人工智能治理与可持续发展实践白皮书

可用 | 可靠 | 可信

人工智能治理与可持续发展实践白皮书

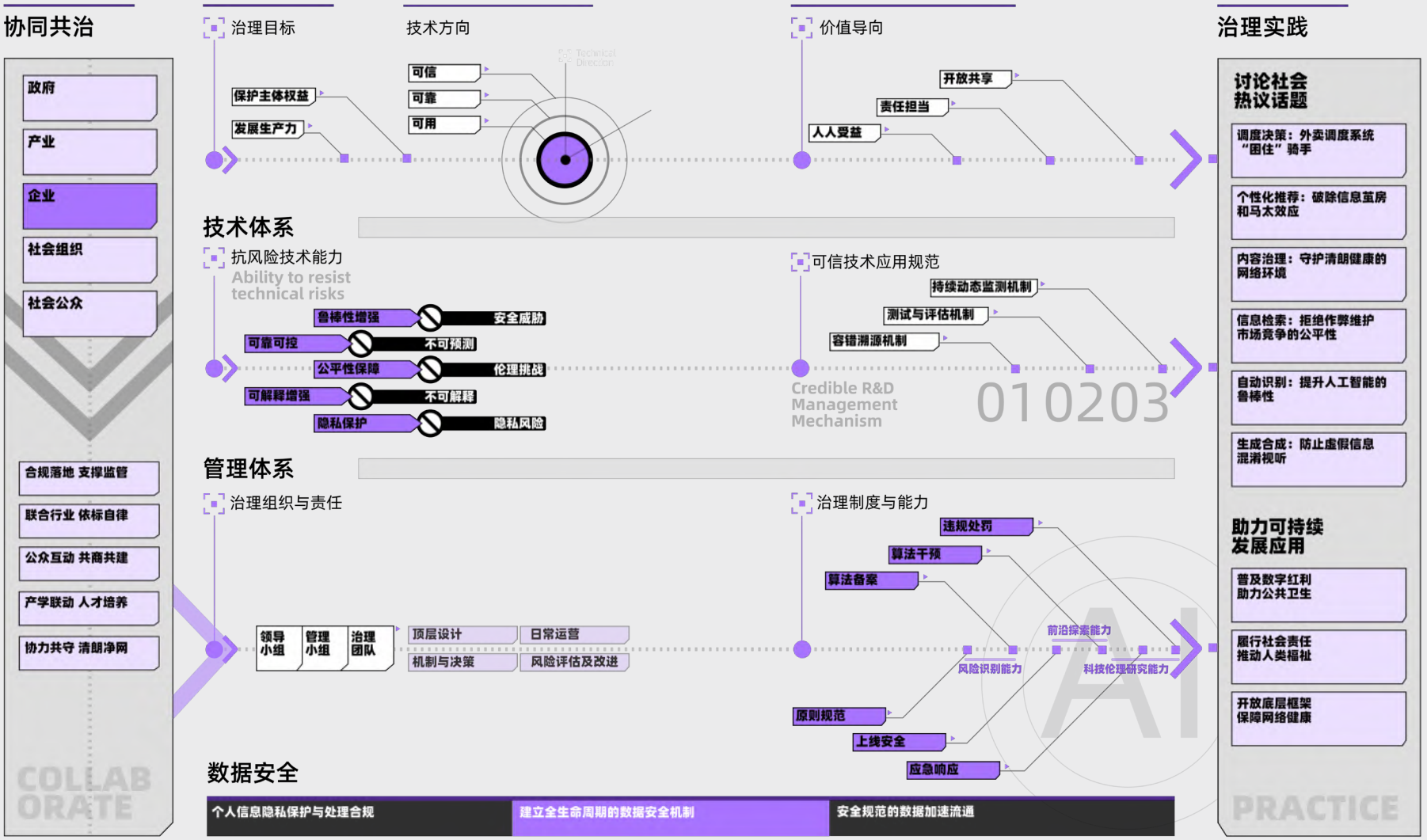
可用 | 可靠 | 可信

ARTIFICIAL INTELLIGENCE GOVERNANCE& SUSTAINABLE DEVELOPMENT

Alibaba Group 阿里巴巴集团

CAICT 中国信通院

企业
面向可持续发展的人工智能治理体系



前言 FOREWORD

当前，人工智能技术蓬勃发展，广泛赋能千行百业，给人类生产生活带来了深刻变化。人工智能在促进社会发展的同时，也存在着风险和挑战。习近平总书记曾多次强调，“要确保人工智能安全、可靠、可控”。开展人工智能治理，发展负责任和可信的人工智能，助力可持续发展愿景实现，正在成为共识。

全球已经形成了大量的人工智能治理与发展原则，下一步的工作重点，是将抽象的原则落实到实践中。企业作为人工智能技术研发和应用的重要力量，亟需探索一套适合自身业务发展的人工智能治理实践体系，将各项治理要求贯彻到人工智能的全

生命周期，以有效治理为人工智能红利的释放奠定基础，从而加速推动可持续发展愿景的实现。

本白皮书全面总结了阿里巴巴在人工智能治理与可持续发展领域的实践，重点针对当前人工智能应用中的热点问题，从数据、技术、管理及多元协同等方面系统性介绍了我们的实践思路和方法，同时辅以若干专题进行阐释，期待为社会各界提供有益参考。

未来我们将持续跟踪相关领域进展，不余遗力地推动人工智能健康发展。由于人工智能治理是一项复杂的系统性工程，我们的认识还有待进一步深化，白皮书中存在的不足之处，欢迎大家批评指正。

人工智能治理与可持续发展实践白皮书编写组

指导委员会

郑俊芳　阿里巴巴集团首席风险官兼首席客户官

程立　阿里巴巴集团首席技术官兼科技伦理治理委员会主席

魏亮　中国信息通信研究院副院长

专家委员会

钱磊　阿里巴巴集团安全副总裁

魏凯　中国信息通信研究院云计算与大数据研究所副所长

朱红儒　阿里巴巴标准化业务副总裁

黄庆明　中国科学院大学计算机科学与技术学院副院长

屠剑威　阿里巴巴集团法务副总裁

吴飞　浙江大学人工智能研究所所长

安筱鹏　阿里研究院副院长

姚玉峰　浙江大学医学院附属邵逸夫医院眼科主任

编写组组长

薛晖　阿里巴巴人工智能治理与可持续发展研究中心主任兼科技伦理治理委员会执行主席

石霖　中国信息通信研究院云计算与大数据研究所内容科技部副主任

编写组副组长

张荣　阿里巴巴集团算法风险治理团队负责人

刘硕　中国信息通信研究院云计算与大数据研究所内容科技部副主任

朱冉　阿里巴巴集团内容安全副总裁

周斌琦　阿里巴巴集团风险治理与生态合作部总经理

袁媛　阿里研究院数据经济研究中心主任

编写单位

Alibaba Group 阿里巴巴集团

CAICT 中国信通院

编写组主要成员

彭骏涛 傅宏宇 刘翔宇 陈岳峰 杨易侗 段雨泽 安政 呼娜英 孙小童

鸣谢

顾伟 孙勇 马宇诗 杜旭阳 朱琳洁 王彦伟 杜东为 田喜清 眭亚楠

王竟亦 黄正行 张佳一 秦鹏达 许皓天 顾斐 洪澄 季婷 刘大鹏

李进锋 杨颖一 庄涛 曹彬彬 廖剑 阮莎莎 黄龙涛 张东杰 金炫

赵伟 赵平 张冰 张怡远 高婷 王志斌 陈磊 李昊 朱丛

何万青 杨定裕 张静 姜歆 何慧亮 周燕 周喆 孙思怡 杨文波

陈德品 陈克寒 俞林峰 刘源 张波悦

关于我们

阿里巴巴人工智能治理与可持续发展研究中心（AAIG）是阿里巴巴集团旗下的人工智能顶级研发团队，致力于利用 AI 技术解决安全风险问题，并推动 AI 技术更加可用、可靠和可信。团队成员在计算机视觉、自然语言理解、数据挖掘与网络安全等领域的国际顶级会议和期刊上发表论文 100 多篇，获得国际国内专利授权 60 余项，申请中专利 200 多项，所研发的人工智能产品涵盖内容安全、业务风控、数字安防、数据安全与算法安全等多个领域，为集团在全球的千万商家和十亿消费者提供更好的安全和体验。

AAIG

联系我们

aaig@list.alibaba-inc.com

目录

零 ZERO

热议中的人工智能

0.1 人工智能的热点问题

随着人工智能技术发展步入快车道，业务数量和覆盖面不断提升，新业务模式和新产品持续涌现。与此同时也产生了各种科技伦理问题和挑战，特别是那些已经深入到日常生活方方面面的各类算法，引起了社会的广泛担忧。下面主要从阿里巴巴视角，梳理了社会热议甚至是争议的一些情景。

1. 调度决策：外卖调度系统“困住”骑手

外卖平台引入人工智能技术和系统，会帮助骑手规划订单的取餐顺序和配送路线。在大大提升配送效率的同时，也引发了“困住”骑手的质疑，例如系统规划的路径或预计的时长存在不合理情况，而骑手囿于系统限制，为了节省时间往往选择违反交通规则，给自身和他人带来安全隐患。

2. 个性化推荐：电商场景下的信息茧房和马太效应

电商平台根据消费者的兴趣爱好、消费习惯等提供智能推荐服务，节省时间成本、大大提升了匹配效率，但也产生了信息茧房和马太效应。一方面可能出现推荐商品重复率过高的情形；另一方面系统天然倾向于推荐实力强服务好的头部商家，导致市场资源更加集中，加剧了两极分化。

3. 内容治理：如何守护清朗健康的网络环境？

部分不法分子在网络平台上，为了骗取流量，发布各式各样的儿童软色情内容，污染了网络空间，败坏社会风气，严重侵害了未成年人权益。这些低俗信息频繁出现，人工智能可以帮助清理顽固的信息污垢，守护清朗健康的网络环境吗？

4. 信息检索：商品的销量和评价真实可信吗？

在电商平台上，商品销量和评价会在一定程度上决定检索中的排序，从而影响用户的购买意愿。这种机制催生了一些作弊行为，部分不法分子蓄意“灌水”，伪造销量数据，虚构大量相似“好评”。如何防止人工智能被黑灰产恶意投放的作弊数据欺骗，保障信息的真实可信？

5. 自动识别：人工智能可以放心使用吗？

2021 年，美国底特律市的一位男子开启自动驾驶模式后，撞上马路中行驶的半挂卡车，不幸身亡。据分析，事故原因是自动驾驶系统将货车的白色车厢错误识别为蓝天白云，因而没有及时采取刹车制动等措施。之前也发生过数次类似事件，这难免让人们产生疑问，人工智能自动识别的结果可以放心使用吗？

6. 深度合成：眼见不一定为实

近年来，利用“深度合成”技术恶搞公众人物的事件不时发生，造成了恶劣的影响。随着“深度合成”技术和工具快速成熟与扩散，其风险不断增大，恶意使用极易造成虚假信息泛滥，可能严重妨碍司法公正、引发信任危机，甚至危害国家安全。

0.2 人工智能的风险初步探析

引发人工智能风险的因素是多方面的，主要涵盖数据、技术和管理三个方面。一是人工智能系统依赖于训练数据，但时常存在数据选择偏差或质量低下等情况；二是人工智能技术自身还存在不可解释、鲁棒性差等一系列缺陷；三是企业原有的管理体系已经难以适应当前人工智能等新技术的发展情况。

1. 数据不完备和滥用风险突出

1）数据不完备

人工智能进行自动化决策时，如果数据不充分、不达标，就会造成结论偏离的情况。例如，外卖平台的决策调度算法缺少对天气变化、交通状况、电梯的等待时间等数据输入，会给出偏少的预估时间，增加骑手负担；电商平台进行商品推荐时，如果仅使用用户的搜索、购买等行为数据，而缺乏对商品功能与用户兴趣点对应关系的认知，则易形成重复推荐的信息茧房。

2）数据投毒

如果训练集中混杂了虚假的数据，还会对算法形成欺骗，在自动化决策中给出错误的结果。比如在平台电商场景中，商品检索排序依据中包含了商品的销量、评价等信息，如果不法商家雇佣刷手虚构交易和评价，就会对排序算法形成欺骗，使其在排序结果中占据更好的位置。

3）数据滥用

技术进步扩大了用户个人信息的边界，互联网平台企业可以在线且及时的采集用户

购买、收藏、浏览等行为，拥有丰富的算力资源和出众的算法能力，如果企业在借助人工智能对用户数据进行加工、使用的过程中不能够严格遵守法律法规，则可能因为数据滥用而损害用户的权益。

2. 人工智能算法存在固有缺陷

当前，以深度学习为代表的人工智能技术在产业界广泛应用，取得了一系列突破，但其在可解释性、鲁棒性、偏见歧视等方面尚存在局限。

1）可解释性不足

深度学习算法的一个的显著特点是训练过程中自动提取特征，通常比人工挑选的特征效果更好，但这一过程目前尚不可控，在不恰当的数据集上算法可能选择错误的特征。例如，部分模型会把猫识别为狗，可能的原因是算法在自动提取特征时将背景作为了识别的依据，而不是动物本身的形态和细节。当待识别图像出现相似的背景时，会出现错误识别的情况。可解释性不足让人们不能理解算法的决策机理，同时也难以预测算法的行为。

2）鲁棒性不足

深度学习算法在训练过程中会对数据的鲁棒特征和非鲁棒特征进行学习，并依据这些特征进行识别。以图像为例，鲁棒特征可以理解为人类能够理解的语义特征，例如形状、纹理等。而非鲁棒特征为模型能够理解的用于对训练数据进行拟合的特征。**非鲁棒特征给模型的安全性带来了极大的挑战。通常的，可以在输入数据中加入人无法感知到的轻微扰动，激活模型的非鲁棒特征，从而导致模型给出错误的结论。**人工智能算法具有脆弱的一面，可能因为外部的恶意攻击行为，或者无恶意的非平常情况而失灵。

3）偏见与歧视

深度学习算法会挖掘训练数据集中不同因素的相关性，拟合数据分布特性，训练数据集本身的偏见与歧视，会被引入到训练出的模型之中。当模型应用于业务，尤其是用于自动化决策时，可能会暴露出偏见与歧视。当前在自然语言处理等领域，算法的开发普遍采用基础模型加精调的模式，基础模型本身存在的偏见与歧视还会传递到多个下游模型里，影响范围持续扩大。

3. 企业人工智能管理体系不完善

人工智能等新技术特有的应用特征对企业的管理措施提出了极大挑战。一方面，过去为了鼓励创新和效率优先，通常让基层拥有较大的自主权；另一方面，人工智能新技术的负面影响通常不会立即显现，也难以全面评估。这就使得原有的体系并不能适应当前人工智能治理原则。

1）算法需要人为干预

由于人工智能算法固有的缺陷，需要对可能出现的错误结果进行干预纠偏。人工智能算法应用于自动化决策时，如果决策由机器单向做出，缺少相应的人类干预手段，会带来很多问题。比如，在商品检索应用中出现大量重复或相似商品，如果不做去重、打散等干预操作，会导致用户难以快速找到自己喜欢的商品。如果决策的对象是人，在客观上不应该仅由人工智能做决定，还应辅以制度、人情、环境和文化等人文因素。再如，外卖平台可能对骑手分配的单量过大，如果没有给骑手提供干预的渠道，会导致骑手劳动量过大，甚至引发安全事件。

2）用户权益保障不足

人工智能算法用于自动化决策，对用户带来明显影响，并不能做到完全技术中立，需要注意保障用户权益。人工智能应用对用户具有较强的支配能力，且具有信息不对称的特点。用户难免担忧这种支配可能伴随着偏见、歧视等不公平的对待，引发

用户面对人工智能时的无力感：不知道人工智能使用了自己的哪些个人信息，不知道人工智能决策的逻辑，决策的结果自己也无法反对。这些问题的原因在于用户权益保障不足，包括知情权、反馈权与选择权、平等权等，导致用户失去对人工智能应用的信任感。

3）主体责任落实不到位

由于人工智能技术门槛高，且在企业中的运用往往呈现出高动态性、高复杂度等特点，使得外部难以理解其运行机制。而**企业作为人工智能系统的设计者和服务提供者，最了解其中的技术细节和可能蕴含的风险，需要主动承担起相关责任，做好人工智能服务目的与运行机理的解释说明，充分评估潜在的风险并做出相应的防范。**例如，人工智能合成内容已经非常逼真，在缺乏足够信息时，人眼和技术手段都很难分辨真伪。这就需要企业在提供生成合成服务的源头做好各项风险管控措施，包括认真履行用户身份验证、内容审核、添加标识等责任，避免新技术被恶意使用或滥用。

壹 ONE

人工智能治理的愿景和框架

1.1 人工智能治理的愿景：可持续发展

近年来党和国家高度重视人工智能研发利用，将其作为新一轮科技革命和产业变革的重要驱动力量。以人工智能为代表的算法应用有效地推动了数字化转型，带来了巨大的机遇，同时也存在着一定的风险和挑战。企业作为人工智能发展和应用的排头兵，具有技术和能力等优势，深入参与治理工作责无旁贷。**本白皮书尝试构建一套面向企业的体系化治理框架，探讨如何应对包括人工智能的各类算法风险，将治理原则落实到生产实践的各个环节。**

1. 治理应以发展生产力、保护正当主体权益为目标

人工智能深度参与重塑生产力与生产关系，治理过程中坚持发展生产力、保障各方主体权益，对于促进行业产业环境公平、稳定、可持续发展至关重要，更有利于实现更大的社会价值。

在确保安全的底线上坚持发展生产力。应首先认识到人工智能仍在快速的发展阶段，相关技术自身具有一定的脆弱性、以及管理体系落后于技术发展等诸多问题客观存在。其次，需要积极探索适配创新科技高速发展的合理模式，避免抑制人工智能的发展与应用潜力。

保护各主体合法权益，持续释放技术红利。人工智能从设计开发到部署应用，涉及到多类主体。治理过程中需要充分保护人工智能全生命周期各类主体的合法权益，兼顾相关方合理的利益诉求，平衡短期利益与长期目标，确保人工智能技术能够让整个社会受益。

2. 治理应以人人受益、责任担当、开放共享为价值导向

随着人工智能与人类社会融合的不断深入，赋予人工智能正确的价值观、道德感、判断力，对于促进人类未来发展至关重要。**遵循人人受益、责任担当、开放共享的价值导向，既是实现可持续发展治理的内在要求，也是打造可用、可靠、可信的人工智能技术的重要指引** 。

1）人人受益

党的十九大报告指出，我国社会的主要矛盾是人民日益增长的美好生活需要和不平衡不充分的发展之间的矛盾。**人工智能企业对于相关技术和产品的研发应当以人为本，让科技进步服务于对美好生活的要求，服务于可持续发展愿景。**

2）责任担当

习近平总书记指出，“行生于己，名生于人。只有富有爱心的财富才是真正有意义的财富，只有积极承担社会责任的企业才是最有竞争力和生命力的企业”。**企业的发展离不开国家、社会、用户的支持，离不开对于各类资源的消耗，因而对社会、经济、环境的可持续发展责无旁贷。**

3）开放共享

人工智能涉及学科复杂广泛、涉及主体众多，治理手段多样。因此解决产业或者技术存在的问题是个人、单一机构无法完成的，需**搭建开放的、科学的平台，集结社会各方力量共同开展治理工作。**一是基础底座的开放共享，如科技领域的科研成果；二是技术软件的开放，如技术开源；三是开放的数字生产力平台，简化个人和组织的创新程序。

1.2 构建可持续发展人工智能治理框架

人工智能治理需要重点关注并回应人工智能带来的三大问题：一是要避免人工智能大规模部署和应用带来的数据滥用；二是管控技术风险，优化完善人工智能技术，克服固有的缺陷；三是要建立有效的人工智能管理体系，识别人工智能生命周期中的问题并作出应对。

1. 保障数据安全是人工智能治理的前提

数据和人工智能是“水”和“鱼”的关系。一方面，人工智能（鱼）依靠数据（水）来生存，比如人工智能算法的训练、迭代离不开海量数据的支撑，数据作为除土地、劳动力、资本、技术外的第五种生产要素，呈现出乘数效应等特点，数据可和其他要素结合使用，催生“人工智能”等新技术，产生更大的经济价值；另一方面，人工智能（鱼）赋予数据（水）更大的价值，比如数据的开放利用离不开基于人工智能的数据挖掘、 数据分析， 人工智能对数据的获取和利用有根本性的影响，通过人工智能处理，数据价值得到大幅提高，能够帮助企业真正实现数据智能。

在缺乏相应治理规范的情况下，部分组织为争夺更多的数据资源，可能滥用人工智能进行过度、违规收集用户数据，恶意爬取其他企业数据等行为，最终导致用户权益及隐私受损、扰乱市场公平竞争秩序，甚至危害国家安全和社会公共利益。

可持续的人工智能治理需要回应人工智能应用带来的数据滥用问题， 通过平衡、包容的治理机制，一方面促进数据和人工智能同向进步，提高数据驱动能力和数据

智能水平，更好地发挥数据在数字经济建设和社会文化发展中的核心作用，另一方面要防范数据安全风险，有效保护各类主体的核心权益。一是遵循合法、正当、必要的原则，在获得用户授权的条件下合理收集用户数据，尊重用户对个人信息的管理要求，保护好用户隐私，二是遵守商业伦理和经济秩序，通过合作或交易机制获取其他企业的数据，维护数据所承载的商业价值，三是保障数据安全，强化数据治理，避免数据篡改、泄露、破坏等带来的安全风险。

