中关村智用人工智能研究院

人工智能应用社会风险评估与《网络安全法》

中关村智用人工智能研究院
执行院长 孙明俊

2021年9月23日

01

人工智能
产业发展

全球的人工智能产业发展——产业开始落地

新冠疫情带来的投资不确定性推动了人工智能产业的再次增长。其他行业的增长乏力，人工智能产业的增长确定性较高

- **人工智能领域并购增长快速。**全球范围内，相对于2019年，2020年全球人工智能投资总额增长了40%。2020年，人工智能领域的私人投资仍较为稳定。相对2019年，2020年人工智能领域的并购增长了121.7%。
- **私人投资领域，美国的人工智能投资居首，且保持稳定的高增长。而按地域划分，美国的人工智能私人投资保持了一个较平稳的增长态；而第二名的中国，私人投资增长乏力。**当然，考虑到中国国家主导的人工智能产业格局，公共投资在人工智能产业发展中起着更重要角色。
- **人工智能走向应用，基础技术投资减弱。**相对于2019年，由于新冠疫情的冲击，大量资金流向了人工智能驱动的医疗领域和教育领域。仔细观察图3，可以发现更多的资金加快流向人工智能应用技术，而基础技术的募资有所减弱。不仅仅在投资上，很多产业技术应用的成熟也表明人工智能产业落地已经开始走近。

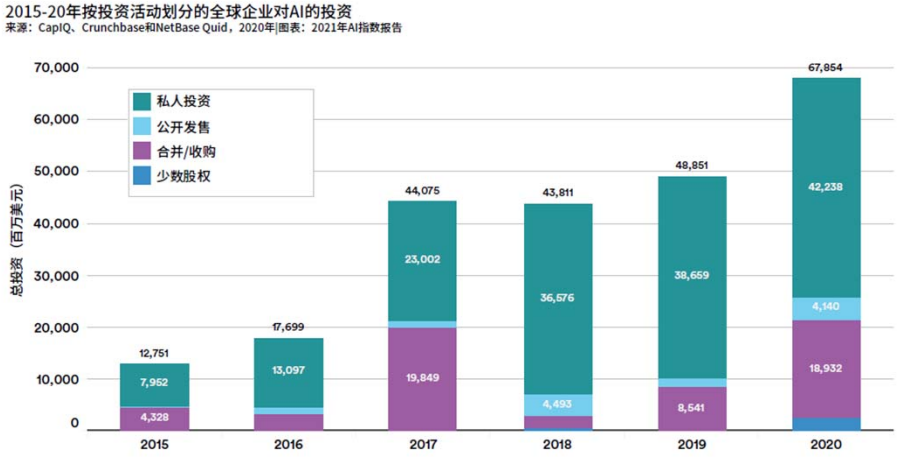


图1

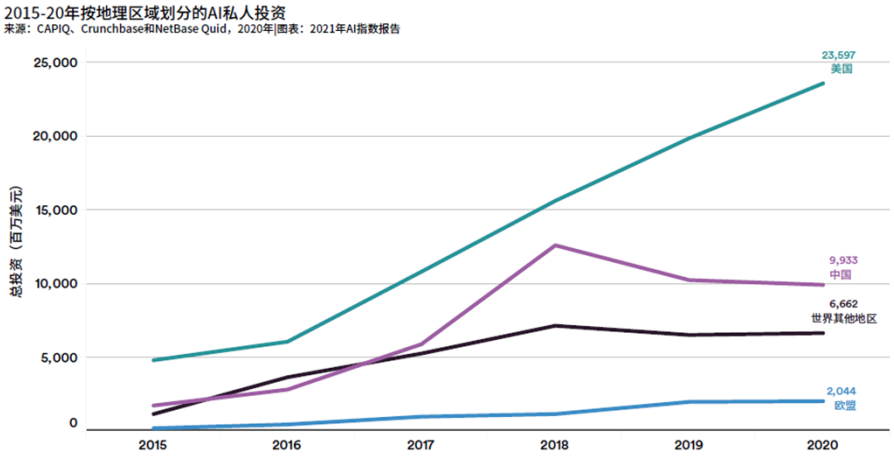


图2

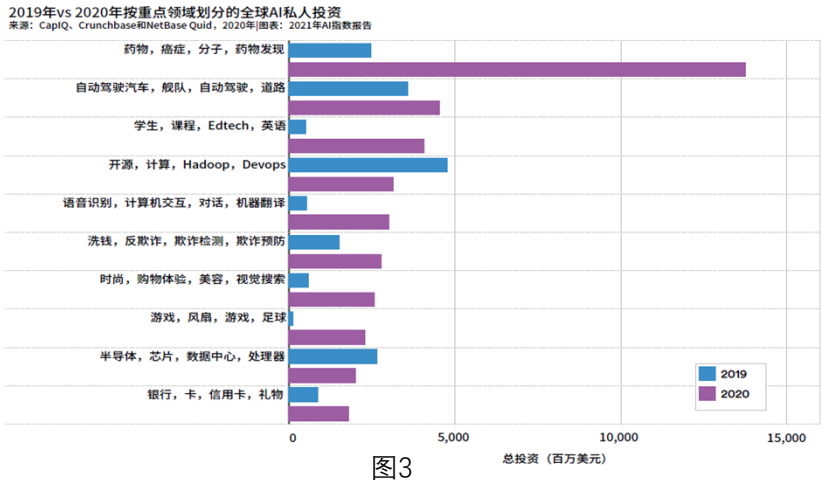


图3

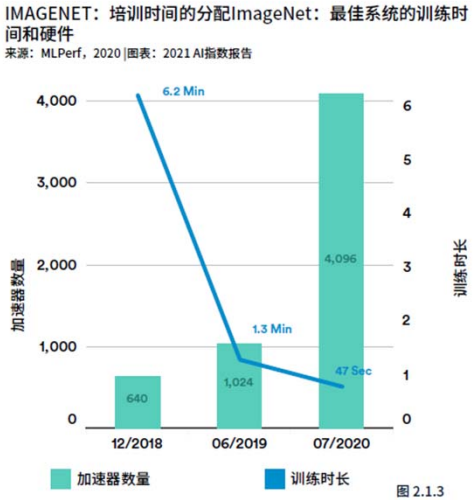
全球的人工智能技术突破——5个关键突破

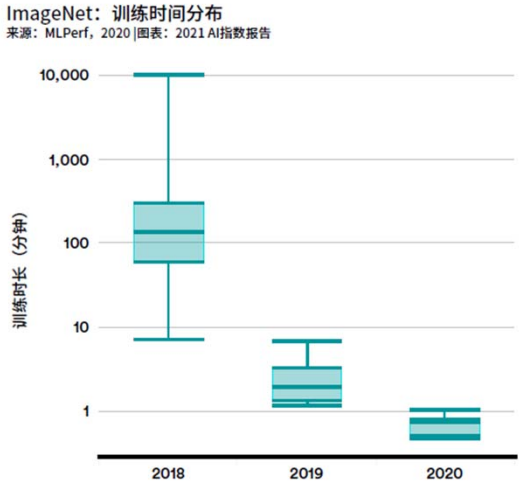
- **技术推动产业落地**：自然语言处理、图像生成技术的产业应用已经能够提供成熟的产品。人工智能技术已经开始部署在虚拟数字人图像采集、模型优化等领域。同时，应用图像处理技术的计算机辅助决策也开始进入制造业。应用图形识别技术对流程和工艺进行优化，同时使用增强现实技术，优化产业工人的生产流程。

