在教育领域，人工智能技术的关键应用之一是为学习者提供高度个性化和有

效的学习体验。因此，对学习者及其学习过程进行监控、理解和评估，对引导学习者在无缝的、有趣的学习体验中走向正确的学习路径和学习内容是至关重要的。然而，部署这些人工智能应用程序的关键挑战之一是学习数据的可用性，尤其是有不同类型的学习数据，包括学习过程数据和学习评估数据。

首先，对于学习过程数据，由于大多数学习都是离线完成的，数据经常会丢失而没有被收集。对于实施 LMS (Learning Management System)的机构，它们只能收集和跟踪一些人工收集的有限的数据，通常所学习的内容和学习进度数据的关系并没有很好的集成。其次，对于学习评估数据，正式的测试和考试是被打分和收集的，它们通常是手工完成的，或者只是获得最终的数字结果呈现。对于非正式的评估数据，比如测验和模拟练习，这些数据对于人工智能机器学习非常有价值，但它们一般没有被收集。

为了涵盖人工智能在教育中面对的数据挑战，我们应该利用不同类型和不同形态的人工智能技术，首先从学习者那里无缝地收集和筛选学习数据，然后使用人工智能从这些数据中学习，从而为学习者建立个性化及高效率的学习体验。

5.2.8 石油行业分析

AI 开源技术在石油领域主要应用于智能勘探、油井监控、生成控制等。目前，石油行业的地质数据早已达到 PB 级以上，人工智能、大数据分析已经率先成功应用在勘探开发领域，随后在管道运输、炼油化工及成品油销售领域开始发挥作用。虽然 AI 开源技术在石油行业的应用逐渐增多，但是也涌现出不少问题：

（1）数据基础和环境不成熟

石油行业相对封闭，支持 AI 分析的数据技术基础和环境还不够成熟，现有的数据采集系统分布杂乱无章，重新布线投资过大，PLC 较慢且不稳定。

（2）数据库适配性差

传统的关系型数据库、嵌入式数据库被广泛的应用于油气管网调控、汽油消费分析等领域，由于数据结构和标准的不同，数据读取时需要适配不同的数据库，将会耗费大量的人力资源。

（3）缺乏共性标准

各种开源软件之间在数据采集、指标口径、分类目录、交换接口、访问接口、

数据质量、安全保密等方面没有关键共性标准。

（4）缺少统一业务平台

现有的开源框架仅被用于单个业务领域的分析挖掘，尚未出现统一的大数据分析平台支持安全生产统一研判、生产安全平稳运行、下游精准销售与客服等各业务领域及全产业链的整体优化。

（5）数据整合困难

实时数据整合困难行业特定数据规范和业务规范复杂且不统一，石油行业信息化长期处于自行发展，存在大量的各类型实时数据库，后续数据整合存在标准难统一、瞬时数据量巨大、与业务数据整合困难的问题。

5.3 AI 数据开放和协同的可行性

数据和代码之间没有严格的分界线。纯自然的信息，往往需要投入人工注意力后才能变成可用于训练模型的数据，这些数据同代码一样，都蕴涵了人的知识和劳动。且数据和代码一样都不具备好的交易属性（容易被拷贝和转卖），可以考虑把商业数据包装成远程的模型训练沙箱，以改善交易属性。

与目前比较流行的数据竞赛类似，数据格式和示例小数据集是公开的，主要区别是把训练过程移到数据所有方的模型训练沙箱中进行，数据使用方不能直接接触数据，但可以拿回训练完成的模型。基础数据共建共享的主要障碍是无法给予数据提供者足够的合理回报，建议采用区块链或其他数据保护技术，让数据提供者开放数据后，仍能保留对数据的所有权、收益权和转让权。

由于 AI 开源数据来源的多重性，数据可能来自于政府、企业、高校、科研机构、个人等不同方面，其开放性需要合理、合法、公正公平的有效保障，才能保证 AI 开源各相关方的利益，主要应从开源数据的顶层设计、法律法规、数据治理、开源数据平台建设等方面进行考虑。

5.3.1 顶层设计

国家应对 AI 开源数据进行顶层设计，应从国家层面进行整体规划，如进行 AI 开源软件及数据相关的实施战略，明确和推进 AI 开源数据的方向、领导组织、技术路线图。确定数据开源过程中利益相关方的角色、职责和协同发展框架，明

确 AI 开源数据的发展生命周期，确保明确生命周期过程中的各个节点和关键过程，进行系统性的设计和引导；通过政府、企业和社区等组织联合推动，建立开源数据制度和环境，通过创新应用推动发展，如政府或企业根据自身资源情况投入基础资源，通过孵化器、人才挖掘、数据使用、政策引导、创新应用竞赛等方式推动各方参与，形成数据开源与创新的生态环境；进而吸引监管部门、社会资本、科研机构、企事业单位的创新网络，聚集人才、资金、基础设施等，营造开放共享的文化。

5.3.2 法律法规

随着人工智能的快速发展，国家已经制定了一系列的发展规划和指导方针，但是在开源软件和数据方面还没有出台相关的法律法规，法律法规的不健全将可能是阻碍 AI 数据开放力度的重要因素。因此应尽快出台相关的法律法规，明确数据开源的标准、责任主体、数据泄露、安全和隐私包含等一系列的问题，如开放哪些数据、开放数据的分级分类标准，什么时候开放、开放频次、数据质量要求、数据开放流程等方面的要求和规范。在数据开源前应明确开源数据的标准和质量、数据格式要求、数据审核机制等，分析供需关系，在保障安全和隐私及合规的条件下依据申请免费或付费进行数据交易，并进行跟踪记录，同时进行阶段性的数据开源情况评价和审计，确保数据开源流程的合理使用和合法合规。

为了提供更好的法律基础来支持数据分享，Linux Foundation 也发起了 CDLA，也称为社区数据许可协议（COMMUNITY DATA LICENSE AGREEMENT）。它是一种专用的协作许可证，用于在个人和组织之间公开、共享和使用数据，且提供了一个法律框架来确保数据的知识产权管理。CDLA 许可证有两种：CDLA 共享许可证（CDLA-Sharing）和 CDLA 许可许可证 （CDLA-Permissive）：

- CDLA 共享许可证规定，如果有人共享他们的数据，下游接收者可以使用和修改该数据，但是下游接收者有义务共享他们对数据的更改或修改。
- CDLA 许可许可证规定数据发布者允许任何人使用、修改和做他们想要对数据的应用，而不需要承担任何共享他们对数据的更改或修改的义务。

在数据开放与分享的过程中，各相关方可基于自身考虑采用这些通用的数据使用许可协议来保证各方的法律权益。

5.3.3 数据治理

数据开源的过程中，涉及了组织、流程和工具等内容，可通过技术和管理相互结合的方式进行统一的管理，在这个过程中涉及的相关方主要有数据开源的统筹者、数据收集方、数据所有者、数据提供方、数据利用者、数据开源者、数据开源管理者等。他们各有定位并相互关联，只有形成协作关系，才能落实好数据开放的方向、流程，以达到较好的成效。同时应对开源组织进行治理，在开源的过程中可从开源组织的战略、组织架构、数据架构、数据生命周期、安全隐私、元数据、主数据、参考数据等范围进行管理。在评价阶段应综合考虑开源数据使用的促成因素、成熟度和审计等关键评价要素。例如，组织在数据开源过程中，应明确数据开源指导委员会，组织协调各方资源，设置开源数据治理工作组，界定数据的边际、质量、标准、使用规范等一系列的内容，并设置相关人员职责进行统一管理等，如设置首席数据官、数据开源科学家等相关职责，协同发展，形成良好的开源数据治理文化。

在数据治理，需要考虑的标准维度包括角色（Roles）、格式 (Data Format)、质量(Data Quality) 、集成 (Data Integration) 、权限 (Data Access) 、流程（Data Process）、受益 (Data Benefits)等。

5.3.4 开源数据平台建设

随着大数据的快速发展，以阿里、腾讯和百度为首的互联网公司积累和存储了大量的用户数据，数据涉及面广泛，但很多数据由于政策、责任、隐私等方面的原因，并未完全开放，造成了很多数据资源的浪费。因此，应从国家层面制定一系列机制来进行开源数据平台的建设，联合政府部门和社会组织共同建立国家、行业及不同区域的开源数据平台。在保障数据安全的基础上，打破相关壁垒，并建立开源数据从收集、整合、传输、存储、开源到销毁等阶段及对应系统、规范的管理机制，从而促进开源数据生命周期的健康发展。可以采用如下措施：