可持续的人工智能治理应当构建全生命周期的数据安全整体能力。数据安全不同于传统网络安全，因数据的价值是流动和使用中产生的，这就需要数据安全覆盖到各种数据处理活动，从数据采集、数据传输、数据存储、数据处理、数据交换到数据销毁整个生命周期。如数据采集活动中要对数据源进行身份鉴别，防范数据伪冒和数据伪造的风险。数据质量管理中要对数据进行治理和标识，保障有质量、已标识的数据作为人工智能的训练数据集。

2. 技术体系应以可用、可靠、可信为原则

人工智能相较于其他技术领域更为复杂、更不可控、更难预测。为**应对人工智能自身缺陷，谨防人工智能滥用，按照可用、可靠、可信的治理原则不断完善人工智能的技术体系。**将愿景、目标、价值导向转化为技术实践，来保障人工智能可持续发展。

01 可用：面向规模化真实场景可落地

可用是人工智能技术发展的前提和基础。人工智能应用的真实场景中往往会面临样本缺少、数据分布不均、重知识、快变异以及多模态等诸多困难。在通用人工智能技术尚未成熟的当下，需要面向真实的垂直场景解决规模化应用的问题。需要企业在解决如上问题的实战中催生出更强大的技术，保障人工智能在迭代发展中，持续应对不断膨胀的业务规模及更加复杂的业务形态。

02 可靠：面向对抗和未知场景更加鲁棒

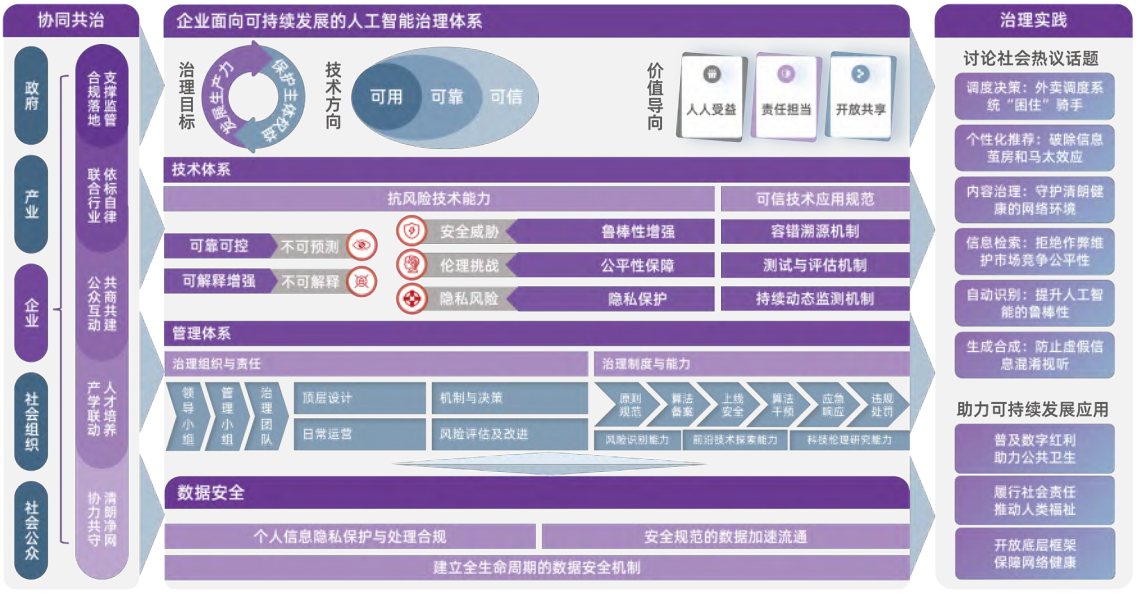
人工智能技术应当在可用的基础上逐渐向可靠的方向发展。由于互联网场景存在巨大商业利益，不可避免催生出一批专业从事信息攻防的黑灰产，针对人工智能的新型攻击手段不断涌现。企业需要在面对对抗和未知风险时，增强技术鲁棒性，保障技术可靠。

03 可信：以人为本，透明、公平和负责

人工智能技术应当以实现全面的技术可信为最终目标。**严格保护隐私数据不被滥用、透明可解释且能够让人类理解和参与决策、**建立在因果推理基础而非统计相关之上、并且是对大多数群体公平公正。人工智能系统需要其设计者赋予其正当的道德伦理观念，坚持将人工智能应用在更有社会价值的场景，才能确保人工智能可持续发展。

3. 面向可持续发展的人工智能治理框架

企业作为落实人工智能治理原则的重要主体，要在风险与发展之间寻求动态平衡，形成覆盖人工智能产品全生命周期的风险管理机制。既要主动调整人工智能的技术发展路线，通过科技手段帮助人工智能更好地服务社会，也需要通过组织机构变革，加强对人工智能的可靠性、可信度、伦理性的审核，确保人工智能匹配人类的伦理价值。**阿里巴巴结合业界“可信人工智能”、“负责任的人工智能”等理念与实践，提出了面向可持续发展的人工智能治理基本框架，积极努力探索从抽象治理框架到具体实践方案的可行路径，助力构建向上、向善的人工智能生态环境。**



贰 TWO

提升数据安全能力，保障人工智能健康发展

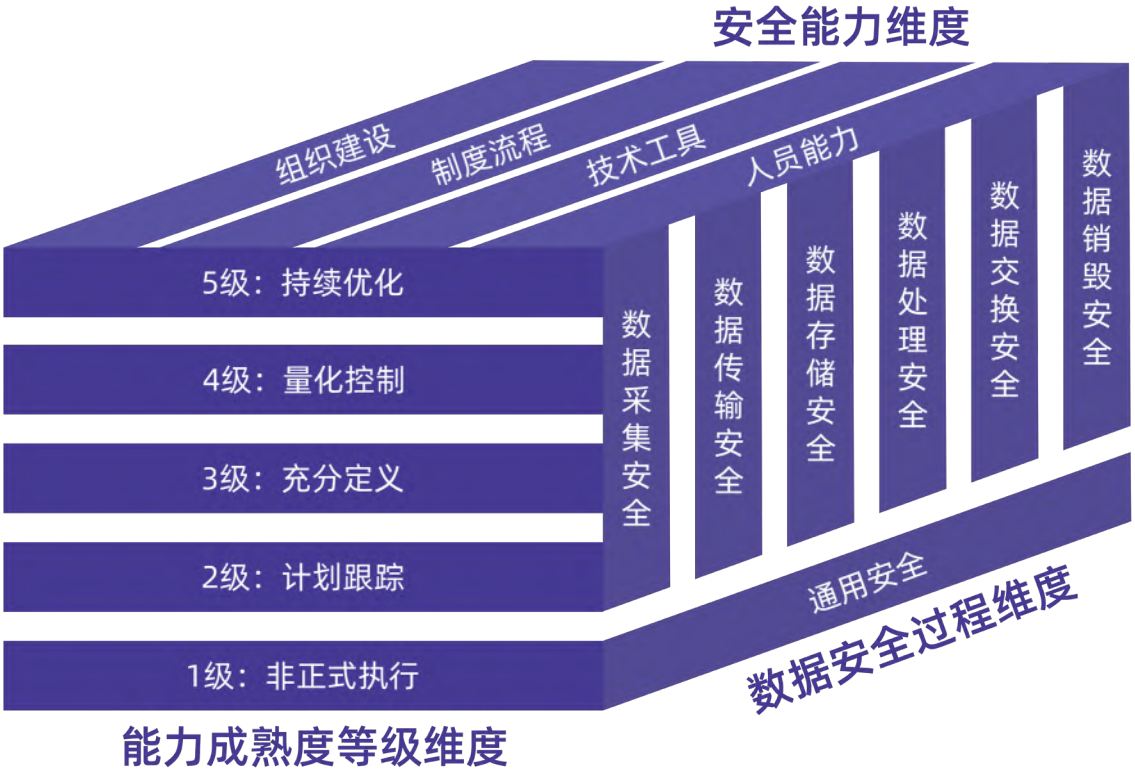
数据已成为国家重要的战略资源、新型的生产要素，数据的开发流通是人工智能高速发展的命脉，也是数字经济高质量发展的新动能。如何在数据隐私保护与开发流通之间保持平衡，最大发挥数据要素的潜在价值，是推动要素市场化配置及人工智能健康发展的重大挑战。一方面，人工智能是建立在数据之上的赋能技术，只有全面提升数据安全能力，才能为人工智能可持续发展保驾护航。另一方面，人工智能技术为发挥数据优势提供新动能，例如，隐私增强技术等新技术实现了“数据可用不可见”、“数据不动算法动”等数据不出域的效果，促进了数据安全和数据流通的协同发展。

2.1 构建全生命周期的数据安全能力

正如前文所述，数据和人工智能彼此如鱼水般紧密结合，在各个环节都相互影响，**企业在人工智能领域具备覆盖全生命周期的数据安全能力是应对数据风险的基础。**随着《中华人民共和国数据安全法》（以下简称“《数据安全法》”）正式颁布，以及数据安全标准体系的持续完善，对数据安全治理提出了更高的要求。企业应全面落实《数据安全法》等法律法规要求，借鉴国家、行业标准作为实践指引，建立以数据为核心、围绕数据生命周期进行的数据安全的完备的能力体系。

只有在各个数据环节中充分考虑安全风险，才能更好地应对由于数据流动性、多样性、可复制性等新要素特点带来的挑战。在数据采集时对数据源进行权属鉴别和记录，防止数据仿冒和恶意数据侵入；在数据分析中建立针对特定数据的脱敏机制，避免原始数据被技术手段复原；在数据处理过程要建立数据正当使用的责任及评估机制，保证收集、使用数据的合法、正当、必要性，避免数据被滥用、非授权使用；在数据资产管理中建立组织的数据质量管理体系，保障数据采集中数据的准确性、一致性和完整性，应对人工智能算法需要有质量数据的需求。

阿里巴巴制定包括《阿里巴巴集团数据安全规范（总纲）》和配套 23 个执行制度或规范；搭建专门的集团数据安全组织； 在技术工具方面，阿里巴巴自研了数据分类分级、敏感数据识别、数据脱敏、数据安全审计、数据流转监控、堡垒机、4A（统一安全管理平台解决方案）等数据安全工具。此外，借鉴多年的数据安全实践，阿里巴巴联合产学研各界力量，合力沉淀出 GB/T 37988-2019《信息安全技术 数据安全成熟度模型》国家标准，该标准从组织建设、制度流程、技术工具和人员能力等各维度，涵盖数据采集安全、数据传输安全、数据存储安全、数据处理安全、数据交换安全、数据销毁安全、通用安全七大环节共 30 个 PA（过程域）进行详细阐述，可作为组织开展数据安全能力建设的依据。目前，该国家标准已得到广泛采用，在全国 20 多个行业、数百家企事业单位获得了深入实践和应用，有效提升了全行业的数据安全水位。



2.2 安全规范的数据流通 加速释放数据价值

2022 年 4 月 10 日，《中共中央国务院关于加快建设全国统一大市场的意见》明确规定加快培养数据要素市场。**数据作为新兴生产要素，能够促进数字经济的持续发展，加速数据流通利于激发数据要素价值。构建数据要素市场需要统筹数据安全与发展，这需要在保护用户隐私、确保数据安全的前提下，促进数据流通，实现数据的高质量利用。**对此，可以从以下三个方面进行探索。

1. 实现数据安全与数据利用的平衡发展

数据是连接用户、企业、社会的纽带和桥梁，数据跨主体流通能够帮助企业了解用户需求，通过“数据从用户来、服务回用户去”的方式，让企业为用户提供更好的数字化服务。数据流通中的安全问题不是绝对的，企业对数据承担的责任也不是无限的。《数据安全法》设立了数据分级分类保护机制，域外国家的立法也确立了企业对数据保护尽职免责制度，完善相关规则可以帮助企业确立在不同场景和条件下对不同类型数据的保护范围和具体要求，在加强数据保护的同时也对自身承担的数据责任形成有效预期，让企业在做好数据安全保障的基础上更放心的将数据投入要素市场，发挥更大作用。

2. 确立个人信息在受信任环境下的合理使用机制

个人信息承载了用户的人格权益，获取和利用个人信息需要尊重用户的选择、获得用户的同意，并不得利用个人信息损害用户隐私和其他合法权益。企业为了更好的提供产品和服务，在生产过程中也会对用户个人信息进行加密处理和进一步利用，**在保证数据安全、保障个人信息权利、尊重用户隐私的前提下，应当允许企业在可信、可控条件下合理地使用个人信息，通过数据分析和数据智能更好地服务用户，提高用户的数字化生活体验。**

3. 认可数据安全技术应用

推动标准建设并推广产业最佳实践。经匿名化处理的个人信息不再具有个人属性，

可以充分流转、利用，因此匿名化标准及其技术实现手段对于促进数据流通利用、发挥数据价值尤其重要。为了发挥匿名化作为个人信息保护调节器的作用，企业需要根据目前隐私增强技术发展的实际情况，充分考虑使用匿名化技术的成本和收益，协同各界共同探索匿名化标准的具体内涵和实现手段。同时，在满足法律法规的要求下，企业可以推广通过技术保护实现匿名化的良好实践，为数据要素市场扩展更多的数据来源。

2.3 隐私增强计算促进数据安全和数据流通协同发展

“隐私增强计算”属于上述数据流通中的新兴安全技术，能够在保障个人信息权利及个人隐私的基础上，实现数据的流动及价值挖掘。

“隐私增强计算”是安全多方计算、同态加密、差分隐私、联邦学习、可信执行环境等一组安全技术的统称，共同点是希望实现不直接接触原始数据，同时完成对数据的计算处理。隐私增强计算用于人工智能训练，可以达到保护训练数据隐私的目的。以安全多方计算（MPC）方案为例，各个机构可以各自在本地部署一个 MPC 计算节点，节点之间以事先商定的 MPC 协议互联。在基于 MPC 的人工智能建模过程中，各方所见的始终只有对方训练数据的秘密分量，无法了解到对方训练的原始数据信息，从而避免了隐私泄露。

隐私增强计算的一个典型人工智能场景是医疗模型训练：科研机构需要从多个医疗机构处汇聚大量医疗信息进行人工智能建模，才能顺利开展致病性分析等工作，但是基因等医疗信息又与个人隐私密切相关，不适合直接跨机构传播。一种可能的解法是同态加密：机构对基因信息进行同态加密之后传递给科研机构，然后科研机构不需要解密即可在基因信息之上进行致病性建模等分析。在国际上，iDASH 竞赛是专门探索解决云环境下海量基因组分析期间的数据隐私和安全性问题的赛事，目前也是国际上在隐私计算方面最高规格的竞赛，足以体现隐私增强技术在基因组数据安全分析场景的重要性，2019 年阿里巴巴在该竞赛中取得了一等奖。

专题：落实法律法规 完善用户隐私保护

隐私保护是以人为中心、构建可持续发展人工智能治理体系的核心要求之一。我国新一代人工智能治理原则、经济合作与发展组织（OECD）和二十国集团（G20）的人工智能治理原则都将隐私保护作为核心原则，我国《个人信息保护法》、欧盟《通用数据保护条例》（GDPR）等法律均对隐私保护措施提出了具体要求。国际众多知名企业也将保护隐私数据作为对用户的责任和商业道德规范，在内部治理和商业决策中对隐私风险和业务发展进行平衡。

数据六大生命周期都涉及到用户的个人信息，数据采集活动中对个人信息的收集需要遵从合法、正当、必要原则，针对敏感个人信息需要采用数据加密方式的数据存储，涉及到个人信息的数据处理、交换等需要告知个人并取得个人的同意等。具体来看，**人工智能发展中需要兼顾个人信息保护，落实以人为本的治理目标，具体包括获得用户同意，允许用户对个人信息的授权使用场景进行控制，以及尊重用户限制和退出人工智能决策的意愿。**

数据采集征得用户知情同意：让用户知情，获得用户授权是实现隐私保护的前提。以人为中心的人工智能治理首先需要帮助用户知悉人工智能的风险、收益和替代性的解决方案，使得用户能够在知情的基础上决定是否将个人信息交由人工智能处理。让用户知悉人工智能系统的存在和运行机制、获得用户同意也是面向用户进行解释的重要环节。通过用户同意，可以保证对用户个人信息的采集和后续人工智能处理的目的和结果符合用户的预期，确保人工智能的决策服务于用户，满足用户的实际需要。

充分保障个人信息主体权利：用户对其个人信息的处理享有知情权、决定权，是隐私保护的重要原则之一。当个人信息存在错误、非必要收集使用或滥用的情况下，用户可以要求更正或删除自己的个人信息。实践中，人工智能系统的数据处理过程

较为复杂，为了帮助用户实现对个人信息的决定权，微软为用户提供了适当的控制选项，帮助用户选择个人信息的使用方式；IEEE 要求未成年人和能力受限人群的个人信息需要通过家长和其他监护人来控制，并建议为用户提供在线代理机器人，利用人工智能技术方式帮助用户进行控制决策。

设置自动化决策退出机制：人工智能广泛应用于信用分析、劳动决策等领域，所做出的自动化决策对个人生活带来诸多影响，如果在分析判断中较多使用个人敏感信息，可能影响针对用户个体决策的公平性。因此个人信息保护原则要求人工智能避免单一追求针对用户个体的精准性，对用户影响重大的决策，应当避免完全通过自动化的方式做出。该机制一直以来是业内热议的焦点话题之一，目前也已达成了一定共识，例如《人工智能北京共识》★要求在未预期情况发生时，建立合理的数据与服务撤销机制，以确保用户自身权益不受侵害。

阿里巴巴始终视用户隐私安全为自身成长的生命线，科技护航隐私安全，合理有度使用个人信息，服务用户，造福社会。一是通过隐私政策、隐私产品、增强告知等方式主动告知用户收集和使用个人信息的规则，在获得同意后，才会收集和使用用户的个人信息。二是用户可以随时查询、修改、删除产品与服务中安全设置、个人资料、个人成长信息、支付宝绑定设置、微博绑定设置、个人交易信息、收货地址、旺旺网页版设置、应用授权等个人信息。高德则开放了历史记录管理功能，用户以管理历史搜索记录，收藏或删除常用的出行路线等方式，实现对人工智能推荐的控制和选择。三是每一类应用 APP 都建立了便捷的个性化推荐退出机制，对于隐私敏感度高的位置信息，高德提供了“足迹设置”，用户足迹设置仅对自己可见，并且可设置开启或关闭足迹地图，也可随时清除出行数据，退出自动化的足迹跟踪，保护自己的隐私。

★《人工智能北京共识》：2019 年 5 月 25 日，北京智源人工智能研究院联合北京大学、清华大学、中国科学院自动化研究所、中国科学院计算技术研究所、新一代人工智能产业技术创新战略联盟等高校、科研院所和产业联盟，共同发布《人工智能北京共识》。

叁 THREE

构建面向可持续发展的人工智能技术体系

人工智能仍是一项新技术，回顾其发展历程，真正大范围地从实验室走向产业实践、广泛应用于我们的生产和生活之中，不过是最近十年的事情。**持续探寻更健壮的技术以及科学管控现有技术的“缺陷”，构建面向可持续发展的人工智能技术体系，致力于推动人工智能技术可用、可靠、可信，其内涵包括提升技术安全和构建技术管理机制两个层面工作。**

人工智能技术安全体系

3.1 提升人工智能抗风险的技术能力

落实人工智能治理的各项要求，需要从基础技术层面，不断提升人工智能鲁棒性、可解释性、公平性等方面的基础能力。

1. 鲁棒性增强技术及实践

人工智能鲁棒性通常用于描述，当输入信息因外部干扰或环境条件发生变化时，人工智能系统仍保持其性能水平的能力。鲁棒性一般可以分为对抗鲁棒性和分布外鲁棒性。其中，对抗鲁棒性指模型防御对抗样本的能力；分布外鲁棒性指模型的泛化能力，即当待识别数据的分布特性与训练数据不同时的识别能力。

增强人工智能系统鲁棒性，一般可以从数据、模型、系统等多个维度开展工作。提升数据质量，进行数据质检确保标注数据的基本质量；采用聚类、主动学习等策略选择更有价值的数据；使用数据增强技术模拟未知输入。提升模型能力，引入对抗样本训练，确保技术安全；尽量复用经过行业广泛验证的最佳模型结构；探索研究自监督学习、半监督学习等技术，充分挖掘无标签数据，提升模型泛化性。系统策略层面，采用模型解耦或者模型集成的方式，同时对模型实施在线更新，全面提升整个系统的鲁棒性。

针对鲁棒性问题，阿里构建了模型鲁棒性评测与防御系统。在鲁棒增强技术方面，阿里从对抗样本检测、弱对抗训练以及网络结构探索等多个方面开展积极的探索与实践，沉淀了丰富的鲁棒防御技术。除了内部实践外，针对鲁棒性评测，阿里联合清华大学、瑞莱智慧一起推出模型对抗鲁棒性和分布外鲁棒性基准平台 ARES（Adversarial Robustness Evaluation for Safety），评测了 ImageNet 上 49 个模型的对抗鲁棒性和分布外鲁棒性性能。此外，阿里开源了业界首个针对视觉模型的鲁棒学习框架 EasyRobust★。EasyRobust 可以便捷地实现数据增强、模型预训练、模型结构设计与选择等，帮助业界快速实施鲁棒性技术研发工作。基于此框架，阿