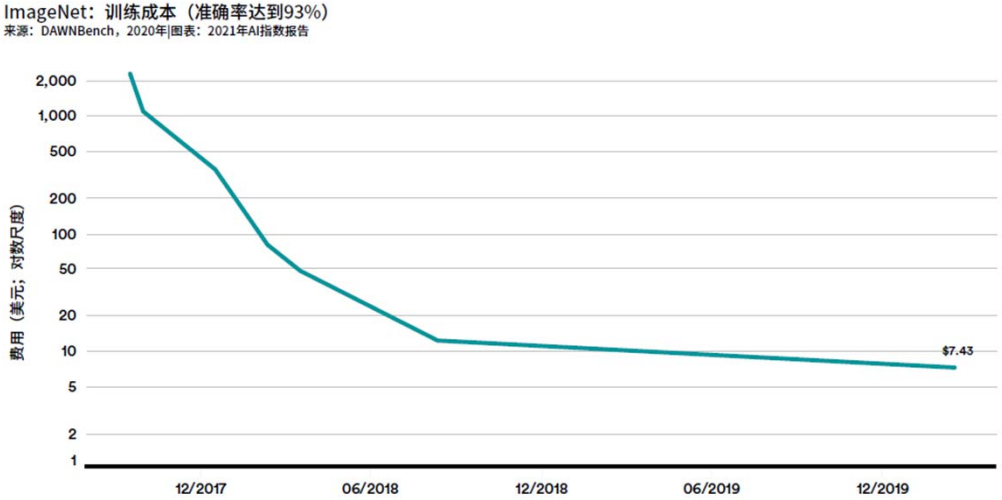
- **计算机视觉的产业化加速**：过去十年，得益于机器学习技术（特别是深度学习技术）的应用，计算机视觉取得了巨大进展。新的数据显示，计算机视觉正在实现产业化。在一些较大的基准库中，算法或模型的性能已经开始趋于平稳。这表明计算机视觉社区需要致力于开发和确定难度更大的基准，以进一步测试性能。各公司正在投入越来越多的计算资源，以比以往更快的速度训练计算机视觉系统。同时，用于已部署系统的技术，如用于分析视频静止帧的对象检测框架，正在迅速成熟，这表明人工智能将会进一步在产业场景中部署。

- **自然语言处理(NLP)超越了现有人类评估基准**：谷歌和微软都在其搜索引擎中部署了BERT语言模型，而微软、OpenAI等公司也开发了其他大型语言模型。NLP的研究进展迅速，以至于它已经开始超过了用于测试它们的基准。在第四届机器翻译大会（WMT19）的竞赛中，Facebook AI使用了一种新型的半监督训练，在几种语言翻译中获得了第一名。Facebook 还引入了一种新的自我监督的预训练方法——若伯塔（RoBERTa），它在数个语言理解任务上超越了所有现有的NLU（自然语言理解）系统。而在某些情况下，这些系统甚至优于人类，包括英德翻译和五个 NLU 基准。
- **推理评估基准的完善**：人工智能知识推理算法的表现，在2018年之前，一直是在当时最佳系统上进行度量。而在2019年，分别计算算法和解算器的性能改善的评估方法被引入SAT竞赛中。通过引入夏普利值，裁判可以确定算法和设施对性能改善的贡献。这种计量方法不仅强化了算法的竞争，也促进了解算器及其运行系统的进步。
- **机器学习正在改变医疗和生物学领域的游戏规则：**随着机器学习技术的引入，医疗和生物行业的格局发生了实质性的变化。DeepMind的AlphaFold应用深度学习技术，在蛋白质折叠这一长达数十年的生物学难题上取得了重大突破。科学家利用机器学习学习化学分子的表示，以制定更有效的化学合成计划。PostEra是一家人工智能初创公司，这家公司在COVID-19流行期间使用基于机器学习的技术来加速发现COVID相关的药物。

例子：AI图像识别——ImageNet







基于机器学习的图像识别技术是人工智能的典型应用。利用训练数据集对人工智能算法进行优化，以实现机器对图像的识别。

来自斯坦福大学和普林斯顿大学的计算机科学家于2009年创建了ImageNet。ImageNet是一个包含超过1400万张图像的数据集，包括200个类别，提供了扩展和改进的人工智能算法训练数据。2012年，来自多伦多大学的研究人员利用深度学习技术，将ImageNet大规模视觉识别挑战赛的结果提升到了一个新的水平。至此，深度学习成为图像识别的重要技术。

ImageNet挑战的图像分类任务要求机器根据图像中的主要对象，对图像进行分类。在给定分类准确性的基础上，自2018年起，算法设计团队不断缩短算法训练时间。与之一同降低的还有算法训练的成本。在2020年，顶尖团队训练一套图像识别算法的成本只有几美元。

在训练算法时，工程团队也部署了更多加速器，这与人工智能日益普及的趋势是一致的。

人工智能的行业部署

2020年各行业和职能部门采用AI的情况
来源：麦肯锡公司，2020年|图表：2021年AI指数报告

产业	人力资源	制造业	市场营销与销售	产品和/或服务开发	风险	服务运营	战略与企业融资	供应链管理
所有行业	8%	12%	15%	21%	10%	21%	7%	9%
汽车与装配	13%	29%	10%	21%	2%	16%	8%	18%
商业，法律和专业服务	13%	9%	16%	21%	13%	20%	10%	9%
消费品/零售	1%	19%	20%	14%	3%	10%	2%	10%
金融服务	5%	5%	21%	15%	32%	34%	7%	2%
医疗保健/制药	3%	12%	16%	15%	4%	11%	2%	6%
高科技/电信	14%	11%	26%	37%	14%	39%	9%	12%

2020年嵌入标准业务流程的AI能力
来源：麦肯锡公司，2020年|图表：2021年AI指数报告

产业	自动驾驶汽车	计算机视觉	对话界面	深度学习	NL一代	NL语言理解	NL文字理解	其他机器学习技术	物理机器人	机器人过程自动化
所有行业	7%	18%	15%	16%	11%	12%	13%	23%	13%	22%
汽车与装配	20%	33%	16%	19%	12%	14%	19%	27%	31%	33%
商业，法律和专业服务	7%	13%	17%	19%	14%	15%	18%	25%	11%	13%
消费品/零售	13%	10%	9%	6%	6%	6%	9%	12%	23%	14%
金融服务	6%	18%	24%	19%	18%	19%	26%	32%	8%	37%
医疗保健/制药	1%	15%	10%	14%	12%	11%	15%	19%	10%	18%
高科技/电信	9%	34%	32%	30%	18%	25%	33%	37%	14%	34%

