中国互联网协会
Internet Society of China

AI技术赋能网络内容安全保障研究报告

中国互联网协会

2018年8月

前　言

人工智能（Artificial Intelligence，简称 AI）近年得到广泛关注，且对于未来网络安全变得愈发重要，机器学习、深度学习等人工智能技术在网络安全领域的应用正在引发新技术研发热潮和新安全产业增长。本研究的主要目的是分析 AI 赋能网络内容安全保障的关键技术及解决方案，通过 AI 技术在网络内容安全的应用场景及发展环境，分析 AI 产生的企业价值和社会价值，研究 AI 技术保障网络内容安全的演进趋势。

在海量数据时代，伴随不断增长的物联网设备，越来越多的数据通过互联网进行传送，使得网络安全形势严峻。而人工智能的高速发展，AI 技术通过机器学习，极大提高了识别安全威胁的准确度及移动安全机制的应急反应速度，促进经济发展的同时有效维护了网络安全。“人工智能+网络安全”成为重要热点技术。机器学习应用于网络安全领域已成为重要趋势。

本研究从 AI 赋能网络内容安全保障的发展概况、关键技术、解决方案、企业价值、社会价值、发展趋势、发展建议等七个部分切入，进行了深入分析论述。第一部分对网络内容安全保障这一定义做出界定，并简要介绍了 AI 保障网络内容安全的政策环境、社会环境和市场格局。第二部分从文本内容检测、视频/图片内容检测、语音内容检测三个方面，剖析与网络内容安全保障所关联的 AI 关键技术。第三部分从 AI 技术赋能网络内容安全解决方案架构为基础，对新零售、传媒、泛娱乐、互联网金融、在线教育、政府/物业安防、信息通信

及其他生活场景等八类应用场景的解决方案进行阐述。第四部分，从建设阶段和维护阶段两个层面分别介绍 AI 技术为企业在网络内容安全保障领域所带来的帮助与价值。第五部分，对产生的正面社会价值进行宏观阐述，重点包括维护国家安全和社会稳定、净化网络空间、保护知识产权、降低维护真实信息的社会总成本、提升互联网内容质量、舆情管控应对、维护公民权益等七个方面。第六部分，详细探讨 AI 保障网络内容安全的发展趋势，从对抗网络的应用、AI 技术仿生学、内容审核中的作用等方面进行趋势预测。第七部分，提出了要营造产业发展良好环境、推动 AI 技术发展、鼓励政府和企业采取人工智能技术进行内容审核，加强数据保护体系建设出发，探索协同治理模式等发展建议。

本研究主要服务于涉及网络内容安全建设与保障的企业界人士、AI 技术领域有关专家及爱好者。2017 年国务院印发《新一代人工智能发展规划》明确指出，人工智能是引领未来的战略性技术，2018 年《政府工作报告》中多次提及人工智能等关键词，人工智能已上升为国家战略。本研究主要侧重于 AI 技术在保障网络内容安全的发展、应用及优势，结合国家“加强新一代人工智能研发应用，发展智能产业，拓展智能生活”的战略政策指导方向，对 AI 技术的网络内容安全应用环境进行宏观分析。研究建议在数字经济的大背景下，我国国家安全和国际竞争形势更加复杂，必须放眼全球，把人工智能发展放在国家战略层面系统布局、主动谋划，牢牢把握人工智能发展新阶段国际竞争的战略主动，打造竞争新优势、开拓发展新空间。同时，切

实做好 AI 技术对网络内容安全的保障工作，建立政府安全监管、市场安全服务、企业主体安全的协同联动机制。“没有网络安全就没有国家安全”，在商业应用迅速拓展的同时，构建良好网络生态环境，人工智能一定要成为手中利器，加强前瞻预防与约束引导，最大限度降低风险，确保人工智能安全、可靠、可控发展，有效保障国家安全。

报告课题组成员：王朔、徐杰、谷勇浩、李勇、吴志鹏、王一飞、陈哲、向坤、尹艳鹏、张威、薛晖、张荣、冯雪涛、韩婷钰。

目 录

一、发展概况

1.术语定义

网络内容安全保障是指通过技术手段对网络内容进行分级、分类、过滤等，依法治理煽动颠覆国家政权、煽动民族宗教仇恨和恐怖主义，以及色情暴力血腥等违法违规内容，维护政治意识形态安全，规范信息传播秩序、保障积极健康的网络生态环境的一系列行为。

2.政策环境

2017 年是中国人工智能发展的关键之年，为抢抓人工智能发展的重大战略机遇，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，国务院于 2017 年 7 月印发《新一代人工智能发展规划》（以下简称《规划》）。相比其它国家的人工智能战略，《规划》包含了研发、工业化、人才发展、教育和职业培训、标准制定和法规、道德规范与安全等各个方面的战略和发展目标，具有鲜明的系统性和全面性特点。

在《规划》指导意见中提出了三步走战略目标。即，第一步，到 2020 年人工智能总体技术和应用与世界先进水平同步，人工智能产业成为新的重要经济增长点，人工智能技术应用成为改善民生的新途径，有力支撑进入创新型国家行列和实现全面建成小康社会的奋斗目标；第二步，到 2025 年人工智能基础理论实现重大突破，部分技术与应用达到世界领先水平，人工智能成为带动我国产业升级和经济转型的主要动力，智能社会建设取得积极进展；第三步，到 2030 年人工智能理论、技术与应用总体达到世界领先水平，成为世界主要人工

智能创新中心，智能经济、智能社会取得明显成效，为跻身创新型国家前列和经济强国奠定重要基础。中国在 2030 年的目标是人工智能产值达到 1 万亿人民币，而相关行业的总产值达到 10 万亿人民币。这一计划还明确了政府将会鼓励招揽全球最优秀的人才，加强对国内 AI 劳动力的培训，并在促进人工智能发展的法律、法规和道德规范方面引领世界。这其中包含了积极寻求全球 AI 领导者的意图。

在《规划》发布之后，工信部又于 2017 年 12 月发布了《促进新一代人工智能产业发展三年行动计划（2018-2020 年）》。该计划可谓对《规划》第一步战略的落实，即希望推动中国的 AI 产业于 2020 年达到世界一流水平。具体而言，主要包括四个方面：（1）培育智能产品。着重在智能网联汽车、视频图像身份识别系统、智能语音交互系统等八大领域率先取得突破；（2）突破核心基础。着重在智能传感器、神经网络专用芯片和开源开放平台等三大领域率先取得突破；（3）深化发展智能制造。着重在智能制造关键技术装备和智能制造新模式等领域率先取得突破；（4）构建支撑体系。着重在行业训练资源库、标准测试及知识产权服务平台、智能化网络基础设施和网络安全保障体系等四大领域率先取得突破。

此后各个省市相继出台了人工智能的相关政策文件。通过对 15 个省、直辖市 2017 年发布的人工智能相关政策文件梳理发现，明确出台人工智能相关规划的省市有：北京市、上海市、重庆市、浙江省、安徽省、江西省、福建省、湖北省。在四个直辖市中，北京市以机器人产业为主攻方向，出台了人工智能产业指导意见和创新路线图；上

海则出台了人工智能专项支持实施细则；重庆提出以智能化为引领的创新驱动发展战略三年行动计划；天津主要在智能制造方面规划布局。在东部省份，浙江、福建明确出台了人工智能产业发展的行动计划，山东出台了智能制造发展规划。西部省份，贵州提出了智能贵州发展规划。中部省份，湖北、江西从促进人工智能发展政策方面出台了相关文件。

在网络内容安全方面，国家互联网信息办公室在 2017 年印发《互联网用户公众账号信息服务管理规定》提出，互联网用户公众账号服务提供者应落实信息内容安全管理主体责任，加强对本平台公众账号发布内容的监测管理，发现有传播违法违规信息的，应立即采取相应处置措施等。规定明确提出，互联网群组信息服务提供者应当落实信息内容安全管理主体责任，配备与服务规模相适应的专业人员和技术能力，建立健全用户注册、信息审核、应急处置、安全防护等管理制度。

国家互联网信息办公室 2017 年 5 月 2 日公布《互联网新闻信息服务管理规定》（以下简称《规定》），自 2017 年 6 月 1 日起施行。国家互联网信息办公室有关负责人表示，出台《规定》旨在进一步加强网络空间法治建设，促进互联网新闻信息服务健康有序发展。

《规定》共六章，二十九条。第一章是总则，对立法目的、原则、适用范围、监管主体作出规定。第二章是许可，对从事互联网新闻信息服务许可的条件、材料、受理、决定作出规定。第三章是运行，对互联网新闻信息服务提供者的日常运行制度作出规范。第四章是监督

检查，对国家互联网信息办公室及地方互联网信息办公室监督执法作出规定。第五章是法律责任，对违反《规定》的行为的法律责任作出规定。第六章是附则，对有关术语的定义和公布实施作出规定。

3. 社会环境

移动互联网、云计算、物联网等新兴技术促使互联网环境更加复杂，互联网上的数据呈现爆炸式增长。每天通过互联网上传的视频、图片、文字数据超过 15 亿条，且数据量还在呈指数级增长趋势。海量大数据的积累，极大地丰富了人们的精神和物质生活，也为经济和社会发展提供了一种新的思路和解决方案。互联网行业自身取得的成绩无需多言，许多依靠传统手段管理运营的行业，在得到互联网赋能后，发展质量和效益都得到了有效提升，焕发出新的活力。而人工智能依托数据的大量提升，也迎来发展的黄金时代。