★EasyRobust：https://github.com/alibaba/easyrobust

里重新思考 Vision Transformer 的鲁棒性设计原理，对网络组成单元进行了鲁棒性分析（包括对抗鲁棒、通用噪声、分布漂移等），通过组合各个鲁棒单元，创新性地构造了鲁棒视觉网络结构 RVT，相关研究工作已被计算机视觉顶会 CVPR2022 录用。

2. 公平性保障技术及实践

人工智能在自动化决策中存在不公平决策行为，呈现出有意识或无意识偏见。产生这些偏见的原因可能存在于数据采集、算法构建、模型应用的各个环节，因此需要构建公平性评估指标加以约束。根据现有研究，公平性评估指标总体可以分为个体公平性（individual fairness）和群体公平性（group fairness）两大类，前者强调对于任何相似的个体，都能给出一致的决策结果；后者侧重于衡量人工智能系统对基于敏感属性（如性别、种族、宗教、信仰等）划分的不同群体之间的偏见程度。

算法去偏（bias mitigation）技术是预防和消除人工智能算法偏见，实现决策公平的核心手段，可以按照介入阶段分为以下三类。其中，预处理公平性技术通过优化算法训练数据来解决训练数据中的偏见问题，例如将训练数据进行映射转换以生成公平的训练数据集。过程中公平性技术主要通过在训练阶段增加公平性指标约束，从而改进和优化智能算法，以在模型学习阶段消除偏见。后处理公平性技术通过直接修改算法的决策结果以使其满足公平性要求。

值得注意的是，公平的概念是抽象的、发展的，公平性指标也随场景的不同而拥有不同的含义和表现形式。阿里针对公平性问题开展了积极的探索，在电商场景下不断实践算法去偏技术。例如，在用户侧，通过过程中约束相关性和发现性等指标并对推荐结果进行重排、打散以减少推荐匹配错误和信息茧房现象的产生，保障各个消费群体使用推荐产品的体验；在商家侧，通过对有潜力的长尾商家和高品质商品进行孵化，缓解平台上头部聚集的马太效应，同时依托各种反作弊技术打击“蹭热

点”式的软性流量劫持，保障流量资源分配的公平。

3. 可解释增强技术及实践

研究可解释性更强的人工智能算法，对于预防和减少用户担忧，提升人工智能可信度具有十分重要的意义。近年来，可解释人工智能的研究备受关注，相关方法可以分为模型自解释、模型事后解释和因果机制解释等三类。

模型自解释指从模型自身在理论层面上具有解释性，决策逻辑能被直接理解，例如在深度学习领域，Joel Vaughan 等通过对岭函数和投影系数的展示，解释输入特征与复杂神经网络输出之间的关系，相关研究还在不断探索中。模型事后解释是针对已训练好的人工智能模型进行解释，包括全局解释和局部解释方法。其中，局部解释方法通过分析样本特征对模型决策结果的影响程度，来帮助人们理解和学习模型针对特定输入样本的决策过程和决策依据。全局解释方法以人们可理解的方式从整体上解释模型背后的决策逻辑和内部机制，将模型的结果进行可视化。因果机制解释是针对由不可观测变量导致的因果倒置等问题，帮助发现和理解背后的因果机制，从而提升决策合理性和可预测性。可简单分为结构因果模型和潜在结果框架两大体系，其中结构因果模型通过建立因果图和一系列结构方程，直观地展示变量之间的因果关系；潜在结果框架旨在估计不同干预（即变量取值）下的潜在结果，以评估变量对目标的因果效应。

总体上来看，人工智能可解释性相关的技术和理论研究工作还处于探索阶段。阿里十分重视相关研究工作，围绕可解释的深度学习模型、因果推理等前沿方向发表多篇学术论文。同时在风险策略智能化、推荐算法自评估等场景中，积极探索模型解释技术的多样化应用，如规则可解释、风险异动归因、关键信息可视化等，为相关工作的优化提供了有力支撑。

3.2 构建全生命周期的可信技术应用规范

应对人工智能技术在实际应用中引发的风险，除了积极推动人工智能技术可信能力的提升，不断减少技术本身的脆弱性，还应该构建更为积极的技术应用规范，规避现阶段人工智能技术“缺陷”带来的问题。

1. 人工智能产品不同阶段的风险挑战

人工智能产品研发中规划设计、研发部署、运营使用环节的风险挑战识别十分重要，对于构建可信研发的实践范式具有重要意义。

在规划设计阶段，难以在初始阶段形成完备的风险分析，与此同时，确保相关理念贯彻执行存在挑战，在设计理念、规范传达给个层级实施人员过程中，存在非正确传达、误解等风险，尤其机器学习场景中固有的不可预测性，传达实施偏差会进一步加剧。

在研发部署阶段，一方面，数据层面可能会遇到数据缺失、重复、不一致、来源不明等问题；另一方面，模型技术层面存在着技术选型不恰当，模型尚未完备训练即开始上线服务，以及模型运行之后的动态更新缺乏足够验证等挑战。

在运营使用阶段，一方面，在人类和人工智能交互时可能出现误用、过度依赖等问题；另一方面，人工智能相关技术存在着被恶意使用的风险。

2. 构建可信人工智能研发管理机制

正是由于人工智能技术在一定程度上的不可预测性，以及其通常是持续更新、持续服务的特性，使得人工智能技术规范与传统技术存在差异。结合人工智能在不同阶段的风险，总体上看需要从三个方面构建可信的技术规范。

1. 构建充分的容错机制和溯源机制。对人工智能软件和硬件等进行冗余设计，同时增加不同层级的人工控制点，确保使用者在最终决策过程中具有决定权和自主权。此外，应留存人工智能系统各项相关记录，确保全过程的可溯源。

2. 强化全生命周期的评估测试工作。在人工智能技术系统研发的全生命周期，设计严谨、准确、全面的人工智能测试与评估流程。围绕人工智能系统的可信要求，设计具体的测试指标与测试方法，以量化的方式评估相应技术系统的性能优劣，确保设计的理念得到明确的贯彻执行。

3. 开展持续的动态监测工作。对人工智能技术和系统在实际应用过程中，进行全方位的监测，确保及时发现、并处置相应的风险问题。

专题：筑牢深度合成全链条治理基石

1. 深度合成基本概念及治理要求

深度合成是指以深度学习、虚拟现实为代表的生成合成类技术，已经广泛用于智能问答机器人、创建软件代码、促进药物研发等。Gartner 将生成式人工智能（Generative Artificial Intelligence）列为 2022 年重要战略技术趋势，预计到 2025 年，由机器生成的数据将从不到 1% 提升到 10%。

深度合成技术可以实现图像、声音、视频的篡改、伪造和自动生成。近年来随着技术不断成熟，生成内容愈发逼真，已经达到人眼难以辨别真假的程度，存在包括攻击个人声誉、干扰司法鉴定、冲击社会信任，甚至影响国家安全的风险隐患。2020 年 1 月 1 日起施行的《网络音视频信息服务管理规定》中，对深度合成提出了相关要求，包括以显著方式标识非真实音视频、部署非真实音视频鉴别技术等。在 2022 年 1 月国家互联网信息办公室发布的《互联网信息服务深度合成管理规定（征求意见稿）》中，从信息标识、备案、安全评估、投诉等各方面提出了具体的要求。

2. 深度合成需多方共同参与治理

深度合成技术治理中涉及到四类主体，需要各方提升相关能力，如图所示。其中算法提供方主要涉及到提供深度合成算法和模型的科研人员、科研单位或企业单位等。服务提供方包括提供深度合成类算法服务的主体单位。服务使用方包括使用深度合成服务的个人开发者或者企业单位。内容平台方包括具备内容发布功能的社交媒体或者自媒体等平台。

深度合成技术当前的效果已经非常逼真，在没有先验信息的情况下，依靠肉眼以及技术手段做出有效辨识的难度在持续加大。对深度合成内容在成本可控的前提下实现有效管控的可行方式，是企业作为内容生产的源头、内容传播的渠道，严格履行主体责任，包括建立相关技术能力、落实管理措施等。

3. 阿里针对深度合成的治理实践

阿里认真落实深度合成治理相关要求，作为算法提供方、服务提供方、内容平台方等，积极研发和部署相关技术能力。

作为算法提供方，联合高校等科研单位等，进行可逆合成、模型水印等相关技术研究，实现深度合成的可溯源和可解释能力。

作为服务提供方，加强深度合成服务管控。一是通过人脸识别、活体检测等技术手段对相关服务使用者进行身份认证。二是主动识别合成数据及其所用素材内容的安

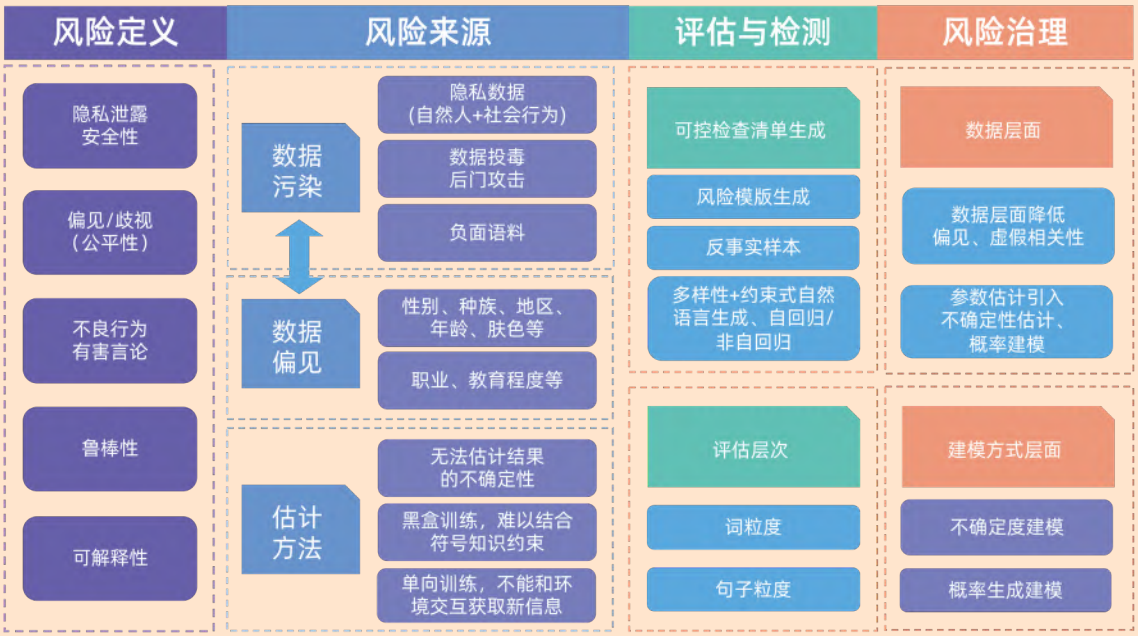
全风险，确保符合个人信息保护有关规定以及内容风险控制要求。三是采用数字水印技术对合成数据嵌入鲁棒暗水印信息，实现对合成数据的有效溯源。四是嵌入明水印标识进行显著标识，明确告知用户该内容属于合成数据范畴。

作为内容平台方，在强化信息内容审核、合成数据溯源的基础上，提升合成信息鉴别能力。阿里与中科大合作的鉴伪技术被斯坦福大学发布的《2022 年人工智能报告》专门提及。此外，积极参与由国家互联网信息办公室、工信部、公安部、国家广电总局和厦门市政府联合主办的第三届中国人工智能大赛——深度伪造视频检测赛题，获得了 A 级证书。

专题：打造安全可控的基础模型

1. 开源预训练大模型存在多方面安全风险

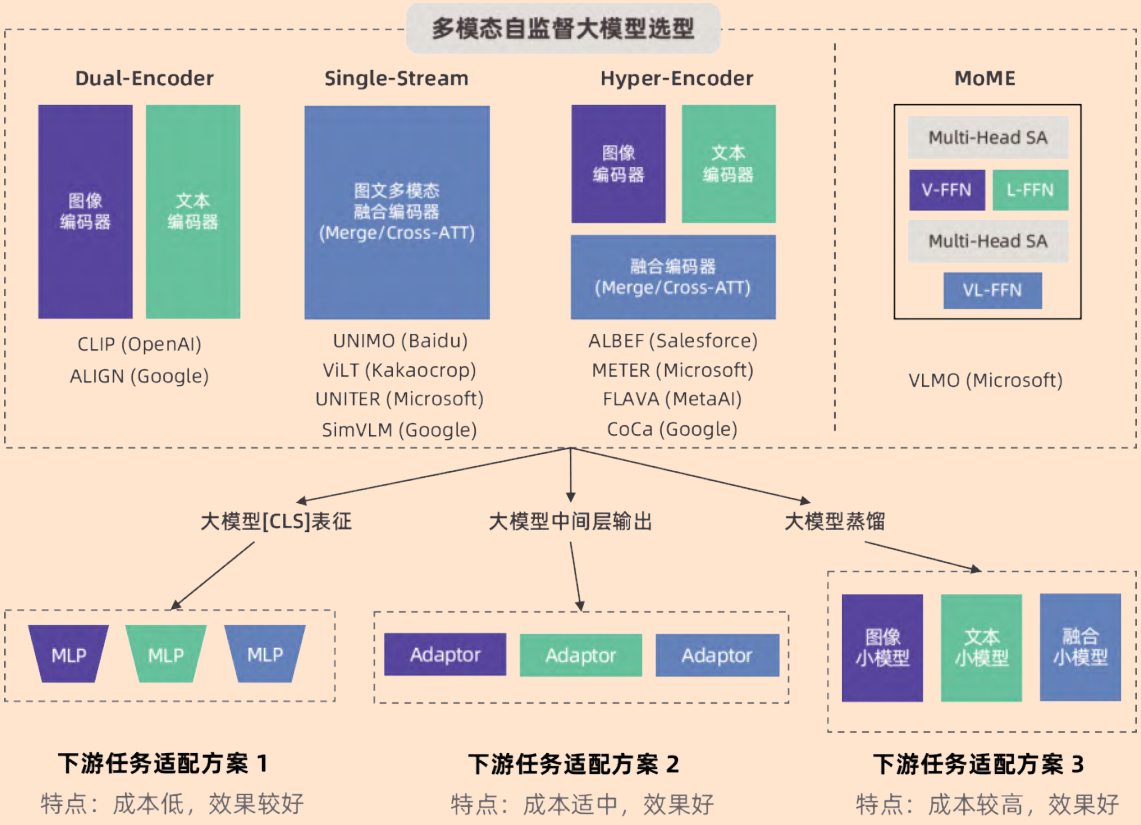
深度学习最开始发展的时候依赖大量的人工标注数据，通过端到端监督学习完成特定的识别任务。由于预训练模型呈现的各种优势，越来越多的企业和研究人员使用开源的预训练模型在各自特定场景下进行微调或者是下游任务适配。随着预训练在越来越多的场景中应用，针对预训练模型的风险越来越受到研究人员的关注。阿里巴巴从隐私泄露与安全性、偏见与公平性、不良行为和有害言论、鲁棒性风险、可解释性这五个方面总结了开源预训练模型的风险，具体如图所示。



2. 阿里针对基础模型研究的安全实践

为了应对预训练模型的各种威胁，构建安全可控的“基础模型”迫在眉睫。目前，基础模型普遍接受的定义是：在超大规模数据上训练（通常指自监督训练）且可以复用到广泛的下游任务中的大模型。随着越来越多的公司和机构加入到基础模型的研究工作中，思考基础模型带来的应用算法模型生产开发方式的变革也变得越来越重要。斯坦福大学的 Percy Liang 教授指出了这场变革的本质：传统的模型开发遵循“Top-Down”思路，先发现问题，然后做数据收集和标注，最后选择合适的算法框架做模型训练；相比之下，基于基础模型的模型开发则是遵循“Bottom-Up”的思路，先选择基础模型的架构，让它尽可能地通用于广泛的下游任务。

阿里安全在 2019 年就开始投入大量精力在自监督预训练技术的研究，在单模态基础模型和多模态基础模型两个方面均取得了不错的成果。不仅沉淀了针对自监督训练优化的代码框架 OODN（Object Oriented Deep Network），而且积极探索在风控业务场景落地的最佳实践，取得了显著的业务效果提升。下面从数据建设、基础模型选型、基础模型训练和下游任务适配四个方面介绍：



1. 数据建设

自监督学习是基础模型的默认训练范式，主要依赖超大规模的无标注或弱标注数据。阿里安全在实践过程中发现，成熟的数据收集和数据清洗链路对基础模型起到非常重要的作用。通用基础模型收集通用领域数据且需要过滤安全性敏感数据，应对预训练模型的不良行为和有害言论威胁；相比之下，安全基础模型同样需要采集通用数据，但是同时会针对性地采集大量安全领域数据（例如色情、辱骂、违禁品等），因此阿里安全算法团队借助安全知识图谱的逻辑梳理和信息扩源，建立针对性的数据收集方案，进行全面且可持续性的数据采集，从而提升模型的鲁棒性。

2. 基础模型选型

网络架构选择也是核心的问题，数据拟合能力和下游任务适配能力是主要考虑因素。当前学界和业界均倾向于选择 Transformer 构建基础模型。首先，Transformer 网络结构易于做模型规模的扩展，且对并行训练的支持友好；同时，Transformer 结构在图像、文本和语音编码均取得了出色的效果。此外，最近研究也表明，Transformer 网络结构相比传统的卷积结构具有更强的鲁棒性。

阿里安全对基础模型选型的标准是架构收敛，希望通过统一的多模态自监督预训练，同时产出图像、文本和图文融合基础模型。上图展示的是当前流行的多模态自监督预训练架构，其中的 Dual-Encoder、Single-Stream 和 Hybird-Encoder 架构的基础模型在阿里安全业务中均有落地应用。

3. 基础模型训练

阿里安全算法团队在自监督训练领域积累了大量的技术沉淀。在自监督损失函数方面，图像领域主要包括对比学习（SimCLR 为代表）、非对比式学习（BYOL 为代表）和图像模块建模（MAE 为代表）；文本领域主要包括掩码语言建模（BERT 为代表）和文本对比学习（SimCSE 为代表）；多模态领域主要包括图文匹配监督（UNITER 为代表）、图文对比学习（CLIP 为代表）、掩码图文建模（FLAVA 为代表）。

4. 下游任务适配

首先，基础模型产出的表征可以直接用于提升风险排查效果。同时，基于基础模型表征建立表征级别的业务模型，在保证防控效果的同时，显著降低了防控成本。对于表征级别模型无法满足的风险场景，实施了基于“大模型 +Adaptor”的业务模型方案，效果提升的同时收敛了大量原本相互独立的 End2End 防控模型，给模型管理和迭代升级的便利性带来了增益。与此同时，阿里安全算法团队在对抗鲁棒性研究方面一直保持学界和业界的领先地位。成熟的鲁棒性评估体系也确保了基础模型在下游应用过程中的安全可控。

专题：构造多方受益的信息流推荐系统

1. 信息流推荐存在“信息茧房”和“马太效应”两大问题

信息流推荐算法通常在用户授权的前提下，利用用户在信息内容上的各类行为表达，理解用户的兴趣需求，为用户尽快找到心仪的内容集合、商品集合，同时也可以帮助生产侧快速获取用户。在电商场景，信息流推荐算法可以大幅提升用户和商品、内容的匹配效率，让用户、商家和平台共同受益。

在广泛应用和快速发展中，信息流推荐也逐渐出现了一些问题。一方面，对于用户可能造成信息茧房问题，具体表现为推荐结果越来越单一和同质化。另一方面，对于生产者可能造成马太效应，即流量在头部生产者的聚集现象越来越明显。

上述问题的根源，在于推荐算法设计阶段，过于关注流量的分发效率、点击率的优化、成交转化率的提升，相对忽视了用户实际体验以及生产侧供给生态的持续优化。此外，深度学习自身的不可解释等固有缺陷也给系统的优化和干预增加了挑战。

2. 淘宝针对电商场景下信息流推荐算法的治理实践

淘宝积极开展技术和机制上的创新，致力于解决电商场景推荐算法的缺陷。应对信息茧房问题，不断提升推荐结果的多样性和新颖性，同时严格遵守《个人信息保护法》等法律法规要求，为用户提供推荐系统个性化退出开关。应对马太效应问题，持续孵化有潜力的中小长尾商家和高品质商品。

1）建模用户负反馈数据，减少推荐用户不喜欢的内容

给用户提供了便利的反馈推荐问题的入口，包含用户长按不喜欢的商品、内容等，在随后出现的浮窗中选择负反馈的细分原因。推荐系统可以使用用户的负反馈信息来建模用户的负向兴趣，减少给用户推荐不喜欢的内容。

负反馈数据通常都比较稀疏，在实践中淘宝提出使用多任务学习的方法，通过其他辅助任务来帮助负向兴趣的学习。在负向兴趣建模中，分别引入用户的近期点击行为、长期点击行为来刻画用户的正向兴趣，引入用户的负反馈行为、近期曝光未点击行为来刻画用户的负向兴趣。长期、大量的线上数据表明，使用该方法能够促使针对整体商品的负反馈明显下降。

