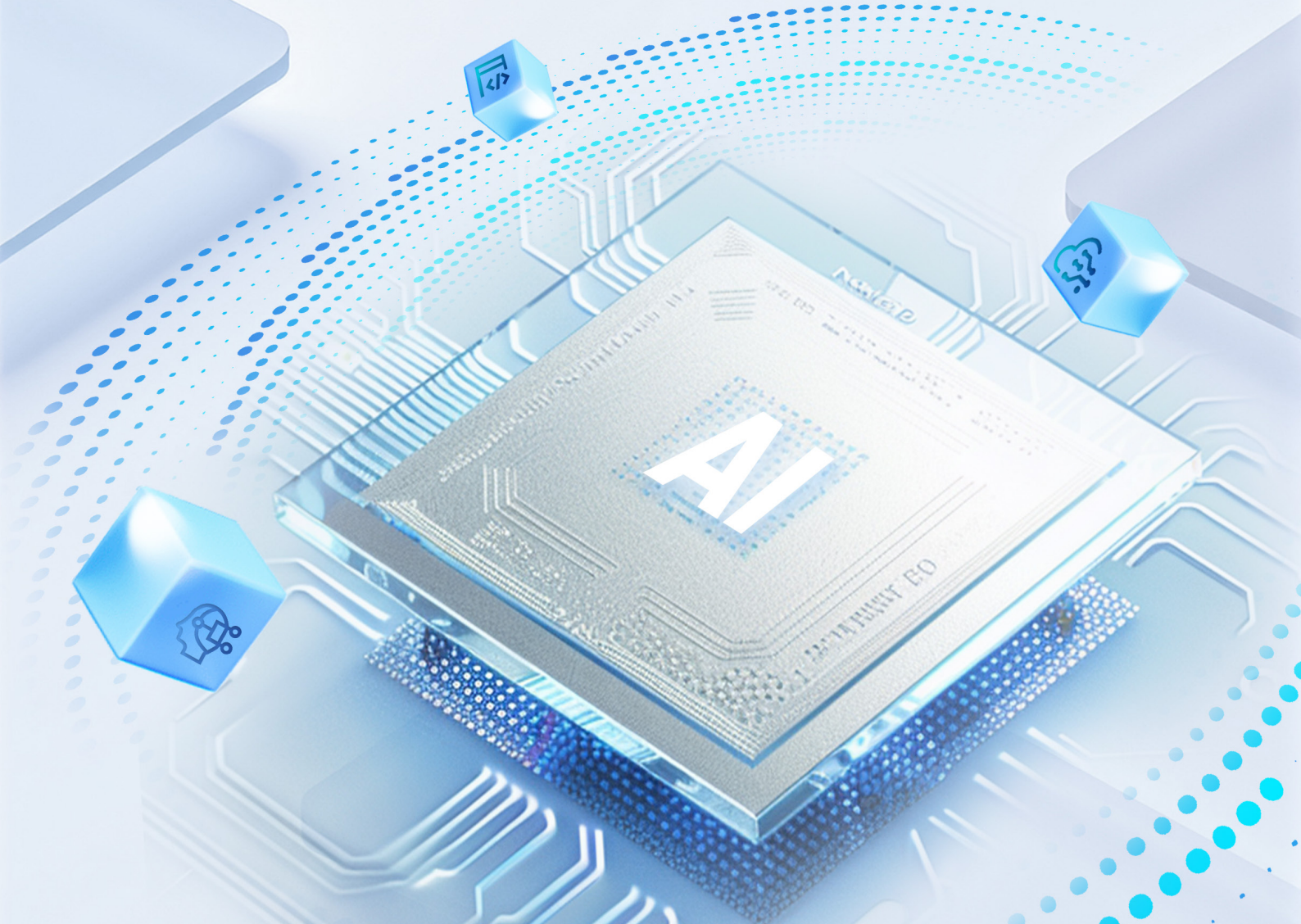
SAISD
上海市人工智能与社会发展研究会
Shanghai Association for AI and Social Development

EY安永
Shape the future with confidence
聚信心 塑未来

可信人工智能治理白皮书

TRUSTED ARTIFICIAL INTELLIGENCE GOVERNANCE WHITE PAPER



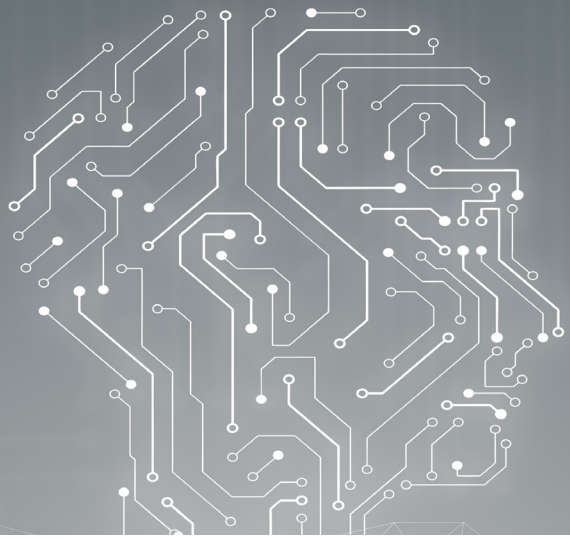
版权声明

COPYRIGHT STATEMENT

出品方

安永（中国）企业咨询有限公司

上海市人工智能与社会发展研究会



前言
FOREWORD

21 世纪人类社会迎来了前所未有的技术革命——人工智能（AI）的兴起。它以迅雷不及掩耳之势，渗透到我们生活的方方面面，重塑着我们的认知和社会格局。它不仅极大地提高了生产效率，还为解决复杂的社会问题提供了新的视角和工具。然而，随着 AI 技术的迅猛发展，其所带来的伦理、法律和社会问题也日益凸显，如何确保 AI 技术的“可信”属性，成为全球关注的焦点。

在此背景下，《可信人工智能治理白皮书》应运而生。本白皮书由安永（中国）企业咨询有限公司和上海市人工智能社会发展研究会联合撰写，旨在深入探讨人工智能的全球发展态势、监管体系、可信原则、关键问题、企业合规要求、风险治理理论、进阶工具以及行业洞察等多个方面。我们希望通过这份白皮书，为政策制定者、企业管理者、技术开发者以及所有关心 AI 发展的读者，提供一份全面、深入、客观的参考和指导。

在这份白皮书中，我们将重点探讨“可信人工智能”的内涵，分析其在算法透明度、数据安全、伦理道德等方面所面临的挑战。同时，我们也将关注企业在 AI 应用中的合规要求以及风险治理这一 AI 发展中的重要议题。本白皮书将详细阐述风险治理架构的构建，以及如何在 AI 的生命周期中实施有效的风险管理。此外，我们还将介绍企业 AI 治理的进阶工具——可信 AI 等级标识，为企业构建和完善自身的 AI 治理体系提供实用的指导。

在提供有效的 AI 治理工具的同时，我们也将聚焦具有启发性的行业实践。我们将深入分析汽车、医药、零售、服务等行业在 AI 应用上的现状和挑战，以及这些行业如何构建和运营自己的 AI 管理体系。通过这些行业案例，我们希望能够为读者提供具体的行业应用视角，以及在实际操作中可能遇到的问题和解决方案。

在这份白皮书的撰写过程中，我们深刻感受到 AI 技术的发展不仅仅是技术层面的突破，更是对人类社会价值观、伦理道德和法律体系的一次全面考验。我们相信，只有通过不断提高对风险的把控能力，预判、分析和应对潜在的治理问题，建立起一套公正、透明、高效的 AI 治理体系，才能确保 AI 技术的健康发展，让它成为推动人类社会进步的正能量。

在此，我们诚挚地邀请您一同走进 AI 的世界，探索其无限可能，同时也思考和面对它所带来的挑战。让我们携手前行，在 AI 的时代中，共同寻找可信 AI 实现的最佳路径。

目录
Comtents

目录
Comtents

目录
Comtents

第一章
全球人工智能发展与监管体系

01
PART

当前，全球人工智能正处于技术创新、重塑行业、重新定义我们与技术互动方式的快速更迭阶段。AI 技术已经走出实验室，逐渐嵌入医疗、金融、零售、制造、交通、娱乐等行业领域的各项应用之中，社会也随之展开深刻变革。据 Fortune Business Insights 预测，人工智能市场预计将在未来十年内出现强劲增长，预计将从 2024 年的 6211.9 亿美元增长到 2032 年 27404.6 亿美元[1]。面对这一潜力巨大的颠覆性技术，全球各国竞相发力，正在共同将人工智能发展推向时代浪潮顶峰。

1.1 全球人工智能发展概述

截至 2024 年，全球共有 69 个国家和地区制定了人工智能相关政策和立法，涵盖了从 AI 治理、隐私保护、数据监管到伦理规范等广泛的主题，反映了各国对 AI 技术发展的重视及其潜在风险的管理需求。

以下国家或地区明确针对人工智能出台了立法或监管要求:

1.1.1 欧盟人工智能发展

欧盟《人工智能法案》是全球首部全面且具影响力的人工智能法规，于2024 年 5 月 21 日获批通过，并于 2024 年 8 月 1 日正式生效。该法案将“人工智能系统”定义为一种基于机器的系统，被设计为以不同程度的自主性运行，并且在部署后可能表现出适应性。该系统基于明确或隐含的目标，从接收到的输入中推断如何生成可影响物理或虚拟环境的输出，例如预测、内容、推荐或决策。《人工智能法案》采取了基于风险的方法，将人工智能系统分为最低风险、低风险、高风险和不可接受风险这四级。其中，认知行为操纵和社会评分等具有不被接受风险的人工智能系统被禁止进入欧盟市场；高风险人工智能系统需要获得批准，同时需要遵守技术要求、风险管理、数据治理、透明度等一系列要求和义务才被允许进入欧盟市场；其他非高风险人工智能系统的提供者则被鼓励制定行为守则，包括相关的治理机制，以促进自愿适用高风险人工智能系统的部分或全部强制性要求。

此外，欧盟委员会正在制定《人工智能责任指令》（Artificial Intelligence Liability Directive），该指令草案提出于 2022 年 9 月 28 日，旨在为人工智能系统造成的、非合同类责任的特定类型损害提供统一规则，解决举证责任难的问题，确保受到人工智能技术伤害的人能够获得经济补偿。该提案共九条，主要规定了高风险人工智能提供者的证据披露义务、过错与 AI 输出之间因果关系的推定两方面核心内容。

1.1.2 美国人工智能发展

美国正在选择一种针对特定行业的方法（sector-specific approach）。2023 年 9 月，美国参议员和众议员提出的《2023 年算法问责法案》草案（Algorithmic Accountability Act of 2023）为受人工智能系统影响的人们提供了新的保护措施。该法案适用于用于关键决策的新型生成式人工智能系统，

以及其他人工智能和自动化系统，要求公司评估其使用和销售的人工智能系统的影响，为何时以及如何使用此类系统创造新的透明度，使消费者能够在与人工智能系统交互时做出明智的选择。

同年 10 月，拜登总统签署了一项人工智能行政命令《关于安全、可靠和可信赖地开发和使用人工智能的行政命令》（Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence），为人工智能安全、安保、隐私保护、促进公平和公民权利、促进竞争和创新建立了新的基准。该行政命令要求制定一份国家安全备忘录，以指导人工智能在军事和情报行动中的安全和合乎道德的应用，确保保护美国人的隐私，并培育一个开放、竞争激烈的人工智能市场，强调美国的创新。

为应对人工智能生成内容的知识产权挑战，美国专利商标局（USPTO）和版权局（USCO）已经发布了相关执法行动和指南。2024 年 2 月，美国专利商标局发布的关于人工智能辅助发明的详细指 南（AI and inventorship guidance: Incentivizing human ingenuity and investment in AI-assisted inventions）明确规定了 AI 辅助创新中的发明权确定问题，即人类贡献必须足够显著才能获得专利保护。2023 年，美国版权局发布的“包含人工智能生成材料的作品”（Works Containing Material Generated by Artificial Intelligence）的政策规定，如果作品纯粹由 AI 生成，除非有显著的人类创作参与，否则不符合版权保护条件。如果人类在 AI 生成材料的选择、安排或修改中作出了重大贡献，则人类创作元素可获得版权保护。