但网络不是一块无主之地，需要加强引导。尤其是在当下社会价值多元化，信息传播速度越来越快，传播范围越来越广的大背景下，各类“灰犀牛”事件发生的概率也在愈发增加。数字鸿沟、不良信息泛滥、数据隐私侵犯、涉黄涉爆等一系列社会风险和隐患，比以往任何时候都要突出。不仅如此，社会各个领域广泛存在各种问题，如就业、教育、医疗、金融、物流等领域的矛盾和问题也交织叠加。这一切，不仅是对国家治理体系和治理能力的巨大挑战，也意味着解决上述问题所带来的巨大社会和经济潜在效益。

4. 市场格局

机器学习、深度学习等人工智能技术在网络安全领域的应用正在

引发全球新技术研发热潮和新安全产业增长。一是人工智能在网络安全领域应用的学术研究如火如荼开展。二是一批致力于“人工智能+网络安全”的企业发展势头良好。三是传统大型 IT 企业向“人工智能+网络安全”战略转向明显。四是“人工智能+网络安全”逐步上升到国家网络安全层面。

阿里、腾讯、百度等互联网企业积累了海量数据，构建“AI+网络安全”生态体系，AI 技术通过机器学习，提高识别安全威胁的准确度及移动安全机制的应急反应速度，促进经济发展的同时有效维护网络安全。其中谷歌利用机器学习技术对安卓系统上运行的移动终端威胁进行分析，在手机系统识别和移除恶意软件。腾讯通过“安全态势感知系统”等五大安全应用系统，基于大数据和深度学习技术，形成基于 AI 的“事前-事中-事后”全链条反诈体系。百度利用多项 AI 技术如自然语言处理、深度学习技术、图像识别技术等搭建内容风控一体化服务解决方案，对恶意网址采取风险标注、搜索降权、广告下线等多种拦截措施。

2017 年互联网企业在网络安全反诈骗领域动作频频，阿里巴巴与广东移动、腾讯与上海公安合作，通过数据、人才、技术等方面的互通，共同推进“AI+网络安全”生态体系的建设。阿里巴巴云盾内容安全是内容安全领域的先行者，依托阿里云、淘宝、支付宝等平台的管控经验，将 AI 技术应用于文本、图片、视频中的敏感信息识别，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各类风险，保障应用的信

息内容安全。为此，本报告以“阿里巴巴云盾内容安全”解决方案为例，进行典型案例剖析，针对由各类网络不良与违法信息传播所引起的社会现象与问题进行论述，对推动网络内容安全保障的相关 AI 技术进行科普，阐述 AI 技术对内容安全保障所带来的积极作用与价值。

二、关键技术

网络内容具有即时性、海量性和多态性等特点。网络内容安全管理面临审核标准差异化、动态化，对抗行为较为突出的特点。传统的基于人工审核、人工特征工程的网络内容分析方法面临极大的挑战。随着人工智能第三次浪潮的兴起，新的 AI 技术如雨后春笋般涌现，逐渐以一种截然不同的方式，应用于网络内容安全领域。

人工智能是计算机科学的一个分支，它企图了解智能的实质，并生产出一种新的能以人类智能相似的方式做出反应的智能机器，该领域的研究包括机器人、语言识别、图像识别、自然语言处理和专家系统等。深度学习是近期人工智能中发展迅速的领域之一，可帮助计算机理解大量图像、声音和文本形式的数据。利用多层次的神经网络，使得计算机能像人类一样观察、学习复杂的情况，并做出相应的反应，有时甚至比人类做得还好。

深度学习的概念源于人工神经网络的研究。基于深度置信网络(DBN)提出非监督贪心逐层训练算法，为解决深层结构相关的优化难题带来希望，随后提出多层自动编码器深层结构。此外 Lecun 等人提出的卷积神经网络是第一个真正多层结构学习算法，它利用空间相对关系减少参数数目以提高训练性能。含多隐层的多层感知器就是一种

深度学习结构。深度学习通过组合低层特征形成更加抽象的高层表示属性类别或特征，以发现数据的分布式特征表示。深度学习是机器学习中一种基于对数据进行表征学习的方法。观测值（例如一幅图像）可以使用多种方式来表示，如每个像素强度值的向量，或者更抽象地表示成一系列边、特定形状的区域等。而使用某些特定的表示方法更容易从实例中学习任务（例如人脸识别或面部表情识别）。深度学习的好处是用非监督式或半监督式的特征学习和分层特征提取高效算法来替代手工获取特征。深度学习是机器学习研究中的一个新的领域，旨在研究如何从数据中自动提取多层特征表示，其核心思想是通过数据驱动的方式，采用一系列的非线性变换，从原始数据中提取由低层到高层、由具体到抽象、由一般到特定语义的特征。深度学习在自然语言处理、图像识别、语音识别等领域展现出了巨大的优势，并且仍在继续发展变化。

1.文本内容检测

（1）自然语言处理

自然语言处理（Natural Language Processing,NLP）是深度学习的一个重要应用领域，经过几十年多的发展，基于统计的模型已经成为NLP的主流，同时人工神经网络在NLP领域也受到了理论界的足够重视。世界上最早的深度学习用于NLP 的研究工作诞生于NEC Labs American，其研究员Collobert 和Weston从2008年开始采用embedding 和多层一维卷积的结构，用于词性标注、分块、命名实体识别、语义角色标注等4 个典型NLP 问题。值得注意的是，他们将

同一个模型用于不同的任务，都取得了与现有技术水平相当的准确率。在一些常见的自然语言处理任务中，基于深度学习的方法已经取得了最佳的结果，诸如命名文本分类(Text Categorization，TC)、实体识别（Named Entity Recognition, NER)、词性标注（Part of Speech Tagging）和情感分析（Sentiment Analysis）等。

（2）文本分类技术

文本分类技术是自然语言处理的重要基础技术，是文本内容检测的基础工作，其作用是根据文本的某些特征，在预先给定的类别标记(Label)集合下，根据文本内容判定它的类别。文本分类的应用非常广泛，如垃圾邮件分类、主题分类、文本数据的意图、情感和情绪分析等。传统的文本分类模式基于知识工程和专家系统，在灵活性和分类效果上都有很大的缺陷。20 世纪 90 年代以来，机器学习的分类算法有了日新月异的发展，很多分类器模型逐步被应用到文本分类之中，比如支持向量机、最近邻法、决策树、朴素贝叶斯等。基于机器学习的文本分类方法首先对文本进行预处理，将文本用模型表示，进行特征提取和特征降维，然后构造并训练分类器，最后利用分类器对新文本进行分类。该方法主要问题在于文本特征是高纬度高稀疏的，表达能力较弱，神经网络很不擅长对此类数据的处理，而且人工进行特征构造和特征提取难度很大。应用深度学习解决大规模文本分类问题最重要的是解决了文本表示的问题，去掉繁杂的人工特征提取过程，端到端的解决问题。

2.视频/图片内容检测

“一图胜千言”，视频/图片作为一种重要的视觉信息载体，具有形象直观、内容丰富等特点，是网络内容最重要的组成部分。随着互联网技术的迅速发展和大规模存储器的普遍使用，以视频/图片为主的多媒体信息正在呈爆炸式增长态势。这对人类社会的影响具有两面性：一方面视频/图片内容成了人们获取信息的重要来源，使生活更加丰富多彩；另一方面色情、暴力、反动视频/图片严重毒害青少年的身心健康，影响社会的稳定团结。

（1）网络视频/图片文字识别

网络视频/图片的文字识别大大难于传统扫描文档中的文字识别，因为它们具有极大的多样性和明显的不确定性，诸如多语言文字、不同的文字大小、不同的字体、多样的文本与背景颜色、多变的光照与亮度、不一致的对比度与分辨率、多方向与形变文本、复杂的背景等。所以，传统的应用于扫描书刊报纸等文档图像的 OCR 技术在网络视频/图片文本识别中具有巨大的局限性。近十年来，国际国内模式识别、文档分析与识别、计算机视觉等领域的众多科研机构（如斯坦福大学、牛津大学、中国科学院自动化研究所、清华大学、北京科技大学等）和大量 IT 工业界巨头（阿里巴巴、腾讯、百度、Google、Microsoft、Amazon 等）都对复杂网络视频/图片文字识别技术进行研究与攻关。

网络视频/图片文本识别技术主要分为两个阶段：首先是对图片中的文字进行检测与提取，输入的是原始图片而输出的是文本区域，即文本检测；然后，对检测出的文本区域进行识别，输入的是文本区域而输出的是结果文字，即文字识别。如果一个系统，输入的是原始

图片而输出直接为最终识别的结果文字，则称之为端到端识别（End-To-End Recognition）技术。当前，绝大部分研究者要么关注文本检测，要么关注文字识别，没有形成良好的端到端识别技术，没有很好的利用识别与检测之间丰富的共享信息和反馈信息。如何对网络视频/图片进行有效的文本检测、识别（特别是端到端识别），一直都是学术界和工业界共同关注的重点与难点。