（1）建立数据基础、数据交换、数据安全、数据服务等方面的数据治理体系，规范平台建设和发展。

（2）建议使用图计算、TensorFlow等开源技术和框架进行开发，为开源平台

和社区建设提供更加开放的环境，确保资源的协同互补。

（3）应鼓励企业在国外优秀开源框架和技术的基础上进行自主创新，实现相对可控，并在保护知识产权的基础上分级分类公开使用，如可先开放责任主体划分清晰的公共数据，涉及个人隐私及商业秘密的应进行授权和监管后进行使用。

5.4 潜在解决方案

为支持新一代数据驱动的人工智能的需求，对目前传统以代码为核心的开源理念需要做进一步升级，主要是延展和覆盖到其他数字内容，包括数据、算法、知识库等领域。目前可以考虑到四种可行性技术架构，分别是中心化模式(Centralized Model)、混合型模式 (Hybrid Model)、去中心化模式(Decentralized Model)以及没有初始数据的引导模式（Bootstrap Model）。四种都有相对优势和劣势，可以参考以下对比分析表：

表 1 人工智能数据开放和协同模式对比分析

数据开放和协同模式	优势	劣势
中心化模式 Centralized Model	1、可以高效统一、收集和集成数据； 2、运营模式和技术很成熟，运营风险低尤其是确保数据安全和隐私。	1、需要强有力的中心主导权和管治能力和体系，尤其是数据标准，也可能会需要很长时间达到共识，造成无法落地； 2、数据贡献者没法把控数据如何被分享和应用，也没法受益； 3、数据被中心组织垄断，影响到行业创新。
混合式模式 Hybrid Model	1、可以把在不同数据中心的孤岛数据利用起来； 2、不同数据中心对数据安全和隐私保持把控。	1、数据范围限制在数据中心里的数据； 2、数据中心需要符合共同的数据标准和技术架构。
去中心化模式 Decentralized Model	1、可以开放给大众参与和贡献数据，也可以受益于数据的贡献； 2、大规模从下至上的数据收集可以提供更全面的数据。	1、可行、合理和高效的共识机制有待建设和验证； 2、区块链技术、性能和数据安全还在初期发展阶段，有待验证，也一定的技术风险。
没有初始数据的引导模式 Bootstrap Model	1、在面对没有初始数据的情况下是唯一可行的方案，让没有数据的企业可以收集和应用到需要数据的深度学习和机器学习的人工智能技术； 2、借鉴不同代的人工智能技术可以实现不同智能功能来解决同一个问题，也可以形成互相监控的价值。	1、需要建立 2 个阶段的系统和引用 2 种不同的技术，所以实现落地会比较复杂； 2、实施时间和成本都会比较高。

5.4.1 中心化模式

中心化模式可以由政府主导，从上到下驱动；可以由政府来推动、管理、运营和管制，并利用市场力量从行业生态圈来推动；可以由行业推动、管理、运营和管制。

对于政府开放数据，可能的解决方法包括：

（1）完善开放数据的政策与标准

首先，完善开放数据的标准与质量，包括数据的格式要求、数据开放审核机制、安全保障和隐私保护、阶段性评估和考核等方面的相关制度。然后，建立“依申请开放”机制，进一步打通需求侧和供给侧，提供具有个性化的数据开放，基

于应用需求推动开放供给。

（2）搭建政府数据共享统一门户

建立中央、省、市各级的统一数据门户，为各级政府构建“一站式”数据管理平台；通过多样应用格式、多种录入检索方式、多类信息条目等手段提高政府收集、整理、发布数据的管理效率；为政府建立起与公众间的线上交互平台，为公众获取、交流、再利用政府数据提供渠道，也为公众充分参与政府事务提供可行路径。

（3）按照统筹在前的原则开放数据

在开放数据之前，应当统筹好中央层面和地方层面纵向数据关系、统筹好不同部门机构之间的数据关系、统筹好开放与保密之间的关系。

（4）建立依据数据价值分级管理机制

一方面，对于气象、公共设施、交通等公众关注度高又不涉及个人隐私、商业秘密的通用性数据，开放后不易引发争议，政府责任容易厘清，一般无须干预。另一方面，对于市场监管、地理空间等可能涉及个人隐私和商业秘密的数据和一些时效性要求较高的数据，需要政府投入更多的数据管理成本，可以在数据开放和使用中加强授权和监管，纳入依申请开放范围，或者通过合作形式引导企业加强数据安全工作。

（5）提供政府数据开放平台的增值服务

通过多种增值服务提升政府数据开放平台的价值，常用增值服务包括：基础服务（基础分析技术、数据清洗技术、可视化技术等基于云计算的公共基础设施）、应用服务（政府数据大赛项目、社会征集项目等）、网络伙伴服务（开发者和政府、投资机构、企业、研究机构等服务团体合作的项目等）。

对于行业开放数据，以医疗行业为例，可能的解决方法包括：

（1）改进医疗数据脱敏方法

除了匿名化之外，采用新的医疗数据脱敏方法，主要包括：扰动（对原始数据进行变换）、泛化（用合适范围内的新值替换原值）、抑制（删除患者标识信息）、K-匿名（K 条记录的准标识符属性值相同）、聚类（将数据分为若干簇，对每簇数据进行扰动和泛化等操作）、分布式保护（将大型数据集沿水平或竖直方向进行分割）。

（2）制定完善的患者隐私保障制度

首先，提高患者对自身医疗数据的知情权，包括原始数据、脱敏数据、数据流向等；其次，制定医疗数据分级管理制度，确定哪些数据可以开放，哪些不可以开放，对谁开放等；此外，制定医疗数据追溯管理制度，对于已经开放给相关机构的数据，要定期跟踪数据的使用情况，以及患者的反馈意见，确保按照相关机构按照协议约束正常使用，避免违规使用。

（3）制定医疗数据开放的互惠制度

对于愿意开放自身医疗数据的患者，医院或相关机构应给予医疗费减免或其它补偿措施；根据患者开放自身医疗数据的程度和次数，建立相关的积分制度，根据分值享受医院或相关机构的后期健康服务。

（4）建立医疗数据平等使用制度

政府应通过资金补贴、政策倾斜等措施，积极鼓励和推动大型医疗机构开放自身数据，鼓励大型医疗机构和中小型医疗机构通过合作形式共享医疗数据，制定合理的分工措施，发挥不同机构的优势，弥补大型医疗机构人力紧张和中小型医疗机构技术匮乏等缺陷。对于和技术细节相关的敏感数据，要建立识别和保护制度，确保大型医疗机构的在科研原创领域的技术优势；同时，对于已经泛化和脱敏后的数据，以及经过临床验证的有效研究结果，可以向其它医疗机构公开，促进技术传播，提高整个社会的医疗水平。

（5）制定医疗数据开放的标准

从国家层面，一方面参考国外医疗数据开放的成功经验，另一方面选择规模较大的几家公立医院作为数据开放试点，发现难点和问题及时处理，同时根据实际经验提供相关准则和最佳做法。从医院的层面，可以依据国家标准和自身情况，对上传数据的格式和内容等进行定期检查和评估。

5.4.2 混合型模式

混合型模式的基本概念是企业各自建立独立去中心数据中心，将数据进行交换和共享，但同时可以采用技术手段来保证数据所有权及其隐私防止数据泄露。

在 B2B 的场景下，此方案特别适合，尤其是企业或组织自身具有大量有价值的数据时，可以让企业建立数据联盟共享和协同数据，引出数据的价值，同时保

持数据的安全和隐私。在落地实现过程，可以根据不同行业的场景及需求，定制化引用混合模式来实现中心化和去中心的平衡。

在各参与方之间的互信和技术手段基础之上，可建立起联盟制的数据交换合作机制。各参与方既可以是数据的拥有者，也可以是数据请求方，或者两种角色兼有。各数据拥有者在本地存放原始数据，对于自有数据享有完全的控制权。对外不提供直接的数据访问入口，只在本地计算并返回结果数据。

在此模型中，数据交换中心功能负责管理、监控和调度运算请求。任何来自数据请求方的请求，在审核通过之后将发送给各数据提供方的数据中心进行分布式计算，并统一搜集并汇总运算后获得的结果，如各类知识或者统计类的汇总数据等。所有原始数据并不会存放到数据交换中心功能，确保原始数据的私有性。