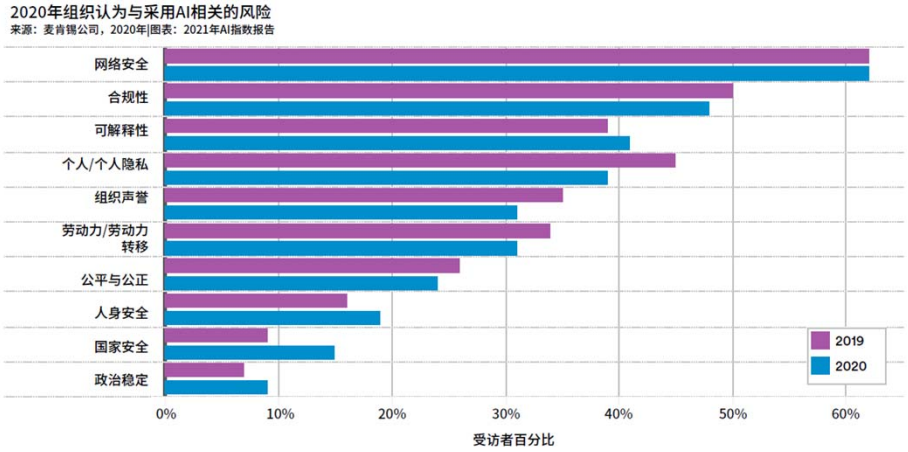
人工智能的产业应用仍相对有限，如图1所示，仍集中于资本密集领域，例如制造业、金融业的风险管理和服务以及高科技产业。

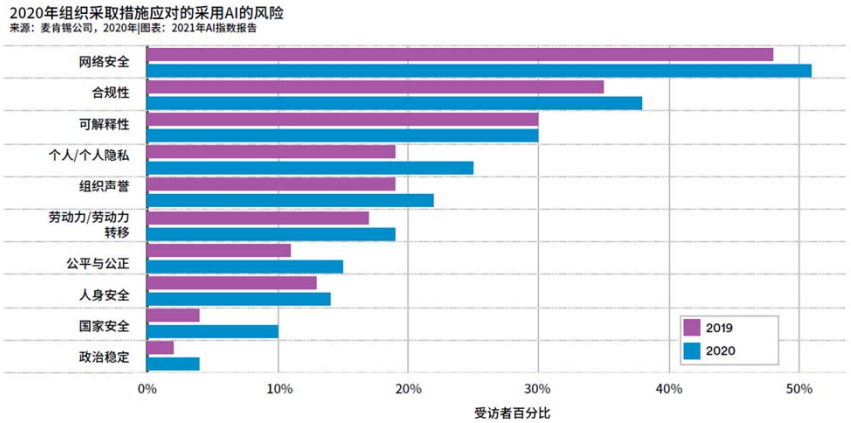
而根据图二，高科技行业之外，在一般行业中，人工智能的计算机视觉、对话交互和机器人得到更广泛的应用。

人工智能逐渐协助人处理复杂工作，并开始在核心商业流程中起关键作用的情形愈发普及。

人工智能的风险与预防

- 在AI加快部署的背景下，很多企业也在探索AI部署可能带来的风险。
  - 网络安全是市场主体的首要关心。网络安全，也即企业的重要数字资产是否得到妥善保护。
  - 合规性，也即AI场景应用是否符合该场景领域的法律法规；
  - 算法的可解释性，即算法的支持逻辑是否可以使用日常语言解释；
  - 隐私问题上，由于2018年，欧盟的《一般数据保护规则（GDPR）》生效。随后，中国也启动了“个人信息保护”立法（2021年8月20日通过了《个人信息保护法》）。世界范围内的数据治理框架立法加速，提高了市场主体在隐私方面的风险意识。因此，隐私方面的，投资也有所提高
- 与之相对应的，各企业也不断提高AI风险的应对措施。利用隐私学习、同态加密和区块链技术实现数据的合规处理和安全流转。



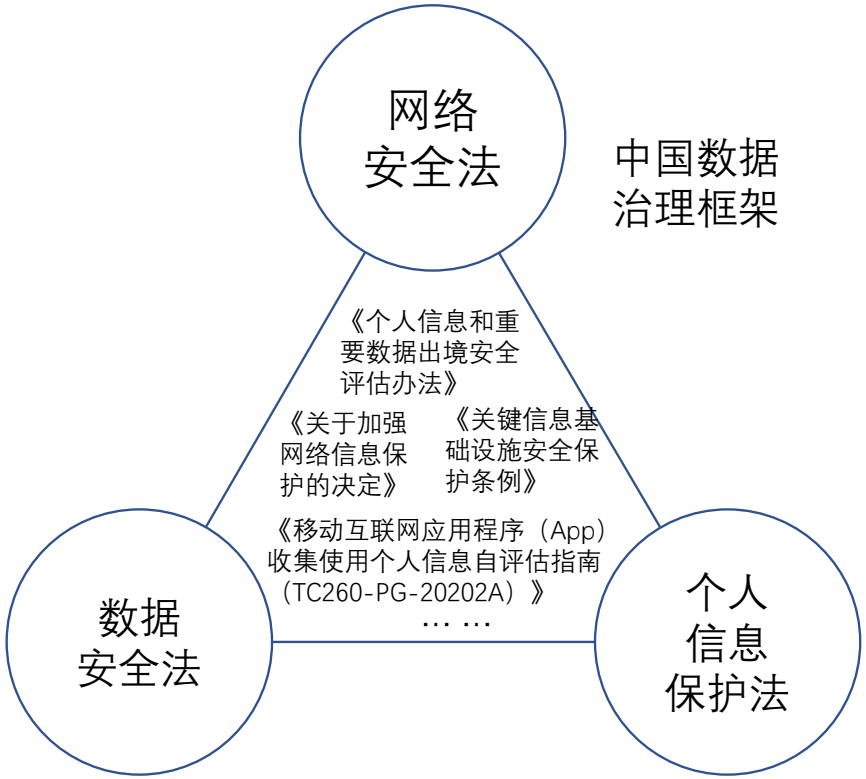


02

《网络安全法》与信息安全技术标准

网络安全法中的基本原则

- 一、网络空间主权原则
  - *第一条　为了保障网络安全，维护网络空间主权和国家安全、社会公共利益，保护公民、法人和其他组织的合法权益，促进经济社会信息化健康发展，制定本法。*
  - 互联网虽然建立一个无国界的空间。但在互联网中活动的个体仍然来自各个国家，仍具备地域属性。因此网络安全监管，仍需要坚持网络空间主权原则。在产业实践上，很多电子商务网站采用了纵向地域限制技术，以减少版权作品的非法使用。
- 二、网络安全与信息化发展并重原则
  - *第三条 国家坚持网络安全与信息化发展并重，遵循积极利用、科学发展、依法管理、确保安全的方针，推进网络基础设施建设和互联互通，鼓励网络技术创新和应用，支持培养网络安全人才，建立健全网络安全保障体系，提高网络安全保护能力。*
  - *第十三条　国家支持研究开发有利于未成年人健康成长的网络产品和服务，依法惩治利用网络从事危害未成年人身心健康的活动，为未成年人提供安全、健康的网络环境*
  - 中国政府已经把数据治理放在国家战略层面。包容型治理是中国信息化发展的原则，在保证网络安全的基础上推动互联网、数字经济有序发展是《网络安全法》立法的应有之义。
- 三、共同治理原则。
  - 网络空间安全仅仅依靠政府是无法实现的，需要政府、企业、社会组织、技术社群和公民等网络利益相关者的共同参与。《网络安全法》坚持共同治理原则，要求采取措施鼓励全社会共同参与，政府部门、网络建设者、网络运营者、网络服务提供者、网络行业相关组织、高等院校、职业学校、社会公众等都应根据各自的角色参与网络安全治理工作。