此外，美国一些州也已经在人工智能相关领域制定监管法律。例如，2024 年 5 月 17 日，科罗拉多州州长签署了参议院第 24-205 号《关于在人工智能系统交互中保护消费者权益》（Concerning consumer protections in interactions with artificial intelligence systems）的法案。该法案将高风险人工智能系统定义为任何在投入使用时能够做出重大决策或在做出重大决策中起到关键作用的人工智能系统，要求高风险人工智能系统的开发者（developer）向部署者（deployer）披露有关高风险系统的特定信息、提供完成影响评估所需的信息和文件，并要求部署者实施风险管理政策和程序，定期进行影响评估，并向消费者披露其交互对象的人工智能系统的某些通知，以避免高风险人工智能系统内的算法歧视。再如，2022 年伊利诺伊州颁布了《人工智能视频面试法案》（Artificial Intelligence Video Interview Act in 2022），对使用人工智能分析视频面试的雇主提出要求。加利福尼亚州、新泽西州、纽约州、佛蒙特州和华盛顿特区也提出了相关立法，规范人工智能在招聘和晋升中的使用。

1.1.3 中国人工智能发展

2024 年 5 月，“人工智能法草案”被列入《国务院 2024 年度立法工作计划》。尽管目前我国尚未出台人工智能专门立法，但已初步建立限制和规范人工智能开发与应用的法律框架，即以《中华人民共和国网络安全法》《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》以及《中华人民共和国科学技术进步法》为基础，由《互联网信息服务算法推荐管理规定》《互联网信息服务深度合成管理规定》《生成式人工智能服务管理暂行办法》等一系列规范文件发挥主要监管作用，同时配有《互联网信息服务管理办法》《网络信息内容生态治理规定》《具有舆论属性或社会动员能力的互联网信息服务安全评估规定》等关联法律法规。

其中，《互联网信息服务算法推荐管理规定》构建了算法监管的机制，确定了算法服务的监管部门，为算法服务提供者设定了算法备案等合规义务与监管要求义务；《互联网信息服务深度合成管理规定》进一步规范了深度合成服务提供者、技术支撑者的行为；《生成式人工智能服务管理暂行办法》

对生成式人工智能服务提供者规定了数据和基础模型的来源合法、数据标注、网络安全、未成年人保护以及评估备案等一系列合规义务，明确了提供者的网络信息内容生产者责任。

随着全球各国逐步加大对人工智能技术的立法和监管力度，各种法律和监管规范不断涌现。然而，仅仅依靠法律和监管并不足以全面应对人工智能带来的技术黑箱、信息隐匿、责任错综复杂等多维度的挑战，与之对应的治理能力明显具有滞后性，可信人工智能治理已然成为全球各国确保前沿技术创新、先进法律制度发展和国家安全的核心议题。

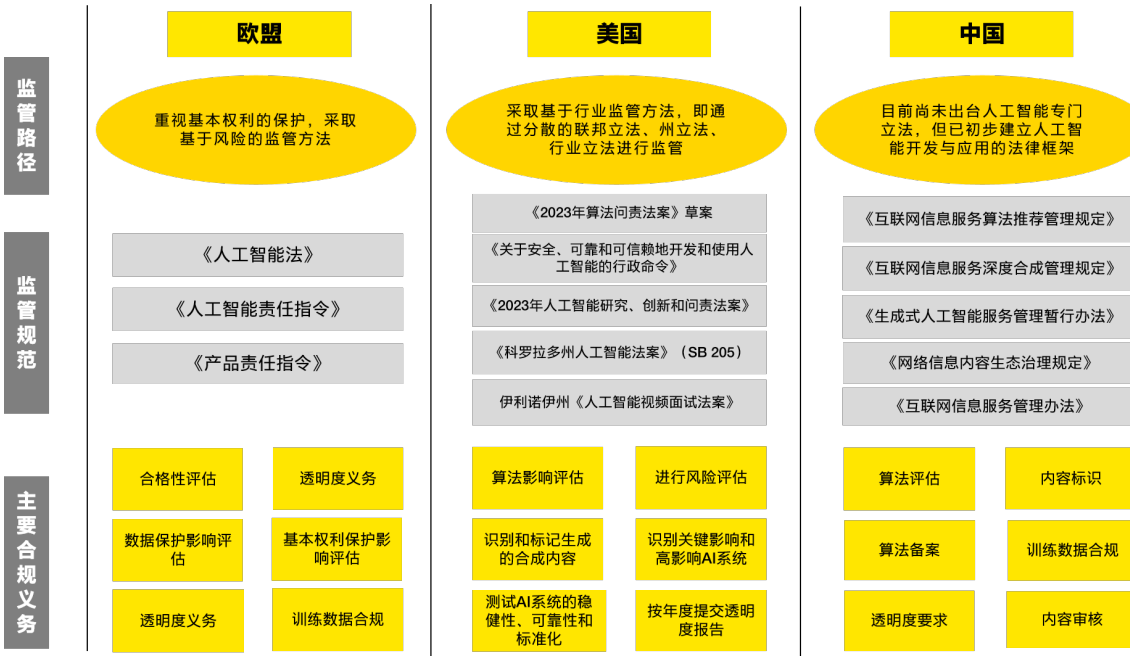


图 1 中美欧人工智能监管区分

1.2 “可信人工智能”概念提出

人工智能的可信已经成为全球各国在发展人工智能中的一致性原则。全球范围内诸多自律性标准或文件都对“ 可信人工智能”提出了具体要求。

图 2 全球自律性文件及“可信人工智能”概念

欧盟《可信人工智能伦理指南》《G20 人工智能原则》《人工智能风险管理框架》和《G7 行为准则》共同关注着人工智能的技术稳健性和安全性、AI 系统透明度、非歧视和非偏见的公平性以及可问责性。在此基础上，本白皮书将“可信人工智能”定义为设计、开发和部署技术可靠、安全，能够输出不带歧视和偏见的公正结果，基本原理和运行过程透明，可问责且以人为本的人工智能系统。

在我们开始探究可信人工智能相关问题之前，需要首先对可信人工智能提供者、部署者等概念进行界定，厘清概念之间相关性。

在人工智能的语境中，提经常被本白皮书采用欧盟《人工智能法》第三条对于提供者（provider）和部署者（deployer）的定义。

- 提供者（provider）：开发人工智能系统或通用人工智能模型，或已开发人工智能系统或通用人工智能模型，并将其投放市场或以自己的名义或商标提供服务的自然人或法人，公共机关、机构或其他团体，无论有偿还是无偿。
- 部署者（deployer）：在其授权下使用人工智能系统的任何自然人或法人、公共机关、机构或其他团体，但在个人非职业活动中使用人工智能系统的情况除外。

第二章

人工智能的可信原则

02
PART

为了确保人工智能技术的健康发展，推动其在促进科技创新和经济发展中发挥积极作用，我们必须着眼于构建可信人工智能体系。可信人工智能的核心在于，其设计、开发和部署过程中必须严格遵循一系列基本原则，包括但不限于安全性、公平性、可解释性和隐私保护。这些原则不仅关乎人工智能技术的内在可信度，更关系到其与人类社会的和谐共融。安永认为相关利益方应积极理解 AI 的核心原则，建立可信人工智能体系，构建 AI 治理与风险管理架构，并随着监管态势、科技发展、社会生产力等变化动态调整，以实现可持续发展进步。

2.1 G20 人工智能可信原则

《G20 贸易和数字经济部长声明》中提出的人工智能原则是在 2019 年由 G20 成员共同商定的一套国际性指导原则，旨在确保人工智能的发展方向能够符合人类社会的共同利益，人工智能应用是可持续、负责任和以人为本的。所有利益相关者，根据各自的角色，秉持以下原则对可信人工智能进行负责任管理。

1．包容性增长、可持续发展及福祉

应积极负责任地管理值得信赖的 AI，以追求人类和地球的有益成果，从而振兴包容性增长，可持续发展和福祉。增强人的能力和创造力，促进和包容代表性不足的人，减少经济，社会，性别和其他不平等，保护自然环境。

2．尊重法治、人权以及民主价值观，包括公平和隐私

应尊重法治、人权、民主和在整个 AI 系统生命周期中以人为本的价值观。这些包括不歧视和平等、自由、尊严、个人自主， 隐私和数据保护、多样性、公平性、社会正义以及国际公认的劳工权利。为此目的，还应实施适合于环境和符合技术水平的机制和保障措施。

3．透明度和可解释性

应致力于负责任地披露 AI 系统信息。为此，应提供适合于客观实际并符合技术发展水平的有意义的信息：促进对 AI 系统的普遍理解；让利益相关者了解与 AI 系统的互动情况；使受 AI 系统影响的人能够理解系统的结果；使受到 AI 系统不利影响的人能够根据相关因素的简单易懂的信息，作为预测、建议或决策基础的逻辑来源，并可质疑 AI 系统产生不利影响。

4．稳健性、安全性和保障性

AI 系统应在整个生命周期内保持稳健、安全和有保障，从而在正常使用、可预见的使用或误用，或其他不利条件下，能够正常运作，避免造成不合理的安全风险。还需确保如果人工智能系统有风险导致不当伤害或表现出不良行为的，可以被覆盖、修复或根据需要安全退役。在技术上可行的情况下，还应加强信息 廉洁，同时确保尊重言论自由。

5．问责制

应对人工智能系统的正常运作负责，并对人工

智能系统的正常运作负责；同时还应确保在人工智能系统生命周期中做出的数据集、流程和决策是可追溯的；人工智能系统的各方参与者有责任根据其角色，以适合于客观实际并符合技术发展水平的方式，确保 AI 系统的正常运行，将风险管理方法应用于人工智能系统的全生命周期。

2.2 欧盟可信人工智能原则

2019 年 4 月 8 日，欧盟人工智能高级别专家组发布《人工智能伦理指南》（Ethics Guidelines for Artificial Intelligence），提出了人工智能系统应满足的七项关键要求，才能被视为可信赖。

具体要求如下：

（一）人类的能动性与监督：人类应对 AI 系统拥有权力并能够监督人工智能所做的决定和行为。

（二）技术的稳健性与安全性：AI 系统应是稳健的、可复原的、安全的，出现问题时可最大限度地减少和防止意外伤害。

（三）隐私与数据治理：AI 系统需充分尊重隐私和数据保护，同时还应有适当的数据质量、完整性、合法访问等相关数据治理机制。

（四）透明度：AI 的数据、系统和商业模式应该是透明的；AI 系统的相关决策、系统功能以及局限性应对相关利益方进行解释和告知。

（五）多样性、非歧视与公平性：AI 系统必须避免歧视性以及不公平的偏见；同时，AI 系统应保持多样性，应该向所有人开放，并让相关利益方参与其中。

（六）社会和环境福祉：AI 系统应造福全人类，并考虑其他生物生存的社会影响，确保是可持续和环保的。

（七）问责制：建立问责机制，确保 AI 系统及其成果是可问责的，能够对算法、数据和设计流程进行审计，还应确保提供适当和便捷的补救措施。

2.3 安永观点：可信人工智能原则

经济合作与发展组织（OECD）、二十国集团（G20）以及欧盟提出的人工智能原则，提倡人工智能应具有创新性、值得信赖、尊重人权和民主价值观，为人工智能制定了切实可行、灵活多变、经得起时间考验的标准。体现了国际社会对人工智能技术发展的共同关注和期望，也为全球范围内人工智能的研究、开发和应用提供了一套价值观和行为准则。为此，安永在全球化的商业服务实践中，将 AI 技术与安永服务能力进行深度结合，不仅增强了客户服务的转型升级，也为可信 AI 的未来发展建立了社会信心。安永强调可信、以人为本的方法实现 AI 的潜力，专注于为所有人创造价值，以实现可持续增长并为更好的商业世界赋予人和社会力量。安永根据国际化的人工智能商业实践经验和对全球人工智能法规政策的洞察，认为具有问责性、合规性、可解释性、公平性等原则的人工智能才是可信人工智能。