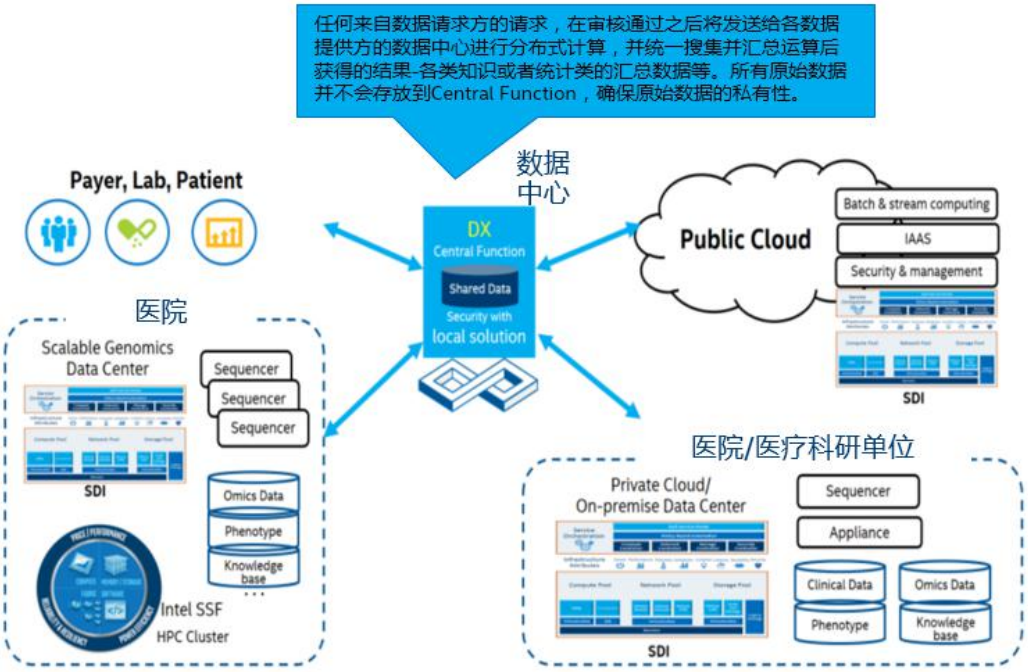

图 17 典型的联邦制数据交换机制

5.4.3 去中心化模式

去中心化模式的基本概念包含以下三个内容：数字内容 + 开放生态圈 + 电子智能合约（Digital Content + Open Ecosystem + Smart Contract）来实现数字内容在开放生态圈的协同管理和管制。例如，建立在区块链分布式平台的开源方案，在 C2B 的场景下，此方案的目的就是建立一个公共消费群体来支持数据分享和协同，可以搭建一个协同平台，从行业生态圈来推动。为了要解决在人工智能领域

更大范围的需求和落地挑战，尤其是数据、算法、知识库等，建议重新思考开源体系，以区块链的去中心化应用 （Decentralized App 或是简称为 DApp）为基础来支持和促进下一代人工智能产业的发展，命名为 “开放生态圈 DApp” (Open Ecosystem DApp) 简称为 OED。

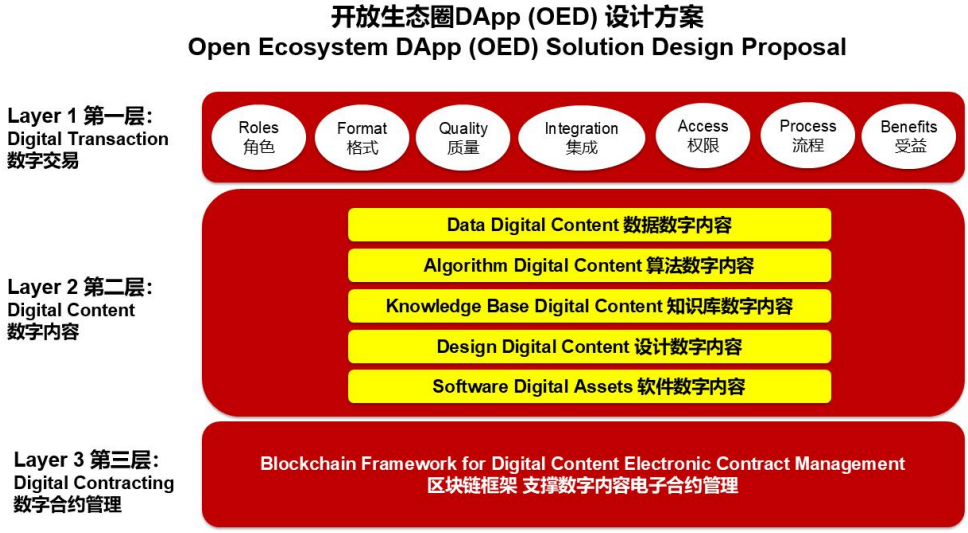

图 18 OED (开放生态圈 DApp) 的技术设计方案

要引出 OED，可以考虑和尝试以下 5 步战略规划：

（1）概念研究 (Ideation)：基于支持人工智能产业发展的战略目的，加速 OED 理念和理论的研究以白皮书为基础，推广给政府、企业、研究所、大学、孵化器层面参与探讨和研究。

（2）平台孵化 (Incubation) ：同时成立孵化基金在全国不同孵化器分享、鼓励、引导和提供种子投资给初创企业，来搭建 OED 的平台技术、运营和应用。

（3）技术实现 (Technology) ：建立在传统软件代码开源机制的基础上（精神、哲学、社区管理、法律框架等），在已有的代码开源平台上研发出 OED 技术平台的解决方案。

（4）运营实现 (Operation) ：同时在传统软件代码的基础上 (包括精神，哲学，社区管理，法律框架) 为 OED 延伸出一套新的“数字内容的去中心化+有目标性的开放式协同”运营机制。

（5）应用实现 (Application) ：在不同行业，分享、鼓励、引导和提供种子投资给初创企业在 OED 的平台推动不同领域的“数字内容的去中心化有目标性的开放式协同”，并以数据内容为第一选择的数字内容。

图 19　OED (开放生态圈 DApp) 的战略规划

提倡 OED 会带来以下 4 大价值：

（1）AI产业蓬勃发展：为了支撑AI产业的发展尤其是数据和算法分享的瓶颈，OED可以建立一个完整开源体系来解决AI的独特需求（数据、算法、知识库）。

（2）数字内容的开源化：支持更开放的知识产权/数字内容的开源式协同。除了软件以外，加入其他的数字内容如数据、算法、知识库、设计资产等，促进生态圈的建设、协同和创新。

（3）自主创新：OED体现出一个宝贵的机会和转折点，可以推动中国企业和人才打造独特的创新技术和平台，从而支撑下一轮的产业智能化和数字化升级。

（4）国际开放合作：推动OED的创新发展， 可以帮助中国突破和超越目前以软件代码为基础开源的格局。中国可以更加有力的参与和影响在国际生态圈的开放合作，同时也支持升级国内知识产权的创新，贡献于国外知识产权的发展。

5.4.4 没有初始数据的模式

AI 应用和部署面临的挑战之一是缺乏足够的高质量数据来支持机器学习和深度学习的模型训练。在这样的背景下，仍然有两种可能的引导方法可以考虑来解决这个问题：

（1）通过自我竞争生成数据

在操作规则有限的情况，可以设想生成初始数据来引导建立小型的 AI 代理，

安排 AI 代理相互竞争，在过程中生成数据并用于训练和改进模型的数据。

（2）通过符号 AI 代理收集数据

在没有明确操作规则的情况下，开发基于人类知识的 AI 代理，尤其是第一代已被证明的人工智能技术如符号推理 （Symbolic Reasoning），这包括函数式编程（Functional Programming）如 LISP、逻辑编程（Logic Programming）和逻辑推理 (Logical Reasoning) 如 Prolog 或基于规则的系统、专家系统（Expert System）是一个很好的选择。虽然这些 AI 代理可能不是完美的，但它们被证明是有用的智能系统，能够驱动与用户的互动，并收集宝贵的数据给基于机器学习和深度学习的新一代 AI 代理。此外，这些 AI 代理还提供了一种非常有用的特性，其中推断的结果是“可解释的”，因此提供了一种有用的工具来互补基于机器学习和深度学习的 AI 代理，因为后者黑箱式的计算方式对推理出的结果是“无法解释的”。

第六章 AI 领域开源与标准的关系

6.1 开源与标准联动的案例

标准和开源都是促进产业发展的手段，在不同场景、阶段发挥着不同的作用。在分析 AI 领域开源、标准与其他产业形态联动之前，先分别介绍在容器、大数据存储文件格式和网络功能虚拟化三个场景下标准和开源联动的案例。

6.1.1 容器

容器是将应用及其整个运行时环境（包括全部所需文件）一起进行打包隔离的技术，让应用在不同环境（如开发、测试和生产等环境）之间轻松迁移的同时，保留全部功能。近些年来容器技术异常火爆，各主流云计算平台都提供了容器服务。容器技术的火爆得益于 Docker 公司，自 Docker 公司开源了自身容器引擎方案后，随后短短几年内，行业涌现出大量基于 Docker 技术的方案和应用，Docker 有成为容器生态事实标准的趋势。

Docker 技术虽然开源，但是主导者还是一家公司。从商业角度来看，一旦这家公司自身的商业策略发生变化，那么对所有基于 Docker 技术提供解决方案的

公司都会带来巨大影响。从技术角度来看，其他容器引擎方案如果跟随 Docker 开源的技术，那么技术创新就会受到 Docker 公司的约束；如果不跟随 Docker，会使整个产业碎片化，实现不了最优化的协同效果。因此业界渐渐产生了担忧。2014 年，随着 Docker 公司的产品拓展到编排调度和操作系统，业界需要容器标准的呼声就越来越高。