了解风险、预防风险、管理风险——网络安全管理的开始

- *《网安法》第三十一条 国家对公共通信和信息服务、能源、交通、水利、金融、公共服务、电子政务等重要行业和领域，以及其他一旦遭到破坏、丧失功能或者数据泄露，可能严重危害国家安全、国计民生、公共利益的关键信息基础设施，在网络安全等级保护制度的基础上，实行重点保护。*
- *“风险意味着预期的变化和偶然”——安东尼·吉登斯《现代化的结果》*
- *“风险是现代化的一个附带问题。”“风险的扩散和商品化并不破坏资本发展的逻辑，它意味着资本进入一个新的阶段”——乌尔里奇 贝克《风险社会》*
- *“3.29 信息安全——信息保密性、完整性和可用性的保存。——《ISO/IEC 27000 信息安全——概览与词汇》*

- 国际标准化组织（International Organization for Standardization, ISO）确定了信息技术中风险管理的内涵。在此基础上，各经济体开发出自有且与外国体系兼容的风险管理方法。
  - 《ISO 31000 风险管理》将风险定义为“目标上的不确定性的意外影响”。
  - 《ISO 27005 信息安全风险管理》定义了，当“**对资产的弱点产生威胁以对组织造成伤害**”时，即认为**风险出现**；
  - 国标《GB/T25069—2010 信息安全技术 术语》 2.3.35 **风险** 一个给定的威胁,利用一项资产或多项资产的脆弱性,对组织造成损害的潜能。可通过事件的概率及其后果进行度量。
  - 欧盟《网络与信息安全指令（2016/1148）》将风险管理视作网络安全的重要组成部分。
  - 法国国家网络安全局（ANSSI）开发“艾比奥斯（ EBIOS ）”网络安全管理工具；德国联邦信息安全办公室（BSI）开发了“信息技术基本保护（IT-Grundschutz）”标准工具。

- 但**风险管理并不能提供一种全面的方法彻底消除风险，而是采取一种可行的方法识别和降低风险，在风险发生后，降低风险引发的负面影响。**——欧盟网络安全局（ENISA）《港口网络风险管理（建议）》

技术标准在人工智能评估中的角色

- 《网安法》第十条　建设、运营网络或者通过网络提供服务，应当依照法律、行政法规的规定和**国家标准的强制性要求**，**采取技术措施和其他必要措施**，保障网络安全、稳定运行，有效应对网络安全事件，防范网络违法犯罪活动，维护网络数据的完整性、保密性和可用性。
- 2021年4月21日，欧盟委员会放出了《制定人工智能协调规则并修改特定联盟法规》，也即《AI法案》。该法案尝试将人工智能定义为一种软件，并尝试将人工智能整合进欧盟现有产品质量标准体系。欧洲产品质量标准体系也即CE标准（*Conformité Européenne*）
- 使用技术标准评估人工智能可能成为一种**可实践**的人工智能评估方法。
  - ISO体系内形成了完整的"信息安全——隐私影响评估——风险管理"的标准框架；
  - IEEE正在开发"P7000"系列人工智能评估方法。
  - 欧盟网络安全局（ENISA）、欧洲数据保护委员会（EDPB）、欧洲数据安全主管（EDPS）都建议了"网络安全——数据保护——风险管理"的方法。尤其地，欧盟委员会建议了"风险方法"作为创新技术影响评估的基础方法。
  - 法国的EBIOS、德国的"信息技术基本保护"管理工具和标准数据保护模型（SDM）。
  - 中国的信息安全标准化委员会在ISO的"信息安全标准系列"之上，还制定了《网络安全标准实践指南 —人工智能伦理安全风险防范指引》
- 基于"网络安全——数据保护——算法审计"的技术标准评估方法可以保证人工智能风险评估与之前标准体系的兼容性。

ISO　IEC　IEEE Advancing Technology for Humanity

全国信息安全标准化技术委员会 NATIONAL INFORMATION SECURITY STANDARDIZATION TECHNICAL COMMITTEE　CE

ETSI MEMBER　acm Association for Computing Machinery

edpb European Data Protection Board　enisa EUROPEAN UNION AGENCY FOR CYBERSECURITY