2）搭建发现性推荐链路，提升推荐系统的多样性

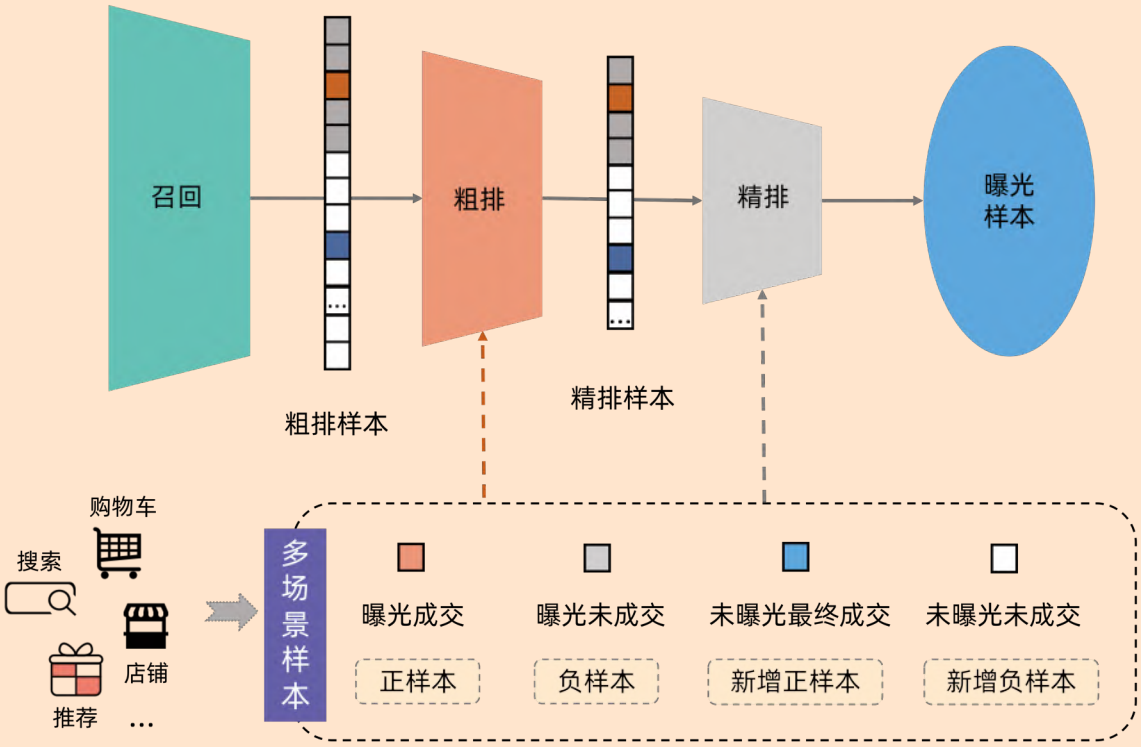
淘宝首页推荐搭建了独立的发现性推荐链路，专门推荐用户近期未点击过的类目商品。用户访问淘宝时，常规链路和发现性链路分别产出推荐结果，并由后续的混排算法综合这两部结果完成最终的排序。

发现性召回的逻辑主要包含四个部分。一是发现性向量化召回，即基于用户短期行为，学习用户的跨品类点击行为，为用户推荐与近期行为相关的品类。二是发现性检索召回，结合用户短期和长期行为与跨类目相似商品索引，构建发现性推荐能力。三是时令召回，在节假日、果蔬上市时间、季节更替等时令节点到来之前，将对应的商品召回。四是基于认知推理的标签召回，基于标签 / 类目构建知识图谱和推理链路，实现拓展用户兴趣的目的。

3）开展全链路无偏学习，刻画用户多样的兴趣分布

相比于传统针对单场景或单任务建模的方式，利用用户在淘宝多场景的数据进行学习，可以更好地刻画用户兴趣分布。但由于不同场景的数据分布差异性较大，导致直接使用多场景数据进行训练的效果往往不理想。针对此问题，提出了信息流推荐全链路无偏学习解决方案，充分利用推荐系统的漏斗型结构以及淘宝多场景的数据，解决了单场景单任务建模遇到数据选择偏差问题和数据稀疏问题，如下图所示。模

型在首页信息流推荐落地后，对打破推荐中越买越推的循环起到了明显的改善效果。



4）开辟新品赛道，助力中小商家快速成长

为了缓解马太效应，淘宝开辟了新品赛道，帮助中小商家解决新品上架后启动流量低的问题，同时建立了分层机制帮助优质新品快速成长。具体实现上，设计了新品冷启动机制，在新品发布后给予一定的初始流量保障。在初始流量保障下，新品就能够顺利度过冷启动阶段，进入正常的优胜劣汰的排序机制。为进一步提升流量的利用效率，淘宝提出了基于多模态迁移网络的算法选品模型，更准确高效地预测新品的成长潜力，挖掘出更多的潜力新品进行孵化。项目上线后在各类目都获得了广泛的应用，并取得了良好的业务效果。

5）通过精准商品潜力挖掘，助力乡村振兴

相对于头部的大商家，后发展乡村的中小商户和企业的商品由于自身的劣势在海量商品中难以被发现，优质商品容易在竞争激烈的市场中被埋没。针对这一问题，首先通过层次图神经网络算法分析市场中商品与商家的经营表现构建商品的基础画

像，与市场中消费者的潜在需求进行匹配，从而发现中小商户发布商品的潜在用户机会。其次针对缺少用户行为的商品通过元学习构建了能够快速适应数据的商品表达，并提出深度兴趣迁移网络从而通过知识迁移实现对于商品成长的精准预测。

基于该技术打造的商家赋能系统帮助湖北恩施、陕西富平、甘肃定西等后发展地区商家快速成长，为淘宝平台上的中小商家日均带来近千万的成交增量，为实现共同富裕添砖加瓦。相关成果形成论文《Hierarchical Bipartite Graph Neural Networks: Towards Large-Scale E-commerce Applications》发表在 ICDE-2020。

专题：维护电商平台信息真实和竞争公平

1. 电商场景下反作弊的核心问题和挑战

伴随电商平台的快速成长，包含虚假交易、刷单炒信、虚假评价、风险流量在内的一系列作弊手段和风险行为成为电商生态所面临的重要威胁，成为影响电商环境真实公平的一大问题。为了维护商家的正当利益、保障广大消费者的合法权益，电商平台需要对作弊行为进行有力的感知识别，在博弈的过程中也需要不断克服各种挑战：

1. 黑产恶意对抗

随着平台反作弊算法能力的升级和管控的深入，黑产分子的行为模式也在循序演进，并且不再机械单一。不法分子通过对动作添加一些噪声扰动，从而规避风险识别模型，例如：刷单过程也往往进行货比三家、嫌疑用户也在刻意增加页面停留时长等。这些黑产对抗行为对反作弊识别模型的鲁棒性提出了更加严苛的要求。

2. 作弊行为越发隐蔽

为了规避线上反作弊系统的识别和人工巡场稽查，部分黑产已摆脱了传统单打独斗、链路单一的行为作弊方式。 开始向手法更为隐蔽、账号风险更低、导购链路更仿真的“地推刷单”模式演进。作弊商家精心谋划的作弊炒信行为的痕迹极易淹没在超大规模的用户浏览数据之中。

总体而言，反作弊动态攻防的特点日益突出，电商黑产也朝向高度规模化、组织化、技术化的方向发展。

2. 淘宝针对电商场景下作弊行为的治理实践

针对当前黑产作弊手段和发展趋势，亟需构建大规模、专业化、强鲁棒、可解释的电商反作弊风控平台，助力维护公平的电商生态。

1）AutoRisk 行为风控引擎

淘宝基于电商场景行为风险重点建设了 AutoRisk 行为风控引擎，主要包括主动发现、风险提纯、风险认知和解释三个环节。在识别能力鲁棒性与隐蔽性团伙发现能力上得到显著提升，并引入自监督异常发现算法大规模应用于未知模式攻击检测领域。同时根据业务特性，对于交易前 - 中 - 后全链路风险进行提前研判和实时化风险预测，全面应用于淘宝电商场景反作弊风控，实现了在异常主动发现和风险提前防控方面的突破。

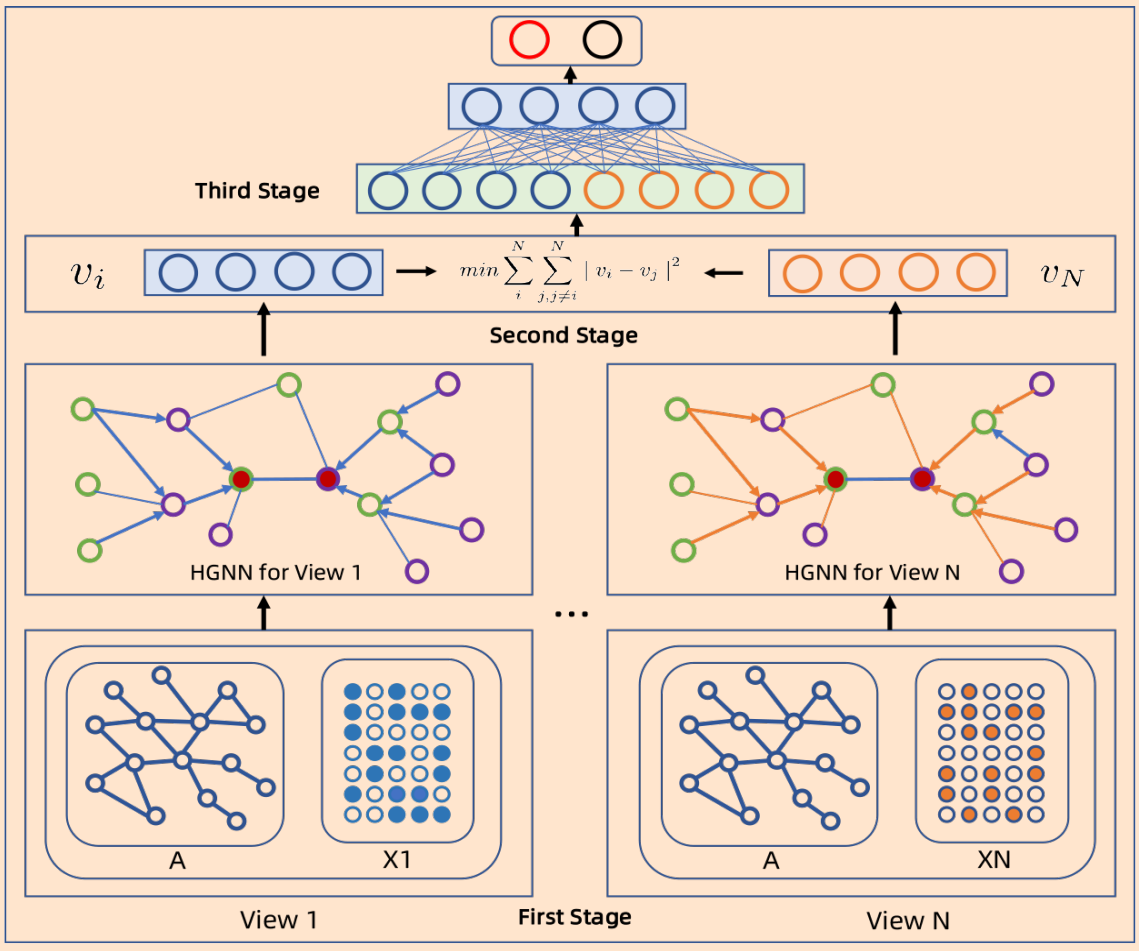
2）对抗训练提升模型鲁棒性

通过对抗训练的方式提升在噪声干扰下的模型鲁棒性，从而削弱或者消除这些扰动对模型准度的影响。在添加对抗样本时需要着重考虑以下几个因素：i）有效性：

添加的对抗样本能够有效地对样本进行扰动；ii）高效性：生成对抗样本的过程应该是高效的，能够在线性时间内完成，并在对抗特征生成上快速实现；iii）可行性：添加对抗样本的方式在现实中必须是黑产可操作的，否则对抗训练样本将与实际不符，不但无法加强模型的鲁棒性，反而会损害模型性能。经过对抗训练的模型在初始样本集和对抗样本集中都展现出了较好的效能。

3）风险团伙挖掘

尽管刷单形式和作弊手段不断升级，隐蔽性越来越强，但在有限的资源约束下，黑产团伙在执行任务时往往会呈现出团伙聚集等特征。因此，淘宝提出了风险团伙挖掘的解决方案。一是基于大规模图神经网络的 FRODO 反作弊系统，其中针对高风险欺诈用户，我们提出了多视图异构图神经网络 Multi-view HGNN 算法，如下图所示。线上实时识别虚假交易，线下主动聚类挖掘风险社群，并形成联动外部系统



的自动化情报网络，实时高效地传递风险信号。最终构建了域内＋域外、线上＋线下、离散行为＋聚集团伙、提前预警＋主动防控的全方位打击网络虚假交易的能力。

与此同时，该创新工作形成学术论文《What Happens Behind the Scene? Towards Fraud Community Detection in E-Commerce from Online to Offline》，获得 WWW2021 物联网研讨会最佳论文奖。二是基于无监督的风险团伙挖掘方案 Phalanx，解决在少样本甚至无样本条件下风险团伙挖掘问题。Phalanx 利用用户在商品上的行为数据构图，基于聚集性和相似性主动发现作弊团伙，包括“自动构图 - 团伙发现 - 团伙解释 - 风险团伙输出”的一站式输出。FRODO 和 Phalanx 等方案的的大规模应用，显著提升了交易公平性，有力保障了商家的正当利益和广大消费者的合法权益。

淘宝也借助此能力协助司法机关对刷单炒信行为进行有力打击，实现全社会范围内的法治提升。例如刷单治理的相关工作推动“全国首例电商平台打假案判决”、“全国首例组织刷单入刑案判决”，被《法制日报》和法制网评选入“2017 年推动互联网法治进程十大事件”。

专题：加强儿童类商品内容治理，守护未成年人健康成长

1. 儿童类商品内容治理日益重要

在电商平台上，部分儿童类商品的商家会发布一些有害未成年人身心健康的商品内容，例如儿童服饰中的贴身衣物图片聚焦敏感部位、让不谙世事的孩子模仿成人做出各种诱惑动作拍照，商品描述中包含低俗暗示性语言等。这些行为破坏了健康文明有序的网络环境，引发了社会的广泛关注。清理信息污垢，守护清朗健康的网络环境，保护未成年人的身心健康和合法权益、促进未成年人健康成长，电商平台责无旁贷。

2. 淘宝针对儿童类商品内容治理的制度规范实践

面向用户，在用户隐私权政策方面，对于 14 周岁以下的儿童，淘宝专门制定了《儿童个人信息保护规则》，在监护人仔细阅读并同意后，儿童才可在监护人的指导下使用淘宝的服务，确保在使用淘宝的服务和进行交易时的安全。同时，在商品展示上也做了约束保障，例如用户搜索“儿童”时，则返回的结果中不包含成人用具等类目的商品；未成年人搜索“白酒”等法律不允许未成年人购买的商品，则返回空结果。

面向商家，淘宝颁布了《淘宝儿童类商品行业规范》，明确严厉禁止涉及未成年人的违法不良信息，包括但不仅限于儿童服饰、玩具和儿童影像、照片等，范围覆盖商品标题、商品文字描述、商品图片和商品评价等内容。同时，淘宝上线儿童类商品发布合规提醒功能，后台会明显提示“请您发布儿童商品时，参照《淘宝儿童类

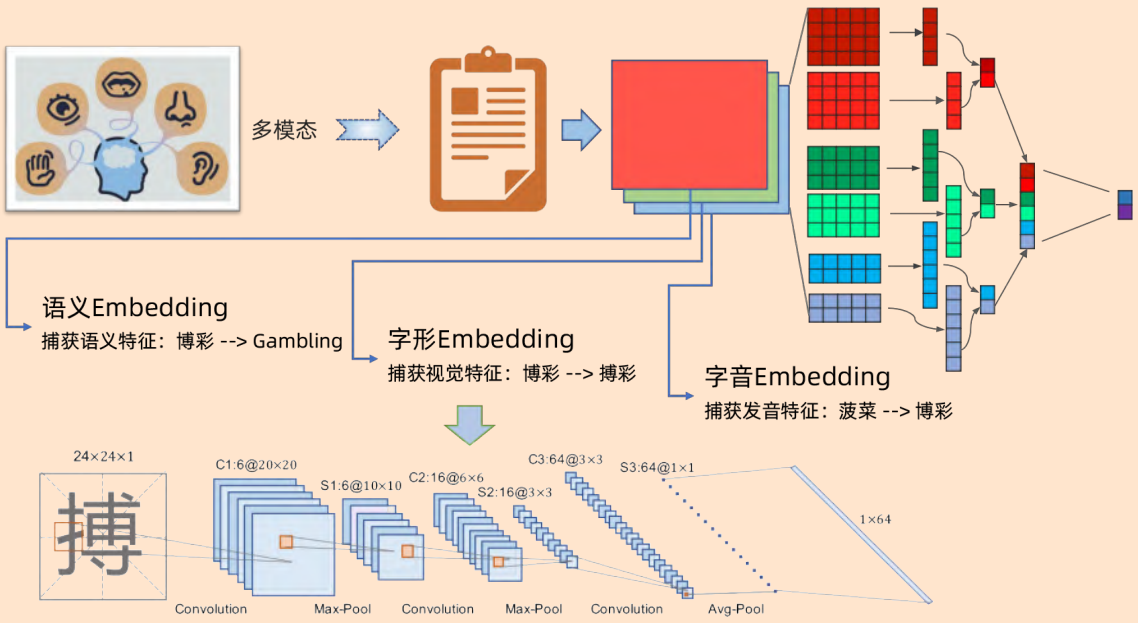
商品行业规范》进行发布，否则可能有下架 / 删除 / 扣分等处罚风险”，并建议儿童服饰类商家尽可能减少使用真人模特，避免贴身衣物聚焦敏感部位等问题。

3. 淘宝针对儿童类商品内容治理的技术能力实践

在完善制度建设和提升商家合规意识的基础上，淘宝进一步发挥在技术能力上的资源和优势，针对儿童类商品内容风险的特点，在文本和图片维度开展了一系列技术能力建设。

1）文本对抗技术

不良商家往往会对发布的文本内容添加精心构造的对抗扰动，例如 “学 生 专 属 -> 學銑專屬”， 试图绕过线上识别系统的拦截。 深度神经网络的固有脆弱性导致其在对抗攻击下给出错误的识别结果；人脑做信息理解时会综合语义、视觉和声音



等多个模态的特征，因此这种只在文本维度构造的对抗攻击往往不会影响人的阅读理解。受此启发，淘宝提出了基于多模态词嵌入的文本对抗防御技术，通过语义嵌入、字形嵌入和字音嵌入分别提取文本的语义、视觉、读音三个模态的特征，然后通过多模态融合技术对齐和融合三个模态的特征以形成语义丰富的鲁棒表征，可以有效增强模型针对文本对抗风险的鲁棒性。

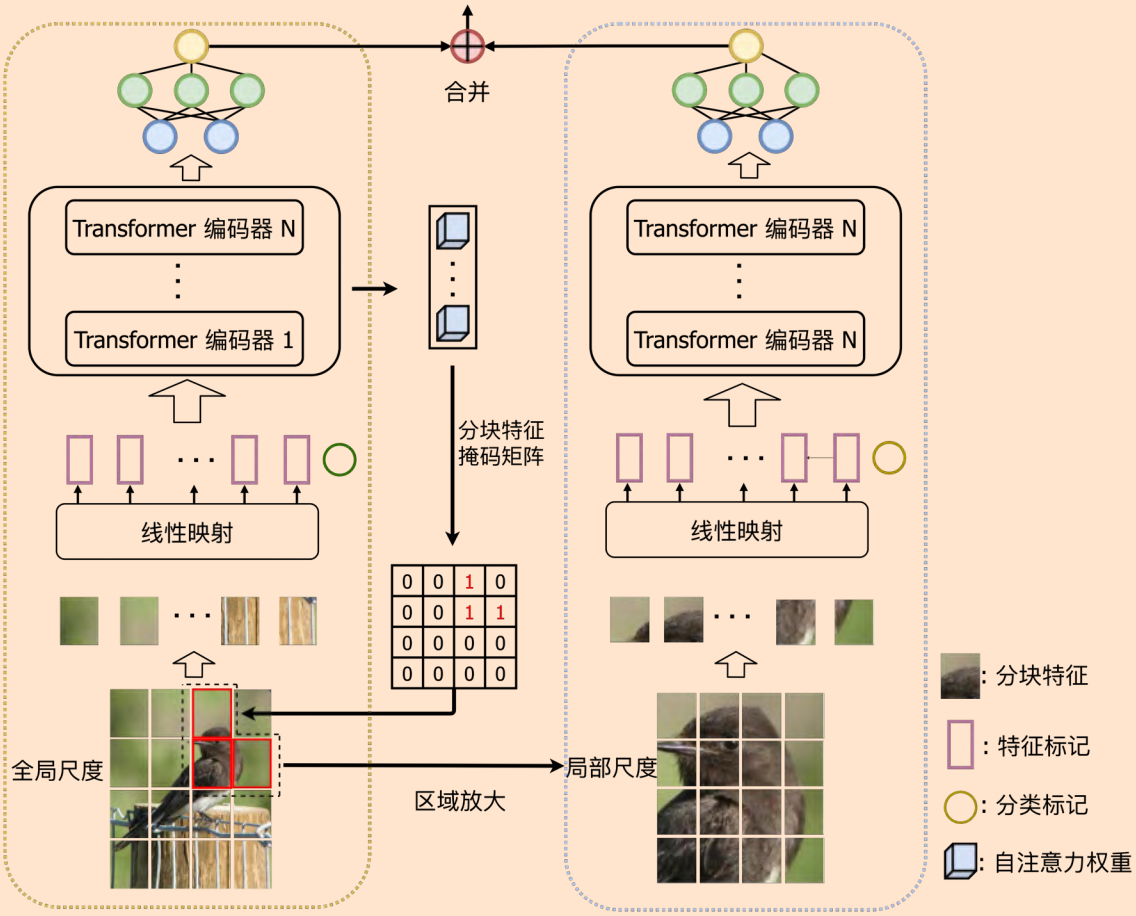
2）领域知识图谱

青少年经常创造一些新的语义符号，成为自身群体的专属语言，例如“cpdd -> 找人组 CP 有意请联系我”。如果不具备这些背景知识，面对青少年群体中的特定文本，成年人很难理解其中蕴含的真实意义，无法识别风险；模型也是如此。因此，我们首先需要搜集青少年群体常用的符号、代称、特定语义等，建立相互之间的联系，并与正常语言做映射，构造青少年语义知识图谱；然后在识别模型的设计上，把输入分为上下文和知识点两部分，模型分别提取文本和知识的表征，实现结构化知识和无结构文本之间的信息共享，借助先验知识加强模型对输入的语义理解，提升识别能力。