OCI（Open Container Initiative）组织就是在这种背景下，由 Docker、CoreOS、IBM、微软、亚马逊和华为等公司共同牵头创立的。这个组织通过开放治理的方式来制定容器行业的标准，由 Linux 基金会负责运营。OCI 组织自 2015 年成立以来，先后推出了运行时标准、镜像标准、分发标准以及相关标准的测试认证工具。一开始标准是基于当时的开源实现提炼而来，但很快在社区就形成了“标准讨论——标准制定——标准测试项目完善——容器方案通过测试认证”的完备流程，标准和开源项目得到了很好的协同。OCI 标准的出现打消了行业的后顾之忧，越来越多的容器引擎方案也先后商业落地，比如 CoreOS 的 rkt 方案、华为的 iSula 方案等等。容器行业百花齐放，同时也加速了技术创新。

可见，云计算容器开源技术驱动了标准的产生，无论是从技术还是商业的角度，都更好利于 Docker 公司和提供容器引擎方案公司进行技术创新和商业应用落地。标准流程和开放治理方式的愈加成熟，推动了开源技术的创新与发展。

6.1.2 大数据文件格式

Apache 软件基金会（ASF）是目前大数据领域最权威的开源组织，其中的 Hadoop、Spark 已成为大数据开源的事实标准。国内也有开源方案转化成标准的案例，如 Apache CarbonData，该开源项目源自华为，团队成员依托多年的数据处理经验和行业理解，发现大数据领域缺乏一个能够支持多场景的高性能数据分析项目，于是创建了 CarbonData。

Apache CarbonData 是一种新的大数据文件格式，使用先进柱状存储、索引、压缩和编码技术实现更快速的交互式查询，以提高计算效率，将有助于加速查询超过 PetaBytes 数量级数据的速度。CarbonData 项目之所以从众多大数据项目脱颖而出，得益于项目的开源运作模式。作为大数据整体方案的一部分，开源模式提高了透明度，便于上下游产品团队参与到 CarbonData 的集成开发中，进而顺

利构建端到端的解决方案。大量社区开发者的参与增加了使用场景，吸引了更多社区用户和企业用户，从而进入良性循环，最后成功在 Apache 基金会孵化并毕业成为顶级项目。CarbonData 也被更广泛的应用在诸如金融、能源、工业自动化、运输等产业。

6.1.3 OPNFV（网络功能虚拟化）

以上两个案例共同特征是先有开源项目，随后通过大量的用户使用与认可成为事实标准，最终形成了相关标准，从而进一步推进了用户的广泛使用及产业生态的繁荣。而以下这个基于电信场景下 OPNFV 的案例，则是开源和标准的联动从分歧走向了融合。

电信产业的生态一直强依赖标准，而且往往是先有标准，再有各家实现，再有对接测试、认证，然后测试、认证的结果会成为客户采购的考量。但在电信网络云化及业务创新的趋势下，传统的“标准-实现-对接测试”模式为技术创新的快速迭代带来了挑战。

2014 年 10 月，由 AT&T、NTT、中国移动、Redhat、爱立信等厂商发起的 OPNFV（Open Platform of NFV）开源社区正式成立，立足为 NFV(网络功能虚拟化)提供了统一的开源基础平台。

OPNFV 社区成员在第一阶段简单延续了开源的理念，认为社区是靠“代码”说话，而标准的制定者不懂代码、不能给社区带来实质的贡献。此外，标准的演进过程太慢会影响项目进展。因此当标准组织(ETSI: 欧洲电信标准协会）给 OPNFV 社区提案时，社区是直接驳回。

但当社区成员真正开始做项目的时候，逐渐发现了标准的重要性，特别是在测试领域。OPNFV 社区没有办法在短时间内开发一套测试用例，因为测试（尤其是性能测试）涉及到指标上的博弈，而这些博弈是一个松散的、开放自治的组织很难快速达成共识的。于是 OPNFV 社区在性能测试方面大量接受并引用了 ETSI 关于 NFV 测试的标准。然而，ETSI 只是制定了一个框架和部分测试标准，运营商在采购和入网测试时缺乏完备的标准，无法利用前期设备采购的那一套内容和机制，导致 NFV 技术落地缓慢。因此目前 OPNFV 也正在着手做社区的认证计划（OVP），希望把 OVP 打造成为一个 NFV 产业的事实标准，用标准更好推动技

术的落地。

6.2 AI 领域开源与标准联动的思考

通过 6.1 节三个案例可以看到，开源作为技术创新的重要方式，可以提高产业内企业间协同开发效率，加速技术的演进和实现；标准则是对行业方案的能力要求、评测规范、互联互通格式做了约定，确保了一致性和落地质量。开源驱动了标准的产生，反过来标准也促进了技术的创新，开源和标准的联动是技术和商业结合的有效路径。

从第四章不同行业 AI 工作流的案例可以看到，这些工作流本质上都是通过数据收集与分析、准备与清理后，进行模型构建、验证与测试，不断迭代，最终完成部署。但它们对外呈现方案却有明显的不同，究其原因还是当前 AI 技术处于蓬勃发展期，开源社区、开源项目之间重叠严重。以深度学习框架为例，在 GitHub 上开源项目就有 10 种以上，都有自己的接口、数据格式、模型格式等，这极大增加了 AI 从业人员及企业的沟通维护成本。人工智能的落地需要依靠产业化，产业化的发展离不开规范，其中最重要的就是标准。如果缺失标准，人工智能的研发和应用将变得混乱，市场很容易发生分裂。

目前，人工智能标准化已经成为不少致力于人工智能产业的企业和国家的重要战场。某些国际、国外标准化组织积极部署开展人工智能标准化工作，但尚未形成完整的标准体系。我国虽然在某些领域已具备一定的标准化基础，但标准化程度不足，分散的标准化工作不足以支撑起整个人工智能领域的发展。目前我国拥有大量参与人工智能建设的企业，在缺乏自身科学的标准体系的情况下会出现沟通不畅，开发低效，冗余生产等问题。在这种情况下，建立统一完善的标准体系，以标准手段促进我国人工智能技术、产业的发展，对加快我国人工智能技术创新和成果转化、提升产品服务和质量、保障用户安全、建立公平开放的产业生态意义重大。

开源作为技术创新的重要方式，可以提高产业内各方协同创新的效率，加速技术的演进和实现。在技术创新层面期望优先制定标准，并要求开源社区遵照标准来实现是不现实的，也是对产业发展的一种伤害。但同时也应看到，如果开源社区完全按照自己的路径发展，只关注与社区积极互动的用户需求，从长远来说

也会形成对 AI 在更广泛行业快速落地的障碍。可能会出现的问题有：

（1）各种开源项目风起云涌，彼此之间没有清晰的界限和接口，不光有重叠，甚至存在矛盾与冲突，用户及行业应用面临如何选择的问题。

（2）行业客户，尤其在中国，需要标准来进行采购背书，如果AI领域标准、测评、认证缺失，将会为落地行业应用带来障碍。

（3）用户的诉求如何影响到社区的技术演进方向是AI开源技术落地各行业的一个障碍。大量的用户是技术消费者而不是技术创造的参与者，让用户参与到开源社区，像社区开发者那样与开源社区有效的互动也是一个挑战。

因此在 AI 领域开源与标准更应该相互协同相互补充，开源更聚焦在技术创新，而标准的制定更面向如何帮助用户选型。标准制定之后如何影响到开源社区也是一个挑战。从历史上的成功案例看，产业内的某个开源参与者（公司或组织）如果既有标准人员也有相关开源社区人员，二者在同一个团队中，承担统一的生态目标，甚至一个专家既有开源社区参与的经验也有标准制定的经验等，他们将在社区与标准协同中发挥重要作用。同时应该建立起开源社区与标准组织的联动机制，这一方面可以通过上述既有标准能力也有开源社区能力的企业、组织来承担，另一方面也可以通过前瞻性的开源组织及标准组织来拉通、协同，如开源组织中的 Linux 基金会以及标准组织 IEEE 及 IETF 等。

6.3 本次标准机遇研究的范围与内容

在界定本研究报告范围之前，我们对 AI 领域的基本层次进行一个分析，得到如下的一个三层结构：行业应用、领域技术与业务使能、基础技术及能力。其中领域技术与业务使能、基础技术及能力这两层合起来一般称之为 AI 平台，而行业应用则是架构在 AI 平台之上，从而形成完整的 AI 应用全景图。下图中也标明了三层结构与前面第三章中 AI 开源全栈（蓝色背景图）的对应关系。