03 人工智能社会风险评估的技术实现

识别与降低人工智能的风险的方法

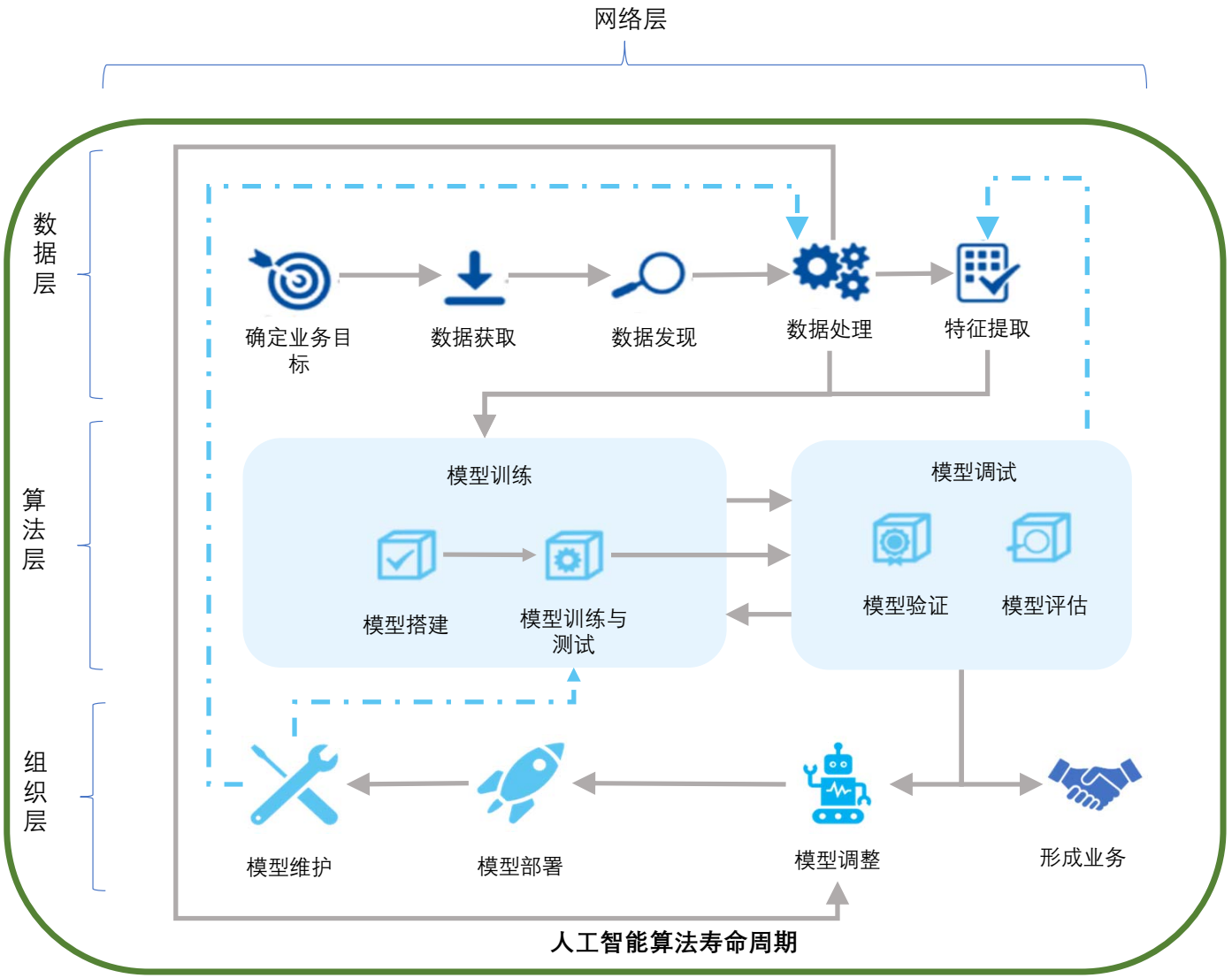
网络层：需要制定完整、可实施的网络安全管理策略。对业务核心算法，应制定设备保护计划和应急处置方案。在人工智能应用数据储存的信息域内，数据（尤其是个人数据）应是默认保护的。

数据层：对数据的安全要求进行识别，确定数据的保护要求，具体包括识别数据库的可用性、完整性和保密性。对数据库的转移链路和访问权限应作记载。

算法层：基于网络层和数据层的风险管理策略，算法应部署必要算法，以实现网络安全和数据保护，具体包括隐私计算、区块链等算法。

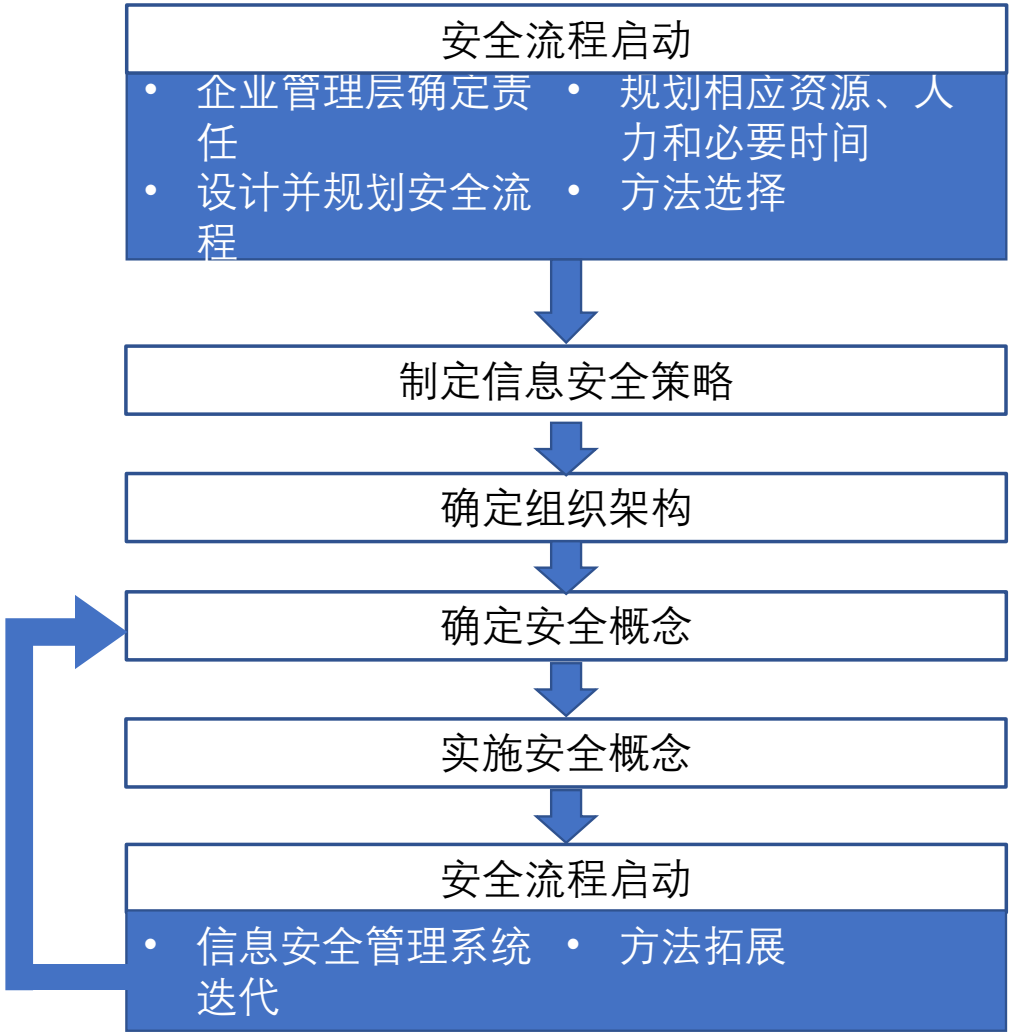
公平算法审计：由于网络层和数据层安全策略的技术实施必然产生相应的基础算力要求，在这种背景下，算力资源成为衡量企业人工智能风险管理策略实施一致性的重要标尺。（Github开源项目洞察报告）

良好组织框架：大型企业良好的组织框架，相对于先进技术研发，显然是成效更划算的风险管理方式。但这对于中小企业可能存在问题，中小企业并没有足够的人手，在人工智能风险管理上，派出专门人手。