OECD & G20：AI 原则	基本措施	安永：AI 原则
包容性增长、可持续发展和福祉	应积极负责任地管理值得信赖的，以人类和地球追求有益的成果，从而振兴包容性增长、可持续发展和福祉	可持续性 公平和偏见
	增强人类的能力和增强创造力	
	促进和保护代表性不足的人	
	减少经济、社会、性别和其他不平等现象	
	保护自然环境	

OECD & G20：AI 原则	基本措施	安永：AI 原则
以人为本的价值观和公平	人应尊重法治、人权、民主和在整个 AI 系统生命周期中以人为本的价值观。这些包括不歧视和平等、自由、尊严、个人自主，隐私和数据保护、多样性、公平性、社会正义以及国际公认的劳工权利。	可持续性 公平和偏见 合规
	为此目的，还应实施适合于环境和符合技术水平的机制和保障措施。	
透明度和可解释性	应致力于负责任地披露 AI 系统信息。为此，应提供适合于客观实际并符合技术发展水平的有意义的信息：	透明度
	促进对人工智能系统的全面理解	
	让利益相关者了解与 AI 系统的互动情况	
	使受 AI 系统影响的人能够理解系统的结果	可解释性
	使受到 AI 系统不利影响的人能够根据相关因素的简单易懂的信息，作为预测、建议或决策基础的逻辑来源，并可质疑 AI 系统产生不利影响。	
稳健性、安全性和保障性	AI 系统应在整个生命周期内保持稳健、安全和有保障，从而在正常使用、可预见的使用或误用，或其他不利条件下，能够正常运作，避免造成不合理的安全风险。	可靠性 安全
	还需确保如果人工智能系统有风险导致不当伤害或表现出不良行为的，可以被覆盖、修复或根据需要安全退役。	透明度
	在技术上可行的情况下，还应加强信息廉洁，同时确保尊重言论自由。	可靠性
问责制	AI 系统的各方参与者有责任根据其角色，以适合于客观实际并符合技术发展水平的方式，确保 AI 系统的正常运行，将风险管理方法应用于人工智能系统的全生命周期。	问责制

安永在全球化的商业服务实践中，将 AI 技术与安永服务能力进行深度结合，不仅增强了客户服务的转型升级，也为可信 AI 的未来发展建立了社会信心，创造指数级的价值。安永强调可信、以人为本的方法实现 AI 的潜力，专注于为所有人创造价值，已实现可持续增长并为更好地商业世界赋予人和社会力量。

图 3 安永所倡导的 AI 原则

第三章

可信人工智能的关键问题

03
PART

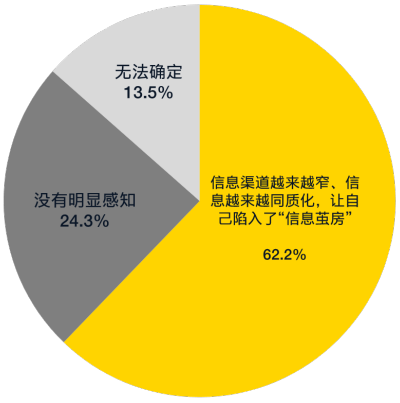
在各国人工智能法律法规的框架之下，企业对人工智能技术的应用依然面临着诸多风险与问题。一方面，数据隐私问题令人担忧，大量的个人数据被收集用于训练算法，然而这些数据的安全性和合规使用却难以得到完全保障。另一方面，算法偏见也不容忽视。由于训练数据的不全面或偏差，算法可能会对某些群体产生不公平的判断和决策。人工智能技术固有的内生安全问题可能被恶意利用，造成难以估量的危害。这些风险对企业如何更合规应用人工智能技术提出了新的挑战。

3.1 算法黑箱与信息茧房问题

2024 年，某大型网约车服务平台推出“车费保镖”服务旨在通过实时监控车费波动，为消费者提供一个公平且透明的出行体验，并在检测到异常收费时提供先行赔付。然而，上海市消费者权益保护委员会对该服务的算法透明度提出批评，指出由于缺乏明确的标准来定义“合理”的车费变动，消费者难以评估服务的实际保障水平，从而使“车费保镖”服务在实际运作中存在透明度不足的问题。

算法黑箱与信息茧房问题因其直接触及人们的日常使用体验和信息获取路径而尤为凸显。算法的不透明性常常让用户在使用搜索引擎、社交媒体和在线购物平台时感到困惑，因为他们无法洞悉背后的推荐逻辑，这无疑影响了用户体验。

中国青年报社社会调查中心对 1501 名受访者进行的调查表明，超过 62% 的受访者感受到“大数据 + 算法”精准推送带来的“信息茧房”[2]。



算法黑箱和信息茧房问题，往往根植于大型机器学习模型的内在特性。与基础算法相比，人工智能模型采用的神经网络算法的复杂性源于其拥有庞大的参数网络和处理高维数据的能力，且决策过程无法仅通过底层算法预测，而是基于训练数据深度学习到的复杂模式和规律。以个性化推荐算法为例，其依据用户行为数据进行内容定制，以此增强用户体验和平台互动性。然而，这种算法可能导致用户行为（如点击、点赞和评论）进一步强化其内部推荐逻辑，形成了一个自我加强的反馈循环。用户难以直观地查看或修改模型的内部机制，也难以理解模型如何完成文本生成、推断、分类等任务。

此外，企业在处理数据时，出于隐私保护和商业机密的考虑，往往不会完全公开其数据处理流程。同时，现行的知识产权法律在保护创新成果的同时，也可能间接限制了算法透明度的提升。

3.2 算法偏见问题

算法偏见是人工智能领域内一个日益显著的问题，由于训练数据的不均衡或算法设计的不完善，这导致系统输出结果可能带有不公平性。这种偏见可能在招聘、司法、金融等多个关键领域显现，对特定群体造成系统性的不公正影响。

2018 年某知名跨国电商企业的招聘算法被发现在筛选简历时，对含有与“女性”相关词汇的简历评分偏低，这一性别偏见问题最终导致了该算法的关闭。无独有偶，多家知名企业广泛采用的 AI 面试工具，被指出在解读面试者表情时可能存在误判，从而导致对面试者性格特征的不准确推断。

算法偏见的形成原因可以归纳为以下三个方面：

- 训练数据的不均衡性：如果训练数据在收集时存在历史偏见或代表性不足，算法可能会继承并放大这些偏见，从而产生歧视性结果。
- 特征选择的偏差：在模型训练过程中，选择哪些特征作为输入变量对结果具有重大影响，不当的特征选择可能无意中引入歧视性偏见。
- 算法设计的不完善：如果算法设计过程中未能充分考虑公平性原则，或设计者本身存在偏见，可能导致算法在处理数据时无法保持公正和客观。

3.3 数据安全问题

数据安全是构建可信人工智能的关键，它不仅保护用户隐私，确保企业遵守法律法规，还构建了用户对系统的信任。数据安全通过防止算法偏见、保障决策公正性、维护数据的完整性和准确性，支撑了人工智能的可信应用，并促进了技术与社会的和谐发展。公众对人工智能的信任建立在数据安全的基础之上，一旦发生安全事件，不仅用户隐私受损，企业的声誉和信任度也会受到严重影响。

3.3.1 底层数据源

2017 年，英国信息委员会（ICO）对伦敦某医院基金会与某科技公司旗下人工智能研究公司的智能医疗合作项目进行了调查，调查结果显示，在开发移动医疗应用的过程中，该人工智能研究公司未能充分向患者披露其个人医疗数据的使用目的和范围，使用超过 160 万患者的医疗记录作为训练和测试的数据源 [3]。数据源获取的合法性问题可能触发一系列复杂问题，包括知识产权侵权、公民个人信息保护不足等。数据作为用户与模型之间的沟通桥梁，其质量和来源直接影响到人工智能模型的性能表现。如果数据来源不明确或获取方式存在法律问题，那么在使用这些数据的过程中，就难以满足现行的数据保护法规要求。这不仅会损害用户的利益，还可能对人工智能行业的健康发展造成负面影响。

底层数据源的安全问题是一个多维度的挑战，它不仅包括数据源获取的合法性，还涵盖了数据语料库的管理、数据投毒攻击、数据源的依赖性和时效性等。庞大的数据语料库可能潜藏着仇恨言论、歧视性言论、虚假信息等有害内容，这些内容若被恶意注入，不仅会误导人工智能的学习过程，损害模型的准确性，还可能对用户的利益造成实质性的威胁。同样，确保数据源的多样性和时效性至关重要，如果算法过度依赖单一或有限的数据源，其鲁棒性将受到削弱，而依赖过时的数据源，使用不再反映现实的陈旧数据也会降低人工智能决策的准确性和可靠性。

3.3.2 数据泄露

2023 年 4 月，全球著名的商业集团因允许其半导体部门工程师使用 ChatGPT 解决源代码问题，不慎泄露包括半导体设备测量资料和产品良率等在内的机密数据。根据美国数据安全公司 Cyberhaven 发布的报告[4]，到 2024 年 3 月，员工向人工智能模型输入的公司数据中，敏感数据的比例从 10.7% 激增至 27.4%。尽管企业有责任严密保护机密信息，但员工可能因缺乏安全意识在无意中将机密数据以提问的形式输入人工智能模型中，致使企业商业机密面临持续泄露风险。

人工智能模型的反向工程是一种高级的数据泄露风险，攻击者通过精心设计的查询，分析人工智能模型的输出结果，从而推断出用于训练该模型的敏感数据。由于人工智能模型可能记住了训练过程中的特定模式和细节，攻击者利用这些信息，可以对模型进行“反向学习”，恢复或重建原始数据集的一部分，对数据隐私和安全构成严重威胁。

3.3.3 出海涉及相关数据出境规制

在全球化大潮中，人工智能系统处理的数据频繁跨越国界，带来了跨境数据传输与处理的挑战。数据主权作为国家对数据控制权的体现，涵盖了数据的收集、存储、处理和传输等生命周期各环节。个人信息出境要求，作为数据主权的重要组成部分，规定了个人信息在跨国传输时必须遵循的规则。为了保护本国数据安全、维护国家利益和公民隐私，各国纷纷制定并实施了相应的法律法规。

我国通过《中华人民共和国网络安全法》《中华人民共和国数据安全法》和《中华人民共和国个人信息保护法》等法律法规，确立了个人信息出境的合规要求。这些法规要求企业在处理数据出境事宜时，必须开展安全评估，确保数据接收方能够提供充分的数据保护措施，并且必须获得数据主体的明确同意。遵循这些法规不仅是企业履行法律义务的需要，也是维护企业声誉、保持客户信任的关键。

3.4 内生安全问题

人工智能技术的内生安全问题根植于系统内部，从基础设施的脆弱性、数据处理的漏洞，到开发框架的缺陷、模型训练和部署的不规范，以及应用层和接口层的攻击风险。这些隐患不仅威胁到系统的安全性和用户隐私，还可能导致合规性问题和声誉损害。安全层的不足和监控维护的滞后可能使系统对新威胁的防御能力下降，而用户交互层的安全问题将直接影响用户信任。若不妥善解决，这些风险可能引发错误决策、信任动摇、数据泄露、安全性下降，以及技术更新滞后，对企业运营和用户权益造成严重影响。

3.4.1 人工智能基础设施安全风险

人工智能基础设施的安全风险包括软件漏洞、依赖性风险、技术过时和网络攻击，这些问题共同威胁着系统的稳定性和可靠性。

软件层面，操作系统、数据库和应用程序中的未修复漏洞可能使系统面临未授权访问和破坏的风险，而维护更新不足会加剧这一问题。对特定软件库或框架的过度依赖可能引入依赖性风险，一旦这些依赖项存在问题，整个系统将受影响。使用过时的技术因缺乏最新安全功能易于遭受攻击。此外，网络攻击如 DDoS、端口扫描和未授权访问会直接影响系统的可用性和稳定性。