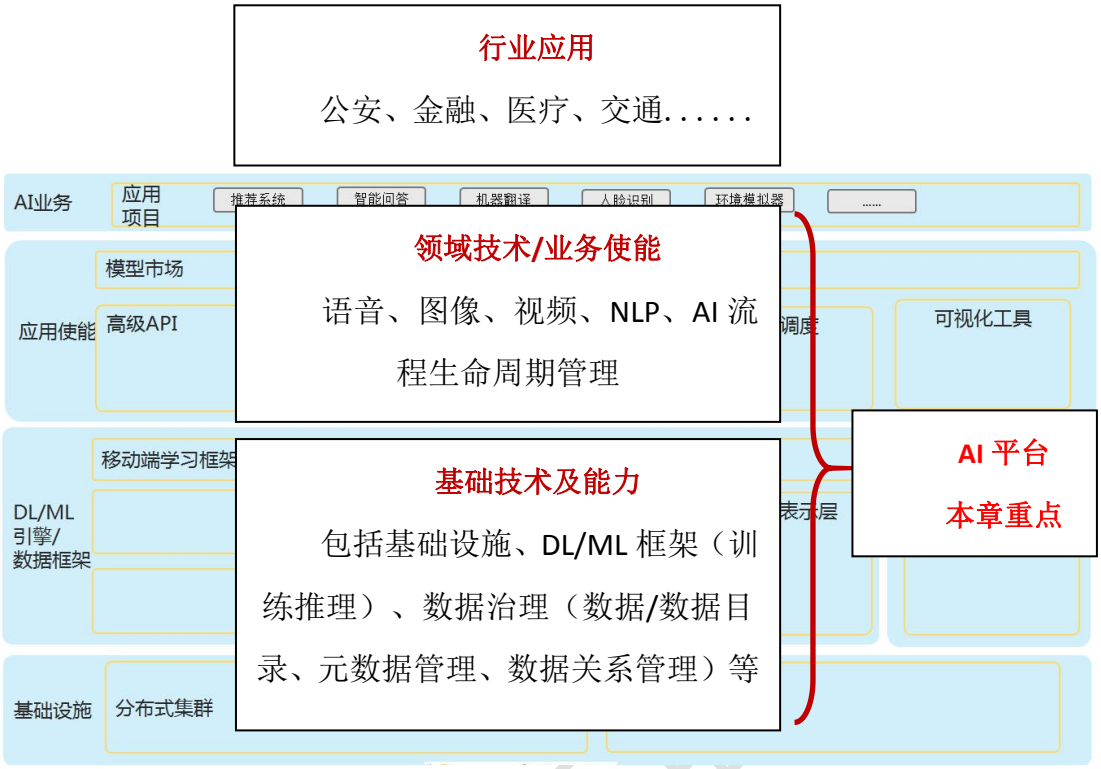

图 20　AI 应用全景图

6.3.1 行业应用标准

由于行业之间存在巨大差异化，行业内的复制以及行业间的能力重用代价很大，需要制定大量标准。但这些标准更适合由不同行业单独制定，因此行业应用的标准建议将不在本研究报告范围内。

6.3.2 AI 平台标准

统一、高效、商业化、标准化的 AI 平台可以降低各行业智能化转型的技术门槛和实际的建设成本，有利于 AI 业务推陈出新。

AI 平台技术是开源技术主导，开源技术距离商业落地还存在一定距离。一方面 AI 平台涉及多种 AI 开源技术，AI 开源技术之间存在兼容性问题，如接口、版本能力等；另一方面开源技术在工程化方面有不足，难以支撑企业级商业应用，如功能完备性、稳定性、扩展性、安全性等。因此需要 AI 平台的架构、关键功能、接口等基础层面，形成统一理解和共识结合。

基础技术及能力是开源创新的主战场，其标准位于人工智能标准体系结构的最底层，支撑标准体系结构中的其它部分，因此需要优先考虑、加大这方面的研究力度。有六类标准成为用户关注的、产业建设急需的标准：

（1）基础设施标准规范

为支撑 AI 算法开发、训练、部署及管理对高性能、低能耗、易管理的算力需求，AI 平台需要规范硬件抽象接口。目前 GPU 领域由于 nvidia 一家独大，其标准封装 Cuda/CUDNN 渐成事实标准，但用于适配 GPU、FPGA、ASIC 等各种异构计算资源的标准还处于缺位状态，存在标准化的需求。

目前在 OpenStack 社区中的官方项目 Cyborg 是唯一的能够提供异构加速硬件通用管理框架的开源项目。Cyborg 提供面向异构加速硬件的基础生命周期管理能力(CRUD 操作)，通过抽象通用的数据模型与统一的管理操作 API，为用户提供统一的异构计算资源使用体验，而无需针对每一种异构加速硬件特别构建管理模块。同时，Cyborg 提供异构加速硬件管理元数据的标准化，使得资源描述和业务需求之间的映射关系更加准确，在调度上可以更加通用与便捷。因此这个社区可以是标准联动、协同的首选对象。

（2）数据标准规范

中国自身拥有大量特定领域数据，需要加快研究这部分特有数据标准构建。

元数据，又称中介数据、中继数据，是描述数据的数据，主要是描述数据属性信息，可以支持指向存储位置、历史数据、资源查找、文件记录等功能，是构建人工智能数据网络的基础元素。元数据概念提出始于 20 世纪 80 年代末、90 年代初，迄今已有 30 多年的历史，由于数据种类和格式的差异，难以统一描述，导致与之相关的元数据标准风格各异。目前，元数据的通用标准多达 30 余种，主要包括：DC（都柏林核心元数据）、CDWA（艺术作品描述目录）、CIMI（博物馆信息计算机交换标准框架）、ONIX（在线信息交换）、GILS（政府信息定位服务）、SGML（通用标准标记语言）、OWL（网络本体语言）。上述众多元数据标准，给人工智能领域的数据融合及交换带来了诸多问题，亟待统一。

目前，在人工智能领域存在若干标准数据集，如机器视觉领域的 MNIST、CIFAR、ImageNet，自然语言领域的 WikiText、SQuAD、Billion Words、Stanford Sentiment Treebank，语音识别领域的 2000 HUB5 English、CHIME，推荐系统领域

的 Netflix Challenge、Million Song Dataset 等。这些数据集在推动人工智能算法优化、工程开发等方面起到了重要作用。但是，数据集的定义、格式、接口等差异很大，导致开发出来的人工智能程序必须不能兼容和互操作，影响人工智能平台的集成式性能。如果各个公司都独自开发不同数据集之间的转换程序，会造成重复劳动，影响工作效率。因此，亟待对上述数据集进行标准化，由于人工智能不同领域的数据特点存在差异，所以按照领域（如机器视觉、自然语言等）制定专门的数据标准，是一种可行的方案。

此外，公开的数据集标注相对粗糙，仅能够满足通用型人工智能算法验证的需要，无法有效验证蓬勃发展的人工智能细分领域应用型算法，特别是和专用场景相关的行业特殊算法的性能。在这种情况下，人工智能公司们必须想尽办法，积累符合自身应用方向，标注得更细致、更准确的数据。在某种程度上，高质量的标注数据决定了一家人工智能公司的竞争力。在这种情况下，催生了人工智能产业链的重要环节——数据标注。从事该项工作的机构类型包括三种：众包平台、公司自标、数据标注公司，不同机构根据所标注数据的特点和流程，各自制定对应的标准。数据标注有许多类型，如分类、画框、注释、标记等，不同类型的标注可以根据行业、需求、目标、语言等要素进一步细分，因此导致不同机构标注的同类数据，甚至相同数据，标注差异都很明显，不利于后继的分析和处理。通过制定统一的数据标注标准，能够减少这种差异。此外，数据标注是十分耗费成本的人力劳动，熟练者平均一天可以标注 40 张图片，前提是只需要为图片中的物体打框、标注类别和前后关系，如果涉及到刻画建筑物边缘等复杂细节，一天标注 10 张已是极限。但需要处理的数据通常以“万”为单位，在工期紧张的情况下，往往是数据标注团队的几百人同时上场，连续标注几周甚至数月，成本极其高昂。在人工智能相同领域的多个公司，其标注的数据存在重复性。如果制定数据标注标准，以及互惠的数据开源制度，减少重复标注工作，可以大幅度降低企业的数据标注成本，使其能够投入更多的精力研究和优化人工智能算法本身。

AI 在数据格式标准化（标注、元数据）中技术标准的机遇与需求:

- 编码规范：采用统一的编码规范（UTF-8）可以减少由于字符集不一致导致的低效处理问题。
- 数据标注规范：同种数据的标注格式规范可以避免由于数据标注不一导

致的不必要的数据预处理流程。

- 标签规范：针对分词、词性、句法等由于标签不一，导致各种模型训练的不一致，因此标签规范也极为重要。下表显示了语音数据集在发音人方面的一个标准，可供参考与借鉴。该标准要求每个地区发音人200个，没有发音障碍，听力正常。年龄、性别以及口音和文化程度分布如下，允许误差5%。