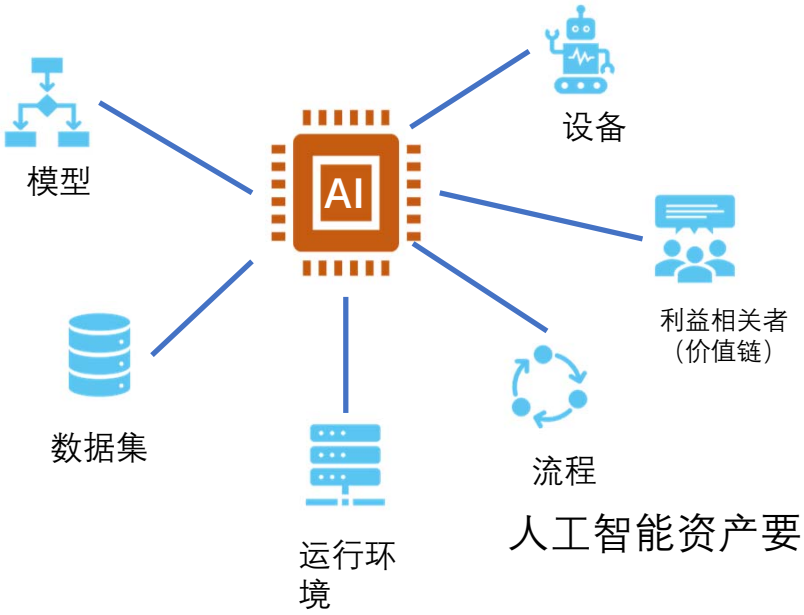


安全流程——以人工智能为例

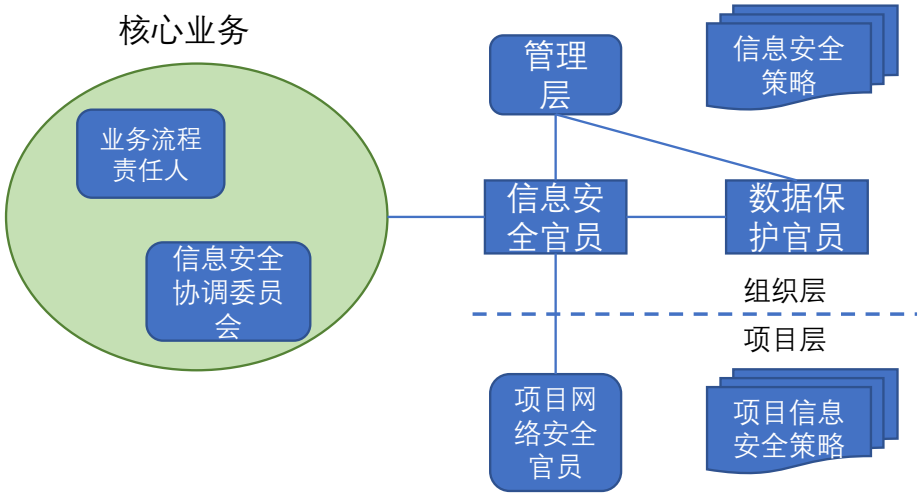
- 网络安全的有效实现=“内控制度+负责人+有效记录+分级管理+技术保障”——《网安法》第二十一条
- 人工智能安全管理的前提是责任明确。因此管理层在机构管理制度上的人工智能安全策略是安全流程的第一步。
- 管理层需要指定专人对人工智能的安全问题负责，这样可以回避责任真空问题，保证机构内部持续有人对人工智能的安全运转付出努力。另外，在机构员工层面，也需要准备相应的资源和人力，以保证管理层能正常履责。当然，中小企业可能并没有这么多员工，很可能一人身兼多职。
- 在方法选择上，《网络安全法》第四十六条第三款，建议了对网络安全事件进行分级。事实上，分级管理是信息安全管理方法中的重要思想。对于一些影响较小的，且极容易被取代的业务流程，并不需要很高级别的保护。



关键流程1：安全策略（Security Policy）评估



人工智能资产要素



典型中型企业的人工智能信息安全组织架构

人工智能安全策略

信息是公司和政府机构的重要资产，因此需要充分保护。 今天，大部分信息都是在信息技术 (IT) 的帮助下生成、存储、传输或进一步处理的。

人工智能的资产要素包括算法模型、数据集、机构的数字化业务流程、算法的运行环境、运行设备和人工智能价值链上的利益相关者。

人工智能介入制造、管理领域的业务流程。利用图像识别和自然语言处理技术，人工智能在生产和流通领域辅助人类决策。但人工智能的安全问题和风险管理往往被忽略，由于人们对机器的盲目信任。这种现实可能威胁机构的生存，无论是政府机构还是企业。而良好的人工智能安全策略可以利用灵活的可拓展组织架构、使用相对合适的资源实现。

关键流程2：记录技术（Documentation）——管理策略

业务流程				
ID	流程描述	流程分类	责任人	实施者
ZYBP001	**生产：**模型搭建、环境配置、算法训练并提供终端产品。包括研发、数据传输、测试以及产品部署。	核心	生产总监	所有员工
ZYBP002	**售前访问**：处理客户需求。通常以邮件、访问形式进行。信息存在于数字介质或传统介质	支持	售前总监	销售部门
ZYBP003	**订单确认：**客户使用邮件或者挂号信寄送邮件。所有收据必须以纸质形式打印。消费者会需要再次确认订单，如果有订制生产流程或者依客户需求。	核心	订单处理	销售部门

范例1：业务流程评估表

- *《网安法》第二十一条第三款 采取监测、记录网络运行状态、网络安全事件的技术措施，并按照规定留存相关的网络日志不少于六个月；*
- 记录技术是人工智能安全评估中的重要组成部分。这种方法可以帮助机构建立完整有效的安全事件分级体系。
- 系统的安全是建立在各个部件的安全的基础上。利用通过建立完整的业务、工具以及二者之间的映射的记录，形成对机构的数字化业务的完整记录，确保业务与工具清晰的责任人和实施者。
- 这种方法也可以保证在发生安全事故时，事故损失被有效评估，故障组件（设备、网络组件、数据）被及时隔离。

数字化工具描述						
ID	工作工具描述	业务平台	设备位置	状态	用户	管理员
ZYA001	**文本处理**	MS Word 2016		运行	所有员工	运营/维护
ZYA002	**即时聊天**	标准化软件		运行	所有员工	运营/维护
ZYA003	**订单系统**	智用数据库	北京市石景山区实兴大街30号院17号楼606	测试	数据库管理员	运营/维护

范例2：结构分析表

业务流程/数字化工具映射					
业务流程/数字化工具	ZYA001	ZYA002	ZYA003	ZYA004	ZYA005
ZYBP001	X	X	X		
ZYBP002	X		X		
ZYBP003				X	
ZYBP004		X			X
ZYBP005		X		X	X