3.4.2 数据安全风险

人工智能系统的数据泄露风险遍布数据处理各环节，尤其在安全防护不足时更为严峻。原因多样：数据加密不足易遭截获；软件漏洞可能被攻击者利用；不安全的系统配置如开放端口和简单密码易于被攻破；内部人员的疏忽或故意行为可能引起数据泄露；员工和用户的数据保护意识不足，易成为网络攻击目标；数据访问控制宽松，敏感信息面临未

授权访问风险；日志和监控机制不足，难以及时检测泄露。这些因素交织，形成数据泄露的复杂风险环境。

3.4.3 模型安全风险

在人工智能领域，模型安全风险是核心问题，包括对抗性攻击、数据投毒、模型窃取、算法后门和模型过拟合等方面。

对抗性攻击利用构造输入误导模型，数据投毒通过篡改训练数据扭曲学习过程，模型窃取尝试逆向工程获取内部参数，算法后门植入隐蔽触发条件，而过拟合导致模型泛化能力差。这些问题往往源于数据管理不足、安全措施缺失、技术缺陷、计算资源限制、透明度和监管不足、复杂性增加、验证和测试不完善以及内部威胁等复合因素。随着攻击技术的发展，模型安全问题变得更加隐蔽和难以防范，要求在数据管理、模型设计、训练和验证各环节加强安全防护。

3.4.4 应用服务安全风险

在人工智能领域，应用服务的安全风险是多方面的，涉及复杂的技术和伦理问题。用户与人工智能服务的互动可能伴随着隐私泄露的风险，用户可能无意中分享了敏感信息，或者人工智能系统可能在缺乏告知同意的情况下收集数据，可能违反数据保护法规。此外，用户反馈对人工智能系统既是改进服务的宝贵资源，也可能被恶意用户可能利用从而进行数据操纵或探测系统缺陷。而用户提问中可能包含的不当言论或虚假信息，不仅违背平台规定，也可能对社会秩序和个人权益造成损害。

在内容生成与决策方面，人工智能模型可能产生带有偏见或不准确的信息。同时，人工智能技术的滥用，如深度伪造（Deepfake）技术或自动化网络攻击，对个人名誉和网络安全构成了严重威胁。人工智能模型的决策安全在医疗、金融、交通等领域至关重要，这些领域的人工智能技术应用一旦做出错误决策，可能会引发包括经济损失、健康风险甚至生命安全在内的严重后果。

3.5 科技伦理问题

当前以人工智能等核心科技的新一轮科技革命和产业变革加速演进，在生命科学、医学、人工智能等领域的科技活动的伦理挑战日益增多，AI 深度生成、基因编辑、自动驾驶等人工智能技术正面临科技伦理的拷问与审视。人工智能大模型的训练、推理、应用等活动依赖大量的数据来训练，其训练决策的过程复杂且不透明，可能会引发如下科技伦理问题：

- 自然人的知情权、选择权、公平交易权以及未成年人和老年人群体的合法权益受损，从而导致大数据杀熟、沉迷消费、服务歧视等伦理问题；
- 特定行业和企业在利用人工智能技术时缺乏科技伦理审查所引发产品缺陷、侵犯用户合法权益、使用人工智能技术进行违法犯罪活动等法律风险；
- 人工智能技术的伦理问题也可能引发社会群体性失业、弱势群体不公平待遇进一步加剧等社会性的伦理问题。

企业若违反科技伦理不仅影响企业声誉，更会面临行政责任乃至刑事责任。2023 年 12 月 1 日正式施行的《科技伦理审查办法（试行）》为规范科学研究、技术开发等科技活动，强化科技伦理风险防控，提供了法律依据。

第四章
企业级 AI 的合规要求

04
PART

在全球人工智能监管日益趋强的背景下，为更好应对算法黑箱、算法偏见和外部攻击等纷繁复杂的内忧外患，相关企业也面临着从准入资质、算法合规义务到生成内容合规义务的一系列要求，以确保企业人工智能相关的发展和应用合法合规。

4.1 资质监管要求

人工智能相关企业的业务可能既包括信息服务和电子邮件、短信服务等增值电信业务，也涉及网络新闻、网络文化、网络出版等内容，因此会涉及诸多资质。基于此，白皮书将相关资质分为互联网信息服务提供者的基础资质和提供具体服务的特殊资质两大类，并重点说明相关企业应当注意的监管要求。

4.1.1 互联网信息服务提供者的基础资质

通过互联网向上网用户提供信息的服务提供者应当首先具备互联网信息服务的基础资质。常见的基础资质包括互联网信息服务经营许可证（ICP 许可证）、增值电信业务经营许可证 - 在线数据处理与交易处理业务（EDI 许可证）和增值电信业务经营许可证 - 移动网信息服务业务（SP 许可证）等。

通过互联网进行电子公告服务和电子邮件、短信服务等增值电信业务的经营性互联网信息服务提供者，应当向省、自治区、直辖市电信管理机构或者国务院信息产业主管部门申请办理互联网信息服务增值电信业务经营许可证。从事在线数据处理和交易处理业务、移动网信息服务业务等电信增值服务的服务提供者还应当提交满足申请条件的材料，以在业务开展前期获得 EDI 许可证和 SP 许可证。

4.1.2 提供具体服务的特殊资质

互联网新闻信息服务、网络出版服务、网络文化活动和网络视听节目服务、生成式人工智能服务等具体服务的提供者，应当相应取得互联网新闻信息服务许可、网络出版服务许可证、信息网络传播视听节目许可证等特殊资质。

以互联网新闻信息服务为例，“互联网新闻信息服务”包括互联网新闻信息采编发布服务、转载服务和传播平台服务。申请主体需为通过互联网站、应用程序、论坛、博客、微博客、公众账号、即时通信工具、网络直播等形式向社会公众提供互联网新闻信息服务的主体，包括提供互联网新闻信息服务的算法推荐服务提供者、移动互联网应用程序提供者等主体。这些企业应当向特定主管部门，即省、自治区、直辖市互联网信息办公室提出申请。同样，从事网络出版服务、网络文化活动和网络视听节目服务的深度合成服务提供者和技术支持者取得网络出版服务、网络文化活动和网络视听节目服务相关许可；生成式人工智能服务提供者则应当取得生成式人工智能服务相关行政许可。

4.2 算法合规要求

在《中华人民共和国网络安全法》《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》三驾马车的引领下，算法合规要求的主要依据来源于《算法推荐管理规定》《深度合成管理规定》和《生成式人工智能服务管理办法》的监管规范，并初步形成了算法合规义务清单。

图 4 算法合规义务清单

本节主要解读算法备案、算法评估和算法透明性要求，聚焦企业的基础合规要点与实践操作指引。

4.2.1 算法备案要求

算法备案制度要求具有舆论属性或者社会动员能力的算法推荐服务提供者、深度合成服务提供者、深度合成服务技术支持者和生成式人工智能服务提供者，按照法律程序就其所使用的算法推荐技术、深度合成技术和生成式人工智能技术进行备案。

其中，“具有舆论属性或者社会动员能力”是指以下两种情形：（1）开办论坛、博客、微博客、聊天室、通讯群组、公众账号、短视频、网络直播、信息分享、小程序等信息服务或者附设相应功能（2）开办提供公众舆论表达渠道或者具有发动社会公众从事特定活动能力的其他互联网信息服务。算法推荐服务提供者是指利用生成合成类、个性化推送类、排序精选类、检索过滤类、调度决策类等算法技术向用户提供信息的组织或个人；深度合成服务提供者是指通过利用深度学习、虚拟现实等生成合成类算法制作文本、图像、音频、视频、虚拟场景等网络信息的技术来提供服务的组织或个人；技术支持者是指为深度合成服务提供技术支持的组织、个人；生成式人工智能服务提供者是指利用生成式人工智能技术提供生成式人工智能服务（包括通过提供可编程接口等方式提供生成式人工智能服务）的组织、个人。

符合前述条件的企业应当通过互联网信息服务算法备案系统填报服务提供者的名称、服务形式、应用领域、算法类型、算法自评估报告、拟公示内容等信息，以履行备案手续。

4.2.2 算法评估要求

根据评估主体的不同，算法评估可以区分为自评估和监管机构评估。自评估是指具有舆论属性或者社会动员能力的算法推荐服务提供者、深度合成服务提供者和技术支持者、具有舆论属性或者社会动员能力的生成式人工智能服务提供者、上线具有舆论属性或者社会动员能力的新技术、新应用、新

功能的应用程序提供者、互联网新闻信息服务提供者、互联网信息服务提供者等主体，就其所应用的技术自行组织开展或委托第三方机构进行安全评估。监管机构评估是指网信、电信、公安、市场监管等有关部门就相关企业所应用的技术自行组织或委托第三方机构开展安全评估。

从评估对象角度，算法评估包括针对服务和技术的评估。针对算法服务的评估内容主要包括算法梳理情况、对算法服务基本原理、目的意图和主要运行机制的公示、安全风险、可恢复性、安全应急处置机制、个人信息收集和使用的安全管理、用户选择的弹窗等设置以及投诉举报反馈机制。针对算法技术的评估内容主要包括算法的通用条款、设计开发、验证确认、部署运行、维护升级和退役下线的具体技术情况。

因此，算法服务提供者应当开展算法自评估活动，就其所提供的服务进行安全评估。算法自评估报告应当覆盖下列信息：第一，算法基础属性信息：选择所需备案算法的算法类型、应用领域、使用场景等，并填写《算法安全自评估报告》和《拟公示内容》；第二，算法详细描述信息：填写算法简介、应用范围、服务群体、用户数量、社会影响情况、软硬件设施及部署位置等信息；第三，算法风险描述信息：根据评估算法特点，确定可能存在的算法滥用、恶意利用、算法漏洞等安全风险，并加以描述。

4.2.3 算法透明性要求

算法监管规范将企业的透明性义务表述为“告知”，即算法推荐服务提供者应当以显著方式告知用户其提供算法推荐服务的情况，并以适当方式公示算法推荐服务的基本原理、目的意图和主要运行机制等；深度合成服务提供者和技术支持者提供人脸、人声等生物识别信息编辑功能的，应当提示深度合成服务使用者依法告知被编辑的个人，并取得其单独同意。

实践中，企业作为服务提供者应当重视相关规则的透明化和可解释化，在提供算法相关服务时，以显著的方式告知用户基本情况，并以可理解的语言表述公示算法的基本原理、使用目的和运行情况。需要取得同意的，应当及时告知用户并征得其明确单独的同意。

4.3 内容合规要求

我国重视内容治理，多部规范就生成内容为企业设定了合规要求，具体包括内容审核义务、内容标识义务和投诉处理机制。

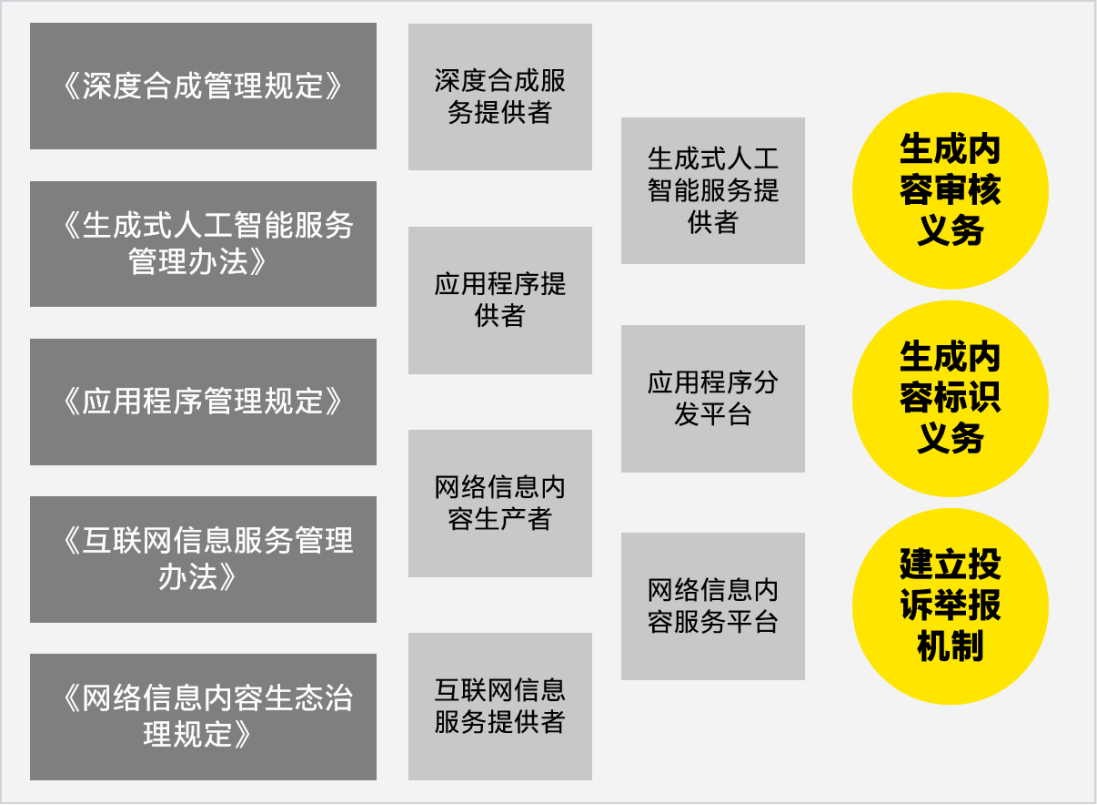


图 5 人工智能生成内容的合规要求

4.3.1 生成内容审核义务

从义务主体角度，深度合成服务提供者、生成式人工智能服务提供者、应用程序提供者、应用程序分发平台、网络信息内容生产者、网络信息内容服务平台和互联网信息服务提供者都应当对网络信息内容负有责任。依据其性质，可以将这些主体分为网络信息内容生产者、网络信息内容服务平台和网络信息内容服务提供者三类。