表 2　发音人要求和分布

年龄	青年（50%）	中年（40%）	老年（10%）
性别	男女各一半	男女各一半	男女各一半
口音	中度二级口音 80%，一级乙等 5%，三级 15%		
文化程度	90% 高中以上学历，10%高中以下学历		

（3）模型标准规范

在人工智能行业蓬勃发展的今天，从业人员拥有大量可选框架来完成他们自定义的模型训练工作。然而在任意一个框架上训练的神经网络模型，无法直接在另一个框架上使用，开发者需要耗费大量时间精力把模型从一个开发平台移植到另一个。为了解决机器学习模型缺乏互操作性的问题，由亚马逊、微软和 Facebook 牵头发布了新的一个机器学习开源社区项目—ONNX，开发者能更方便地在不同框架间切换，为不同任务选择最优工具。

基本每个框架都会针对某个特定属性进行优化，比如训练速度、对网络架构的支持、能在移动设备上推理等等。在大多数情况下，研发阶段最需要的属性和产品阶段是不一样的。这导致效率降低，比如选择切换到最合适的框架，又或者把模型转移到另一个框架导致额外的工作，造成进度延迟。使用支持 ONNX 表示方式的框架，则大幅简化了切换过程，让开发者的工具选择更灵活，能够通过 ONNX 导入导出模型的框架。

ONNX 的工作原理是：实时跟踪某个神经网络是如何在这些框架上生成的，接着，使用这些信息创建一个通用的计算图，即符合 ONNX 标准的计算图，这样在计算方面，虽然更高级的表达不相同，但是这些框架产生的最终结果都非常接近。目前 ONNX 支持下框架如下：

表 3　ONNX 支持框架(截止到 2018 年 9 月)

Framework/tool	Installation	Exporting to ONNX（frontend）	Importing ONNX models（backend）
Caffe2	part of caffe2 package	Exporting	Importing
Pytorch	part of pytorch package	Exporting,Extending support	coming soon
Cognitive toolkit(CNTK)	Built-in	Exporting	Importing
Apache MXNet	Part of mxnet package docs github	Exporting	Importing
Chainer	Chainer/onnx-chainer	Exporting	coming soon
TensorFlow	onnx/onnx-tensorflow	Exporting	Importing(experiment)
Apple CoreML	onnx/onnx-coreml and onnx/onnxmltools	Exporting	Importing
SciKit-Learm	onnx/onnxmltools	Exporting	n/a
ML.NET	Built-in Convert to ONNX-ML	Exporting	n/a
Merioh	Pfnet-research/menoch	n/a	Importing

由于目前 ONNX 所定义的模型格式还未成为各深度学习框架中的原生格式，用户基数还不够庞大，因此推动标准化时机未到。建议当前还是以参与社区为主，待大多数主流框架采用后，仿照这章开始容器格式 OCI 的案例，在社区形成相关规范并择机标准化。

（4）评测标准规范

行业用户智能化转型，需要有规范指导用户选择有较好工程能力、商业化能力的人工智能系统/平台提供者，指导用户识别商业化发行版本人工智能平台。人工智能平台通用测评规范规定，硬件平台和软件平台基础能力的测试方法，覆盖基本的功能、运维、安全、可用性、兼容性等能力，适用于人工智能平台产品的评估、验收等。

人工智能平台通用测评规范相关要求，包括：测试对象、测试环境、测试数据、测试模型、测试指标、测试用例等，从功能测试方法和测试指标、性能测试方法和测试指标，以及其他商业化能力测试方法和指标（如可用性、可扩展性等方面），来衡量人工智能硬件平台、软件平台、服务等基础能力。

评测规范最先需要设计的内容是 AI 平台测试项的测评指标，尤其是衡量方案和传统领域存在巨大不同的地方。以深度学习芯片 Benchmark 评测为例，

Benchmark 是衡量处理器系统性能的标尺，但在深度学习处理器领域目前还没有公认的测试基准。虽然传统的处理器有相应的 Benchmark 测试基准，但无法满足深度学习处理器的评测要求，主要原因在于深度学习处理器与传统通用处理器之间存在巨大差异：

- 深度学习处理器架构与传统通用处理器存在差异；
- 深度学习处理器从功能角度看又分训练和推理，二者评测方式和衡量指标也不完全相同；
- 深度学习处理器往往针对应用会做特定的优化等。

（5）AI平台接口规范

AI 平台是一个开放、多样化的平台系统，由具有不同功能且相对独立的模块构成，不同层面的互联互通是极为必要的，这样不仅有助于提高 AI 平台的通用性和兼容性，降低应用开发和迁移的成本，也对于帮助用户跨平台开发上层应用有重要意义。目前数家企业在其 AI 引擎之上提供 API 服务，但产业尚未形成统一规范。

AI 平台接口规范主要规定了 AI 平台中各个层次之间的接口，在基础能力、常见参数、实现方式、维护更新等方面进行规范，形成统一接口。人工智能硬件平台接口能力，包括异构服务器架构、操作系统、计算架构、算法部署等各层之间的接口能力。人工智能软件平台接口能力，包括深度学习平台、机器学习、强化学习、特征检索引擎、任务调度、资源管理、算法仓管理等各层之间的接口能力。

AI 平台接口规范还需要考虑与非标准接口的兼容性，由于众多企业及开发者存在自身业务需要，将其底层能力同 AI 平台整合，因此接口的复用性显得极为重要。此外，针对物联网、大数据等业界标准接口的整合也是 AI 平台需要关注的内容。

（6）AI算法体系规范

现阶段 AI 算法正在蓬勃发展，由于算法应用领域不同，数据类型相互异构，导致 AI 算法并不具备清晰的体系。AI 算法按照能力范围划分，可以划分成三个维度：

- 普适型算法

普适型算法主要覆盖公共领域的算法，将算法作为基础能力提供方，直接将AI作为能力进行输出。普适型算法主要聚焦于覆盖性及其鲁棒性。其中，方言识别、人脸识别就是典型的普适型算法。它们并不依赖于任何的业务场景，并可以直接将其能力作为服务输出。从业人员可以针对普适型算法的基础上通过自身场景和数据不断完善与优化。

- 应用型算法

应用型算法主要面向具体的应用层，针对不同的数据维度及应用需求设计合适的应用算法。应用型算法主要聚焦于同业务与应用紧密贴合的算法，强调易用，精准，贴合，闭环。其中，商品推荐、目标检测、多轮对话都是应用型算法。从业人员可以根据研究现阶段的应用型算法，分析贴合其场景的相关算法，并针对自身场景不断优化算法，进而更好服务于自身的应用场景。

- 支撑型算法

支撑型算法主要针对特定领域（语音、文本、图像）的底层技术算法，介于普适型算法和应用型算法之间。支撑型算法并不能独立于应用场景而直接输送能力，但并不完全依赖于具体应用。其中，中文分词、声纹识别就是典型的支撑型算法。从业人员可以根据特定领域选择对应支撑型算法，进而辅助自身业务应用的搭建。

6.3.3 安全标准

人工智能安全性威胁一方面是来自于外部因素如黑客攻击，另一方面是源于人工智能模型自身设计中存在的漏洞可能被利用。由于人工智能在诸多领域都发生着深刻的变化，如果不具备有效的安全防护能力，将带来极大的隐患。另外值得注意的是，如果AI的训练数据被污染，或者在推理中被噪音数据干扰，有可能会导致模型输出的结果违背设计的初衷，倘若被恶意操纵还将可能带来负面影响。

人工智能技术使得商业自动化的程度越来越高，随之带来的IT风险集中，网络安全隐患变得突出。企业的规模和发展的阶段不一，其安全防控水平也参差不齐，差异较大，风险的洼地效应非常明显。因此，在外部层面，建议深入研究人工智能安全发展战略，针对不同环节和应用场景，明晰不同主体的责任、权利

和义务，制定有关的政策、措施、法规和标准，划定人工智能安全发展的边界，操作与流程指引，安全评估及安全防范，准入与退出机制等。

人工智能主要依靠软件和算法驱动，难免出现技术漏洞和人为缺陷。在内部层面，建议开展人工智能算法、产品和系统的安全要求和测评评估标准工作，搭建 AI 模型漏洞检测、抵抗非法访问的能力评估、防止算法模型被反向试探泄露信息的能力评估、针对抵抗样本和干扰攻击的能力评估。应针对当前主流算法和应用启动研究性标准项目，构建 AI 安全性评估度量的模型，进行评估模型和方式的标准化。