范例3：业务流程/数字化工具映射

关键流程2：记录技术（Documentation）——设备记录

《网安法》第三十三条
建设关键信息基础设施应当确保其具有支持业务稳定、持续运行的性能，并保证安全技术措施同步规划、同步建设、同步使用。

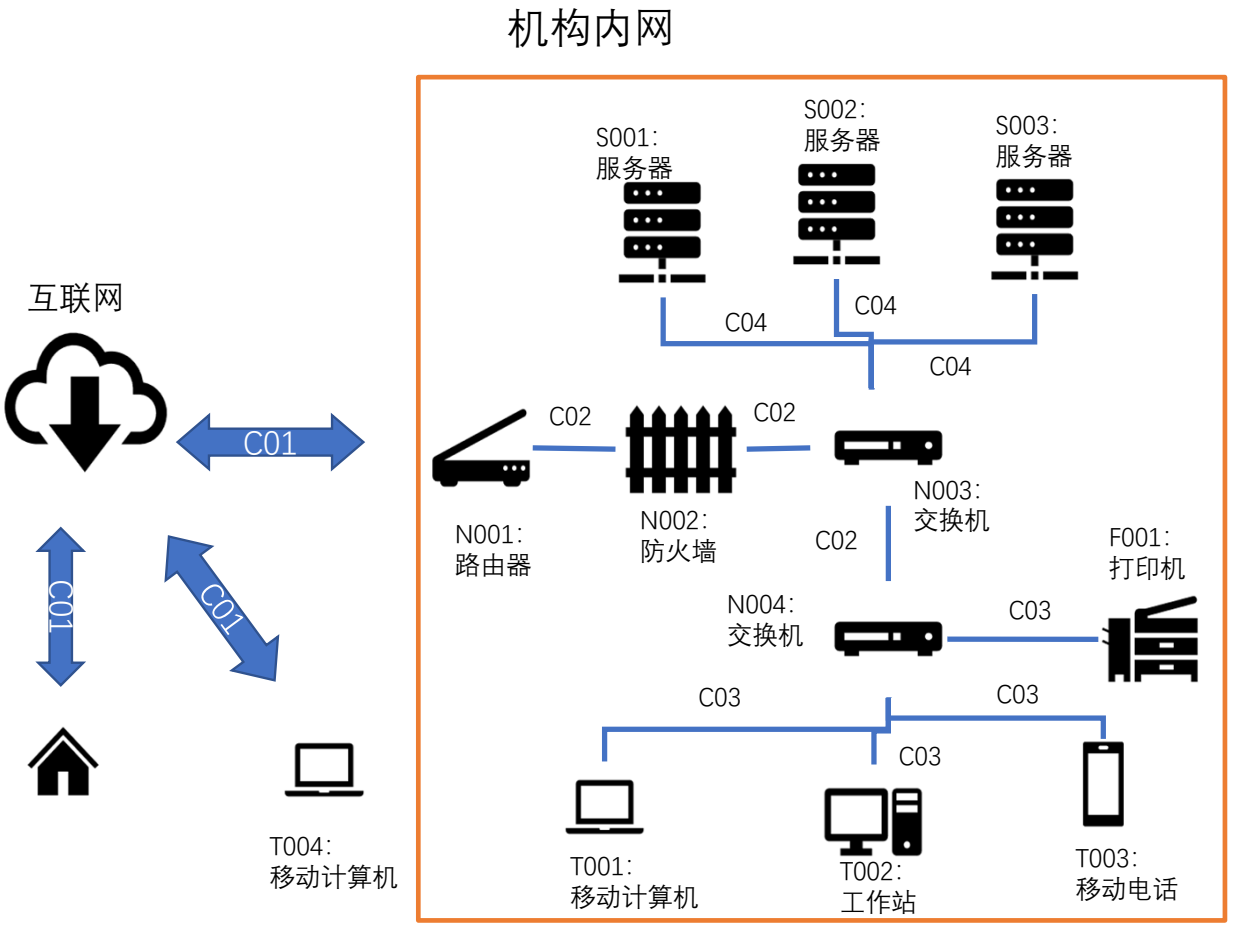
对人工智能运行的设备和基础网络进行编号并登记是确定设备保护要求的重要工作。

根据关键商业流程涉及到的设备和工具，以及不同设备在网络中的位置，确定关键信息基础设施和关键链路，进而设计专门保护措施。

设备标记方法

CXX：网络链接
NXXX：网络中的节点
SXXX：服务器
TXXX：终端设备
FXXX：关键设备

在这张图表中，链路C02就是关键信息链路，是业务正常运转的基础。同理，节点N001、N002和N003是关键信息基础设施。而相对于N001，N004的重要性相对较低。N004的瘫痪，只导致机构内部设备无法访问服务器和内部联系。而外部仍是能访问机构服务器的。



关键流程3：定义安全——风险识别

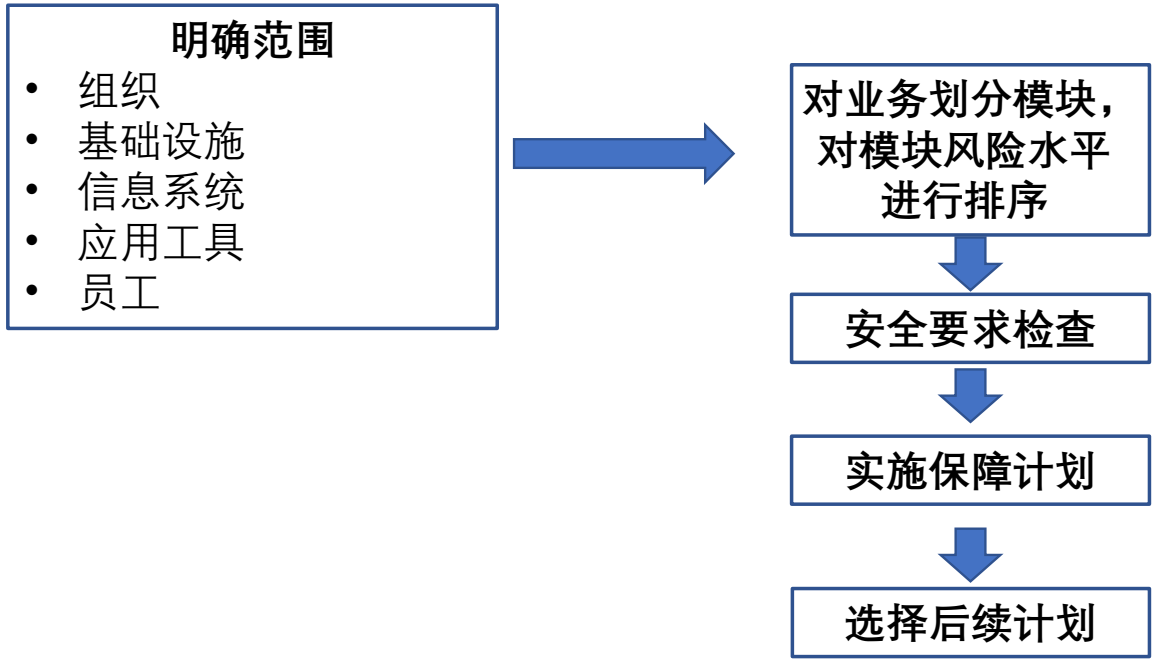
安全并不是个绝对的概念，人工智能的网络安全应是风险管理问题。也即识别风险类别、确认风险严重性和发生可能性，并提供相应的保障计划。

在现有的方法中，首先我们需要做的就是对业务进行**模块划分**，形成“业务流程——设备——数据”的模块，确认关键模块。在安全检查中，对业务的各子模块进行模拟压力测试，确认风险类别、严重性以及可能性。