3）细粒度图像识别技术

《淘宝儿童类商品行业规范》针对图片定义了明确的风险类型和详细的认定标准，例如在某些服装下的特定姿态、聚焦部位等，这给识别模型带来了多标签、易混淆的挑战。传统的粗粒度分类已经无法满足要求，需要模型输出更细粒度的标签对内容进行管控。在细粒度图像识别领域，区域注意力的定位和放大是一个核心难题。我们提出了多尺度循环注意力的 Transformer（RAMS-Trans），具体框架如下图所示。它利用 Transformer 的自注意力机制，以多尺度的方式循环地学习判别性区域注意力。方法的核心在于动态图像块建议模块（DPPM）引导区域放大，完成多尺度图像子区域的整合。DPPM 从全尺寸图像区域开始，通过每个尺度下产生注意力权重强度作为指标，迭代放大区域注意力，生成从全局到局部的新图像块。 该方法在多个公开数据集上均取得了目前最好的效果，相关成果形成论文《RAMS-Trans: Recurrent Attention Multi-scale Transformer forFine-grained

Image Recognition》被 ACM Multimedia 2021 接收。



4）单域多模态识别技术

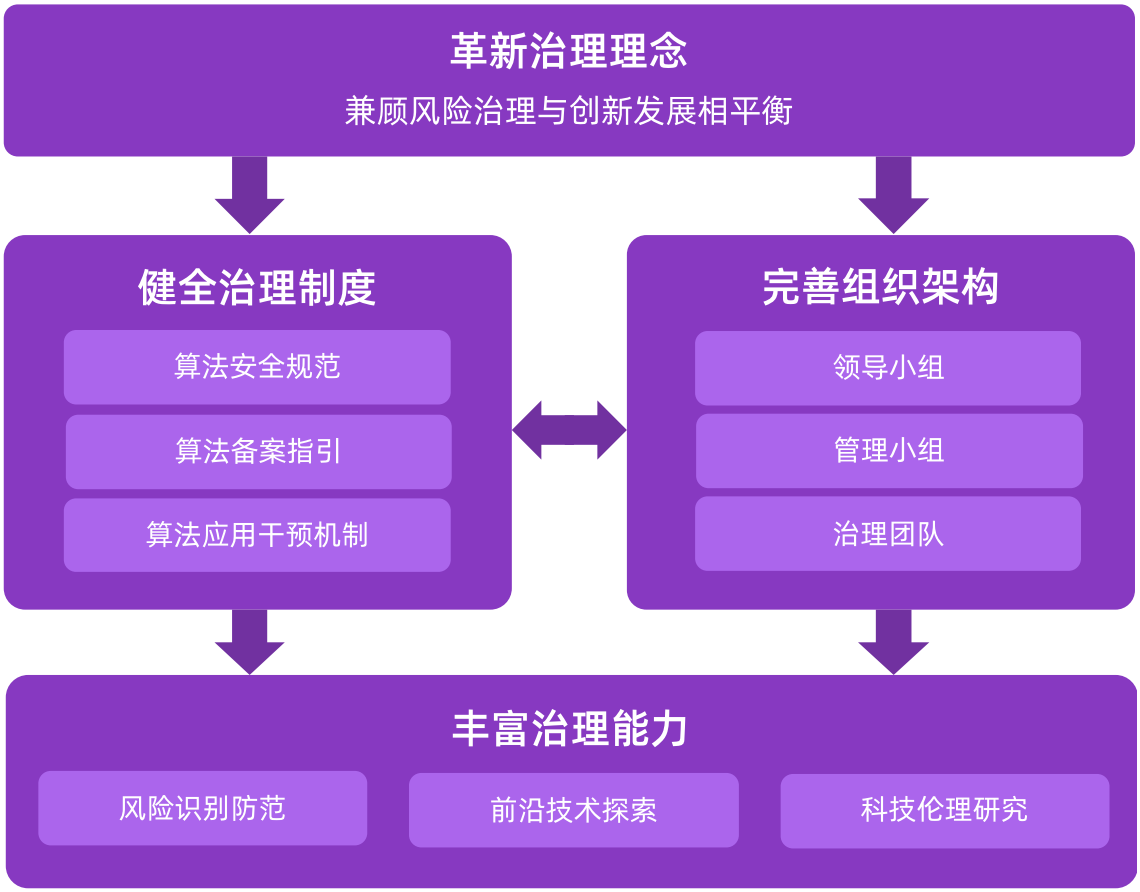
单域多模态指图片内容场景中的信息多模态，例如图片上加入文字信息，在商品图片和商品评价场景中广泛存在。这种情况下，针对单一图像维度往往难以明确识别风险，需要模型综合多个维度的信息来判断。目前多模态研究工作主要集中在不同模态信息的融合位置上，比较典型的如早期融合和晚期融合。我们发现对于不同的样本，其对应的图片重要性以及文本重要性是不同的，因此需要根据不同样本模态信息的重要性来动态选择权重进行多专家融合，从而达到更优的性能。我们设计了两个路由机制分别接受文字和图片信息，根据模态信息的数据分布确定其重要性。通过重要性融合产生参数来指导动态网络进行融合。路由机制可以选择性使

用，因此兼容了多模态和单模态的任务。此外，框架的高度模块化也使得在工业实践上也具有很强的适配性和可用性。相关工作整理为论文《DRDF: Determining the Importance of Different Multimodal Information with Dual-Router Dynamic Framework》被 ACM Multimedia 2021 接收。

肆 FOUR

构建全方位的人工智能管理体系

企业是人工智能技术、产品或服务的研发和使用的核心主体，也是人工智能治理落地实践中最重要的主体。**企业落实人工智能治理与可持续发展的要求，需要从治理理念、治理组织、治理能力等方面入手，构建全方位的人工智能管理体系。**人工智能管理体系框架图如下所示：



4.1 革新治理理念：兼顾风险治理与发展创新

人工智能与传统技术的风险治理存在明显差异。就技术发展速度而言，传统领域的技术发展速度比较平缓，可能十年，甚至更长时间才有新一代技术出现；而且新技术的应用、扩散也需要较长的时间。相比之下，人工智能技术的发展速度非常快，几乎每年都更新换代；得益于发达的互联网，新技术的扩散速度也很快，往往几个月内就在业界普及，并触达海量用户。就治理的成本与时效性而言，对于传统领域技术，参与风险治理的主体有足够的时间对成本、收益和风险做考察与思考，再反复酝酿治理的具体手段。然而，对于以深度学习为主要形式的人工智能，社会各界尚不能完全认识其本质原理，也就无法准确计算其成本、收益与风险。

企业需要秉持可持续发展的人工智能治理理念，兼顾风险治理与创新发展相平衡。首先，企业要做到合规，在法律法规的范围内开展经营活动；其次，企业需要在促进技术创新的领域开展有益探索，在合乎科技伦理的范围内鼓励新技术新应用的尝试。

4.2 健全治理制度：建立合规机制与规范行为

我国针对人工智能治理、算法治理出台了相应的伦理原则、法律法规等文件，对于企业规范开展人工智能相关业务具有重要的指导意义。当前，企业内开发、应用人工智能等算法十分普遍，其中牵涉大量的人员和业务，对内部管理提出了极大的挑战。不断完善人工智能算法安全治理制度，不仅方便与监管要求对标，将治理落到实处，同时也能够助力企业提升管理效率、降低管理成本，保障业务健康发展。

具体而言，企业需要规范算法研发应用的具体行为、加强算法研发应用的风险管控、明确算法合规管理机制、落实算法安全主体责任、建立算法应用干预机制、定义算法安全事件和响应机制等。人工智能算法安全管理制度应包括的主要内容如下：

1. 明确算法设计、研发及应用的基本原则

树立算法正确导向、算法公开透明、避免算法滥用、保护用户合法权益等基本原则 。

2. 落实算法备案要求

明确算法备案的主体责任、公示要求、协同机制、填报事项等，积极落实算法备案要求。

3. 算法上线安全规范

- 建立算法上线前测试及灰度过程。
- 重要算法上线时，人工智能风险治理团队进行安全评估；
- 业务部门应建立算法上线监控机制。

4. 算法应用干预机制

针对生成合成类、个性化推送类、排序精选类、检索过滤类、调度决策类等 5 大类算法，结合企业业务特点，对算法结果的干预责任、操作人员资质、干预原则、合规动作等做出具体要求。

5. 算法安全事件及响应

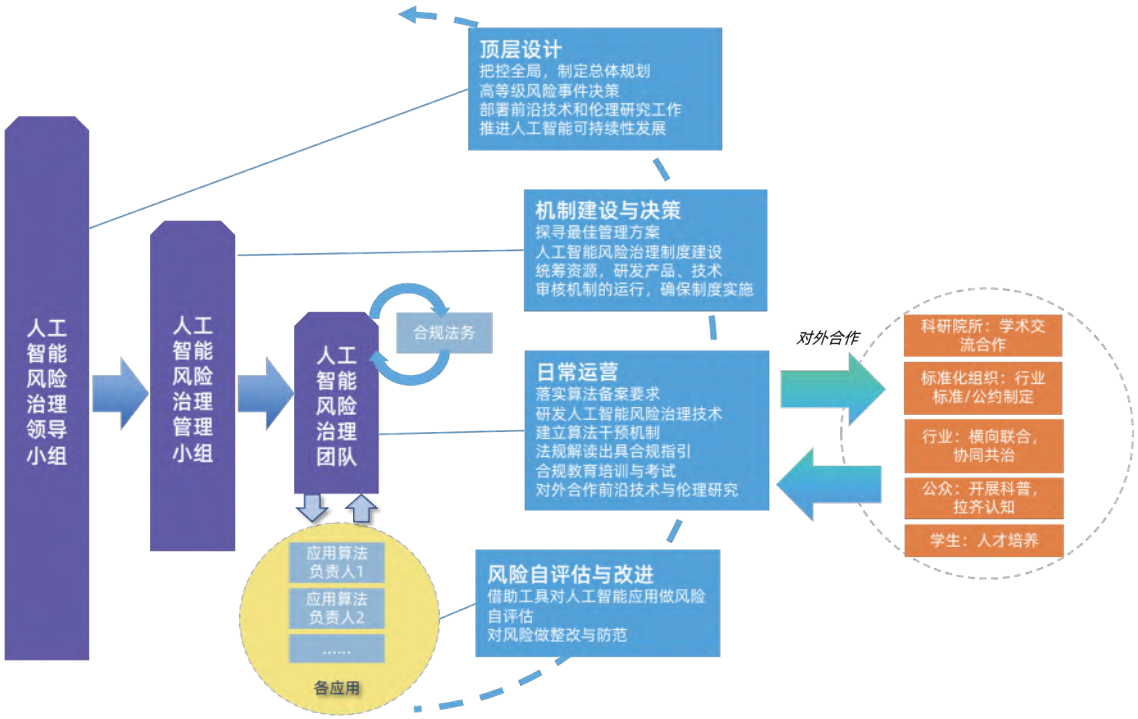
根据算法安全事件的性质、严重程度、社会影响力等因素，对算法安全事件进行分级，不同级别的事件建立适合组织状况的响应机制。

6. 违规处罚

对于违反制度规定的员工，按照具体情形给予处罚。

4.3 完善治理组织：明确责任归属与岗位分工

企业中人工智能风险治理的运营管理是一个涉及面广，参与者众多的复杂的系统性工作。**企业内部应当设置纵向包括领导小组、管理小组、专职风险治理团队在内的三层组织，横向多部门联动的工作方式，建立责任清晰、分工明确的风险治理体系，目标是确保人工智能安全可控、标准规范、全面高效，形成行业最佳实践。**企业人工智能内部组织安排如下图所示：



1. 构建担纲顶层设计的领导小组

企业主要负责人必须主管人工智能风险治理工作，以避免实践中可能遇到的问题，典型如影响力不足、权限不够，与业务部门冲突时难以协调，无法推动落地等。由于人工智能风险治理需要在法律和伦理的约束下展开，在技术标准、行业公约的指引下进行，在各业务的技术实践中落地，这涉及到法律、伦理、技术、业务的全面融合，因此相关领域的主要负责人都需要参与并担负领导责任。

组织顶层设计是设立领导小组，成员包含企业风险管理负责人、企业法务负责人、企业技术负责人、企业科技伦理治理委员会负责人，以及安全技术的负责人。领导小组应主要负责下述工作：（1）把控全局、制定人工智能风险治理的规划；（2）针对高等级风险事件进行决策；（3）布局伦理道德和前沿技术的相关研究；（4）推进人工智能技术助力社会可持续性发展。

2. 构建横向部门联动的管理小组

领导小组之下设立管理小组，由人工智能风险治理团队负责人、数据安全负责人、相关法务、从事人工智能伦理治理研究的专业团队，以及各算法负责人等组成。管理小组应主要负责下述工作：（1）研究人工智能风险治理的实践案例，探寻最佳管理方案；（2）制定企业内人工智能研发和应用全生命周期的安全管理规定；（3）组织资源，通过体系化、产品化的方式促进治理措施在应用场景落实；（4）研发安全检测和日常运营的产品，并对人工智能安全事件开展监控、调查和分析工作；（5）定期审核产品或措施的有效性，确保制度的实施。

3. 构建专职日常运营的治理团队

人工智能风险治理团队在管理小组的领导下，负责日常运营与管理，主要工作包括：（1）落实算法备案要求，建立企业内算法风险大图和算法安全评估制度；（2）建设人工智能风险治理产品与工具：牵头研发安全风险监控、安全能力评测、风险事件预警等产品，提供合规解决方案；（3）建立算法干预机制，充分保障用户的合法权益；（4）前沿技术、伦理研究：开展对外合作，探索人工智能的前沿技术，展开人工智能伦理的研究；（5）解读相关政策法律：与法务部门合作，对法律法规及时进行解读，识别业务的合规风险，给出风险治理指引。（6）合规教育培训：搭建合规教育培训的渠道，积累业务中算法风险治理的最佳实践，联系法律法规组织培训与考试。

4.4 丰富治理能力：结合风险防范与前沿探索

当前，我国出台一系列法律法规、政策规范，要求企业全面构建人工智能治理能力。国家新一代人工智能治理专业委员会于 2019 年、2021 年相继发布《新一代人工智能治理原则——发展负责任的人工智能》《新一代人工智能伦理规范》，旨在将伦理道德融入人工智能全生命周期。2021 年 12 月，中央全面深化改革委员会审议通过了《关于加强科技伦理治理的指导意见》，在顶层设计层面加强科技伦理治理。《个人信息保护法》《网络安全法》《数据安全法》等法律要求企业开展活动时落实数据安全、保护个人信息的相关要求。2021 年 10 月，国家互联网信息办公室等九部委联合发布《关于加强互联网信息服务算法综合治理的指导意见》，明确提出强化企业主体责任。企业应构建完善的人工智能管理能力，切实防范人工智能发展过程中的各项风险。落实上述要求，企业需要从全生命周期的风险识别及防范能力、前沿技术探索能力、伦理研究能力等三个层次全面构建完备的人工智能治理能力。企业人工智能治理能力构建如下图所示：

1. 完备风险识别防范能力

企业应当将治理理念落实到工程技术实践中，在人工智能系统全生命周期构建完备风险识别及防范能力。一是以算法备案为契机，建立算法风险大图和算法安全评估制度。当前，我国构建了以双新评估、算法备案、算法检查为主的算法监管制度架构。一方面，企业需要以算法监管为切入点，积极进行算法备案，深入摸排算法应用的基本情况，全局把握风险。另一方面，企业需要积极开展算法安全评估，审核算法的具体内容，对应用场景的功能做价值导向判断，对算法的原理做合理性分析，对可能存在的算法滥用、自身缺陷做风险评估防护，把风险拦截在萌芽状态，保障用户的正当权益。二是落实合规要求，建立风险检测能力。企业在算法应用的全生命周期，应积极落实各项合规要求。如建立算法设计与训练的安全审查，算法测试与上线的稳健性监控，以及日常运行的风险感知及记录存档。三是形成算法干预机制，对结果进行纠偏。由于人工智能现阶段存在因为数据使用不当、自身缺陷等原因造成自动化决策结果有偏，需要算法服务提供者对其进行有力的干预，充分保障用户权益。例如，对于推荐、检索类算法，需要加入内容打散等干预功能，并优化规则的透明度和可解释性；对于决策调度类算法，应当通过完善算法运行机制保障劳动者的合法权益。

2. 提升前沿技术探索能力

治理人工智能的风险，需要从技术底层入手理解其本质和内涵，并根据研究结果寻找有效的治理方案。在预测、分散、降低风险发生的基础上，为人工智能的使用与发展开辟足够的空间，从而实现人工智能的价值最大化。底层原理的分析，以及对风险的预测、规避等都属于人工智能的前沿技术，企业应加大投入进行深入研究。

3. 重视科技伦理研究能力

从世界范围来看，人工智能技术和应用发达的国家和地区，都高度重视其伦理问题。科技伦理的适度约束，将会促使人工智能新成果获取公众支持和业界认可；过于严苛的约束，则有可能制约技术发展速度。**企业作为落实人工智能治理的重要主体，需要加强人工智能伦理研究，在技术创新和伦理之间寻找到平衡点，实践“高科技 + 好科技”的文化理念，对新生事物加以约束防范风险，为利用人工智能促进社会可持续发展提供保障。**

专题：如何构建行之有效的算法透明

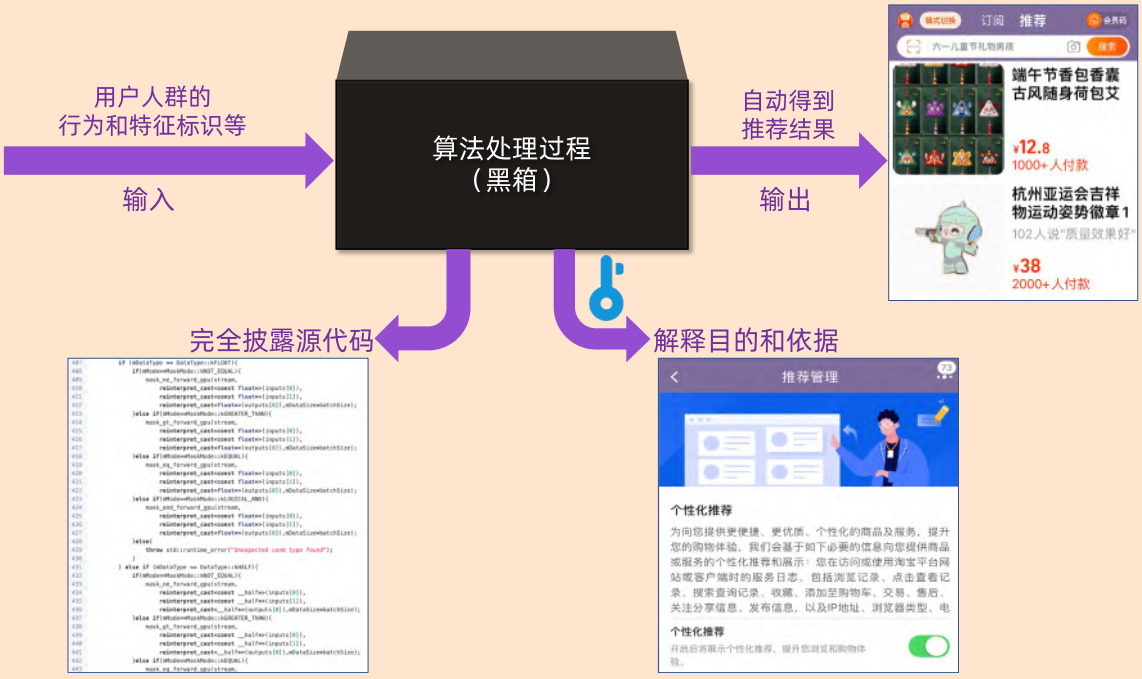
1. 打开算法黑箱需要构建算法透明机制

算法服务提供者通常使用算法来处理数据，但对于用户来说，看到的只是数据（或个人信息）被收集、以及依据数据决策的部分结果，看不到整个数据处理过程、算法工作原理、算法决策结果，算法服务提供者也没有对此进行解释或公开，因而形成“黑箱”。