另外，在数据方面存在网络信息安全风险，建议应继续加强与数据安全管理、隐私保护测等支撑类安全标准的研制联动，以进行更好的数据安全管控，防止后台数据滥用。

6.3.4 应用智能化水平评估

人工智能已经成为智能机器人、智能金融、智能家居、自动驾驶等不同场景下核心技术能力，尽管市场上许多产品都加上了“智能化”的标签，但是其智能化程度可能还处在较低水平。另外，人工智能产品与服务的智能化水平要达到什么样的要求，才能真正满足不同场景下用户的需求，目前还没有系统的评价体系。目前，产业界逐渐开始对单项产品如智能音箱、智能摄像头进行智能化水平的评估，以智能音箱为例，评估其智能化水平在不同的应用场景下，检测其感知能力、交互能力、灵敏度等。

由于产品定位的不同很难用同一个评价体系来评估，且单项产品的智能化水平并不能代表在具体的场景中的智能化水平。故需建立一套基于不同业务场景下的智能化水平分级评估体系，以进行客观、公正的评测，并根据评估结果，给出不同的等级，以帮助产业树立产品和服务的标杆，为用户提供可信赖的参考依据。

建议从若干重点且较成熟的领域入手，加强智能家居、智能汽车、智能机器人、智能工厂、自动驾驶等领域下的软硬件、系统、数据、测试等标准化工作，尽快展开智能化水平评估和分级工作，以促进产业的开放协同与公平竞争。

6.4 制定人工智能标准中要考虑的因素

人工智能作为一种通用信息技术创新，为各领域创新发展注入了新的活力，但也给风险管理、业务安全、行业秩序等方面带来了新挑战。首先，人工智能的分析方法更注重相关性逻辑而非强因果逻辑，不同模块和因素之间相互关联、渗透，导致风险更加错综复杂，传导性也可能更强、更广泛。其次，人工智能基于信息技术将业务流变成了信息流，在提升效率的同时也打破了风险传导的时空限制，使得风险传播的速度更快。最后，人工智能创新产品的结构复杂、可能涉及多重嵌套，其风险容易被其表面所掩盖，难以识别和度量，风险的隐蔽性更大，传统的风控措施或将难以奏效。

因此，建议在未来制定人工智能开源技术标准时，重点关注以下几方面的因素，引导人工智能开源技术健康有序发展。

6.4.1 伦理与社会关注

随着人工智能的迅猛发展，其带来的伦理道德、法律以及社会关注等相关问题也日益引起争议。在其发展的过程中，可能引发的一般性伦理与社会问题，其中包括失业问题、隐私问题、算法偏见、安全问题、机器权利和道德代码等。但是，人工智能领域的伦理标准与法律制度问题并不健全，当前我们务必包含对人类伦理价值的正确考量。因此，设定人工智能技术的伦理要求，要依托于社会和公众对人工智能伦理的深入思考和广泛共识，并遵循普适共识原则。针对人工智能带来的快速变化，制定的标准也应具备包容性与开放性。为了应对人工智能带来的伦理社会挑战，世界各国及相关国际组织都开始致力于推动人工智能伦理与社会问题的研究，积极建立相关规范与指南，推动社会各界就人工智能的伦理与社会监管达成共识。不同的国家具有不同的国情，对待社会伦理道理也不尽相同，需要关注中国语境下的人工智能伦理、道德和法律的独特性，希望技术、法律和伦理在人工智能平台上共同发展。

现阶段人工智能领域更多是技术工程师在参与，缺乏人文领域，如哲学、伦理学、法学等其它社会学科人员的参与。有关人工智能伦理与法律标准的研究需要加强，需要学术界和研究机构在人工智能伦理以及社会关注方面持续投入，需

要从人工智能伦理道德主体地位、伦理问题及其治理问题、伦理标准应用原则、伦理原则制定等方面进行研究。同时需要结合政府，企业，高校的相关资源，通过产学研结合，建立一个利用人工智能技术造福于社会、保护公众利益的法律和伦理标准化环境。

与本报告同时进行编撰的还有《人工智能伦理风险分析报告》，建议对此问题有兴趣的读者可以进行进一步了解。

6.4.2 监管与治理因素

开源人工智能技术作为信息技术带来的创新，更强调是前沿信息技术对各个领域的支持。例如针对合规业务，人工智能需要具备辅助、支持和改进作用，其核心是帮助传统业务实现“三升两降”，即提升效率、体验、规模，同时降低成本和风险。但是，开源人工智能技术本身并不能凭空创造出一个新的产业，以及脱离原有的产业监管，技术的运用仍需遵循原有行业的内在规律和秩序、遵守现行法律和行业监管要求。人工智能应当扮演好辅助决策的功能，不能本末倒置，扰乱原有市场秩序。

在金融行业中，如果使用人工智能技术的目的是规避金融监管，进行监管套利，例如在无相应的金融业务牌照的情况下，进行违规智能投顾交易或基金智能推荐销售，或以人工智能模型进行风控开展违规现金贷业务，或违规进行 ICO 或发行虚拟货币以激励人工智能应用的用户参与知识库标注或贡献数据等，皆属于扰乱金融稳定行为，并不具备健康发展的可持续性。在服务型行业中，如果将人工智能技术用于诱惑、误导用户，或过度包装营销自身产品，在用户缺乏感知的情况下完成相关操作均属于欺诈行为，并不有利于服务型行业的长远发展。

因此，建议在标准的制定中，也应重点关注监管与治理因素。一方面，建议官方机构与监管机构引进沙盒机制和监管科技，不仅利用人工智能开源技术的创新业务在风险可控的前提下先行先试，又能确保做好穿透式监管，提高监管的规范性和风险监测识别的能力；同时，也建议从业机构在通过人工智能开源技术提供创新产品和服务的同时，向监管机构提供相应的监管解决方案，并提供自身安全性保障，规避隐形黑盒带来的监管盲区。以提高创新业务的合规透明度，共同应对“不当或违规使用人工智能技术”所带来的风险。

6.4.3 把握开源与标准平衡，促进创新与产业发展

技术的标准化有其最佳时间点，即“标准化窗口”，过早的标准化会抑制创新，而过晚的标准化则不利于产业发展。AI 行业处于快速发展阶段，创新项目、创新技术层出不穷，过早建立单一的标准很有可能影响创新的节奏。因此对于 AI 领域的标准制定，一方面我们应当从“小”做起，聚焦基础、成熟的内容，然后再逐渐扩大，慢慢完善整个体系；另一方面应从“粗”做起，制定平台框架、平台接口而不是限定具体的实现细节。同时，我们需要借鉴同样快速发展的汽车领域标准制定的思路，适当加快标准更新的节奏。当 AI 领域出现大突破和技术性革命的时候，需要推出新的标准，以适应产业的发展。

对于 AI 方案，产业界可能存在多种实现的思路。标准的制定需要充分考虑这些方案，做到兼容并举，将共性内容提炼成为推荐性标准，而差异化内容则根据对行业的影响程度加以区分，以不同的标准级别纳入标准体系。

在平衡创新和产业发展方面，开放容器标准（OCI）组织的做法值得借鉴。OCI 组织在容器标准的制定上一直坚持“最小集优先”原则，对于行业公认的运行时与镜像格式等方案，大胆采纳，较快地推出 1.0 标准版本，迭代速度快；对于存在异议的分发格式，经历了长达两年的论证、直到确实没有差异化方案出现的情况下才纳入标准体系。另外，由于不同体系结构下，容器的规范有所差别，OCI 也根据行业重要性将这些体系结构的规范纳入到标准里面，确保了差异化方案的并存。同时，这种标准的讨论也是“公开、透明”，并通过投票机制保证技术创新与产业发展间的平衡。

结 语

人工智能毫无疑问正处于一次发展热潮中，2018 年 NIPS 的论文提交量增长了 50%以上，达到了约 5000 篇。基于深度神经网络的技术已在视觉和语音识别等领域取得了实实在在的应用，并颠覆了传统算法与技术的效果。更通用的机器学习也拥有了大量的真实世界用例。但与此同时，我们应该警惕人工智能在发展高峰期产生过多的泡沫。1980 年代曾有过一次人工智能泡沫，2000 年左右又有一次互联网科技泡沫，而在当前的这场人工智能浪潮中，技术与商业的组合泡沫已经开始产生。在 2016 年到 2017 年期间，很多人工智能初创公司可以通过比拼雇佣科学家的量级和发布论文的数量就可以吸引巨额投资，而进入 2018 年后，大家却看到缺少商业应用能力的人工智能公司将越来越难生存。于此同时我们也看到大量的商业场景及企业又在呼唤人工智能的“民主化”与落地，这些矛盾如何解决值得我们深思。