（2）人脸检测

所谓人脸检测，就是给定任意一张图片，找到其中是否存在一个或多个人脸，并返回图片中每个人脸的位置和范围。人脸检测是一种非接触性技术，具有可视化、符合人的思维习惯的特点，得以在商业、安全等领域广泛应用。人脸检测的研究在过去二十年里取得了巨大进步，特别是 Viola 和 Jones 提出了开创性算法，他们通过 Haar-Like 特征和 AdaBoost 去训练级联分类器获得实时效果很好的人脸检测器，然而研究指出当人脸在非约束环境下，该算法检测效果极差。这里说的非约束环境是对比于约束情况下人脸数单一、背景简单、直立正脸等相对理想的条件而言的，随着人脸识别、人脸跟踪等的大规模应用，人脸检测面临的要求越来越高：人脸尺度多变、数量冗大、姿势多样包括俯拍人脸、戴帽子口罩、等的遮挡、表情夸张、化妆伪装、光照条件恶劣、分辨率低甚至连肉眼都较难区分等。人脸检测算法以往被分为基于知识的、基于特征的、基于模板匹配的、基于外观的四类方法。随着近些年 DPM 算法（可变部件模型）和深度学习 CNN（卷积神经网络）的广泛运用，人脸检测所有算法可以总分为两类：1）Based

on rigid templates：代表有 boosting+features 和 CNN；2）Based on parts model：主要是 DPM。

基于深度学习的人脸检测方法可以作为第一类方法的代表。往往一个简单的卷积神经网络在人脸检测就能获得很好效果，同时有文献验证了深度卷积神经网络的第一层特征和 SIFT 类型特征极其相似。DPM 算法是由 Felzenszwalb 于 2008 年提出的一种基于部件的检测方法，对目标的形变具有很强的鲁棒性，目前已成为分类、分割、动态估计等算法的核心组成部分。应用 DPM 的算法采用了改进后的 Hog 特征、SVM 分类器和滑动窗口检测思想，在非约束人脸检测中取得极好效果。

随着 DCNN（深度卷积神经网络）的发展，基于深度学习的方法获得了的长足进步，可见未来人脸检测算法主要的发展将围绕 DPM 和 DCNN 展开。同时将 DPM 和 DCNN 结合的方法也将是研究趋势。虽然深度比传统算法在识别率方面有显著的提升，但是需要更多的计算资源。同时，深度网络由于参数众多，网络结构多样，在未经调优的情况下得到的结果往往还不如传统方法。此外，深度网络还需要大量的样本做训练以避免过拟合。要改进这一局面，不仅需要学术界的不断加深理论层面的支持，更需要工业界在不同的场景下的使用经验来丰富深度学习的研究素材，使得深度学习的潜力得到更全面的释放。

（3）特定标识检测

视频/图片中的标识无处不在，且包含着重要的语义信息，比如电视台标、节目标识、招牌、车牌、交通标识、枪支、刀具、旗帜等

等。标识检测对视频/图片内容理解具有重要的意义。快速准确地检测出视频/图片中的标识对基于内容的视频检索和过滤都具有重要的作用。例如，通过视频台标检测能够获得电视台名、节目取向等信息；通过节目标识或栏目标识检测能够获得视频所描述的节目内容；通过视频/图片中内嵌的各种广告牌、车牌等标识检测能够获得更加丰富的语义信息；通过检测枪支、刀具、蒙面人和恐怖组织旗帜能够识别暴恐视频/图片。目前在学术界中，标识识别性能受到复杂背景、光照变化、姿态变化、透明镂空等众多因素的影响，仍然是一个具有挑战性的研究课题。

常见的标识检测和识别方法有基于颜色直方图、基于形状、基于机器学习、基于局部不变特征等。目前，基于深度学习的标识识别方法取得了较大的进展。从 R-CNN、Fast R-CNN、Faster R-CNN 一直发展到目前的 Mask R-CNN，使得标识/对象检测性能和效率有了非常显著的提升，实现了端到端的、像素级的标识/对象检测。

（4）有害视频场景检测

视觉场景特征提取与分类识别是计算机视觉领域的研究热点之一。现有研究主要通过层次化场景结构描述、关联文本描述词性分析、非线性静止子空间分析等方法，从自然场景、室内场景等互联网视频中提取不变特征，并将其划分至预定义类别。然而，暴力、血腥、爆炸等有害视频在光照变化、相机运动、内容复杂度、场景分辨率等方面比一般的互联网视频更复杂，其场景往往具有更低的类内相似性和更高的类间多样性。传统面向自然、室内等场景的分类识别技术往往

难以在速度和精度等方面充分满足有害视频场景分类识别的需求。因此，迫切需要研究面向有害视频场景的特征提取与分类识别技术。

在海量的互联网视频中，有害视频的占比极小，但场景复杂程度极高。为了高效、精准地处理海量互联网视频并从中识别出有害视频，识别速度、识别准确率和识别召回率均需要达到极高水平，这对面向有害视频的场景特征提取与分类识别技术提出了极高挑战。随着大数据时代的来临和深度学习的发展,使用深度学习方法解决场景识别问题已经成为场景识别领域未来的发展方向。

3.语音内容检测

随着人工智能的发展，人与计算机之间的自由交互也变得越来越重要，语音识别则是其中的重要一环。语音识别的终极目标，是计算机真正能够理解人类语言甚至是方言。语音识别是语音内容检测以及语音意图理解等工作的基础。尽管近 50 年来语音识别一直属于热门研究领域，然而构建能够理解人类语言的机器仍旧是人工智能领域最具挑战性的问题之一，要实现这一目标非常困难。

从 2009 年深度学习被引入语音识别领域，短短几年时间内，其在 TIMIT 数据集上基于传统的混合高斯模型（GMM）的错误率就从 21.7%下降到 17.9%，引起业界广泛关注。Google 在应用深度学习后将语音识别模型的错误率降低了 20%，改进幅度超过了过去多年的总和。深度学习在语音识别领域取得的成绩是突破性的，几乎所有的关于语音的研究都已转向深度学习。之所以能有这样的技术突破，是因为深度学习可以自动的从海量数据中提取复杂而且有效的特征，不需

要人工提取，提升了模型准确度。

尽管语音识别的表现和应用出现了巨大的飞跃，目前还面临一些重大挑战：1）噪音的敏感性问题。一个语音识别系统在非常接近麦克风而且不嘈杂的环境中运行得很好。然而，如果说话的声音比较远或者环境很嘈杂能迅速降低系统的效能。2）语言可扩展。世界上大约有 7000 种语言，绝大多数语音识别系统能够支持的语言数量大约是几十种，给系统扩展带来了巨大的挑战。3）硬件资源消耗问题。深度学习与语音识别相结合，因此对 CPU 和内存的占用量不容小觑。研发面向深度学习的专用语音识别芯片势在必行。

三、解决方案

1.解决方案架构

阿里巴巴云盾内容安全基于深度学习技术及阿里巴巴多年的海量数据支撑，提供图片、视频、文字等多媒体的内容风险智能识别服务，能有效降低色情、暴恐、涉政等违规行为，保证业务健康发展，为互联网内容的健康性、有效性、合规性提供技术支持。伴随人工智能在技术上不断突破，与各个行业进行了深度融合,形成如图 3.1 所示的安全产品架构。

图 3.1 云盾内容安全整体产品架构

云盾内容安全对于直播、视频、图片、文本、语音等应用场景中可能出现的多重问题，实现多风险统一检测模式，一次性发现所有可能风险，优化审核模式，降低成本。

接入方式采取在线 API 调用，本地化部署及 OSS\CDN\视频云 SAAS 化开通三种接入方式。其中内容检测 API 提供文本、图片、视频等多媒体内容安全检测的接口服务；OSS 违规检测提供便捷易用的结果展示平台。支持每日增量扫描和存量扫描，无死角覆盖风险。CDN 违规检测可导出违规图片源站地址，批量删除。站点检测内容安全针对信息内容安全检测及管控服务，提前预警，提供违规网页地址及快照查看功能；私有化部署提供专属的内容安全解决方案。

云盾内容安全提出了 CNN、RNN 和 Attention 与 Self-Attention 三大算法架构，基本上覆盖了现有的 AI 算法和运用。

云盾内容安全的体系结构由三网二图所构成。其中三网包含感知网，负责识别文本、图片、视频中的内容风险；商品网，负责多任务

多模态侦测，识别商品中的风险；行为网，负责识别账号内全链路行为的风险。而二图是指商品图，即商品 KNN 图；关系图，账号、商品、行为等联动关系。

云盾内容安全服务于阿里巴巴经济体内外横跨多个行业领域的业务，包括电商、新闻、社交、直播、金融、娱乐和搜索等。先后发布五大功能应用：包括敏感人脸搜索、相似图片搜索、视频鉴黄服务、OCR 证件识别及声纹检测。

云盾内容安全敏感人脸搜索针对教育、学校以及部分工厂企业，公安对于违法犯罪份子以及敏感人物进行 1：N 识别服务，客户可增删减建立和管理自己的人脸库，对进入监控环境的人进行敏感和危险人物核对使用。

云盾内容安全相似图片搜索为图片版权保护提供侵权判定检测服务，用户可自建图片库，通过客户自有舆情和关注渠道图片比对判定是否侵权。

云盾内容安全国际视频鉴黄服务为国内直播短视频出海用户提供海外视频违法检测能力，为国际用户提供海外视频鉴黄服务，解决在所在地的内容合规问题

云盾内容安全-OCR 证件识别通过 OCR 图文结构化识别来辨别证件真伪，为需要证件核验使用，通过机器方式判定而非人工，提升识别率，降低人力成本。

云盾内容安全-声纹 1:1 对比，对直播、视频内违法人员及敏感人物等信息进行识别。

阿里巴巴云盾内容安全整体技术架构提供从前端智能交互到后端数据处理的完整闭环，通过建设业务中台和数据中台，实现共享服务体系和统一的数据集散平台，提供弹性、平滑、稳定、安全的云计算基础设施，针对新零售、传媒、泛娱乐、教育、金融、政府/物业安防，运营商，其他生活服务等八个应用场景提供完整的行业解决方案，并对 UGC 智能审核、内容版权保护、智能安防以及内容管理提出通用解决方案，如图 3.2 所示。