其中，网络信息内容生产者就其制作、复制、发布相关信息承担责任；网络信息内容服务平台主要就其提供的发布、下载、动态加载等网络信息内容传播服务承担责任，即信息内容管理主体责任；网络信息内容服务提供者则根据情况承担不同的责任，例如，生成式人工智能服务提供者依法承担网络信息内容生产者责任，深度合成服务提供者则承担信息内容管理主体责任，采取技术或者人工方式对深度合成服务使用者的输入数据和合成结果进行审核。

目前，内容审核义务主要由网络信息内容服务平台承担。2021 年 9 月 15 日，国家互联网信息办公室发布《关于进一步压实网站平台信息内容管理主体责任的意见》，明确要求网络信息内容服务平台充分发挥网站平台信息内容管理第一责任人作用。因此，网络信息内容服务平台企业应当健全内容审核机制，依据意见落实总编辑负责制度，明确总编辑在信息内容审核方面的权利和所应承担的责任，以建立总编辑对全产品、全链条信息内容的审核工作机制。

4.3.2 生成内容标识义务

内容标识义务主要围绕合成内容展开。算法推荐服务提供者负有对未作显著标识的算法生成合成信息进行标识的义务；深度合成服务提供者对使用其深度合成服务生成或者编辑的信息内容，以及所提供的特定深度合成服务可能导致公众混淆或者误认的，都负有标识义务；生成式人工智能服务提供者负有对图片、视频等生成内容进行标识的义务。

内容标识的对象包括算法生成合成信息、深度合成的信息以及生成内容，据此分析，标识的目的在于说明相关内容非真人产出的情况。因此，企业应当重视内容输出时的标识要求，根据具体情况采用显式水印标识或隐式水印标识，对人工智能生成内容标识“由 AI 生成”或“人工智能为您提供服务”等字样。

4.3.3 建立投诉举报机制

投诉举报机制是保护用户权益和体现透明度的有效路径。投诉举报不仅是针对算法推荐内容、深度合成内容和人工智能生成内容的内容合规的重要路径，同样也是应用程序合规的基本要求。企业应当设置便捷的用户申诉和公众投诉、举报入口，公布投诉举报方式，公布处理流程和反馈时限，及时受理、处理公众投诉举报并反馈处理结果。

第五章
人工智能风险治理

05
PART

5.1 风险治理架构

人工智能风险治理架构旨在构建一套周详而立体的体系，强化对人工智能潜在风险的管控效能，并提升人工智能技术的可靠性与责任感。该架构由若干紧密交织的层面组成，每一层面均针对人工智能治理的特定维度进行深入考量和精细规划，从而确保整个架构的协同运作与高效实施。通过这种多维度的治理机制，我们期望实现对人工智能技术发展与应用的全方位引导，确保其在遵循伦理准则和社会责任的前提下稳健前行。

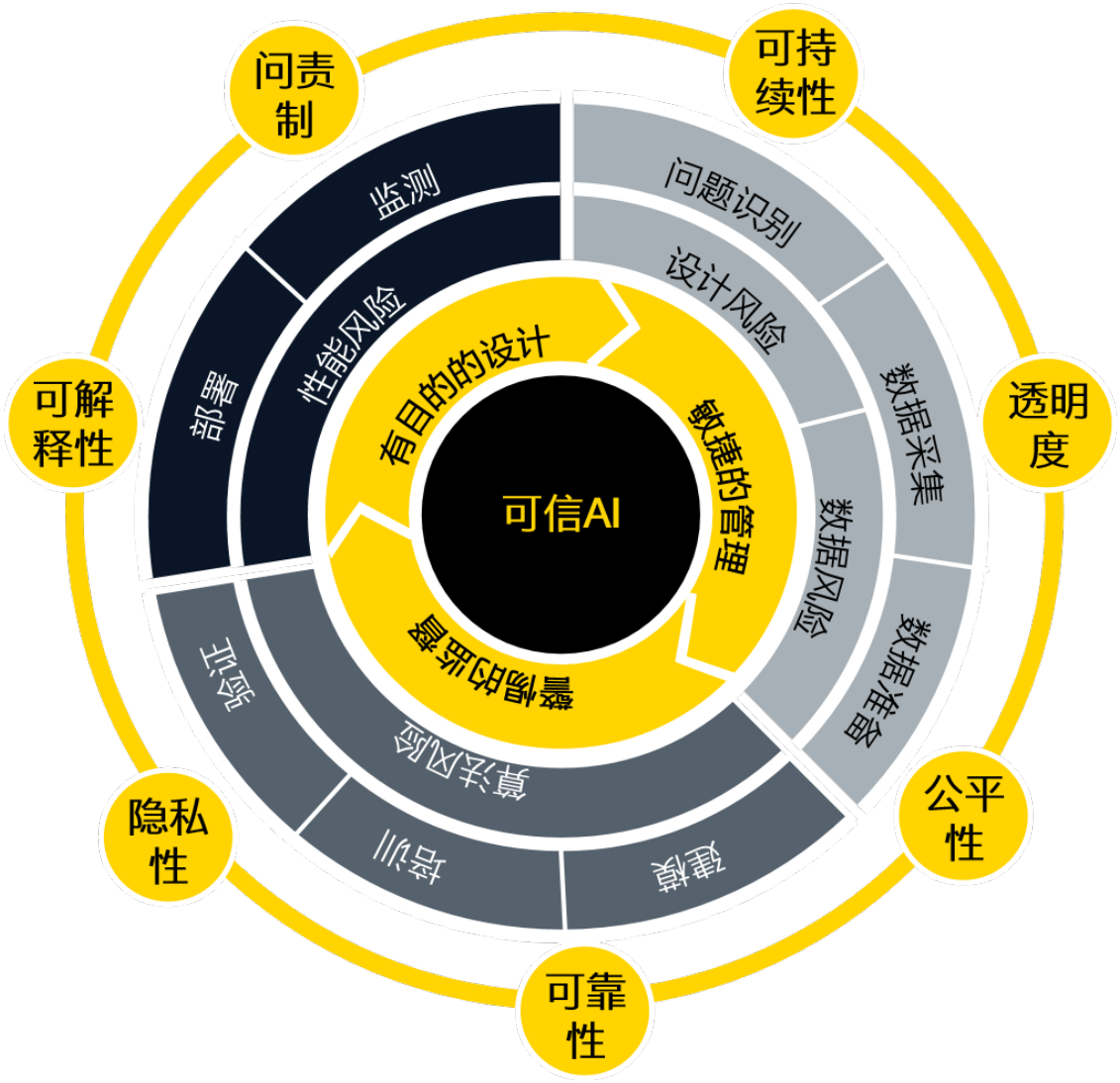


图 6 人工智能风险治理框架

5.1.1 风险治理架构对企业的指导

人工智能技术的崛起为企业提供了巨大的机遇，但同时也带来了复杂的风险挑战。因此，一个全面的人工智能风险治理架构对于企业来说至关重要，它不仅关系到企业的合规性、安全性、社会责任和商业连续性，还直接影响到企业的运营效率、决策质量、创新能力和市场声誉。人工智能风险治理架构是企业应对人工智能相关风险的基础，它包括风险识别、评估、监控和缓解等多个环节，涵盖了数据管理、模型开发、技术应用和伦理合规等多个方面。

该治理架构为企业提供了以下指导和建议：

- 在数据管理上建立严格的政策和流程，确保数据的质量和安全，同时符合数据保护法规。这不仅提高了数据的可靠性，也为人工智能模型的开发和部署提供了坚实的基础。
- 确保人工智能模型的开发过程遵循公平性、透明度和可解释性原则，避免算法偏见和歧视。这要求企业在模型设计、训练和验证阶段进行严格的风险评估和测试。
- 建立实时监控系统，对人工智能系统的性能进行持续评估，及时发现并应对潜在风险。这不仅提高了系统的稳定性和可靠性，也为风险管理和决策提供了实时数据支持。

人工智能风险治理架构对企业的影响是多维度的，它不仅塑造了企业运营的稳健性，也是推动创新和维护声誉的关键因素。企业需要认识到风险治理的重要性，并积极构建和实施这一架构，以实现人工智能技术的健康发展和商业价值的最大化。通过有效的风险治理，企业不仅能够提升自身的竞争力，也能够为社会的可持续发展做出贡献。

人工智能风险治理架构为企业提供了一个结构化的决策支持系统。通过风险识别和量化，企业能够更准确地评估人工智能项目的风险和收益，做出更加科学和合理的决策。此外，该架构还强调了风险沟通的重要性，通过提高企业内部和外部的风险透明度，增强了公众对企业决策的信任，为企业树立了积极的社会责任形象。

同时，人工智能风险治理架构为企业在人工智能领域的创新尝试和探索提供了一个安全的环境。通过平衡风险管理和创新需求，企业能够在保障安全的前提下，实现可持续的创新发展。这种平衡策略不仅有助于企业在市场中保持领先地位，也是企业声誉和品牌忠诚度提升的基石。

积极的人工智能风险治理还使企业能够在面临危机时快速响应和有效应对，减少对企业声誉的负面影响。这种危机管理能力是企业声誉保护的关键，有助于企业在复杂多变的市场环境中保持稳定发展，赢得消费者和社会的长期信任。

5.1.2 风险治理架构对创新的推动

在技术创新方面，人工智能风险治理架构也有着深远影响。它首先确保了技术创新活动在安全可控的范围内进行，避免潜在的伦理和法律风险。通过制定相应的政策和标准，人工智能风险治理架构能够引导技术开发者遵循最佳实践，从而提高技术创新的质量和可靠性。

进一步地，人工智能风险治理架构通过识别和评估技术创新的潜在风险，促进了技术创新的健康发展。该架构赋予了相关部门前瞻性地采取措施的能力，以减少负面影响，保护公众利益，并通过维护市场秩序和增强消费者信心，为人工智能创新发展注入了持续的活力。

此外，人工智能风险治理架构在全球化背景下尤为重要，其促进了跨国企业间的交流与合作。通过共同制定和遵守统一的治理架构，各国企业可以加强信息共享和经验交流，共同应对全球性挑战，

推动全球技术创新的进步。

综上所述，人工智能风险治理架构对技术创新具有积极的推动作用，它不仅能够保障技术创新的安全性，还能够促进技术创新的健康发展和国际合作，为全球人工智能技术创新的繁荣贡献了力量。