我们处在一个技术与商业持续创新的时期，机遇也伴随着不确定性。我们一方面为人工智能给人类生活带来的改变一次次欢呼，同时又需要心存敬畏并保持清醒。人工智能像很多新技术新产业一样，发展道路是曲折的，但前途是光明的。如何一方面推动技术的快速发展与推广，同时又警惕对“泡沫”不加控制最后“捧杀”了人工智能，这些都需要我们不断地去探索。

我们希望更多方能参与人工智能生态圈的建设。开源、开放、协同是构建人工智能生态圈的可持续性发展的实施路径。

我们鼓励充分发挥参与者的想象力。人工智能科技的发展是动态的、已有很多历史性的创新。尽管本报告重点关注当下的创新突破：Machine Learning 和 Deep Learning，但推动与促进人工智能产业发展，“奇点”在未来。

我们号召自主创新。人工智能是人类史上的重大发明，未来如何演变是个未知数。在开源，开放和协同的基础上，每个国家、每个企业、每个人都有大量自主创新的空间和机会。通过创造更多国际间的交流与合作，为下一代人工智能的健康发展贡献力量，形成人和机器更好的融合，让人类生活得更美好！

附录 A

表 A.1 AI 开源项目社区活跃度指标统计

Domain	Project	star	fork	commits	contributors
芯片使能	Vulkan	3,100	604	1,466	52
	LLVM	3,050	1,535	167,738	548
	RISC-V	299	123	768,245	∞
	DLA	215	310	6,377	202
分布式集群	KubeFlow	4,228	542	598	86
	FfDL	293	92	70	12
大数据支撑	YARN	32,631	1,857	2,097	437
	hadoop	7,743	4,902	19,637	138
	Flink	4,107	2,744	14,352	426
	OpenStack	2,793	1,308	137,915	1,651
	Hive	2,202	2,065	12,608	166
	HBASE	2,144	1,621	15,600	192
学习框架	TensorFlow	106,738	66,181	37,593	1,581
	Spark	18,322	16,592	22,425	1,270
	mxnet	14,855	5,407	7,580	571
	PyTorch	17,627	4,135	12,597	731
	Caffe2	8,197	2,035	3,678	193
	PaddlePaddle	7,373	2,014	16,998	148
	mahout	1,483	879	3,969	29
	BigDL	2,578	595	2,384	50
	Analytics Zoo	83	72	669	34
知识图谱	Scikit-learn	29,744	14,762	23,094	1,140
	YAGO	316	38	1,319	8
	Open IE	101	29	23	1
	FreeBase	32	14	5	1
	NELL	5	4	27	1
强化学习	OpenAI Gym	13,081	3,224	816	103
	OpenAI Baselines4910	4,910	1,403	126	40
	PySC2	4,832	685	250	28
	DeepMind Lab	6	4	31	1

中间表示层	TVM	1,787	402	1,490	123
	NNVM	1,537	281	358	60
端侧推理框架	NCNN	4,335	1,080	427	39
	paddle-mobile	3,990	757	1,322	17
	Core ML	1,139	113	291	27
	TensorRT	144	38	43	2
	TensorflowLite	26	9	7	1
	Caffe2go	2	4	39	1
高级 API	Keras	32,233	12,108	4,660	703
	Sonnet	6,720	912	328	19
	TensorLayer	4,146	1,017	2,065	69
	GLUON	2,209	222	13	6
标准模型/算法	Tensor2Tensor	4,627	1,115	2,120	88
	xxNet	355	81	12	1
	ModelZOO	136	50	43	6
开放数据集	Stanford NLP	5,071	1,848	14,845	85
	WordNet	91	18	252	3
分布式调度	Angel	3,548	899	1,081	27
可视化工具	Facets	4,539	547	125	17
	TensorBoard	2,067	483	2,010	124
	VisualDL	1,740	267	504	13
推荐系统	SVD Feature	0	0	3	1
问答系统	DrQA	2,491	479	35	5
语音识别/机器翻译	Kaldi	4,233	2,161	8,130	199
	FairSeq	3,022	523	24	8
	Sockeye	579	167	388	24
人脸识别	FaceNet	5,388	2,302	566	30
	SeetaFace	3,130	1,429	36	5
	DeepFace	1,055	530	25	1
其他	Detectron	15,619	3,046	85	15
	Pattern	6,421	1,256	1,430	20
	Aerosolve	4,410	557	1,064	23
	DSSTNE	4,169	698	331	33
	DeepDetect	1,620	397	1,660	15
	Open Cog	1,543	621	27,177	92
	CaffeOnSpark	1,246	377	260	7
	Numenta	16	19	2,236	16

注：数据来源： https://github.com/

数据截止日：2018/08/08

附录 B

表 B.1 第五章技术术语表

英语短称	英语名称	中文名称
ASCII	American Standard Code for Information Interchange	美国信息交换标准代码
XML	Extensible Markup Language	可扩展标记语言
RDF	Resource Description Framework	资源描述框架
CSV	Comma Separated Values	逗号分隔值
JSON	JavaScript Object Notation	JavaScript 对象表示法
TXT	ASCII text data format	ASCII 文本数据格式
DICOM	Digital Imaging and Communications in Medicine	医学数字影像通信
VTK	The Visualization Toolkit	可视化工具包
GDRP	General Data Protection Regulation	通用数据保护法规
LIM	Logistics Interoperability Model	物流互联互通模型
ERP	Enterprise Resource Planning	企业资源计划
PLM	Product Lifecycle Management	产品生命周期管理
SCM	Supply Chain Management	供应链管理
CRM	Customer Relationship Management	客户关系管理
EMS	Environment Management System	环境管理系统
MES	Manufacturing Execution System	制造执行系统
	Made in China 2025	中国制造 2025
LMS	Learning Management System	学习管理系统
PB	Peta Bytes	十亿兆字节
CDLA	Community Data License Agreement	社区数据许可协议

CDLA-Sharing	Community Data License Agreement-Sharing	CDLA 共享许可证
CDLA-Permissive	Community Data License Agreement-Permissive	CDLA 许可许可证
	Centralized Model	中心化模式
	Hybrid Model	混合模式
	Decentralized Model	去中心模式
	Bootstrap Model	没有初始数据的引导模式
	Digital Content	数字内容
	Open Ecosystem	开放生态圈
	Smart Contract	电子智能合约
	Blockchain	区块链
OED	Open Ecosystem DApp	开放生态圈 DApp
DApp	Decentralized Application	去中心化应用/分布式应用
	Symbolic Reasoning	符号推理
	Functional Programming	函数式编程
LISP	List Processor	功能推理人工智能编辑语言
	Logic Programming	逻辑编程
	Logical Reasoning	逻辑推理
Prolog	Programming in Logic	逻辑推理人工智能编辑语言
	AI Agent	AI 代理
	Expert System	专家系统
	Explainable	可解释的
	Unexplainable	无法解释的

表 B.2 第六章技术术语表

英语短称	英语名称	中文名称
ASCII	American Standard Code for Information Interchange	美国信息交换标准代码
OCI	Open Container Initiative	开放容器标准
	iSula	华为符合开放容器标准实现
	rkt	CoreOS 符合开放容器标准实现
ASF	Apache Software Foundation	Apache 软件基金会
	Apache CarbonData	一种新的大数据文件格式
NFV	Network Functions Virtualization	网络功能虚拟化
OPNFV	Open Platform of NFV	网络功能虚拟化开放平台
OVP	OPNFV Verfied Program	OPNFV 项目认证
ETSI	European Telecommunications Standards Institute	欧洲电信标准协会
CUDA	Compute Unified Device Architecture	GPU 通用并行计算架构
cuDNN	CUDA Deep Neural Network	深度神经网络 GPU 加速库
	Cyborg	通用硬件(GPU\FPGA 等)加速框架
DC	Dublin Core Metadata	都柏林核心元数据
CDWA	Categories for the Description of Works of Art	艺术作品描述目录
CIMI	Consortium for the Computer Interchange of Museum Information	博物馆信息计算机交换标准框架
ONIX	Online Information exchange	在线信息交换
GILS	the Government Information Locator Service	政府信息定位服务
SGML	Standard Generalized Markup Language	通用标准标记语言
OWL	Ontology Wed Language	网络本体语言
MNIST		手写识别数据集

CIFAR		普适物体识别的数据集
ImageNet		计算机视觉系统识别项目
WikiText		Salesforce MetaMind 设计的大型语言建模语料库
SQuAD		斯坦福大学阅读理解数据集
	Billion Words	大型、通用词语表征语言建模数据集
	Stanford Sentiment Treebank	标准情感数据集
	2000 HUB5 English	英语的语音数据集
CHIME		包含噪声的语音识别数据集
	Netflix Challenge	Kaggle 挑战赛数据集
	Million Song Dataset	Kaggle 混合推荐系统数据集
ONNX	Open Neural Network Exchange	开放式神经网络交换格式
	Benchmark	基准检测
UTF-8		统一的编码规范