图 3.2　云盾内容安全整体解决方案

2. 新零售

新零售是企业以互联网为依托，通过运用大数据、人工智能等先进技术手段，对商品的生产、流通与销售过程进行升级改造，进而重塑业态结构与生态圈，并对线上服务、线下体验以及现代物流进行深度融合的零售新模式。

新零售线上线下一体化中的重要环节，重点在于打通线下门店与线上平台间的数据连接。基于物联网技术，通过店内智能硬件建立与

消费者之间的触点，增强用户体验，实现门店数字化、消费需求场景化。图 3.3 所示为新零售对 AI 的需求。

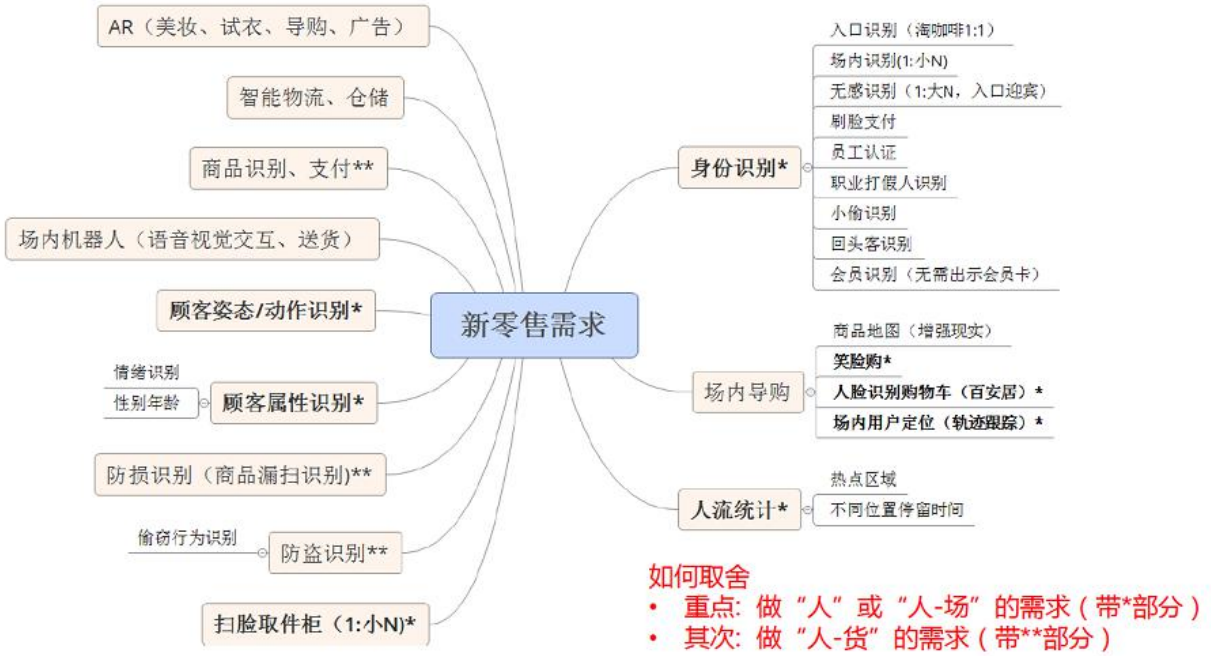

图 3.3　新零售 AI 需求关系

针对行业特点可以通过解决方案对线上业务内容，包括会员、商品、商铺、互动、推广等文本、图片、视频、音频的全方位风险过滤。同时对线下门店/商场提供线下监控视频的人脸识别、物体识别等安防保障。具体应用如下：

（1）会员：通过人脸识别（1:N）技术，识别出会员、熟客、黑名单的信息，配合数据营销系统进行销售分析。

（2）支付：通过人脸识别（1:1）技术，安全快速高效地进行支付，减少结账等待时间，提升用户体验。

（3）客流：使用人脸 / 人体识别与跟踪技术，做人流量分析，轨迹跟踪判断用户的个人喜好。收银排队监控，可用于员工作息排班。

（4）防盗损：使用人体的检测与跟踪、动作识别等技术，检测盗窃、偷吃等异常行为。

3.传媒

传媒产业是指传播各类信息、知识的传媒实体部分所构成的产业群，它是生产、传播各种以文字、图形、艺术、语言、影像、声音、数码、符号等形式存在的信息产品以及提供各种增值服务的特殊产业，如图 3.4 所示。