5.1.3 风险治理架构对市场的维护

人工智能风险治理架构在维护市场公平竞争中也发挥着至关重要的作用。首先，它有助于确保人工智能技术的合规性，防止滥用人工智能造成市场垄断或不正当竞争。通过制定严格的伦理标准和监管措施，该架构能够限制企业利用算法优势排挤竞争对手，维护市场的公平性。

其次，人工智能风险治理架构通过增强透明度和问责制，加深了市场参与者之间的信任。透明的决策过程和明确的责任归属使消费者和其他利益相关者更容易认可人工智能市场的公平性，这对于保持市场秩序和稳定至关重要。

最后，人工智能风险治理架构还有助于应对数据安全、隐私侵犯等潜在风险和挑战。通过预防和解决这些问题，该架构不仅保护了消费者权益，也维护了整个市场生态的健康发展。

综上所述，人工智能风险治理架构通过确保技术合规性、促进透明度和问责制、以及有效应对风险挑战，共同营造了一个公平、透明和健康的市场环境。这些措施不仅维护了市场的竞争秩序，也为所有参与者创造了一个更加稳定和可预测的商业氛围。

5.2 生命周期风险治理

图 7 人工智能生命周期及对应风险模型

管理人工智能生命周期是确保技术可信、合规和有效的关键。问题识别是人工智能项目启动的初期阶段，主要工作包括收集和分析用户及业务需求，梳理合规要求以初步评估风险，设计系统架构，规划必要资源，制定项目计划，为项目的顺利实施打下坚实基础。

在明确人工智能的设计需求后，就要开始准备训练人工智能所需的数据。这一阶段的工作重点在于合法合规地收集数据，并通过清洗和标注提高数据质量，同时进行特征工程以增强模型的预测能力。在开展模型训练前，要选择合适的算法，进行调优，并在后续训练和验证过程中进一步评估模型性能，确保其准确性和泛化能力。

进入模型部署阶段，工作转向确保模型在生产环境中的稳定运行，包括制定和实施部署策略。在模型稳定运行后，则侧重于监控模型性能，收集用户反馈，并及时响应任何性能下降或偏差。随着环境变化，模型可能

会经过变更完成迭代，这一过程需要优化模型以适应新数据或业务需求，同时进行影响评估和测试以确保变更的有效性。最终，若模型不再使用，还需考虑知识保留和数据存档，确保平稳过渡和历史数据的安全性。

管理人工智能生命周期中的风险是实现监管机构所提倡的可信人工智能的关键。在整个人工智能生命周期中，风险无处不在，企业需要从生命周期的各个阶段入手，建立全面的风险治理机制，确保人工智能系统在设计、开发、部署和运行的每个环节都符合监管要求，减少风险发生的可能性，提高系统的可靠性和用户的信任度。这不仅有助于保护用户隐私和数据安全，也有助于企业在快速变化的市场中保持竞争力，实现可持续发展和可信人工智能的目标。

5.2.1 风险管理左移

风险管理左移旨在将风险评估和缓解提前到项目生命周期的起始阶段，其核心在于尽早识别潜在风险。企业需要在规划和设计阶段深入分析项目需求，评估技术可行性，预测资源需求，并识别潜在合规性与伦理问题。这种做法有助于企业制定全面的预防策略，减少后期返工和成本，确保项目顺利推进。

在项目启动初期，企业必须进行全面的合规性审查，确保符合法律法规、数据保护、隐私标准、知识产权和伦理准则。这一过程旨在预防法律和合规风险，确立合法的项目设计和实施框架，增强各利益方的信任。同时，企业应制定合规策略和控制措施，保障项目持续遵循规定，为项目的顺利进行和成功交付奠定基础。

当企业考虑与第三方合作人工智能项目时，需要对其资质、信誉和技术能力进行严格评估和审查，并签订服务水平协议、合同等明确各自的责任和义务。同时，企业还需建立一套有效的监督机制，确保第三方服务提供商的行为符合企业的风险管理政策和标准。

高质量数据集的构建是保障可信人工智能准确性、公平性的基础。企业需确保数据的合法合规采集、数据源的可信度和透明度以及数据集的多样性和代表性；同时，还应开展数据清洗、预处理和标注工作，辅以机器 / 人工审核和校对，提升后续模型训练的效率。在进行数据集管理时，企业应详细记录数据的来源、构成情况以及预处理操作，并进行文档化记录，确保数据的可追溯性和透明度。为提高模型的时效性和预测能力，企业还应加强数据集的可扩展性和可维护性，定期更新数据集的同时完善数据集的版本控制。此外，数据安全与隐私保护是数据集管理的另一核心，企业需通过加密和访问控制等措施来保护数据，并定期开展合规性与伦理审查。

5.2.2 敏捷管理模式

敏捷管理模式强调快速响应和持续改进，风险治理需与这一理念相匹配，确保在整个生命周期中，风险得到及时识别、评估和控制。

首先，在进行模型选择时，需要评估不同模型的理论基础、性能指标和适用性，选择在准确性、泛化能力、鲁棒性、可追溯性和可解释性方面表现最佳的模型。同时，考虑到潜在的风险，如数据偏见、模型复杂性和对抗性攻击等脆弱性，要选择在源头上最小化这些风险的模型。此外，模型选择还应考虑合规性和伦理性，确保所选模型符合法律法规和行业标准，且避免对特定群体产生不公平的影响。对于训练、优化模型所使用的算法，企业需要深入理解它们的优缺点和适用场景以选择最佳方案。在此基础上，设计算法原型并进行初步测试，以评估算法在实际应用中的可行性和性能。同时，要在算法开发的早期阶段就进行机制机理的审核，确保算法的目的、采用的技术、干预的方式等均符合预期目标和伦理标准。

在模型训练阶段，有效的模型训练管理确保了

人工智能模型的高效开发和性能优化。这一过程包括资源合理分配、自动化超参数调优、实施训练指标监控等，并采取正则化和数据增强策略防止过拟合。同时，维护数据质量、实施版本控制、保存模型状态、记录详细日志，以及确保训练过程的安全性，都是提升模型可复现性、可追踪性和透明度的基础措施。

在人工智能模型部署前进行测试是确保其安全性和可靠性的一项关键措施。通过功能、性能、安全性测试和合规性检查，可以对模型的准确性、稳定性、透明度、可解释性、隐私保护和公平性开展综合评估。其目的在于发现并修复模型中的漏洞，保证其在面对各种可能的攻击和滥用时的鲁棒性。同时，制定灵活的部署策略至关重要。可以通过模块化部署等方式确保模型与现有系统的兼容性。通常建议企业实施隔离部署，将人工智能系统或关键组件部署在隔离环境中，以控制交互风险。此外，通过自动化监控和弹性扩展机制动态调整资源，可以应对用户和数据量的增长，保障模型性能和系统稳定性。

企业应及时收集和分析用户的反馈信息，并根据反馈内容对模型进行敏捷迭代和优化，使其更符合用户需求和市场变化；同时，定期使用新数据对模型进行重新训练，以适应业务需求变化，确保系统的稳健运行。在迭代过程中，企业应考虑对模型变更的全面控制，从变更请求的提出到审批、实施、测试，直至最终部署。迭代管理流程要求所有变更活动都必须经过严格的审批，以确保变更的必要性和合理性。同时，它还包括对变更影响的评估，确保变更不会对现有系统造成不利影响，减少因变更引入的风险。

若模型不再使用，企业需确保人工智能系统安全、合规地结束生命周期。这包括数据和模型存档、法律合规性审查、数据保护措施、技术债务清理，以及及时通知用户和利益相关方等。这些措施应保障模型退役的有序性，最小化对业务的影响，维护企业责任和声誉。

5.2.3 审慎评估及监控

审慎评估及监控能够使企业及时发现人工智能系统使用过程中的潜在风险并加以应对，保护企业和用户的利益，同时促进人工智能技术的健康发展。

企业应定期开展模型风险评估，并进行模型风险的分类与分级，识别不同风险类型，根据其潜在影响的严重性和发生概率进行分级。在综合性的模型风险评估后，要在模型训练阶段缓解潜在的风险，并在部署前完成模型核验。为提高模型的可靠性和用户的满意度，必要时可依赖第三方机构开展风险评估和核验工作。此外，企业可依据风险评估结果对不同人工智能模型进行分级管理，确保高风险模型受到更严格的监控和控制，而低风险模型则可以采取相对宽松的管理措施。通过这种基于风险的分级管理，企业可以更有效地分配资源，优化风险控制策略。

企业还应建立全面的监控体系，实时监测人工智能模型状态，并通过日志分析快速定位问题和风险点，以及时发现并修复模型中的缺陷或优化模型的性能和效率。同时，加强输入验证和输出管控，使用自动化测试和人工审核相结合的方式对数据的真实性、完整性和准确性进行严格检查。其中，输入验证既能避免使用者输入敏感数据，也能预防提示词注入等攻击；而输出管控核心在于关注模型结果是否符合预期标准，同时监控训练数据的泄露情况。通过建立追溯机制和审计机制，确保模型接收正确数据、输出有效内容。

在依赖关系评估方面，企业需要识别和理解模型与其他系统组件之间的相互依赖性，彻底审查模型在生产环境中的使用情况，包括它所服务的应用程序、工作流以及与哪些数据源和下游系统相连接。

通过这种评估，可以确定模型变更或退役可能对业务流程、用户体验或数据完整性造成的潜在影响，使所有相关方都意识到即将发生的变更，并采取相应的预防和应对措施。

此外，企业还应建立应急响应机制，针对人工智能系统相关突发事件的预防和准备，如数据泄露、系统故障等。企业应制定应急预案，以便在发生风险事件时能够迅速作出反应，并定期对应急预案进行检查和更新，确保其始终保持有效性。

5.3 人员风险治理

人员相关风险治理关注的是确保人工智能系统的开发者、使用者和监管者具备足够的知识和能力来理解、管理和监督人工智能系统。这包括对人员进行适当的培训、教育和认证，以及建立有效的监督和问责机制。

对企业内部员工进行人工智能伦理和风险意识的培训，可以培养他们在日常工作中主动识别和防范潜在风险的能力。培训内容通常包括数据隐私保护、算法公平性、安全合规等关键议题。此外，还应开展人工智能使用技巧培训，宣贯使用过程中的注意事项，提高团队的适应性的同时避免员工对人工智能系统的滥用。

为确保员工在使用人工智能时的行为得到有效监督和问责，企业应确立明确的 AI 使用政策和指导原则，这些原则需要明确界定员工的责任和行为标准。同时，企业需要部署监控系统来跟踪员工对 AI 系统的操作，利用日志管理工具记录关键活动，以便在需要时进行行为分析和异常检测。所有操作都应有详细记录，确保可追溯性，为审查和问责提供依据。企业还应考虑建立一个安全、匿名的报告机制，鼓励员工报告任何不当行为或安全问题，并确保这些问题得到迅速公正地处理。

此外，企业需要向员工强调人工智能使用过程中的知识保留。这包括捕捉和保存与模型相关的技术和业务知识，详细记录模型的开发背景、功能、性能、使用案例以及任何相关的业务逻辑和决策过程。通过文档化这些信息，企业可以为未来可能的模型重建或优化提供参考。知识保留还涉及对团队成员经验和见解的收集，以及对模型在生产中表现的评估和总结。这有助于新团队成员的学习，以及在类似项目中避免重复过去的错误。

最后，通过跨部门的协作，网络安全、人力资源、法务和 IT 等部门应共同参与到人员风险治理中，形成一个全面的治理体系，以维护人工智能使用的安全性和合规性。

第六章
企业 AI 治理进阶工具

06
PART

除了了解人工智能相关风险和管理思路，如何将 AI 治理工作体系化，通过科学的方式管理 AI 风险，对标法律法规及行业标准设计落实管控措施，持续监控和改进 AI 治理体系，并获得外部专业机构的认证，对于企业来说尤为重要。通过体系化建设，企业不仅能够充分利用 AI 技术的潜力，还能够负责任地管理其带来的挑战和机遇。