为打破摸不着、看不懂的算法“黑箱”，算法服务提供者需要设计并实施合理的算法透明机制，对算法工作机理和运行结果等做出解释，保障用户知情权，构建用户理解和信任算法的基础。

2. 实现算法透明的具体路径

算法透明是一项复杂工程，业界主要存在两种不同的实践路径。一种认为需完全公开算法源代码在内的所有算法要素，另一种不要求完全公开源代码，而是向用户解释算法如何运作和决策，侧重算法运行原理和运行效果的透明。



完全披露源代码并不能有效解决算法透明问题。公开源代码是一种纯粹技术性的公开，没有做到对算法使用数据的情况、算法的运行原理进行有效解释，因而不能达到用户知情的目的。

算法透明的目的是打破算法“黑箱”，让用户和监管了解算法服务提供者如何使用算法处理采集的数据（包括用户人群的行为和特征标识等）、了解算法运行机制，进而对算法运行结果达成理解。因此，实现算法透明应当是使用合适的方式对算法的工作原理、目的意图、运行机制、决策逻辑等进行充分解释和说明，通过企业自控、监管监督、用户反馈的方式，构建用户理解和信任的算法透明体系。

3. 阿里构建算法透明的实践方案

我国《互联网信息服务算法推荐管理规定》（以下简称“《算法推荐管理规定》”）等法律法规对算法透明度和可解释性作出了相关规定，阿里深刻认识到算法治理对

于维护网络空间健康发展的重要意义，根据要求**一方面建立对监管透明的算法管理体系，另一方面构建对用户透明的算法运行机制：**

1) 建立面向监管透明的算法管理体系——启动清朗专项，维护清朗网络空间

全面构建算法合规制度。阿里积极对内解读、宣贯《算法推荐管理规定》等法律法规，修订完善集团算法相关规章制度，包括《阿里巴巴算法安全规范（总纲）》《阿里巴巴算法干预机制指南》等，提升全员算法合规意识。

积极履行算法备案义务。**阿里集团层面成立算法备案小组，对各业务线开展备案培训，开展备案工作。**努力确保与监管在算法治理上的信息对称，积极配合监管开展的安全评估和监督检查工作，提供相应的技术支持。

完善算法安全审查机制，落实算法全生命周期管理。阿里对照《算法推荐管理规定》等法律法规，逐条梳理，完善自身的算法审核机制，规范算法流程管理，摸排检查集团算法应用合规情况，启动内部清朗专项，按照“清朗 - 算法综合治理”要求，进行算法自查自纠。

2) 构建面向用户透明的算法运行机制——以淘宝和饿了么为例

面向所有用户进行通用算法解释，保证用户知情权。淘宝在隐私协议中对所利用的算法部署的情况、算法类型、收集的信息范围、使用目的，以及对用户带来的潜在

影响等情况进行了明确说明。针对特定应用以及用户量不大的小众应用，设计了单独的产品页面，用户通过访问产品页面，可以方便地了解算法的相关情况。针对大众普遍关心的话题，比如骑手调度算法，饿了么发布了《2022 蓝骑士发展与保障报告》进行了详细阐述。为了提高用户触达，帮助用户更为快捷、便利地了解算法运行情况，饿了么还计划未来通过官网展示、官方微博和微信公众号发布等方式将算法规则进行公开，让更多的社会公众了解并进行监督。

对于使用个人信息进行决策的场景，赋予用户便捷退出个性化推荐的权利。淘宝 APP 在隐私页面二级目录设置了退出个性化广告和个性化推荐的选项，用户可以轻松地关闭基于个人信息的商品服务广告和推荐服务，同时对于关闭后因推送精度下降影响用户体验进行提示，帮助用户清楚地做出决策。

建立用户负反馈机制，让用户了解用于算法决策的信息，并可进行有效控制。对于个性化推荐的商品进行效果反馈，淘宝提供了非常方便的操作方式**（比如用户可以长按推荐列表中的商品选择负反馈的具体原因）**，帮助优化算法推荐决策，屏蔽用户不喜欢的商品信息，让用户对算法推荐的商品有最终选择权。对于个性化广告，赋予用户选择和关闭自己不感兴趣的商品品类标签，满足用户对于商品广告的自主需求。

建立人机协同的用户投诉申诉反馈机制。淘宝用户数量巨大，用户反馈频次高、类型多、问题广。通过使用深度学习算法，淘宝为用户提供智能问答对话、智能语音客服服务，一方面及时解决用户大部分的投诉建议问题，另一方面也甄别用户权益受损的案件，通过人工进行进一步解决。

算法透明实践需要全方位考虑社会各主体的期待和利益。阿里在实践中不断完善面向社会各方的算法透明机制，未来也将持续更新和丰富治理措施，确保算法可控可信。

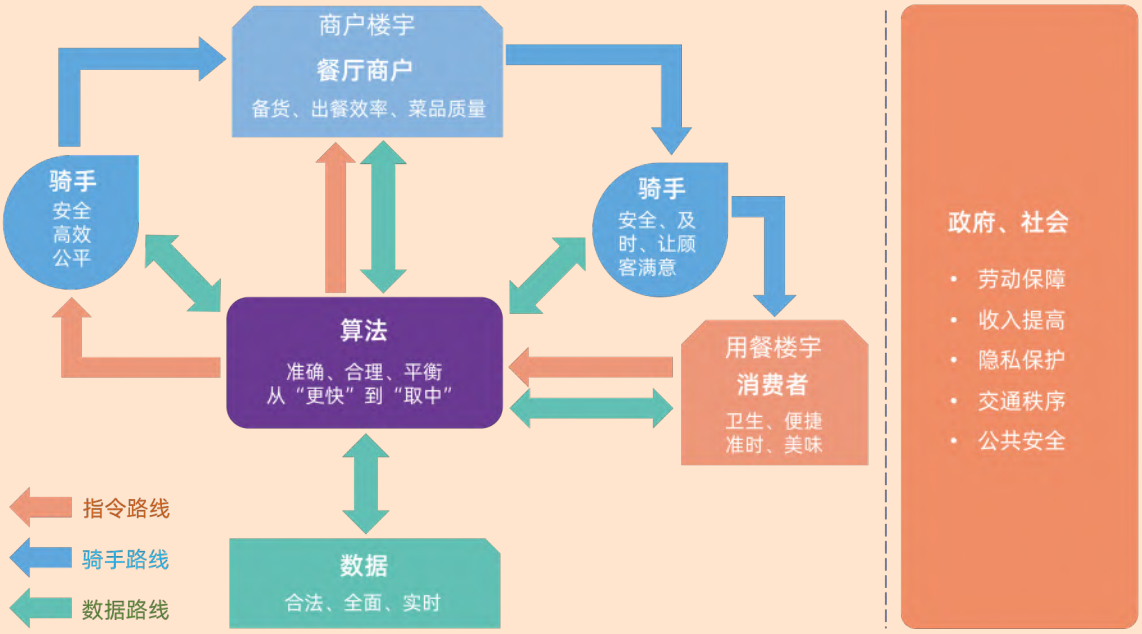
专题：调度决策：落实即时物流系统“算法取中”

1. 调度决策算法影响劳动者权益的成因

调度决策类算法是根据系统的资源分配策略所规定的资源分配算法，被广泛应用于外卖、网约车等行业。以外卖配送系统为例，算法的核心目标是合理利用骑手运力匹配用餐高峰期大量的订单需求，需要在较短的时间内实现订单分配、骑手调度和餐饮送达。目前我国外卖行业的订单峰值主要集中于午餐和晚餐时段，传统的点对点、一人一单的调度方式无法满足高峰时段多样化的用餐需求，而配送调度算法高效地连接了商户和用户，通过优化配送路线、提高背单量的方式分配运力，在提高配送效率的同时也增加了骑手收入。外卖配送调度算法的广泛应用极大推进了零工经济的发展，促进了灵活就业，提高了生活效率，但是通过算法进行的配送调度决策会被引发系统忽视对骑手劳动权益保护的质疑。

调度决策算法对劳动权益保护不足主要有两个原因。一是算法可能出现异常情况识别和处理能力有限的情形。一方面，外卖配送周期较短，场景复杂，安全事故的可预见性较低；另一方面，对配送影响较大的天气情况、交通状况等数据的精准度和实时性匹配欠缺。场景过于复杂，加上数据本身不够完整、精细，限制了算法对安全风险的识别能力。二是对于产生较大劳动压力的算法决策缺乏人工干预机制。不同于传统劳动。不同于传统劳动环境中劳动者与管理者存在较为充分的沟通、交流和反馈机制，通过算法进行配送调度决策主要是基于对骑手行为数据的分析和比对，在欠缺及时沟通的情形下，生成的决策可能存在不合理因素，导致骑手劳动权益受到影响。例如，系统设置的骑手背单量上限可能忽视骑手劳动饱和度的边际效应，骑手被要求多单配送造成心理压力增加，出现骑手违反交通规则等行为。又如，预估时长是外卖配送服务完成质量的核心指标，但存在商家出餐时间、交通异常情况等诸多不确定因素，骑手只能调整自己的配送速度、增加配送强度来对冲商户延迟

或交通拥堵造成的超时风险。可以看出，“骑手被困在系统中”舆论质疑的出现，既有算法对数据使用不当的原因，也有对算法做出的有偏决策结果缺乏干预的原因。



2. 饿了么持续通过算法优化保障骑手权利的实践方案

当前，为更好保护劳动者权益，饿了么开启“蓝骑士保障计划”，通过持续优化业务方案和算法规则，**在商业逻辑合理与价值观合情之间取中，兼顾效率与劳动者保护，付诸努力切实保障劳动者安全与公平。**

1) 提升算法规则劳动者参与度，进行区域化区分，促进调度算法决策机制客观化、人性化

首先，在超时、差评等考核上，饿了么已逐步取消对于骑士的逐单处罚，改为一定时间周期的率值考核。并且当用户做出差评后，会再次人工评估差评是否成立，若骑士服务并无问题，则不会下发差评评价。由于采用率值考核，若骑士整体配送服

务优异，出现个别差评不会造成较大影响。其次，设立明确的补贴机制。提供补贴的场景覆盖法定及其他节假日、特殊天气或环境、较难配送的订单等。此外，根据骑手工作城市的经济状况和物价水平，对骑手的收入进行调整，保障骑手的收入与当地生活水平相匹配，提升骑手的劳动获得感。

2) 提高算法异状识别处理能力，建立异状反馈和决策退出机制

骑手安全是调度决策算法治理的重点要求。为使风险防范前置，饿了么的算法会参考一些历史数据，如：历史路段交通事故情况、历史气象条件、骑士近期疲劳程度，并结合一些实时数据，如：实时天气数据、骑士实时工作时长、实时骑行速度等，综合判断骑士在当前条件下的配送安全系数。当安全系数低于一定范围，饿了么会在派单过程中进行安全防护。比如，当识别到骑士骑行速度过快、综合判断可能接单压力过大时，调度系统会暂停为骑士新增派单；在一些复杂的配送场景，例如暴雨、沙尘天气、道路临时管制、商户出餐慢、联系不上顾客等，饿了么会为骑士自动匹配灵活配送时间；当调度系统感知到局部运力压力过大，如大促爆单等情况，也将自动触发保护方案。不仅如此，骑手也可以通过人工报备的方式，申请匹配灵活配送时间，在突发异常时保障安全。

3) 识别影响劳动者权益的核心决策指标，通过人工干预机制平衡风险和效率

对于背单量，饿了么的调度系统会结合骑手服务能力、局部区域配送压力情况、天气状况等因素调整建议背单量上限，若骑士觉得当前压力过大，或者背单量不足，可以通过人工干预的方式，自主调整背单量上限，以适应实际配送需要。对于配送时长，饿了么也在试点增加人工复核校验，通过站点组成的地面网格实时反馈当下情况，完善补时机制，如遇特殊情况，站长还可以进行人工干预，设置更灵活的配送时间。

未来，饿了么也将继续跟踪科技更新暴露出来的新问题，通过技术创新、管理创新等方式治理外卖配送生态，主动广泛听取社会组织和公众的意见建议，推进调度决策算法应用的完善。

专题：如何获取消费者对电商平台价格和用户权益的信任

1. 大数据杀熟引起定价机制的信任危机

大数据杀熟，最初的含义是在互联网产品中针对同一件商品或服务，老用户看到的价格比新用户更高。后来这个概念被泛化，只要是用户购买同一个商品或服务的到手价不一样，都被一部分用户认为是大数据杀熟。用户产生这种想法，其原因在于：互联网产品采集用户信息并进行统计分析的能力强大，既然能够提供精准的个性化推荐服务，也能够分析用户的购物意愿、预期交易价格，然后将商品或服务设置为不同价格。与此同时，如果商品的价格 / 优惠机制比较复杂，理解成本较高的话，用户也可能产生猜疑和误解。

解决信任危机，首先需要保障用户的知情权，以清晰、易于理解的方式向用户展示说明价格、优惠的构成与计算逻辑；再是企业内部算法管理工作上，设计机制、建立技术手段防止数据滥用，在定价、优惠发放上设置卡口，保障用户权益的公平性。

2. 淘宝价格机制公开的实践方案

对于电商平台而言，用户最终购买价格由以下因素构成：商品定价、店铺优惠、平台优惠、支付优惠等。实践中，企业需要严格遵守相关法律法规，从以下方面保障用户权益的公平性，给出清晰的展示和说明，消除用户的误解。

1) 价格产生机制透明

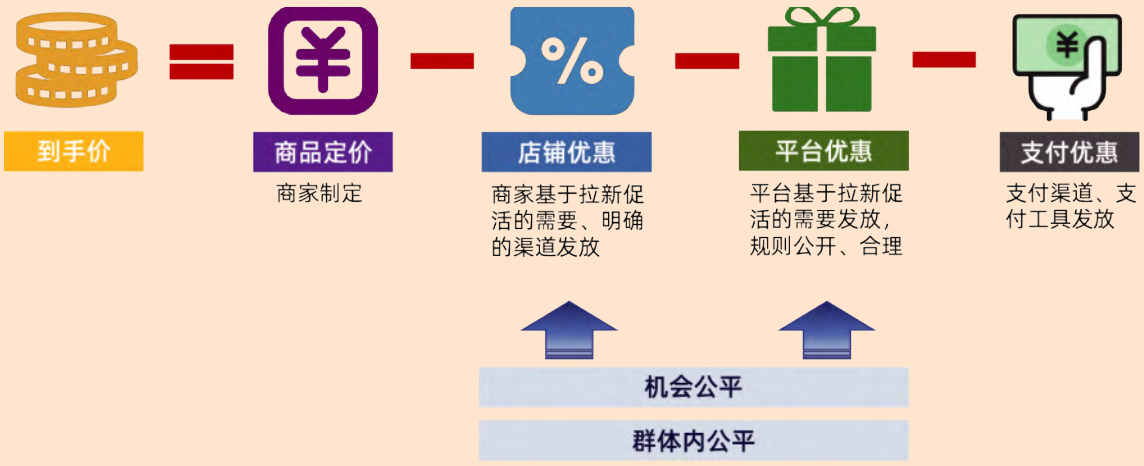
电商平台上，第三方店铺中的商品由商家根据成本、毛利目标以及市场反应等多个因素进行定价，平台并不参与。基于商家有运营会员体系和分层运营的商业诉求，在遵守法律法规的前提下，商家为店铺会员、粉丝等身份的消费者提供折扣、优惠，但是定价策略必须公开，并且明确标识用户身份和对应的价格，同一种身份的消费者看到的价格保持一致。比如，店铺会员看到的价格应明确显示为“会员专享价”。身份的获取，需要有明确的规则，比如通过关注店铺、购物等行为成为店铺会员，就可以享受到“会员专享价”。以上措施确保在商品价格的设置上不会出现针对用户的不合理的差异化定价。

2) 券后价清晰展示

平台、商家会开展多种多样的营销活动让利于消费者，同一时间可能存在多种优惠叠加到同一商品，比如在享受单品优惠的同时可能还享受店铺满减优惠、平台满减优惠等。为了展示对于用户最有利的优惠组合，以及减少用户对券后价计算的疑惑，在搜索、推荐、商品详情、购物车等核心页面清晰地展示券后价，更进一步还可以给出券后价计算公式，明确展示使用的具体优惠项，降低消费者的理解成本。

3) 权益发放规则公开、合理

购买价格（到手价）= 商品定价 - 店铺优惠（单品优惠、店铺满减优惠、店铺或品牌会员权益等）- 平台优惠（平台满减优惠、平台品类优惠、平台会员权益等）- 支付优惠（充值膨胀、支付工具优惠）。计算公式如图所示：



上述计算公式中的店铺优惠、平台优惠和支付优惠都属于用户权益。其中，支付优惠由支付渠道和支付工具提供，电商平台不参与。用户权益的发放需要有明确的规则和合理的理由。店铺基于拉新促活的需要、明确的渠道（直播、返利）等进行优惠的发放；同一渠道优惠发放逻辑需透明，基于同一拉新促活活动类型的消费者获得的面额一致。平台优惠发放的范围和面额差异化逻辑需公开，并让用户有平等的查看、获取权利。平台会员权益规则公开、合理，基于同一拉新促活活动类型的消费者能够领取、兑换的平台权益一致，比如芭芭农场阳光兑换优惠券时兑换比例一致等。

针对不同用户，可能在获得的优惠结果上不一致，但在具体实施时遵循以下原则以保障用户的权益公平性：

机会公平

比如，所有用户都可以通过参与互动游戏、领取平台资产等方式获取权益，并在下单时享受；有的红包发放为了增强趣味性，在不使用用户个人特征的前提下，实行机会均等的金额随机。

群体内公平

比如，购买了同一种付费会员身份的用户，都能够享受特定商品相同的折扣优惠。

在同一时间内公平

比如，优惠金额可能随时间发生改变，但改变不针对用户个人特征，同等条件的用户享受的优惠在同一时间内相同。

除此之外，还有一些遵照法律法规的用户权益发放方式：（1）按照法律法规为照顾特殊群体而进行的优待。比如，在用户主动披露信息的前提下，对儿童提供购票优惠等。（2）按照法律法规和既有商业惯例针对特定群体的优惠。比如在说明规则的前提下，新用户在合理期限内享受特定优惠等。

伍 FIVE

联动多主体落实协同治理要求

5.1 严格落实主体责任，支撑政府提升监管治理效能

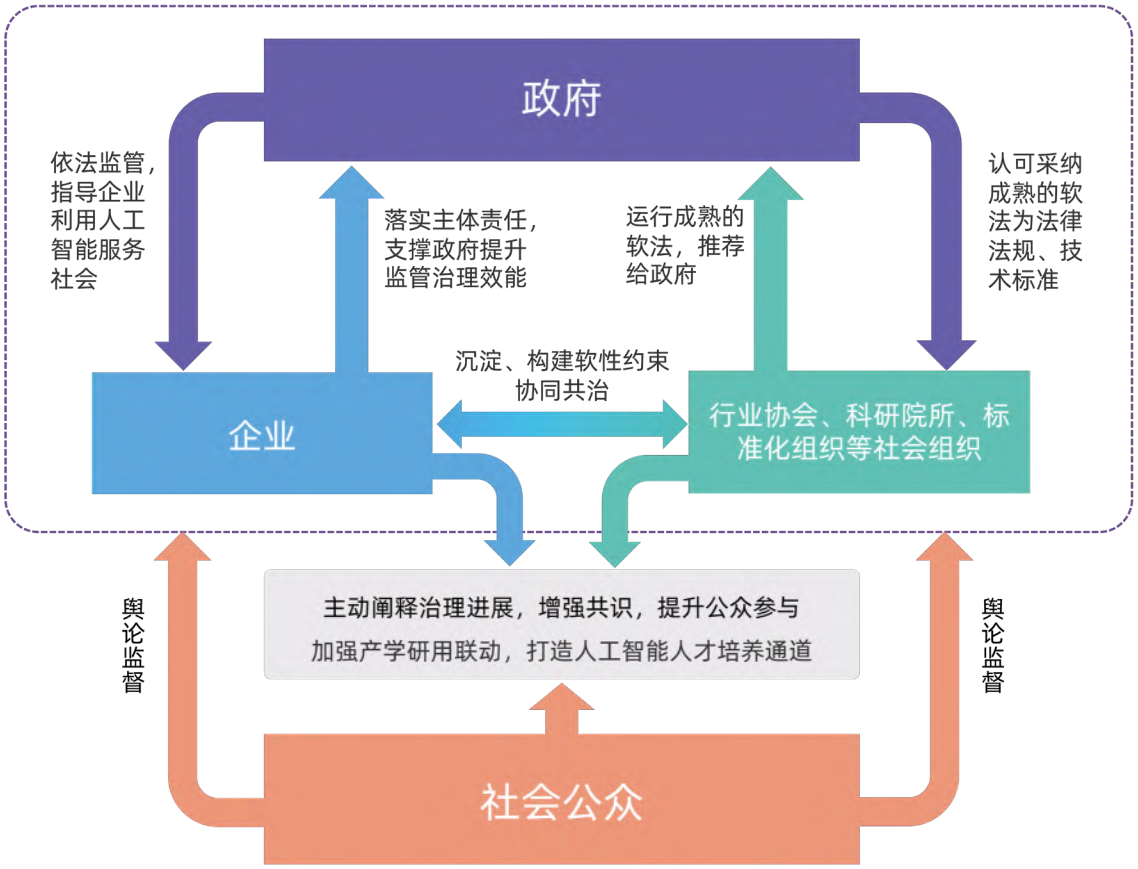
5.2 积极参与标准制定，联合行业组织共促行业自律

5.3 主动阐释治理进展，持续提升社会公众参与程度

5.4 加强产学研用联动，打造人工智能人才培养通道

5.5 联合产业治理力量，守护清朗健康网络生态环境

人工智能治理是复杂性、系统性很高的社会议题，涵盖政策、技术、产业、法律、传播、伦理、安全、国际关系等诸多领域，需要社会各界共同探索合乎人类发展需要的人工智能治理模式。**在治理过程中，政府机构、标准化组织、企业、学校、科研机构等主体的利益诉求、专业程度、角色分工均有所不同，只有加强交流和创新合作，才能实现人工智能的健康发展，促进全球社会、经济和环境的可持续发展，推进人类命运共同体的构建与实现。**企业作为人工智能治理落地实践中最重要的主体之一，应积极推进与其他主体的良性互动，共同推动人工智能多元主体协同治理。