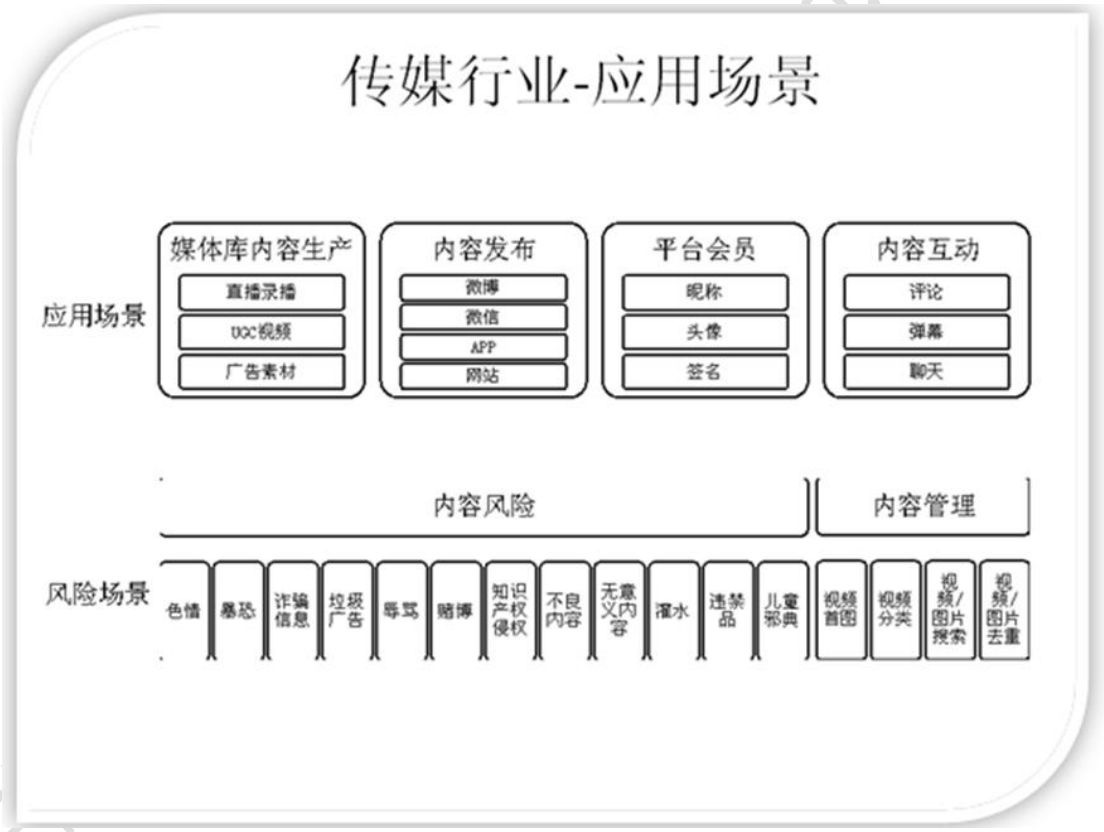

图3.4 传媒行业应用场景

针对媒资内容生产过程的内容获取、媒体入库经由数据采集、智能检测（含去重）、人工审核、审核质量管理多轮审核机制进行提效，优化媒体库内容。

通过图片/视频指纹、分类标签、字幕提取等技术，实现对融合媒体内容的结构化管理，支持有效的风险回查管理、版权保护监控。

在内容分发过程中，对多端分发的内容，包括直播、互动（评论、弹幕等）进行实时监控。兼容本地化、专有云、混合云等多种部署方式，如图3.5所示。

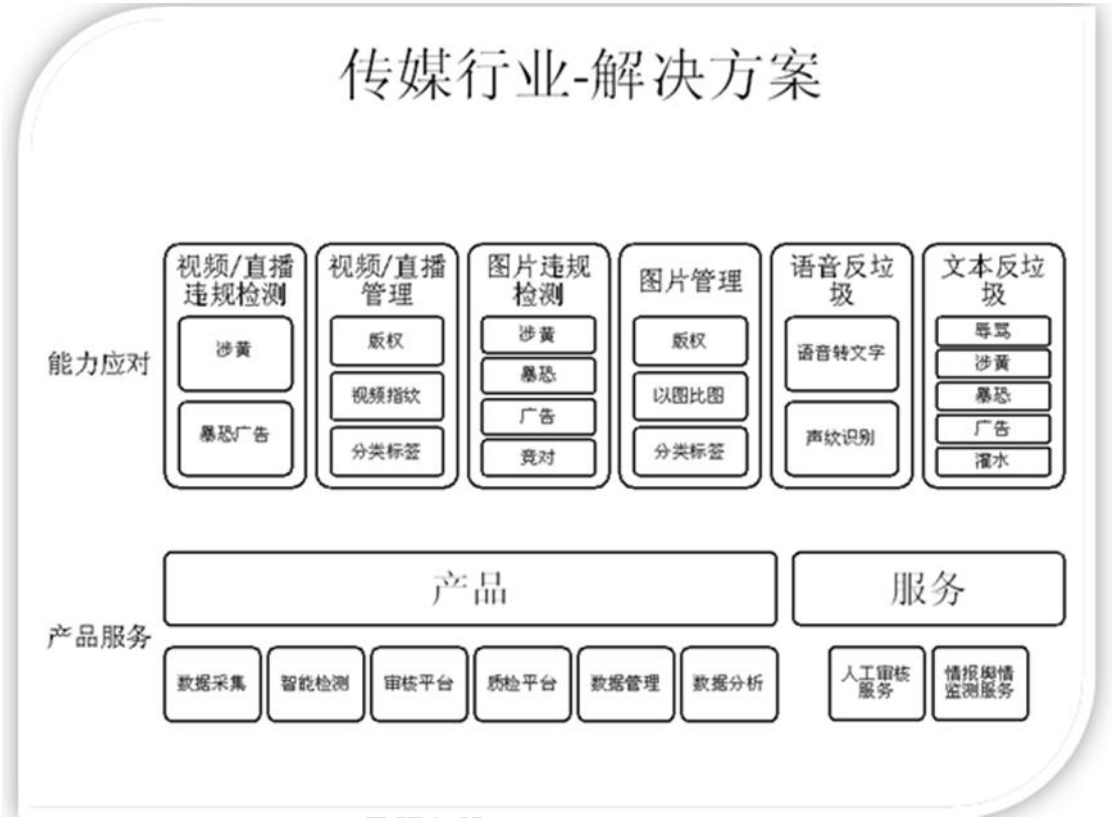

图3.5　传媒行业解决方案

传媒行业可细分为传统媒体和网络媒体。

（1）传统媒体

传统媒体的主要介质是电视台、广播电台。电视台内容安全 AI 应用包括媒体资源库违规内容检测、回捞、媒体资源库结构化、内容防篡改、违规内容排查和版权保护。

1）媒体资源库违规内容检测、回捞包括：画面内容合规检测：对画面内容进行涉政暴恐、色情低俗、非法台标、违禁品等风险检查；对广告内容检查：各省市电视台下属均有电视购物频道，除常规违规

内容检测外，利用自定义文本，NLP 等技术，识别违规广告话术加入广告识别，依据广告法识别违规内容；语音内容合规检测：对节目语音内容进行涉政、辱骂粗口、违规广告等风险检查。

2）媒体资源库结构化包括：利用物体识别、场景识别能力，对视频进行打标签，分类，将整个媒资库结构化，用于后续的大数据分析、用户画像及视频的精准推动等智能算法等，以实现万物（物体、场景标签）识别。

3）内容防篡改包括：使用 MD5/视频指纹技术，防止节目被篡改、编辑、插入、裁剪等，保障原视频质量。

4）违规内容排查包括：使用视频指纹、基于深度特征的视频检索技术，对媒体库内容建立索引，对监管下发的违规内容计算视频指纹 / 特征向量后，在索引库中对比，可快速进行资源排查，筛选违规内容，提高管控能力和效率。

5）版权保护包括：利用视频指纹、图像检索等技术，反查互联网或其他竞品平台，是否存在作品侵权行为。

目前存在私自设立广播电台，发送虚假广告、违规广告等行为。使用语音转文字+文本分类模型的方式，监管机构可以自动化完成遍历各个频段收集广播信号，判断违规内容。

（2）网络媒体

网络媒体包括两微一端（微博微信、各新闻客户端 App）、社区论坛等，其内容安全 AI 应用包括：

1）用户身份认证：会员进行注册、找回密码等重要操作时通过

人脸识别（1:1）、活体、声纹等技术手段核验身份。

2）网站安全：使用站点检测技术，识别网站是否存在被篡改等恶意行为。

3）PGC 内容合规检查：使用图像识别、视频分析、文本反垃圾等技术对发布内容做涉政、色情低俗、广告、引流等风险识别，防止违规内容主动传播、竞品打广告、引流。

4）UGC 内容合规检查：使用图像识别、文本反垃圾等技术针对注册会员的头像、昵称，以及评价内容进行敏感人物、涉政、色情低俗、广告、辱骂检测。

5）为资源商赋能，一些互联网公司会将阿里的 CND、OSS 等资源转卖，可把对图像、视频内容风险的检测能力对其赋能。

4.泛娱乐

泛娱乐是包含游戏、文学、动漫、影视、戏剧等多种文化创意领域的互动娱乐新生态，以直播、短视频为代表的娱乐视听产品涌现，UGC 信息海量增长，泛娱乐产品交互形式丰富。泛娱乐是大量产生 UGC 内容的渠道之一。

针对多媒体、资讯、游戏等在线泛娱乐内容，包括点播/直播视频、视频字幕/弹幕、资讯、账号信息、社区论坛、用户互动聊天、内置电商等场景中的视频、图片、语音、文字进行全维度风险检测，包括涉黄（色情漫画/邪典）、暴恐、敏感人脸、广告、Logo 竞对、不良场景、语音反垃圾、文本反垃圾等，泛娱乐解决方案如图 3.6 所示。

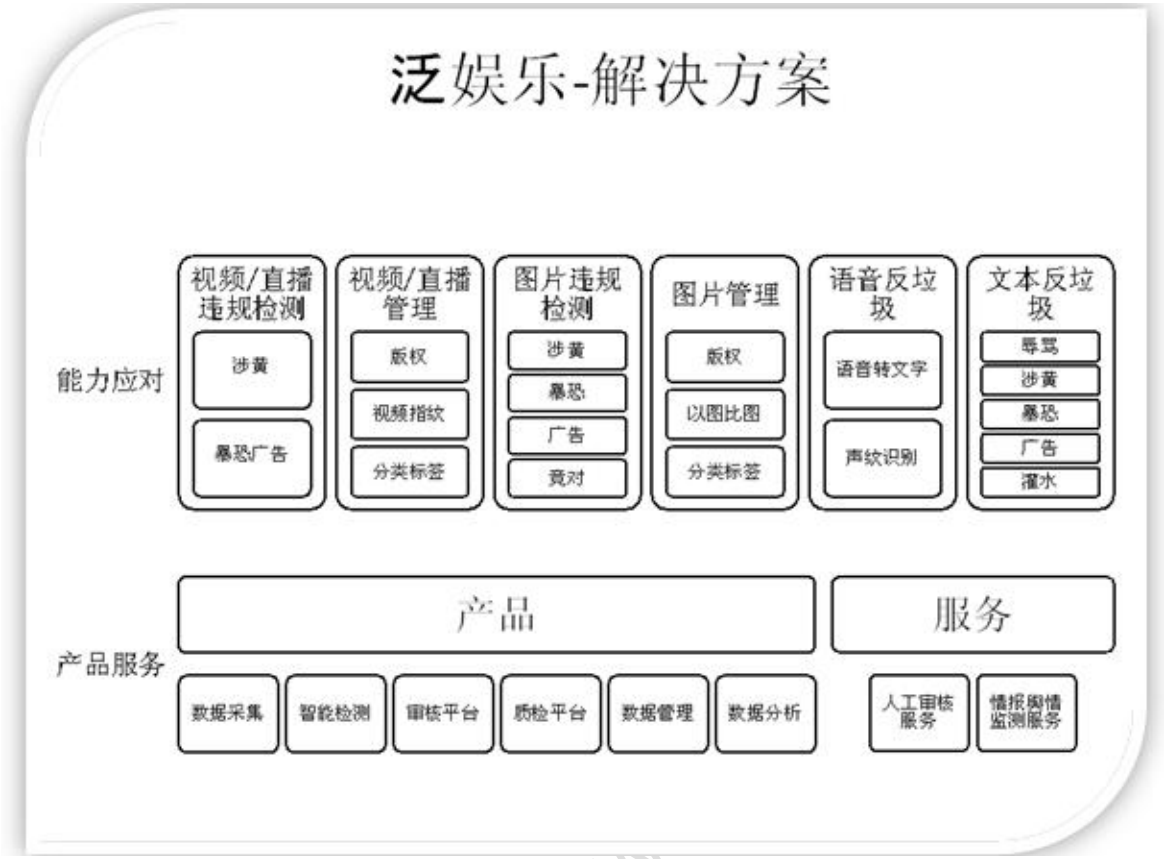

图 3.6　泛娱乐解决方案

泛娱乐主要应用服务领域以短视频内容安全、直播内容安全与游戏内容安全三类为主。

（1）短视频内容安全 AI 应用

1）内容合规检测：利用图像识别、视频分析、语音、NLP 等相关技术检测视频中的涉政暴恐、色情低俗、敏感人物、广告（二维码、小程序码）、违禁品、特定行为（吸毒、赌博等）、辱骂等风险。

2）版权侵权检测：利用 LOGO 识别技术，对短视频中 LOGO 检测，如台标、商品品牌等，防止侵权违规。

3）版权保护：利用视频指纹或者图像检索技术，反查互联网或其他竞品平台，是否存在侵权行为。

4）媒体库违规排查及去重：使用视频指纹、基于深度特征的视

频检索等技术，对媒体库内容索引，对监管下发的违规内容计算视频指纹及特征向量后，对比索引库，可快速资源排查，筛选违规内容，提高管控能力和效率。同样也用于媒体库中图片或视频去重，可节约存储成本，降低排查及结构化难度。

5）媒体库资料结构化：利用物体识别、场景识别能力，对视频打标签，并分类，将整个媒体库结构化，用于后续的大数据分析、用户画像及视频的精准推动等智能算法等。

6）坏帧及静帧检测：检测短视频是否为无效短视频，无画面，或画面静止不动，避免此类低质量视频透出给观众。

（2）直播内容安全 AI 应用

1）主播行为违规检测，具体涵盖了画面合规检测（通过图像识别技术，检测主播是否存在衣着暴露、不良着装、性暗示动作、吸烟吸毒、赌博及危险驾驶等违规行为。通过图片画中画检测、静帧、无意义直播等，及时给予提示，降低 CDN 消耗）。