6.1 AI 治理国际标准

6.1.1 ISO 42001 的意义与价值

ISO/IEC 42001:2023 人工智能管理体系（AIMS）标准由国际标准化组织（ISO）和国际电工委员会（IEC）合作开发，并于 2023 年正式发布，是世界上第一个人工智能管理体系标准。该标准旨在为组织提供 AI 治理的框架，以负责任和有效地管理其人工智能技术。该标准框架不仅覆盖了 AI 的研发、部署、运营及监控等各个环节，而且重视组织文化、组织能力及组织治理在 AI 管理中的核心作用。

多体系标准融合实战

ISO 42001 以 ISO 22989 为基石，沿用了以往 ISO 标准中相关核心概念以及 AI 系统生命周期模型，进一步为组织提供建立、实施、维护和持续改进人工智能管理体系（AIMS）的指导。同时 ISO 42001 在管理方法上借鉴了 ISO 9001、ISO 27001、ISO 27701 的一些原则，组织可以根据自身的需求和业务特点，同时采用这些标准来提升其信息安全管理、隐私保护以及人工智能管理的能力，实现多体系融合治理。此前的人工智能标准包括：

ISO/IEC 22989:2022– 人工智能概念与术语

ISO/IEC 23053:2022– 使用机器学习的人工智能系统框架

ISO/IEC 38507:2022– 组织使用人工智能的治理影响

ISO/IEC 23894:2023–AI 风险管理指南

多国 AI 法律符合性

ISO 42001 是一部兼顾不同国家地区有关人工智能法规要求的标准，涵盖了欧盟等多国家或地区人工智能监管在 AI 风险管理、主体权益保护、数据来源与质量、AI 伦理道德等方面的要求。ISO 42001 也是目前业内公认最具权威性的人工智能体系建设指导标准，获得 ISO 42001 认证是企业对于欧盟人工智能法案等 AI 法律标准合规性的重要参考。

强化 AI 治理能力，传递企业信任

根据 ISO 42001 建立人工智能管理体系，可有效控制和管理人工智能风险，系统化实施 AI 管理控制措施，降低 AI 法律合规和伦理风险。通过持续优化覆盖研发、部署、运营及监控等环节 AI 治理体系，提升 AI 产品在风险管理、主体权益、数据保护、基础安全等方面的能力。获得 ISO 42001 认证能证明组织的 AI 治理能力与 AI 产品可信，维护企业声誉和品牌形象。

6.1.2 ISO 42001 内容简介

图 8 ISO/IEC 42001 人工智能管理体系（AIMS）标准框架

ISO 42001 由正文和四个附录组成，其中正文部分共 10 个章节，详细规定了组织通过 PDCA 模型建立、实施、维护和持续改进 AI 管理体系的要求。正文还强调了组织应理解其内外部环境，明确其在 AI 领域的角色和责任，并基于此制定相应的 AI 管理策略和目标。此外，正文中还包含了对 AI 风险的评估、处置和系统影响评估的要求，旨在帮助组织识别和管理与 AI 相关的潜在风险，确保 AI 系统的开发和使用能够符合伦理和法律要求。

ISO 42001 附录中描述了 AI 管理体系的 9 个控制域、11 个组织控制目标、39 个控制项，及对应的控制措施和实施指南，供企业在建立和运行 AI 管理体系时参考，帮助组织满足管理目标，控制与 AI 系统设计和使用相关的风险；同时，介绍了一系列可能与 AI 相关的组织目标和风险源，帮助组织识别和管理与 AI 系统开发和使用相关的潜在风险和机会。附录也强调了 AI 管理系统集成到更广泛的组织管理系统中的重要性，提供了与 ISO 27001 信息安全管理体系、ISO 27701 隐私信息管理体系、ISO 9000 质量管理体系等其他管理领域标准结合使用的指导，以确保组织在多个管理维度上的一致性和整体性。

通过这些规定和指导，ISO 42001 标准能帮助组织负责任地实施 AI 技术，同时确保符合法律法规要求，满足利益相关方的期望，并促进组织的战略目标。

6.2 AI 治理可信等级管理

对于企业来说，AIMS 建设既能帮助企业满足监管合规要求，也能强化自身 AI 治理能力，增强客户信任。然而，对于很多企业来说，“从 0 到 1”一步到位建设 AIMS 难度大、成本高、耗时长。大多企业 AI 治理基础相对薄弱，不具备相对完善的风险管理、系统开发生命周期、信息安全、数据安全、隐私合规等管理框架作为 AI 治理基础，相关组织结构和人员也存在不足，难以“一步到位”完成 AIMS 体系建立和运行。相比“一步到位”的体系建设，与企业发展需求相适应的“分阶段”治理道路更具优势。

6.2.1 可信 AI 治理等级标识

“可信 AI 治理等级标识”标准对标融合国内外主流的人工智能法规和标准要求，包括 ISO/IEC 42001-2023 人工智能管理体系、欧盟人工智能法案、中国人工智能相关管理办法等，形成三个等级标准（基础级、提高级、体系级），共计 41 个控制项。企业可根据其 AI 系统的风险等级、产品类型、企业 AI 角色（AI 提供者、AI 部署者）等要求，选择适合自身的可信 AI 治理等级目标。

I 级 - 绿色徽章（基础级）

基础级对应的控制项和控制要求为国内外主流人工智能法规和标准要求中通用的、共有的、基础的 AI 管理要求。基础级要求适用于所有 AI 角色，通过基础级评估可获得绿色徽章和证书。

II 级 - 蓝色徽章（提高级）

对于风险等级为中、高的 AI 系统，可进一步进行“可信 AI 治理等级标识”提高级的建设和认证工作。II 级标准对 AI 提供者和 AI 部署者有不同的控制要求，企业可根据要求遵守不同的能力提升路径。通过提高级评估可获得蓝色徽章和证书。

III 级 - 黑色徽章（体系级）

对于风险等级为高的 AI 系统，可进一步进行“可信 AI 治理等级标识”体系级的建设和认证工作，以期满足 ISO 42001 的全部控制项要求。III 级标准要求主要源自 ISO 42001 的正文部分，帮助企业体系化运行 AI 治理体系。通过体系级评估可获得黑色徽章和证书，并获得 ISO 42001 的认证证书。

6.2.2 可信 AI 治理等级标识服务

安永提供可信 AI 治理等级标识咨询服务，帮助企业进行可信 AI 治理建设，而后将会引入权威的国际认证机构 DNV 进行等级审核，根据评测的等级颁发可信 AI 治理等级标识，为企业的 AI 可信提供市场背书。可信 AI 治理等级标识具有如下特色:

广泛的合规适应性

可信 AI 治理等级评估框架融合了国内外主流的人工智能法规和标准要求，包括 ISO/IEC 42001-2023 人工智能管理体系、欧盟人工智能法案、中国人工智能相关管理办法、标准（《生成式人工智能服务管理暂行办法》《互联网信息服务算法推荐管理规定》《互联网信息服务深度合成管理规定》以及算法模型备案要求），具备广泛的合规适应性，为 AI 企业和产品出海奠定了扎实的合规基石。

权威的风险治理框架

AI 治理和风险管理基于安永和 DNV 两家全球性专业机构的方法，根据企业是否为 AI 提供者、部署者等角色、AI 产品类型等要素，可以直接匹配适用的人工智能治理要求，按照不同的风险控制项进行评测，更能体现 AI 生态中不同环节的风险特性。

阶梯式等级划分

不同的 AI 角色和不同风险等级的 AI 产品需承担不同程度的合规风险管理责任。“可信 AI 治理等级标识”服务将人工智能管理体系要求划分为三个级别，更好地满足不同程度的 AI 风险管理需求，也为企业铺设了一条“由基础逐步迈向进阶”的可信 AI 建设道路。

“评估 + 背书”双重价值

评估覆盖 AI 产品及 AI 企业总体治理水平，深入业务逻辑、训练数据、算法模型以及基础安全多个方面，帮助企业评估 AI 产品整体风险、提升整体治理水平；通过评估后，由国际权威认证机构 DNV 颁发“可信 AI 治理等级标识”，为企业带来“评估 + 背书”双重价值加持。

6.2.3 可信 AI 治理重点领域解析

AI 系统影响评估

AI 系统影响评估是指一种正式的、有文档记录的过程，通过这个过程，组织识别、评估并解决其在开发、提供或使用利用人工智能的产品或服务时个人、群体以及社会的影响。

在 ISO 42001 中，6.1.4、8.4 及 A.5 专门提出了关于 AI 系统影响评估的规定和控制要求。ISO 42001 要求组织应定义一个流程，用以评估由人工智能系统的开发、提供或使用可能对个人、群体以及社会带来的潜在后果；同时，人工智能系统影响评估的结果应记录在案。除了 ISO 42001 外，欧盟《人工智能法案》第 27 条同样对影响评估提出要求，其规定高风险人工智能系统提供服务之前，部署者应评估使用该系统可能对基本权利产生的影响，并将评估结果通知市场监督管理机关。

因此，AI 系统影响评估是可信 AI 治理的一个

重点工作。安永根据相关要求及行业经验，形成了 AI 系统影响评估方法论。首先，企业应识别 AI 相关资产，包括基础设施、应用、算法模型、软件框架、数据集等，并根据适用的法律法规、标准和行业实践形成评估指标体系。其次，企业应识别 AI 相关的风险，风险源包括但不限于透明度、可解释性、公平性、鲁棒性、机器学习相关风险、AI 系统硬件问题、AI 系统生命周期管理、训练数据、隐私保护、人身安全等。然后，企业应分析相关风险的影响，包括对个人的影响（如个人权利、身心安全等）、对群体的影响（如群体歧视）、对社会的影响（如教育公平）及对组织自身的影响（如财务损失、声誉等）。随后，评估风险发生的可能性和影响程度，综合得出 AI 风险等级。最后，确定风险责任方，制定风险处置策略和计划，并开展相关整改。

AI 管理责任共担

AI 管理责任共担模型是指在使用人工智能（AI）技术时，不同参与方（如 AI 服务提供者和部署者）共同承担相应的责任。这个模型强调了在 AI 系统的开发、部署和使用过程中，各方需要明确自己的责任范围，并采取相应的措施来确保系统的安全、合规和伦理性。

在 ISO 42001 中，“A.10.2 责任分配”控制项要求组织应确保在其人工智能系统生命周期内，在组织、其合作伙伴、供应商、客户和第三方之间进行责任分配。同时，欧盟《人工智能法案》第 28 条 人工智能价值链上的责任，规定了在特定场景下任何分销者、进口者、部署者或其他第三方均应视为高风险人工智能系统的提供者，均应承担第 16 条规定的提供者义务。故组织应确定自身及相关方的 AI 角色，明确自身及相关方的责任和义务，确保履行对应的安全、合规、伦理义务。

在不同的服务模型中，责任共担的划分可能会有所不同。例如，在基础设施即服务（IaaS）模式中，服务提供商负责物理和虚拟架构的安全，而用户则负责其上部署的应用程序和数据的安全。在平台即服务（PaaS）模式中，服务提供商可能提供预训练模型和基础架构，而用户负责模型的特定训练和数据的安全。在软件即服务（SaaS）模式中，服务提供商通常承担更多的责任，包括应用程序的安全性和知识产权保护，而用户则负责使用这些服务时的数据安全和合规性。下图为微软 AI 责任共担模型 [5]，可有助于划分和明确各自的职责。这同样能使得企业能够以更快的速度创建和部署较新的 AI 应用，同时保持安全与合规。