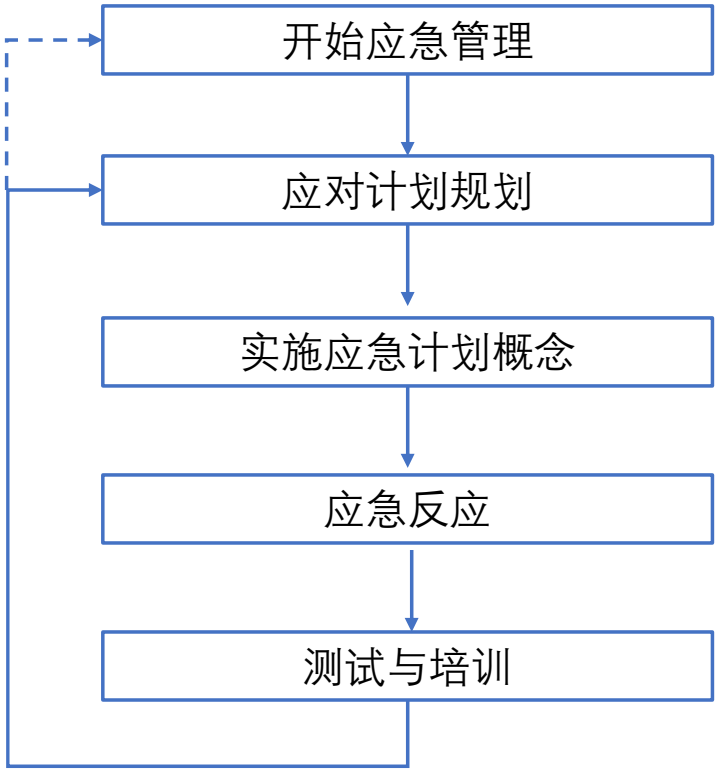
对于关键业务流程，应形成**保障计划**。确保事故发生时，机构的关键业务不受影响。

在完成了上述流程之后，一个机构的业务也就**完成了网络安全的定义**。其核心要点是在风险事件发生时，保证关键业务不受影响。

由于网络威胁的不断演进，网络安全计划的可拓展性也是在定义安全需要考虑的问题。及时对漏洞进行修补并随着业务扩大，隔离单个事故对整个系统的影响是人工智能应用的安全性拓展的重要议题。对于一些开源的人工智能算法，开源社区的更新减速意味着算法的漏洞可能被更广泛地发现。这对开源算法的使用者造成严重威胁。



关键流程4：业务持续性管理



- *《网安法》第五十五条 发生网络安全事件，应当立即启动网络安全事件应急预案，对网络安全事件进行调查和评估，要求网络运营者采取技术措施和其他必要措施，消除安全隐患，防止危害扩大，并及时向社会发布与公众有关的警示信息。*
- 算法黑箱使人工智能部署时社会反应存在极大不确定性。虽然机构制定了安全策略，并实施了保障计划，但对业务持续性进行管理仍是人工智能应用需要考虑的问题。
- 在应急计划规划中，首要目标仍然是保证人工智能相关资产的安全，将事故隔离在业务流程之外。而且在制定应急计划时，在将“流程——设备——数据”模块化之后，需要对易损设备制定事故替代计划。当然这里，进行成效分析（Cost-benefit Analysis）也是一个很重要的工作。如果某个设备一旦损害，需要直接更换，但可以在短时间内完成，且设备储存成本较高。那么事故发生时，产生的业务暂停就是可接受的，只是在应急计划中需要将维修时间纳入考虑。

数字化解决方案

- 社会风险评估采取“专家系统+自动化评论+关系数据库”的技术路线。
- 深度学习聚类→风险预警（图2）
  - 对人工智能应用的策略、数据、算法进行特征抽取，并作无监督聚类。对具备相似特征的人工智能聚类，在一个人工智能业务出现风险时，及时向相似项目告警。
- 私有云+纵向隐私计算
  - 评估数据储存在私有云上。同时探索纵向隐私计算在风险预警上的可能性。

图1：专家系统

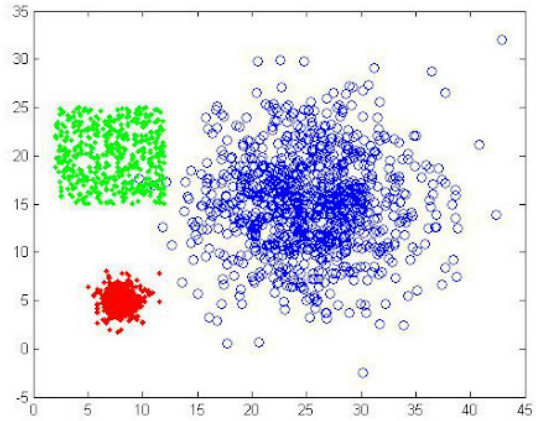
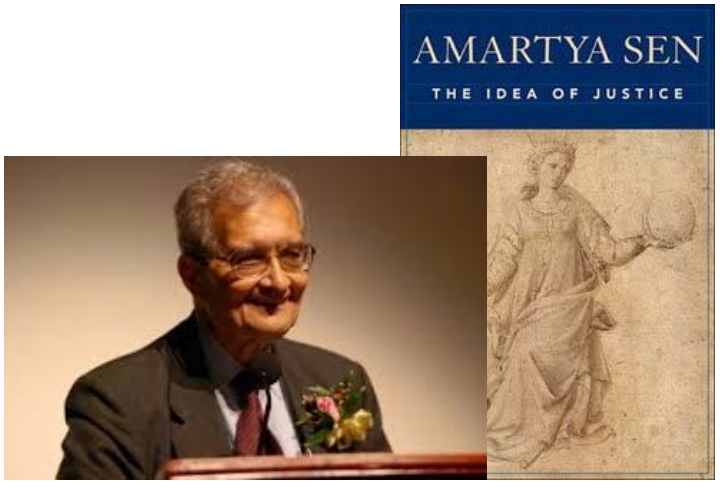


图2：自动化聚类分析

评估机构的第三方角色

- 非盈利性
  - 《民办非企业单位登记管理暂行条例》第四条 “民办非企业单位不得从事营利性经营活动。”也即举办人在民非法人终结时，不得将法人资产余额分配。
  - 这种属性决定了民非法人更多地忠诚于行业利益而非投资人的私人利益。进而，对整个行业以及个体企业的真实报告和信息处理效率是民非法人的核心竞争力。

- 缩短学术-产业价值链条
  - **协同创新**。阿玛提亚 森（Amartya Sen）将发展定义为，在不改变投入总量的前提下，尽可能满足所有个体的需求。因此，**获知市场需求**是风险管理的重要组件。人工智能应用社会风险评估在开发众包平台，一方面，利用精英学术资源，提高学术投入的社会接受水平；另一方面通过扩大自动化评估的业务面，及时反哺学术研究。



阿玛提亚 森（Amartya Sen）与《正义观点》

感谢聆听！