2）语音违规检测（通过语音识别、NLP 技术，检测直播过程中主播是否存在涉黄、涉政等违规行为，用于解决直播画面正常，但主播大肆宣传低俗、涉政等内容风险）。

3）主播身份验证：通过人脸识别（1:1）技术，确认直播主播和开播注册主播为同一人，降低代播风险。

（3）游戏内容安全 AI 应用

1）游戏直播、违禁游戏内容识别：通过图像检索技术识别未准许进入中国境内的游戏，防止出现在游戏直播中。

2）游戏人物识别：识别典型的游戏角色，尤其是敏感人物卡通化。

3）游戏内社交行为识别：对头像昵称、聊天、弹幕等使用图像识别、文本反垃圾技术检测敏感人物、涉政暴恐、色情低俗、辱骂、广告等风险。

5.互联网金融

采用 OCR 技术为电子化办公提效，支持包括证件、银行卡、票据、文档的自动图文识别。采用人脸识别，结合活体认证，实现智能身份验证，且支持包括营业厅的 VIP 识别等重点人脸监控。兼容本地化、专有云、混合云等多种部署方式。金融行业应用场景如图 3.7 所示。

图3.7　金融行业应用场景

金融行业 AI 内容安全应用场景包括身份验证、证件及票据识别。

（1）身份验证：开户、借贷放款时使用人脸识别（1:1）、活体识别、声纹识别（1:1）等技术做强身份验证，确保资金安全。

（2）证件及票据识别：使用 OCR 技术，对用户提交的证件、票据等做信息的结构化提取，大幅度节省审核人力。

6.在线教育

随着互联网时代的发展，更多的传统行业从线下发展到线上。通过智能审核解决方案对线上业务内容，包括人员账户信息、课程描述、课程音视频内容、人员互动聊天评论等文本、图片、视频、音频的全方位风险过滤，对平台的课程内容进行版权保护。对智慧校园监控提供线下监控视频的人脸识别、物体识别等安防保障。同时涉及身份验证，以及施教者和受教者远程视频中，语音、动作、展示的内容。在线教育应用场景如图 3.8 所示。

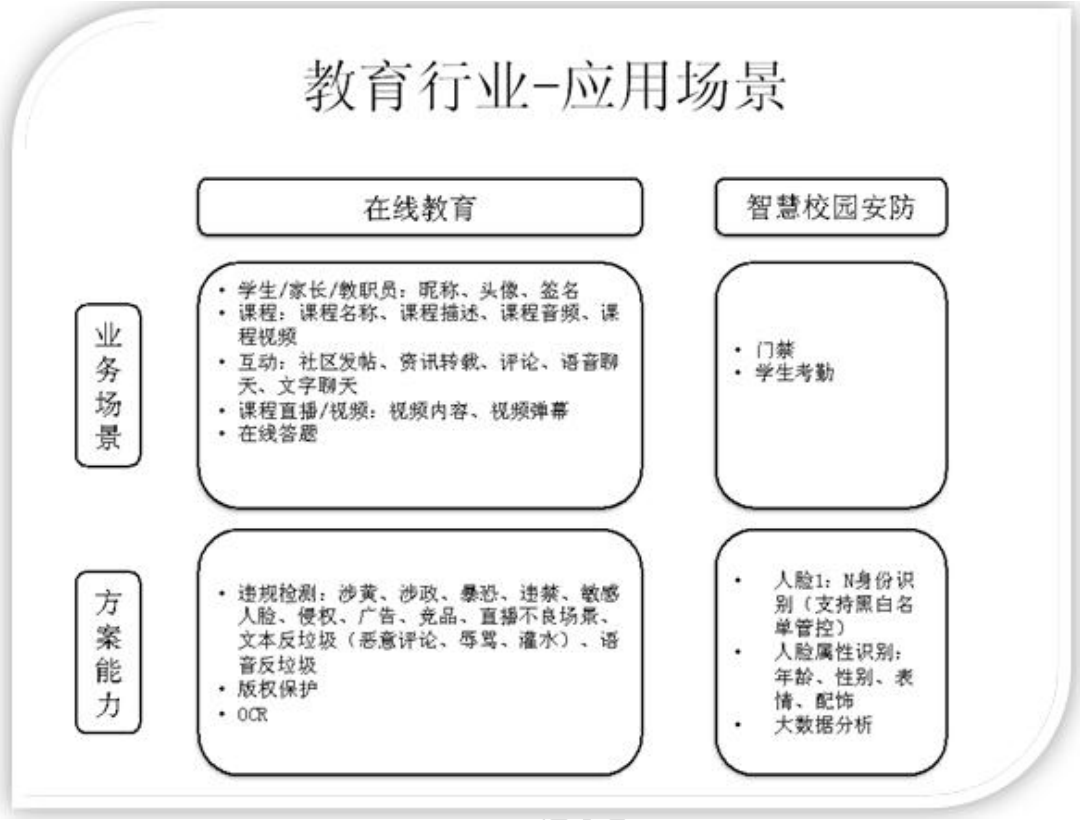

图3.8　在线教育行业应用场景

教育行业 AI 内容安全应用场景包括身份验证及内容合规审查。

（1）身份验证：通过人脸识别（1:N）技术，识别出会员、熟客、黑名单的信息，配合数据营销系统进行销售分析。

（2）内容合规：对于远程视频过程中语音、文本、画面，使用语音反垃圾检测语音中的辱骂粗口、色情低俗等风险；使用文本反垃圾检测交互文本中的色情低俗、辱骂、广告、引流等风险；使用图像识别技术检测画面中的色情低俗、广告、引流等风险。

7.政府/物业安防

（1）政府安防

1）使用人脸识别技术做门禁、身份验证、查找黑名单中危险分子。

2）使用视频分析技术做禁戒区（线），判断是否有人进入禁区；检测拥挤、追逐、打斗、跌倒、翻越等异常事件；识别爆炸、燃烧、烟雾等危险场景。

（2）物业安防

物业安防是指通过接入监控视频，对视频内容进行人脸识别、物体识别、场景行为识别，实时发现安防风险。通过智能终端进行身份核验，支持自定的人脸库管控。通过内容风险大屏辅助监控中心进行实施控制。物业安防应用场景如图3.9所示。

图3.9　物业安防应用场景

8.信息通信

通过对短信/彩信进行智能风险过滤。提供私有化部署，支持数据采集、数据管理、智能检测、人工审核介入管理、审核质量检验、

数据分析在内的全流程覆盖方案。信息通信业务应用场景如图3.10所示。

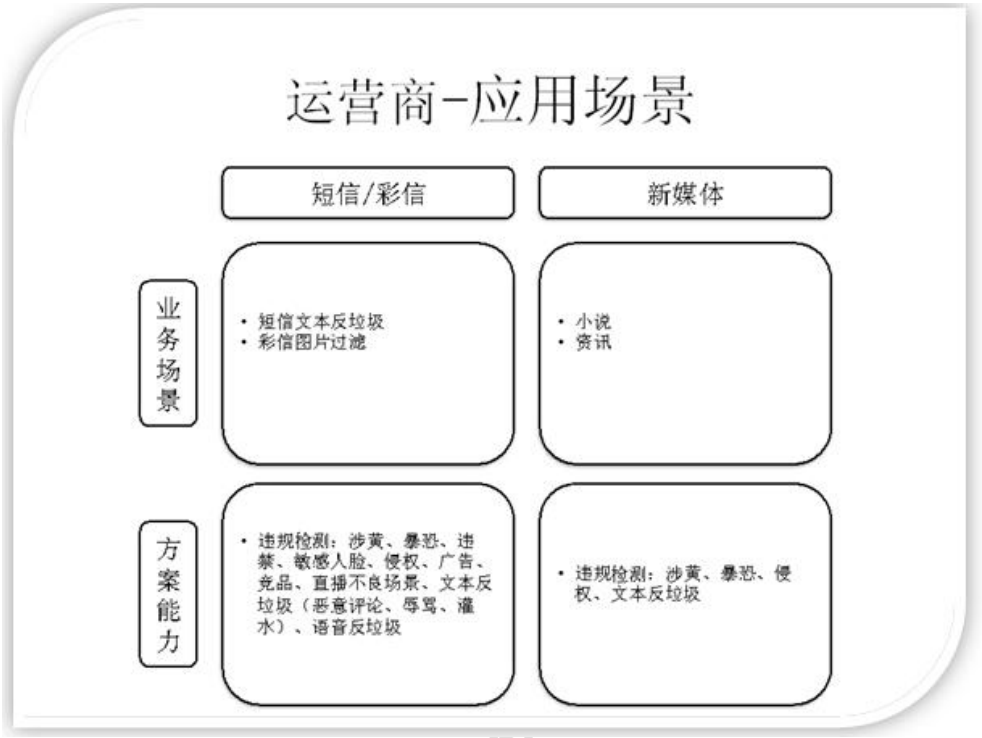

图3.10　通信运营商应用场景

9.其他生活场景

生活中常见人员密集，公众安全要求级别高的场景包括：养老机构、酒店、机场、车站、网约车、景区等。

（1）养老机构：使用人脸识别（1:1）做身份验证；使用视频分析做跌倒等意外事件检测。

（2）酒店：使用人脸识别（1:1）做身份验证。

（3）机场、车站：使用人脸识别（1:1）做身份验证、安检认证（针对工作人员）；使用 OCR 技术做证件 / 票据识别；使用视频分析技术做人流密度、燃烧、烟雾检测等。