5.1 严格落实主体责任，支撑政府提升监管治理效能

在蓬勃发展的人工智能技术的推动下，人工智能治理逐步成为全球关注的焦点，各国政府在治理进程中纷纷颁布法律法规，推动相关规则构建。人工智能作为新技术，更新迭代速度更快，影响波及面更广，企业作为新技术的研发和使用方，应从传统治理的角色中积极转变，充分发挥能动性，在严守法律法规底线的同时也要为新技术立法提供支撑。

企业作为人工智能风险触发的首要主体，需要根据法律法规的要求，严格落实法律责任。

2022 年 3 月 21 日，中共中央办公厅、国务院办公厅印发了《关于加强科技伦理治理的意见》，规定：“企业等单位要履行科技伦理管理主体责任，建立常态化工作机制，根据实际情况设立本单位的科技伦理（审查）委员会。从事生命科学、医学、人工智能等科技活动的单位，研究内容涉及科技伦理敏感领域的，应设立科技伦理（审查）委员会”。意味着中国政府对加强科技伦理治理作出了顶层设计。

企业应探索落实《关于加强科技伦理治理的意见》，设立科技伦理治理委员会，完善企业内科技伦理体系，伦理指导人工智能治理工作，治理中的实践经验反哺伦理体系建设；有效防控科技伦理风险，不断推动科技向善；积极沉淀行业通用的伦理框架和准则，向社会贡献最佳实践方案。企业应积极贯彻落实《网络安全法》《数据安全法》《个人信息保护法》等法律法规的要求，促进个人信息合理利用，保障网络和数据安全。企业应切实落地《互联网信息服务推荐算法管理规定》，优化算法推荐服务机制，促进算法应用向上向善。

企业作为人工智能技术研发和创新应用的一线，需要支撑新技术法律法规的制定，支持政府治理随着技术发展不断更新迭代。企业投入大量资源开展人工智能技术研究，发掘出多种多样的新应用场景。与此同时，企业可以主动向政府介绍新技术的

发展，以及新实践的开展情况，使政府及时全面了解其中的技术原理、在网络空间和现实社会中可能造成的影响，避免政府对创新发展状况的掌握存在滞后，形成信息不对等的情况。企业积极主动寻求指导，还可以帮助政府及时准确地定义责任主体，明确治理方式，为人工智能风险治理主体之间的良性互动奠定技术维度的基础。基于此，阿里在人工智能技术应用过程中，一直积极与政府部门保持联系，通过多种方式向政府部门提供实践经验。

5.2 积极参与标准制定，联合行业组织共创行业自律

标准是“软法”工具箱中核心组成部分，它强调凝聚共识和企业自治，具有高度的灵活性，能够适应复杂化、动态化的社会现实，弥补法律的缺陷和空白，在推动科技创新互动发展、构建人工智能治理格局中发挥着不可或缺的作用，同时也是我国人工智能治理与发展形成具有国际影响力的生态圈和产业链的重要抓手。

人工智能亟需借助标准优势提高治理效能。在多方共治的新格局下，治理复杂度也呈倍数增加，标准是对于复杂问题达成一致的高效解决方案，是协同人工智能产业力量，实现多方共治的重要工具。明确清晰、适应人工智能发展的治理标准化体系，是企业高效贯彻国家方针、充分发挥主体作用的基础。人工智能标准为企业治理提供技术方向。在治理实践中，一方面企业应积极参与人工智能标准的研制，确保人工智能技术的安全性、准确性、鲁棒性，积极进行企业自律；另一方面，企业应与行业组织开展有效对话，促使各方凝聚共识，进而推动行业对于标准共同遵守和执行，最终将标准潜移默化地融入人工智能的软法治理当中，形成行业的自律风气。标准是我国人工智能治理走向全球治理的有力支撑。标准能够打造人工智能竞争新优势、开拓发展新空间。在企业全球化进程中，尤其是提升国际关注和认可，发挥我国人工智能产业独特优势塑造影响力等方面，标准是强大的助推器。

以标准化实践为例，阿里巴巴积极参与到国家、行业等标准化工作中，从人工智能治理全局思考、垂直领域的解决方案到针对性热点问题的探索，主动贡献好的实践输出给业界；同时结合实践现状，与业界共同探索新一代人工智能的标准解决方案，为促进行业共识、提升标准水位贡献力量。在国际标准方面，阿里巴巴标准化部在 IEEE 成立人工智能服务鲁棒性工作组，开展诸如人工智能鲁棒性测试，图像、语音识别等垂直领域鲁棒性等国际标准工作，结合业务一线的海量实践结果，协同国内外、产学研诸多专家共同探讨解决人工智能治理热点问题。

5.3 主动阐释治理进展，持续提升社会公众参与程度

企业需要积极发挥技术前沿的优势，主动向社会公众阐释人工智能治理进展，提升社会公众的参与程度。得益于互联网和人工智能技术的快速发展，社会公众可以低成本在网络空间中获取信息、发表意见并进行高效率的传播。企业可以通过利用各类平台，积极阐释技术治理的进程，广泛激发社会共同治理的活力。

阿里巴巴持续跟踪分享人工智能新技术、治理新观点、可持续发展新风向，积极面向公众阐释算法的技术原理和社会效果，得到社会良好反响，促进公众广泛参与治理。

1.《追 AI 的人》

2021 年 9 月，阿里巴巴人工智能治理与可持续发展研究中心（AAIG）联合高校和产业界发起了一档 AI 治理交互栏目——《追 AI 的人》，关注并分享人工智能新技术、治理新观点、可持续发展新风向。目前已联合清华大学、南开大学、浙江大学、科

幻小说作家、阿里研究院等举办了 14 期直播，主题包括《AI 与安全治理的恩怨情仇》《算法黑箱与算法透明》《国际 AI 视觉大赛冠军方案分享》《刑法介入人工智能风险规制的慎思》《数字人的 AI 心》《科幻、人工智能与伦理》《揭秘安全领域的老司机—风险知识图谱》《模型不可解释，预测不稳定？不妨因果推理融入机器学习试试》《人工智能训练师标准与技术展望——AI 职业技能标准构建新职业生态》《人工智能应用与数据保护》《谁偷了我的 AI 模型？一个“警察抓小偷”的故事》《深度学习对抗攻防 - 人与算法的无间道》《AI 前沿技术对抗中的”天使”与“恶魔”》《对人工智能产业发展四大要素的保护——数据与知识产权的挑战与实践》等内容。

该栏目吸引了超 100 万人次观看学习，并得到国务院新闻办下属中国网、江苏卫视下属荔枝新闻、每日经济新闻、微博直播、B 站直播、知乎直播的全程支持，在观众中持续引发热烈反响，产生了多种形式的互动。

2.《这个 AI 不太冷》

2021 年 11 月，阿里巴巴集团安全部、阿里巴巴人工智能治理与可持续发展研究中心（AAIG）及知乎科技联合发布人工智能与社会生活发展的科普栏目——《这个 AI 不太冷》。当前，第一季已经全部上线，栏目邀请科幻小说作家、电视台评论员、B 站 UP 主、脱口秀演员、中科院、中科大、阿里专家汇聚一堂，就常见的生活场景和热议话题，进行观点讨论，共同揭秘真实 AI。栏目共分为 3 期，第 1 期《“人工智障”，是算法翻车还是人翻车？》、第 2 期《虚拟真的能和人类共存吗？》、第 3 期《“算法偏见”是概念炒作吗？》。栏目总曝光达 3.3 亿次，视频播放量近千万，话题 6 次登上知乎、微博、B 站全国热榜；获得多家权威媒体的关注与转发（如中国青年报、人民网强国论坛、环球网科技频道、36 氪、蓝鲸财经、半月谈等媒体；安徽网信办、内蒙古团委、陕西政法、乌兰浩特政法等政务账号）；得到科幻作家郝景芳、前《瞭望东方周刊》执行总编、新华社采访中心副主任韩松、NHK 制片人刘庆云等人点赞；上榜学习强国，成为社会广泛参与到人工智能治理当中的优秀实践。

5.4 加强产学研用联动，打造人工智能人才培养通道

人工智能相关问题的解决需要专业化人才，在产学研用中必须谋势而动、顺势而为，不断推动人工智能与人才培养的深度融合。企业作为拥有前沿技术和优质人才的社会主体，通过联动产学研用，积极发挥技术创新作用，促进前沿学术发展，打造技术影响力，建立人才培养的良好通道，为人工智能治理奠定基础。

通过举办比赛，激发人才活力。2019 年，阿里巴巴联合清华大学积极开展“安全 AI 挑战者计划”系列比赛，提供场景、技术、数据、算力等支持，为广大安全爱好者提供数字基建安全的试炼场，在高难度的真实环境中提升技术，培养真正有安全实战能力的安全基建人才。当前，已经成功举办了八期“安全 AI 挑战者计划”赛事，吸引了全球 26 个国家 / 地区、900+ 校企、40000 余名选手报名，曾被张钹院士称为全球顶级的人工智能竞赛。通过实战比赛，以赛促学，助力诸多青年走上了人工智能安全的科研道路；每年 10 万美金激励青年、线下举办阿里安全课堂，线上发起系列直播，同时在高校铺设 20 个根据地，搭建企业和学生的直达通道等。

构建多种渠道，加强人才培养。阿里巴巴将通过在全国甚至全球高校设立分舵的方式，从产业界的实际人才需求出发，推进学科发展，让今天的高校学子，在学校能做到“穷理以致其知”，毕业后可以“反躬以践其实”，把人生理想融入为实现中华民族伟大复兴的中国梦的奋斗中。此外，阿里巴巴还将和乡村振兴团队合作，建设人工智能与安全线上课程，让广大乡村青年享有平等的起跑线，一台能联网的手机就能够与全球顶尖的院士、学者教授、工程师无障碍交流。

5.5 联合产业治理力量，守护清朗健康网络生态环境

当前，“网络水军”在网络空间肆虐，“黄、赌、毒”等违法不良信息仍不时出现，严重伤害了用户的合法权益，危及清朗健康的网络生态环境。在问题的解决方面，企业作为单一主体的治理模式难以获得好的成效，需要联动行业广泛积极参与，形成治理合力。

2020 年，阿里安全、阿里巴巴公益、百度安全联合发起“绿网计划”公益项目，配合浙江省市场监督管理局的指导，关注“守护生命健康”、“野生动植物保护”、“未成年人保护”、“人身公共安全”和“个人信息安全”等五大领域。2021 年，

阿里安全、阿里巴巴公益与抖音安全中心联合发起“绿网计划 2.0”，进一步将项目覆盖的领域扩大，新增“防范网络诈骗”“人身公共安全”“个人信息安全”“禁毒”“禁赌”及“拒食野味”等多个领域，响应全国“扫黄打非”办公室“新风 2021”集中行动安排部署，积极落实“净网 2021”、“护苗 2021”工作，旨在帮助未成年人解决遇到的各类网络安全问题。

“绿网计划”作为网络生态环境治理的创新举措，体现了平台企业共促网络生态向好的治理能力与责任担当。截至 2021 年，阿里安全通过风险感知、公益页面引导发起的守护生命公益行动，联合阿里集团 CCO 客服、阿里健康、线下警方、专业机构已救助 4400 余名因抑郁症等产生轻生倾向或行为的人，在公益行动中拨打宣导页面心理咨询热线的就有 34000 余人。目前，“绿网计划”已向全网进行了 2650 多万次正向的科普宣传教育，累计 78 万余位网民自发成为公益守护官。覆盖淘宝、闲鱼等 14 个 APP，非正常搜索词扩增至 15000 多组， 9 家公益合作机构加入，形成了包括阿里、百度、抖音等在内的内容安全治理公益组织——绿网计划联盟。2020 年 12 月，“绿网计划”公益行动入选 2020 年市场监管领域社会共治优秀案例榜首。

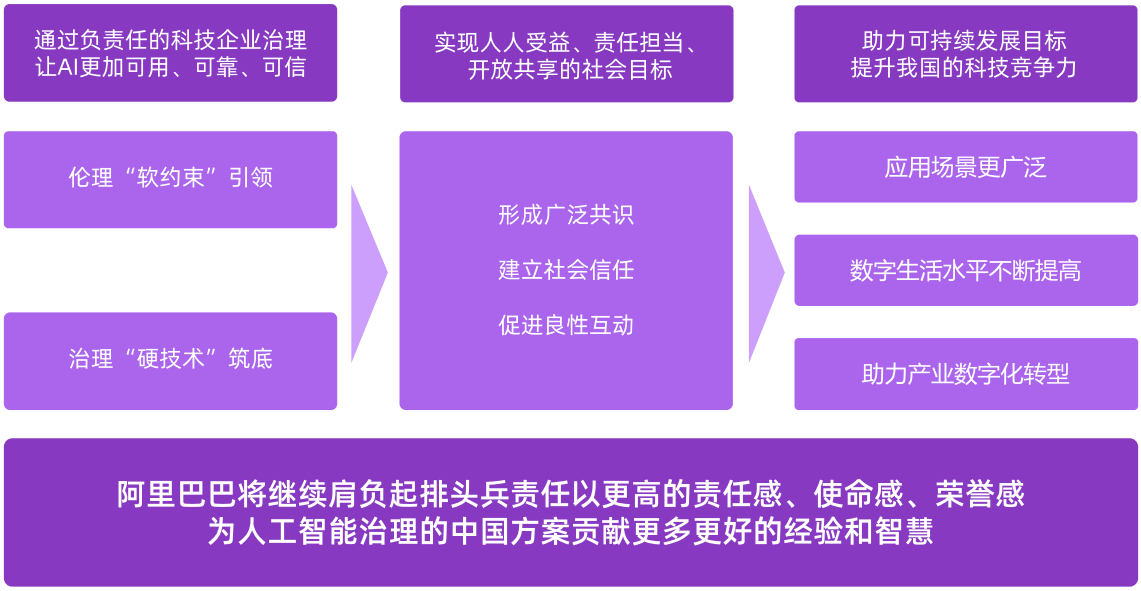
“绿网计划”的意义在于促进网络治理从事后走向事前，从惩戒转向预防，用“智理”实现“治理”，体现了产业对不良行为的坚决抵制，对网络生态的坚定捍卫。产业联合通过平台数字化的技术手段及完善的运作机制协同相关部门进行案件侦破和社会治理，实现了案件的快速响应和闭环处置，达成了公益宣传的引导作用，有力守护了清朗健康的网络生态环境。

陆 six

总结与展望

附　人工智能助力可持续发展的丰富实践

通过实践可持续发展的人工智能治理方法论，我们期待通过践行人人受益、责任担当、开放共享的价值理念，开发出更加可用、可靠、可信的人工智能技术，使得人工智能产业能在下一个五年内蓬勃且健康的发展， 成为助力社会发展的重要引擎。面向未来的人工治理应该进一步促进技术发展，推动更广泛的产业应用，实现符合伦理要求、助力可持续发展的社会价值要求。



第一，治理“硬技术”筑底，提升人工智能治理水平，让人工智能创造更为先进的生产力。**人工智能是创造先进生产力的重要工具，人工智能的发展需要技术不断进步，让人工智能变得更强、更快、更智慧，也需要与之相匹配的治理技术，让人工智能技术更透明、更安全、更好用。**企业应当不断提高人工智能的治理水平，打造可用、可靠、可信的人工智能技术体系，构建兼顾发展与创新的人工智能管理体系，将治理能力转化为人工智能企业的核心竞争力，为提高数字经济全要素生产率、全面提升我国科技的国际竞争力作出贡献。

第二，伦理“软规范”引领，构建人工智能伦理规范，通过人工智能实现人们对美好生活的向往。**人工智能是服务人类的工具，机器智慧不仅要求像人类一样思考，还需要具有人类的道德感和正确的价值观，在面对效率和安全、精准与隐私、客观与公平等冲突和矛盾的时候，能够基于人类的伦理规范，进行有效平衡、做出正确的选择。**我国的人工智能伦理规范正在逐步建立和完善的过程中，企业应当以人工智能伦理为牵引，结合内外部力量构建伦理治理机制，让伦理指引贯穿于人工智能研发、部署和应用的全生命周期之中，进一步弥合数字鸿沟、提高包容普惠、防范滥用误用，让更多的人能够用上更有温度的好科技，为实现可持续发展贡献更多的实践智慧和解决方案。

第三，就人工智能治理形成广泛共识，建立社会信任，形成良性互动。人工智能技术的发展和应用在推动社会生产生活进步的同时也带来了诸多影响，对人工智能的不解和误解造成了人们的不接受、不信任，一定程度上限制了人工智能的发展。人工智能治理需要提高透明度，在技术研发、产业实践、社会应用各方面进行准确、专业的解释，帮助人们理解人工智能的技术特点和应用规律，了解人工智能如何让数字生活更加便利，也让人们相信人工智能从业者严守合规底线，对于人工智能带来的风险，不同背景、不同学科的专业人员正在群策群力、攻坚克难，在最短的时间内寻找到最好的解决方案，让人们对人工智能更放心、有信心。

第四，促进人工智能更好的应用于更广泛的场景，提高用户的数字生活体验，助力产业数字化转型。我国拥有海量数据和丰富应用场景优势，人工智能推动产业实践，在电子商务、交通出行、生活服务、文化传播等领域不断创造新的应用模式，为人们提供更为高效、便捷、丰富的数字化服务。**面向未来，企业需要进一步利用人工智能技术推动产业发展，促进人工智能技术与实体经济深度融合，一方面让用户有更好的数字化生活体验，提升人们在数字经济发展中的获得感，另一方面发挥人工智能硬科技优势，加强人工智能技术在生产领域中的应用，赋能传统产业转型升级。**

纵观全球，世界各国对人工智能治理还没有统一认识，治理方案尚处于起步阶段。而我国已经抓住了人工智能发展的重大历史机遇，提前布局、主动谋划，在核心原则、基本要求、法律规范等方面全面提出了引领世界的人工智能治理中国方案。作为科技企业的代表，阿里巴巴的人工智能技术近年来获得长足发展，不断取得重大突破，离不开党的高度重视、国家的大力支持、人民的充分认可。在人工智能治理方面，阿里巴巴也将继续肩负起排头兵的责任，以更高的责任感、使命感、荣誉感，进一步提高治理技术，完善伦理规范引导，充分协同各方力量，不断提升人工智能治理水平，为人工智能治理的中国方案贡献出更多更好的经验和智慧，助力我国在人工智能治理能力的国际竞争中持续领先。

附：人工智能助力可持续发展的丰富实践

人工智能是新一轮科技革命和产业变革的重要驱动力量，将进一步释放历次科技革命和产业变革积蓄的巨大能量，创造新的强大引擎。作为新一轮产业变革的核心力量，人工智能正逐渐重塑生产、分配、交换和消费等经济活动各环节，催生新业务、新模式和新产品，从衣食住行到医疗教育，在社会经济各个领域深度融合和落地应用。如何更好地服务人类，成为“好科技”，正是本文尝试探讨的核心问题。

在人工智能的实践应用中，应当践行人人受益、责任担当、开放共享的价值导向，助力实现可持续发展愿景。

1. 人人受益：普及数字红利，助力公共卫生

信息技术发展日新月异，给人们生活带来极大便利。特别是近年来移动互联网迅速发展，智能手机已成为人们享受数字红利最主要的渠道。但追剧、网购、手机导航等等对普通人习以为常的事，对视障人士来说却障碍重重。

《中国落实2030年可持续发展议程国别方案》提出，到2030年，要“确保机会均等，减少结果不平等现象”。帮助视障人群享受数字技术发展的红利，减少与普通人之间的不平等，首先要解决的是“看”和“读”的问题。

阿里巴巴与清华大学联合研发了人机交互新技术 Smart Touch，为视障人群更好地“玩手机”提供了新可能。具体做法为给手机贴一个低成本的硅胶薄片，薄片两边各有三个按钮，可触发“返回”“确认”等简单命令，帮助视障人士轻松完成购物、支付等操作，从而减少视障人群在数字技术使用上的障碍和不平等。

其核心原理为，通过人工智能和智能交互技术分析理解不同手机界面的语意，将供视障用户使用的传统读屏软件， 替换为基于语义和逻辑的模式， 结合 Smart Touch 交互技术，提供基于语音和触觉的多模态交互方式，解决滑动操作次数多且复杂等问题。

针对视障人群读写验证码和密码困难的问题，蚂蚁集团研发了两款无障碍创新产品——“挥一挥”空中手势验证码和“划一划”屏幕手势密码。前者主要功能是利用手势交互完成人机识别校验，替代滑块拼图、文字识别等依赖视觉能力的传统验证码产品；后者是用户提前设置一个图形手势作为密码，在需要校验的时候在屏幕上划出手势进行校验，可在登录，解锁，支付等环节替代传统密码或者刷脸等方式。

目前 2 款产品均已在支付宝 APP 上线，在支付宝登录、注册等 10+ 个场景上线应用，帮助视障人士更好地享受数字技术带来的红利。

第二次全国残疾人抽样调查结果显示，我国有听力残疾患者 2780 万人★ ；据教育部统计数据，2021 年（历届）各种形式的高等教育在学总规模 4430 万人★ 几乎所有人都能在周围中找出几名大学生，但同样量级的听障人士却如同隐身。

《中国落实 2030 年可持续发展议程国别方案》提出，到 2030 年，“所有男女，包括青年和残疾人实现充分和生产性就业，有体面工作”。对于听障人群来说，体面工作的一个很重要的前提是解决“听”的问题。

由中国残疾人信息和无障碍技术研究中心、深圳市信息无障碍研究会、浙江省盲人学校及钉钉携手发起成立“智能办公硬件无障碍联盟”，旨在探索办公环境的信息无障碍建设，让残障人士和弱势群体平等地参与社会事务、寻求平等的工作机会。

如，面向听障用户，钉钉探索利用语音转文字、人工智能实时字幕，把视频会议、直播授课中的声音，实时转换成文字字幕，让听障人士也能参与到正常的沟通、交流之中。此外，再结合阿里巴巴电商平台推出的一系列帮助残障人士解决就业的扶持政策和绿色通道，用“授人以渔”的方式对残疾人提供最大帮助，帮助听障人士体面工作。

★ 国家统计局：第二次全国残疾人抽样调查结果正式发布 [EB/OL] . http://www.stats.gov.cn/tjgz/tjdt/200612/t20061205_16908.html,2006-12-05.