（4）网约车：在注册、上岗环节使用人脸识别（1:1）做身份验证。载客途中使用语音转文字+文本分类技术做语音内容合规检查；

使用视频分析技术做扭打抢夺等异常动作识别。

（5）景区：进场、离场使用人脸识别、活体技术做身份验证，判断是否购票，以及查找黑名单中危险分子。

综上，与市场同类产品相较，云盾内容安全有以下特色：

（1）丰富的产品线：另有业务安全、端上安全、云上安全、数据安全等协同输出，提供全面安全解决方案。

（2）顶尖的平台和算法：业内顶尖的安全运营和算法团队，由阿里云提供计算保障。

全链路的服务体系：云盾内容安全提供各维度内容风险识别、样本管理 / 打标、策略配置、动态算法升级与规则优化等全链路的管理能力。

四、企业价值

作为互联网生态中为用户交付服务的主体，向公众提供互联网服务的企业是网络内容安全技术的运用者，也是 AI 新技术在其中应用的受益者和成本承担者。网络内容安全管理，对于企业既是一项建设工作，也是一个维护过程。以 AI 技术推动网络内容安全保障也是同样的道理，引入智能技术需要一系列项目化管理手段，基于 AI 的内容安全管理信息系统投产后则有持续的维护要求，而 AI 对于网络内容安全的企业价值便体现在建设和运维这两大阶段里。

1.建设阶段的价值

（1）达到监管要求，确保合法经营，实现有效投入（达标）

企业做业务价值评价时，既需要考虑自身作为一家经营单位的财

务盈亏，也需要关注触碰合法合规红线的严重影响，这种影响有时可能是一票否决性或影响企业持续经营的。所以做好网络内容安全管理的建设，首先就是在保障企业达到监管要求的标准和不碰红线。内容安全管理技术日新月异，与恶意或无意破坏内容安全者之间博弈时，技术门槛、技术难度水涨船高，缺少 AI 技术的内容安全管理难免存在死角盲区，而放过漏网之鱼的这些死角盲区可能会使内容安全管理体系功亏一篑，所以用 AI 做内容安全当前已成为企业达标、实现有效投入之必要手段。

（2）建设高效率的内容安全管理信息系统（高效）

AI 除了能发挥“魔高一尺、道高一丈”的场景下抬升技术能力的作用，对内容安全管理效率的提高也是明显的。模仿人、代替人是 AI 的重要逻辑，网络内容安全管理中人工工作被释放得越多，整体管理工作的效率就越高。这种效率，不仅是单位处理时间的缩短，也包括能够让计算机系统发挥并行处理复杂事务的长项，让 AI 支撑下的计算机系统消除人工处理的各种单点和瓶颈，不知疲倦地并行运算，这样将使效率提升程度倍增，而效率对于企业就是生产力。需要说明的是，随着 AI 配套硬件技术的发展与成熟，以往 AI 复杂处理中 AI 自身的效率难题，已被极大克服。

（3）高起点确保侦测质量，保障监测效果（高质）

基于 AI 的内容安全管理是站在新技术的高起点上，侦测质量即识别可靠性、处理可靠性上较传统内容安全管理技术有较大进步。除了 AI 带来的技术精度能提升质量外，AI 对管理体系的作用也是明显

的——体系化是质量管理能够规模化、可持续的基础，而具体操作时则是需要明确目标后进行 PDCA[1]的循环。不论是上述的“体系化”还是“PDCA”，AI 助推内容管理的充分数字化都是质量精细管理的基础，从这个逻辑讲，内容安全上 AI 模式对人工和传统方式的替代，从管理角度上对网络内容安全管理的质量也将大有益处。

（4）对内容安全管理中人工、非智能技术手段起部分替代作用，降低成本（节省）

当前，互联网产生的内容是海量级别的。相比 AI 技术，现有的人工审核方式无论是成本还是效率都处于显著劣势。如果把这部分工作交给机器完成，就可以让解放更多人力资源，使其可从事更富有创造性和发展前景的岗位。

根据技术创新的规律，最先掌握新技术的领先企业通常能够在短期内获得超额利润。当新技术普及后，相应生产力水平则能整体得到提升。这一规律同样适用于 AI 在网络内容安全管理保障方面的应用。

综上，从“多”（范围达标）“快”（高效率）“好”（高质量）“省”（低成本）的角度全面论述了建设阶段的企业价值，而多、快、好、省的兼顾则有赖于 AI 的技术进步性。

2.维护阶段的价值

（1）维护企业形象和品牌价值，保护无形资产，创造无形收益

AI 武装后的网络内容安全管理体系，对企业品牌的突出价值主要体现在维护阶段。AI 为网络内容安全所带来的数字化和自动化，

[1] PDCA：即戴明环，P 指 Plan（计划），D 指 Do（执行），C 指 Check（检查）、A 指 Adjust（调整）。在 PDCA 体系下，质量管理形成完整的反馈体系，具有持续优化改造的能力。

对合规合法红线值守得更加完备，在较之建设阶段生命周期更加漫长的维护阶段，能让人工智能发挥出“路遥知马力”的持续价值，以更快速（快）、更稳妥（稳）、更准确（准）的高含金量阻击，消除负面影响对企业品牌的风险隐患。

（2）内容安全领域 AI 技术的产业化和产品化，有助于运维标准化，避免个性化困扰

AI 技术作为网络内容安全保障的一种大趋势，随着应用企业的增多，将逐渐呈现产业化和产品化的特点。越形成产业规模的技术、越具有产品式成熟度的技术，维护阶段对于企业就更容易实现运维标准化。不仅配套技术服务容易采购，且服务标准统一，避免了过多个性定制带来的难以维护的后果。这将使企业的运营管理具有更好的持续性，基于 AI 技术的网络内容安全保障本身也将发挥更大的价值。

五、社会价值

从社会价值方面来看，AI 技术引入到网络内容安全保障后，将会在以下几个方面提升社会价值。

1.维护国家安全和社会稳定

统计数据显示，网络上反动、涉恐及高危篡改等敏感信息占比并不高，但其潜在危害却是最大的。如何尽早发现并识别上述内容，将隐患遏制在萌芽状态，是国家的基本诉求。

借助 AI 技术，可以快速定位反动和暴恐信息（包括但不限于旗帜、标语、人物、场景），为有关部门提供快速、精准的情报支持。例如，此前有媒体报道，我国某市火车站在配备带有 AI 技术的安防

系统后，很快便将混入车站的罪犯识别并成功抓捕；国外某厂商可以借助 AI 技术在 45 秒内准确定位枪声来源。可见，机场、港口和边防口岸若全部配备相关安防系统，无疑将极大提高本国的安全系数和治安管理水平。各国之间若能在相关数据方面实现共享，必将更有力地打击跨境贩毒、洗钱、人口贩卖等犯罪行为，也将更有利于地区局势的稳定。

2.净化网络空间

根据 12321 网络不良与垃圾信息举报受理中心的统计数据显示，2018 年 1-5 月，共收到的举报短信合计达 9.3 万件次，平均每月 1.86 万件次；骚扰电话举报达 14.5 万件次，平均每月 2.9 万件次；不良网站举报为 14.7 万件次，平均每月 2.94 万件次。

事实上，上述数据只是从一个可量化的角度反映了网络空间受到污染的状态。在日常生活中，大多数人会有更为直观的体验，那就是经常会遇到各类骚扰电话、短信和网页弹出广告；父母们会担心孩子上网时被动接触色情信息等等。面对这些问题，普通网民往往无能为力。

对此，利用 AI 技术对图片、视频、文本进行智能识别，实现智能鉴黄、识别不良场景，将有效净化网络空间；通过深度学习算法和实时更新的亿级图像样本库,可对图片与视频进行识别以及色情程度量化；结合行为分析和时间序列对比技术，针对在直播和视频中需要监管的不良场景（如抽烟、画中画、赌博、斗殴等）进行精准识别。上述技术的广泛应用，将有效改善不良与垃圾信息的泛滥情况，避免

网络资源的污染和浪费，减少对人们生活的困扰。

3.保护知识产权

在加入 WTO 及对外开放的承诺当中，我国都表示要进一步加强知识产权保护。中国推进知识产权保护，是提高我国经济竞争力最大的激励，这不但关系到我国的经济发展，也关系到我国的国际地位和国际影响。

在电子商务领域，我们的政府和相关企业因假货泛滥而受到诟病。同样，文学和视频（包括影视）领域由于侵权成本低，使得内容生产者的权益无法得到有效保障。随着技术进步，一个较为有效的手段是，通过 AI 智能识别技术，对商品 LOGO 进行识别，可精确判断真货与假货。同样，对原创视频进行视频指纹匹配，可帮助用户识别原创视频和转发及伪造视频，有效保护原创人员知识产权。

值得一提的是，2018 年是视频行业的风口，无论是运营平台还是内容生产方，在收获大量用户及利润的同时，也必将面临大量的侵权行为。上述技术或许能在某种程度上保驾护航。更重要的是使产业生态链形成良性循环，让合作机制和内容生产愈加流畅。

4.降低维护真实信息的社会总成本

在当前，互联网产生的内容是海量级别的。如果不引入 AI 技术，那么对于内容的审核将消耗大量的人工审核成本，而且这样的工作相对枯燥。如果把这部分工作交给机器完成，就可以让人力放到更加富有创造性和发展前景的岗位上，使之得到有效配置。特别是对提供互联网信息服务的企业而言，聘请大量的审核人员来进行人力审核，是

一个不小的负担。

社会的发展有赖于各个不同部门的分工合作。这就需要减少在合作当中的摩擦成本。而网络的内容对于社会各个组织部门及其成员都有着很大的影响，对于社会的合作效率有着一定影响作用——一些不正确的网络内容，会造成社会对于某些事件或者某些人物的固化印象，从而使得社会合作不能够有效的运行。AI 技术的应用，可以最大限度地消除虚假信息的影响，进而增强社会各部门及其成员的互信，达到降低社会总成本的效果。

5.提升互联网内容质量

随着移动互联网将内容赋权，互联网内容的产生门槛不断降低。自媒体，如公众号、头条号等等，都可以成为媒体内容的生产平台，但是这些内容的生产者往往没有受过正规的新闻训练，同时内容审核和发布流程不健全，无法确保内容的质量，导致在海量内容不断产生的背景下，发生劣币驱赶良币的的现象，从而使得互联网信息内容质量降低。而且，劣质内容越多，越会挤压优秀内容的成长空间，引发社会的激励机制出现扭曲。现如今，由于互联网信息内容知识产权得不到保障，新闻调查内容越来越少，大量的不能确定真实来源和没有第一新闻落点的内容越来越多。AI 技术可以在浩如烟海的互联网信息中，找到最初来源，倒逼从业者提高思想觉悟，提升业务技能，强化内容质量。

6.舆情管控应对

作为公众表达舆情和传递声音的重要窗口，网络舆情具有信息丰

富、表达快捷、渠道多元、传播急速等天然优势。许多热点事件，如山东假疫苗、快手社会摇等事件，都在极短时间内引起大量公众关注，甚至有些事件还存在线上动员、线下活动的精心组织。

众所周知，不断发酵的网络舆情，使得一些问题得到有关部门重视，并成为其了解民情民意的重要途径之一。同样，舆情未得到妥当处理，则直接影响当事方的公信力和形象。有时一个“以讹传讹”的假新闻，也会火速刷屏网络，最终“走形”。

采用深度学习算法，结合海量及实时更新的舆情样本库，可以针对热点舆情事件快速更新、甄别、以及管控。特别对于是不良视频、刺激性画面和文字，用 AI 技术第一时间进行识别，能够迅速把事件的负面效应控制在最小范围。

7.维护公民权益

面对生活中充斥的各类不良及垃圾信息，我们还需要进一步思考，为什么这些信息会“精准”抵达每一个人？例如，有购房意向或需求的人，会接到房屋中介、金融贷款等方面的电话、短信骚扰；有业务往来时，则会收到发票办证、保险理财等方面的信息骚扰等等。公民个人的数据隐私，往往在不经意间就被泄露出去；更有甚者，如恶意发帖这样的黑灰产业，使得违规信息在被搜索时获得大量传播，以极其隐蔽的方式绑架了公众权益。

对此，我们一方面需要对数据存储端进行管理，另一方面需要为客户端提供更有效的应对工具。数据存储端方面，可依托 AI 技术，对网站中存储的个人敏感信息在展示前进行警告或者脱敏、过滤，防

止个人隐私信息泄露。同时，通过文本和语音反垃圾技术，对高危、垃圾、灌水、定制类信息进行智能识别，减轻平台及监管部门压力。通过系统化的应对，最大程度防止用户数据被恶意利用。

六、发展趋势

1.强对抗网络的应用会越来越深入

通过对抗网络，人工智能可能会具有更强的价值观属性，对于内容的理解会更加深入。抛开复杂的函数语言，GANs 在原理本质上酷似博弈论中的二人零和博弈，即非此即彼的胜负游戏。这场游戏中甲的存在价值就是无休止的挑战、质疑和审判，从而迫使乙不断调整方案，尽一切可能逃出甲的刁难。GANs 对这个原理的实现方式是让两个网络相互竞争。其中一个叫做生成器网络（Generator Network），它不断捕捉训练库中的数据，从而产生新的样本。另一个叫做判别器网络（Discriminator Network），它也根据相关数据，去判别生成器提供的数据到底是不是足够真实。

通过对抗网络，可以不断优化人工智能内容审核的质量，通过对有效结果不断的趋近，使得内容审核水平不断提高。

2.AI 技术的仿生学演进愈加清晰

未来人工智能的发展需要心理学家或生物学家的共同合作，加强对人脑工作机制的学习。人脑是具有所谓强人工智能的唯一实例，若进一步提升 AI 的性能，研究人脑的工作机制是必由之路。比如，现有的很多深度学习算法依靠提前设计完成且无法改变的神经网络架构，这与人脑有很大不同：人类从婴幼儿成长到青壮年，人脑的神经

元是逐渐增加的，架构也是逐渐拓展的；人脑在我们休息的时候，尤其是深度睡眠期间，会对冗余的神经元进行修剪，使人脑的推理过程更加节能和高效。现在一些前沿的神经网络生长和修剪算法已经朝着这个方向迈出了一步。

与此相似，大部分机器学习（包括深度学习）算法和推理过程是分开进行的，而人脑的学习和推理是互相嵌套、交叉迭代进行的。反向传播算法一直以来是深度学习算法的核心之一。被称为深度学习之父的 Geoffrey Hinton 教授受神经学研究的启发，提出了胶囊网络，并号召大家摒弃反向传播。

可见，现在深度学习里的一些里程碑式的算法都或多或少从神经科学，心理学等学科的研究得到了启发。

3.AI 技术在内容审核上的作用更加突出

AI 从人类形成的各种文化产品上更加深入地了解人类的价值观和法律规范。设计者可以在提供有违规特征的内容让 AI 系统进行深度学习，也应提供合规以及优秀产品的特征让 AI 自主学习，提升其价值观和道德意识；打开 AI 的“黑盒子”，对其的自我决策机制进行进一步深入分析研究。

这里的思维逻辑是，如果判定内容的性质以对内容的价值观进行判定为前提；内容的价值观是人类在多年的认知和思考当中积累下来的，而且不同的年龄段价值观也需要分级分类；在用户注册时，根据用户身份信息和历史数据，应该适用不同的内容安全分级机制——类似金融业征信机制；系统应将更多的文化因素植入到内容审核当中。

审核流程将前移，不再等到敏感内容出现以后才进行审核，而是在用户注册和上传相关内容时，就会通过 AI 技术预先进行审核。

当然，AI 技术并非无所不能，在复杂、疑难的情况下，仍需要人工进行交叉比对。

七、发展建议

1.营造产业发展良好环境

AI 技术本身并非新生事物，也不是解决一切问题的万能钥匙。此外，AI 技术在落地应用方面还有较长路要走，还处于典型的“幼年期”。

面对尚处于早期培育阶段的产业现状和诱人的产业前景，政府需要给与更多的耐心与宽容，在确保国家安全和遵循社会伦理的前提下，避免过早出台“红旗法案”，将新生事物扼杀于摇篮之中。

特别是当前中国经济正处于转型升级的关键阶段，经济的新旧动能转换，政府需要对新生事物给予充分发展空间，同时做好监管和规范工作，使其在公平竞争中不断发展完善，为不断涌现的成长型企业创造平等竞争的环境。

2.推动 AI 技术发展

AI 技术特别是深度学习技术在图像、语音、自然语言处理等多个方面取得了突破性的进展，甚至在某些领域已经超越了人类。未来需要不断提升和推动国内企业在 AI 领域的自主创新和关键问题的技术攻关能力。首先是 AI 自主学习问题，通过元学习（Meta learning）等研究，让人工智能自己学会思考，学会推理；二是有效解决 AI 成

本问题，通过模型压缩、剪枝、量化、软硬件协同开发等方法加速神经网络模型计算、缩短 AI 的训练时间；三是 AI 模型的可解释性，可解释性对于 AI 模型的验证和改进以及 AI 技术的推广具有十分积极的意义，是未来 AI 领域重要的研究方向；四是 AI 模型安全评估，尽管 AI 在很多方面展现出了相比传统机器学习方法更好的鲁棒性，我们也要警惕专门针对深度神经网络的攻击模型。在人工智能的应用越来越广泛、越来越深入的今天，必须对其安全性保持足够的重视，训练能够应对各种攻击的、更安全的 AI 模型也将会是未来 AI 领域重要的研究内容。

3. 鼓励政府和企业采取人工智能技术进行内容审核

当前，内容审核的主要手段以人工智能和人工审核为主。与人工智能不同的是，人工审核的成本跟随着用户量及数据上传量而成正比例增加，这不利于成本和效率的最优配置。建议政府部门和企业采用人工智能为主、人工审核为辅的方案，不仅可大幅降低经济及时间成本，还可显著提高部门办事效率和企业效益。

4. 加强数据保护体系建设

用户个人信息保护是人工智能领域一个非常重要的方面。人工智能需要大数据的支持，但是当前数据量都集中在各大互联网巨头公司手中，相互之间存在一定壁垒。因此，主管部门应当组织互联网企业加强合作，将各企业采集的个人信息进行整合，由国家统一存储；当企业调用时应取得必要授权且对个人信息中重要属性进行模糊化处理。

5.探索协同治理模式

依据《网络安全法》和有关法律法规要求，需要完善管理机制，形成统一监管、分工负责的管理模式。第一，网信、通信、公安、广电等党政部门牵头建立监管机制和联席工作机制。第二，有关企业应当建立定期内容审核上报机制，并向社会进行公开。第三，优化用户和媒体的维权机制，党政部门、行业协会和企业向社会公布举报途径。第四，高校、智库等研究机构应密切跟踪国内外人工智能发展的新情况和新趋势，不断提出新成果。