★ 教育部：2021 年全国教育事业统计主要结果发布 [EB/OL].http://www.moe.gov.cn/fbh/live/2022/54251/mtbd/202203/t20220301_603465.html,2022-03-01.

中国工程院院士钟南山指出，以人工智能为代表的新一代信息技术蓬勃兴起，并迅速向医疗卫生、健康等行业渗透和融合，给各国经济发展、社会治理、人民生活都带来重大而深远的影响。

抗击新冠疫情。新冠病毒是基因组序列最长的病毒之一，临床诊断需要将患者样本与该病毒基因序列进行比对以确定诊断结果。医院普遍采用核酸检测方法，但只能检测到局部的病毒基因，无法判断新冠毒株的类型；同时病毒有很强的变异性，这种部分匹配的方法还有可能导致漏检。在核酸检测方法对新冠病毒初筛的基础上，用全基因组检测方法作复检，能够进一步确认新冠病毒的类型，以及在此基础上的进化溯源分析辅助确认新冠的流行病学情况。

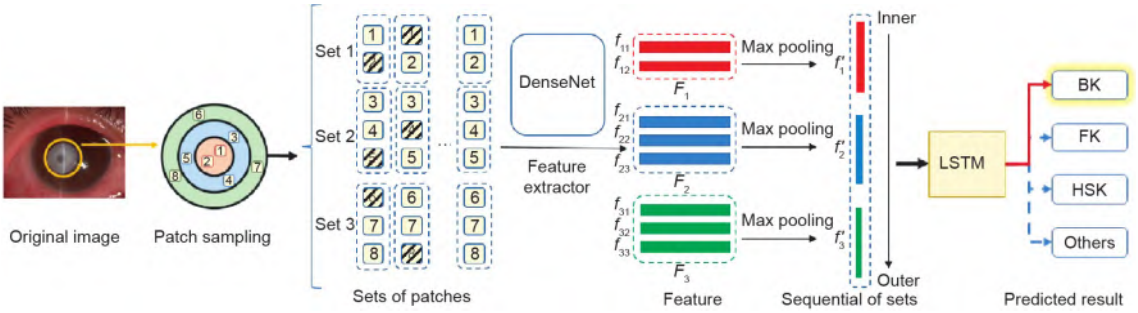
阿里巴巴达摩院采用人工智能算法助力抗击新型冠状病毒肺炎疫情。2021 年 2 月，浙江省疾控中心上线自动化的全基因组检测分析平台，平台利用达摩院研发的人工智能算法，可以高效地对病毒样本进行全基因组序列分析比对，将原来数小时的疑似病例基因分析缩短至半小时，且能精准检测病毒变异情况，大幅提高了疑似病例的缺人速度和准确率。

全基因组智能分析系统在浙江省疾控中心、武汉金银潭医院等数家医疗机构落地应用；同时系统在疾控中心、海关等近 10 家政府机构落地， 帮助研究人员快速开展新冠等病原体微生物的检测和研究工作。

协助各类角膜病的诊断。角膜病是眼科中最重要的病种之一，也是最重要的致盲性疾病。浙江大学邵逸夫医院和浙江大学计算机科学与技术学院研究团队借助人工智能算法来协助角膜病。对于其中占比最高的感染性角膜炎，传统的诊断方法需要资深医生根据患者病情来判断，但不同地区医疗水平参差不齐、不同医生诊断水平参差，角膜病的整体诊断率不高。

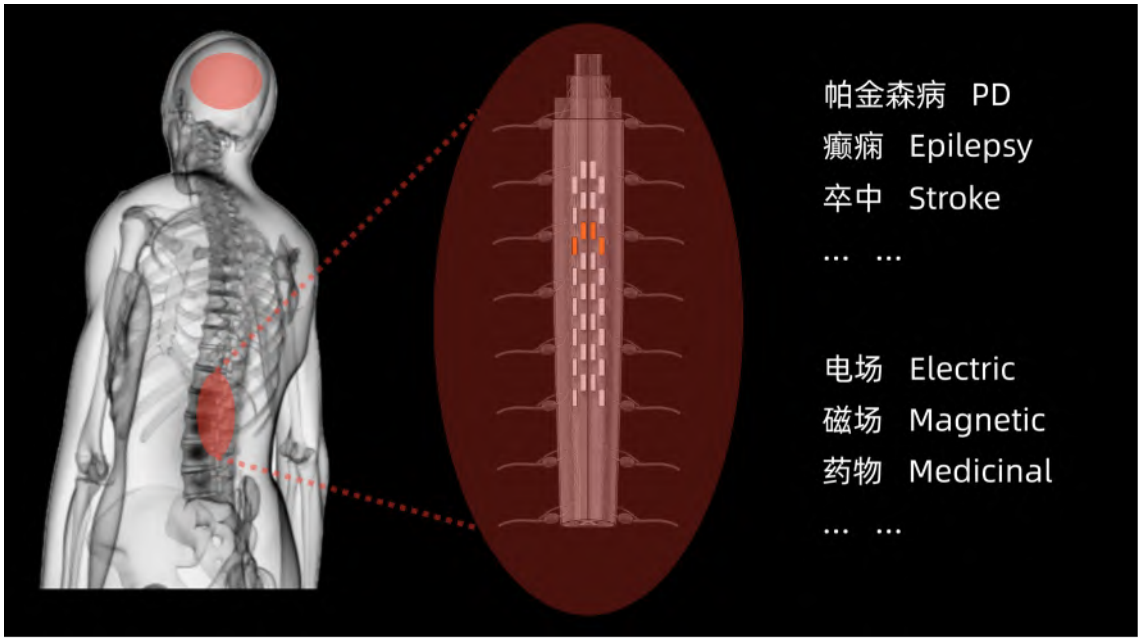
研究团队借助裂隙灯拍摄了大量角膜病图像，形成了特有的角膜病，以此为基础研发出“角膜病序列深度特征学习和识别算法”。人机对比结果显示，没有额外医疗信息的眼科医生平均诊断准确率为 49.27%, 提供了相关病史后平均诊断准确率提升至 57.16%，均远低于人工智能算法诊断准确率 84.78%。

将智能诊断算法镶嵌到眼科检查设备中，改造现有的检查设备，让专用设备智能化，将使得科技创新在更广、更深、更普及的层面惠及广大人民群众的优质医疗资源需求。目前，原型机已进入临床应用实验阶段。



在线强化学习治疗脊髓损伤。交通事故、重体力活动、体育事故是导致脊髓损伤的主要原因。严重脊髓损伤患者表现为双下肢甚至四肢瘫痪，以及呼吸循环障碍、泌尿系统感染、慢性疼痛等各种严重并发症。目前，对于脊髓损伤导致的瘫痪，植入脊神经调控设备，调控感觉 - 运动神经通路，是最先进的临床治疗方案之一。由于患者的生理和病理差异，每个人对调控疗法的承受能力不同，最优调控策略因人而异，医生的临床经验不足以支撑在庞大的策略空间里快速求解。因此，治疗过程中的关键问题是寻找有效的神经调控模式与人机交互方式。

清华大学神经调控国家工程研究中心的学者开发了高密度电极阵列和用于控制复杂多电极阵列的在线学习理论和方法，在保障患者安全的前提下，通过在线安全探索，高效优化神经调控诸多变量。相关研究成果已经成功帮助下肢瘫痪患者实现重新站立和恢复行走功能，并恢复高位截瘫患者手部抓握功能。该研究将人工智能算法应用于临床治疗一线，为瘫痪患者带来新的希望。



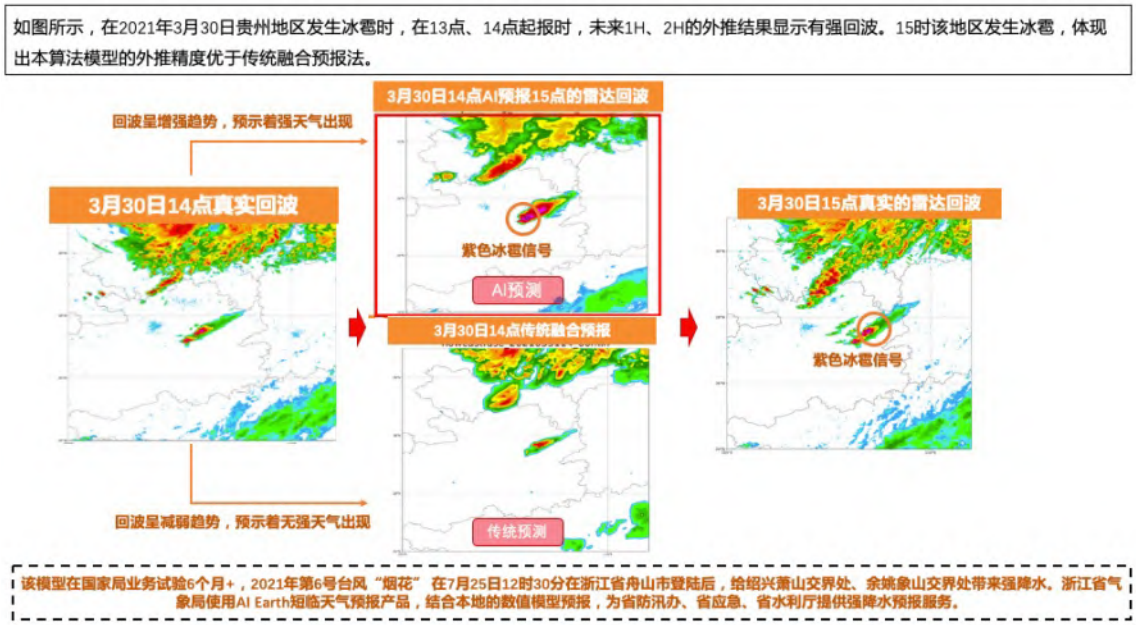
2. 责任担当：履行社会责任，推动人类福祉

过去，基于雷达数据的线性外推方法一直是业界预测短临强对流天气的主要途径，

该方法难以对强对流系统的生成、发展、减弱和消散等过程进行预测，预报精度有限、预报时效短。因此全球各地气象机构纷纷探索用人工智能来解决强对流天气预测难的问题，用人工智能的方法从海量历史数据中，提取对流系统生消演变规律，提升预测能力。

阿里巴巴达摩院与国家气象中心强天气室联合研发的雷达反射率临近预报人工智能算法在预测精准度和精细度上双双实现突破。算法采用时空分离的卷积神经网络，利用达摩院自研的方向自注意力卷积，对大气的聚合消散过程进行建模，准确高效地提取时空特征；设计了同化模块，有机结合地形数据、雷达、卫星数据等多源观测数据，可将预报时效延长至 3 小时。此外，该算法采用全新的生成对抗训练方法，显著提升预报图像的清晰度，可实现全国范围内的雷达回波实时预报。

算法实现了全国范围雷达回波的未来 0-3 小时精细化预报，将预测精度缩小到最小 1 公里范围，可辅助预报员预测临近时段内突发性的强对流天气，有效降低强天气现象造成的经济损失和社会危害，帮助提升城市治理“软实力”。



2020 年，我国提出“二氧化碳排放力争于 2030 年前达到峰值，努力争取 2060 年前实现碳中和”的双碳目标。阿里巴巴集团积极响应国家碳中和目标，围绕国家“双碳”目标制定了自身碳中和路径：将在 2030 年实现自身运营层面的碳中和，同时将协同生态上下游，实现范围 3（阿里巴巴数字生态参与者产生的温室气体排放）碳排放强度比 2020 年降低一半。

随着阿里集团业务全面上云，低成本高效能成为集团技术资源的核心目标。“双碳”战略指引下，技术资源作为高能耗服务是首要攻克的难关。而深入运用人工智能技术“配置 - 检测 - 优化”的反馈闭环，将有效实现运营优化和成本优化。

1. 资源配置

对业务预算、历史用量、应用画像以及健康分等维度进行数据分析和建模，预测业务需求量；同时结合整个容器平台的节点分配率、额度使用率等来预估整个容器平台的需求量，为资源采购提供数据支持。

2. 账单检测

针对各账号使用的费用序列进行周期性检测、平稳性检测等，分解出不同特征序列。采用自学习融合算法，对不同特征的序列分别进行建模预测，通过分类器去学习不同时序特征得到最优模型。

3. 根因分析与成本优化

对异常情况进行分层根因分析，输出分析结果，如基线值、真实值、差值等，从而有力支撑了成本优化的工作。

运用人工智能技术运营后，相较去年，机器投入财务成本下降了 50%，有效支撑了双碳目标的实现。

近年来，以电信网络诈骗为代表的新型违法犯罪活动愈演愈烈，严重危害人民群众财产安全，严重扰乱正常生活秩序，已成为影响人民群众安全感和幸福感的突出问题，对构建社会主义和谐社会带来严峻挑战。

多破案不如少发案，加强犯罪预警和防范，是减少群众被骗的有效手段。针对电话反诈骗电话辨识度不高、屡被拒接、拦截率低的问题，公安部刑侦局联合阿里巴巴推出“钱盾反诈机器人”，用人工智能手段识别电信网络诈骗，通过来电显示“公安反诈专号”，向潜在的电信网络诈骗受害人拨打电话，发送短信、闪信提醒信息，提升反诈劝阻成功率。

“机器人”通过大数据综合分析，将反诈预警数据进行智能化打标分层，对潜在受害人进行人工智能智能语音的精准宣防与劝阻，有效降低民警人工劝阻的工作量，真正实现人工智能在反电信网络诈骗领域的责任担当。钱盾反诈机器人目前具备 12 种典型诈骗场景劝阻能力，同时不断研发提升拟人交互能力、语义分析能力和预警分层能力，截至 2021 年 12 月已累计预警 353 万余次，劝阻成功率超 96%，

劝阻金额超 10 亿元。

3. 开放共享：开放底层框架，保障网络健康

过去 60 多年的发展中，人工智能研究领域不断扩大，扩展到机器学习、自然语言处理、语音识别、影像分析与理解、智能搜索、知识推理等诸多领域。然而，现有人工智能算法存在专用性强、场景不通用等问题，不同场景算法对应不同需求，而场景具有无限性，导致算法需求量巨大。

若面向无限场景的算法均从头设计、开发，既要求开发者具备很高的技术能力，又将会产生大量从 0-1 的重复劳动。在这种情况下，阿里巴巴与 Cape Privacy、OpenMined 联合建立了基于 TensorFlow 之上的 TF-Encrypted 开源安全多方计算框架（简称 TFE），向开发者开放。TFE 的架构设计自下而上分为三层：TensorFlow 原语、多方计算协议、机器学习模型，用户无需耗费精力去接触底层的安全协议，只需专注于从 1-N 的模型构建。

TFE 具有三大主要优势。一是友好性，其 API 与 TensorFlow 保持一致，熟悉

TensorFlow 机器学习建模的开发者均可快速迭代出一个底层融合了安全多方计算的模型版本；二是可扩展性，可以方便地切换支持新的多方计算协议，新的机器学习算法层亦可在现有运算上构建，而不需要接触底层密码学原语；三是性能高，已支持的多方协议均采用当前最优的算法实现，并充分利用了 TensorFlow 后端提供的分布式计算相关优化。

随着互联网的飞速发展，互联网应用逐渐展现出用户群体庞大、用户数据海量、信息传播迅速、影响范围广等特点，也随之产生了大量有害内容，如抹黑英雄或国家形象、宣传暴恐思想、低俗或垃圾广告等信息，对网络空间的安全与秩序产生了极大危害。为推动互联网平台的治理与发展，需要及时对网络平台中的风险内容进行研判与滤除。然而风险内容的管控往往存在数据量大、对抗性强、风险场景复杂等特点，需要体系化、规模化以更加智能的技术手段进行治理。

在此背景下，阿里巴巴集团基于多年的安全技术积累，依托淘宝、阿里云等平台的管控经验，构建了核心安全智能算法服务产品——绿网。绿网深耕自然语言理解、图像识别、OCR、语音识别等业界前沿技术，为企业用户提供成熟的、轻量化接入的内容安全解决方案。帮助企业、开发者在复杂多变的互联网环境下快速发现文本、图片、视频、语音中的各类风险，保障应用的信息内容安全。面对复杂的客户生态，绿网针对不同的用户提供层次化的服务能力：1）算法运营层，能够为算法团队提供有效的算法生命周期管理工具。2）业务运营层，以更好的产品化工具支撑业务安全运营快速搭建防控策略。3）提供标准化的能力和场景化方案，支持在线检测、私有化独立部署等服务模式，帮助客户实现快速管控。

目前绿网的日均算法调用量已达百亿次，帮助各个行业的用户显著降低了各种违规风险内容，携手构建了更加风清气正的互联网内容生态。

参考文献

[1] 国务院 . 关于印发新一代人工智能发展规划的通知（国发 [2017]35 号）[Z].2017-07-20.

[2] 国务院 . 中国落实 2030 年可持续发展议程国别方案 [Z].2016-09-19.

[3] GB/T 37988-2019, 信息安全技术 数据安全能力成熟度模型 [S].

[4] 全国人大常务委员会 . 中华人民共和国个人信息保护法 [Z].2021-8-20.

[5] 国家互联网信息办公室 . 儿童个人信息网络保护规定 [Z].2019-08-23.

[6] 国家互联网信息办公室 . 互联网信息服务算法推荐管理规定 [Z].2022-01-04.

[7] 国家互联网信息办公室 , 工业和信息化部，公安部，市场监督管理总局 . 关于加强互联网信息服务算法综合治理的指导意见 [Z].2021-09-29.

[8] 公安部网络安全保卫局，北京网络行业协会，公安部第三研究所 . 互联网个人信息安全保护指南 [Z].2019-04-28.

[9] 国家市场监督管理总局 . 互联网平台落实主体责任指南（征求意见稿）[Z].2021-10-29.

[10] 国家互联网信息办公室 . 互联网信息服务深度合成管理规定（征求意见稿）[Z].2022-01-28.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei.

Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition[C], pages 248-255. Ieee, 2009.

[12] Ilyas, Andrew, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. "Adversarial examples are not bugs, they are features." Advances in neural information processing systems 32 (2019)[C].

[13] Gartner.2022 年重要战略技术趋势 [EB/OL].https://www.gartner.com/en/information-technology/insights/top-technology-trends，2021-10-19.

[14] 国家互联网信息办公室、文化和旅游部、国家广播电视总局 . 网络音视频信息服务管理规定 [Z].2019-11-29.

[15] Mehrabi N, Morstatter F, Saxena N, et al. A survey on bias and fairness in machine learning[J]. ACM Computing Surveys (CSUR), 2021, 54(6): 1-35.

[16] Kusner M J, Loftus J, Russell C, et al. Counterfactual fairness[J]. Advances in neural information processing systems, 2017, 30.

[17] Vaughan J, Sudjianto A, Brahimi E, et al. Explainable neural networks based on additive index models[J]. arXiv preprint arXiv:1806.01933, 2018.

[18] Pearl J. Causal inference in statistics: An overview[J]. Statistics surveys, 2009, 3: 96-146.

[19] Rubin D B. Causal inference using potential outcomes: Design, modeling, decisions[J]. Journal of the American Statistical Association, 2005, 100(469): 322-331.

版次　2022 年 7 月第 1 版

印次　2022 年 7 月第 1 次印刷

ARTIFICIAL INTELLIGENCE GOVERNANCE & SUSTAINABLE DEVELOPMENT

可靠信用