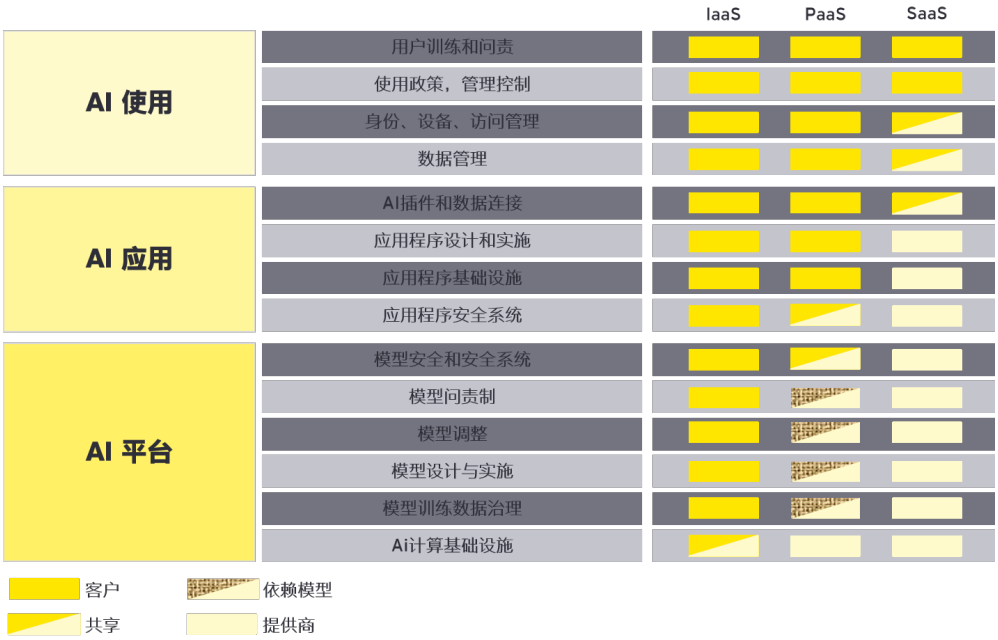


图 9 AI 管理责任共担模型

制度体系

AI 治理体系的建设与运行离不开 AI 管理制度和流程。ISO 42001 的 5.2 和 A.2.2 均规定组织应制定 AI 制度和文档。同样，欧盟《人工智能法案》第 17 条 质量管理系统规定，高风险人工智能系统的提供者应建立质量管理体系，该体系应以书面政策、程序和指令的形式加以记录，该体系应涵盖合规策略、AI 系统设计和验证程序、AI 系统开发和质量保证程序、检查和测试程序、数据管理程序、风险管理程序、事件报送程序、资源管理程序、问责程序等，并应将该质量管理体系文件在人工智能系统投放市场或提供服务后的 10 年内留存。除此之外，国内《互联网信息服务算法推荐管理规定》第七条及《互联网信息服务深度合成管理规定》第七条均提出了 AI 相关管理制度建设要求。

安永根据国内外法律法规、标准的要求，总结出 AI 治理制度体系基本要素，包括但不限于 AI 管理政策、AI 资产和资源管理制度、AI 系统影响评估制度、AI 系统开发制度、AI 算法机制机理审核制度、AI 算法检测制度、AI 系统安全事件应急制度、AI 算法违法违规处置制度、数据安全和个人信息保护制度、用户权益保护制度、AI 分类分级管理制度、知识产权保护制度等。

第七章
行业洞察与 AI 治理情况调研

07
PART

为了更好地了解人工智能在各行业应用情况和各行业的可信人工智能治理发展，进一步探索符合中国国情的人工智能发展路径、现状、挑战与风险管理体系，安永于 2024 年 6 月发起了覆盖多个行业的可信人工智能治理情况调研，旨在透过行业现状捕捉共性问题，聚焦企业应集中精力应对的人工智能治理难题。我们的调研覆盖信息通信、高科技制造、汽车、能源、金融、房地产、医药、零售、仓储物流、文化教育、现代服务业等多个行业的上百家企业，并将具有人工智能应用代表性的行业现状进行了整理，凸显垂直领域的可信人工智能治理问题

7.1 汽车行业调研

人工智能技术的快速发展和应用，为汽车行业注入了新的活力。实现了从“软件定义汽车”到“人工智能定义汽车”的新发展，实现了从辅助驾驶向更高级别的智能辅助驾驶演进，实现了更智慧的人机互联的驾驶体验，同时也正为汽车行业塑造着新的竞争格局和技术发展极。

7.1.1 汽车行业人工智能技术的应用

众多车企积极拥抱人工智能技术，并将人工智能技术运用于车辆研发、智能工厂、市场营销、用户服务等多个方面。

在车辆研发方面，车企通过人工智能算法对车辆的空气动力学进行优化，计算出风阻最小的车身设计方案；此外，借助人工智能模型对不同路况的电池数据、里程数据、操控数据等进行分析，以评估车辆在不同情况下的表现，为电池研发提供指导。

在智能工厂方面，车企通过人工智能系统对机械机器人、自动化设备、看板等进行控制和调度，实现生产线的高效运行；并利用人工智能图像识别技术对车辆零部件和整车进行质量检测，识别出车身凹陷、涂装不均匀等瑕疵产品。

在市场营销方面，车企利用生成式人工智能可以生成个性化的广告文案、视频脚本等内容，为数字化营销渠道、经销商营销渠道等定制不同的营销内容。

在用户服务方面，人工智能技术在汽车的智能辅助驾驶系统中发挥着重要作用，为车主提供自动泊车、智慧灯光、自适应巡航、车道偏离预警等服务；此外，通过人工智能等数字技术为车主的车辆提供进行远程诊断、故障监测等健康体检服务，在售后维修时还可通过“透明车间”服务实时跟踪车辆维修进度。

7.1.2 汽车行业对人工智能技术应用的普遍认知

汽车行业对人工智能技术应用秉持积极而审慎的观念，认为人工智能技术将成为各家车企抢占的新高地。车企在应用人工智能技术，最看重的因素是技术应用与实际收益之间的投入产出比；根据调研情况，在过去十二个月中在生产和营销方面的人工智能技术应用投入较高，认为人工智能技术对业

务发展带来了正向的影响，有利于提升企业的竞争力。同时，目前各个业务部门对人工智能的认知程度存在一定的差异，主要对技术成熟度、场景精准度、投入产出比、风险可控等方面存在一定的不同认识。

7.1.3 汽车行业人工智能技术应用的挑战

车企对人工智能技术的应用受政策和法律环境、技术与产品属性、消费者文化习惯、环境路况等多种因素影响；其影响具体体现在业务战略、数据管理、数据安全与隐私保护、伦理等方面。

在业务战略方面，其外部受人工智能的多种因素影响，内部各业务部门存在不同认识，可能在制定业务战略时无法充分发挥人工智能技术的优势以及难以预测人工智能的应用方向。在数据管理方面，人工智能技术依赖于对数据的分析处理，汽车搭载多种传感器、摄像头等设备收集海量数据，受到相关设备技术、行车环境等影响，其数据可能存在噪声、误差或缺失等问题；此外，对海量数据统一标注规则也极为重要，将直接影响人工智能模型训练效果。在数据安全与隐私方面，随着网络与数据安全、个人信息、人工智能等政策法律文件的颁布，以及消费者个人信息保护意识的提升，车企需要将法定义务和保障消费者个人信息权益的工作融入需求设计、产品研发、用户服务等业务活动中，若出现违法情形，可能面临行政罚款、民事侵权甚至刑事责任等法律风险。在伦理方面，人工智能技术的应用可能对社会公平、劳动者就业、驾驶决策等方面带来影响。

综上，人工智能技术对汽车行业的挑战是现实存在的，需要从政策立法进行引导和规范，同时压实企业责任和保障消费者权益入手，实现人工智能技术在汽车行业向善发展。

7.1.4 汽车行业人工智能管理体系的搭建与运营

人工智能技术在汽车行业的应用愈发成熟，与之同步的人工智能管理体系进入探索阶段。多家车企已经根据 ISO 42001、欧盟《人工智能法案》以及我国发布的关于人工智能服务安全、数据标注安全等标准，从人员组织、制度流程、技术开发等方面逐步开展人工智能管理体系的搭建与运营工作。在人员组织方面，部分车企成立人工智能伦理委员会等类似虚拟机构，成员包括技术专家、法律专家、安全专家、伦理学家等，负责审查相关汽车产品在人工智能技术应用方面的伦理道德、法律合规、系统安全等问题。在制度流程方面，众多车企确立了“以人为本”“保护隐私”“安全可靠”等原则，并将原则补充在现有制度规范、实施流程、技术文档等文件中；某合资车企已经建立了人工智能研发和应用管理制度，从数据来源、数据标注、模型训练、内容安全、功能测试等环节进行管理和监督。在技术应用方面，在智能座舱、智能语言、智能辅助驾驶等人工智能应用场景中进行针对性的管理工作，具体体现在提高算法的公开透明、增强模式的可靠安全、提供人类监督和干预路径等方面，并对以上内容进行安全评估。

7.2 医药行业调研

医药行业，作为全球健康领域的核心，担负着药品和医疗设备的研发、生产、分销及销售等关键职能，对提升公共健康水平起着至关重要的作用。作为一个研发密集型行业，医药行业积极拥抱基因疗法等创新技术，以推动治疗手段的革新。尽管面临高风险且受到严格的法规监管，成功研发的新药能带来显著的经济回报，使其成为一个对社会贡献巨大的行业。人工智能技术的引入正在彻底革新传统的医疗模式，为人类健康带来更多可能性。

7.2.1 医药行业人工智能技术的应用

医药企业对人工智能技术持积极态度，并期待其带来的变革。随着数据科学和机器学习技术的成熟，医药企业已逐渐认识到人工智能在提高研发效率、优化临床试验、个性化治疗、疾病预测和患者监护等方面的巨大潜力。

在药物研发领域，人工智能的应用主要集中在预测分子结构、筛选潜在药物、优化临床试验设计等方面。通过机器学习算法，人工智能技术能够分析大量的化合物数据，预测其生物活性，从而加速新药的发现过程。这不仅提高了研发效率，还降低了研发成本，使得新药更快地进入市场，惠及患者。

精准医疗是人工智能技术应用的另一个重要领域。通过分析患者的遗传信息、生活方式和环境因素，人工智能技术可以帮助医生为患者提供个性化的治疗方案。这种个性化的治疗方法能够更精确地针对患者的具体状况，提高治疗效果，减少不必要的副作用。精准医疗的实现，标志着医疗模式从“一刀切”向“量体裁衣”的转变。

在医疗影像领域，人工智能技术已经被用于辅助诊断，如通过深度学习算法分析 X 光片、CT 扫描和 MRI 图像，以识别肿瘤、骨折等病变。人工智能的高准确率和快速响应能力，极大地提高了诊断的效率和准确性，帮助医生更快地做出治疗决策。

此外，人工智能在患者监护方面也发挥着重要作用。它们可以实时监控患者的生命体征，预测病情变化，及时提醒医护人员采取必要的干预措施。这种智能化的监护不仅提高了护理质量，还减轻了医护人员的工作负担。

7.2.2 医药行业对人工智能技术应用的普遍认知

医药行业引入可信人工智能至关重要，主要因为人工智能能提高诊断的准确性和医疗流程的效率，同时促进个性化医疗和加速药物研发。人工智能在改善患者监护和支持临床决策方面也发挥着重要作用，有助于确保患者安全和提升医疗服务质量。它还能帮助应对医疗领域的复杂性和不确定性，优化医疗资源分配，并支持公共卫生监控。可信人工智能系统的持续学习能力使其能够不断从新数据中学习，改进性能。

医药行业在人工智能技术的选择上，既包括本土化工具，也包括海外引进的技术。本土企业开发的人工智能技术和平台在药物研发、医疗影像分析、电子健康记录分析等方面得到广泛应用。同时，医药企业也与国际领先的人工智能研究机构和企业合作，引进和学习海外先进的人工智能技术。

从模型角度分析，人工智能的应用正通过一系列先进的模型，推动着从药物研发到临床决策、患者监护和医疗数据分析的各个方面的革新。监督学习模型通过分析标记数据来预测药物分子的活性或进行疾病诊断，无监督学习模型则用于探索数据中的内在结构和模式，这对于发现新的生物标记物和疾病亚型至关重要。神经网络（NN）模型，在图像识别和序列数据分析中发挥关键作用。自然语言处理（NLP）模型能够理解和处理自然语言，被用于分析医疗文献、电子健康记录和患者反馈，以提取有价值的信息，这对于临床决策支持和患者教育非常有价值。

7.2.3 医药行业人工智能技术应用的挑战

尽管人工智能技术在医药行业的应用带来了显著的益处，但也面临着数据隐私、算法透明度、法规遵从和兼容性等挑战。

患者的健康数据通常包含敏感信息，如何在人工智能系统处理这些数据时确保数据隐私和安全是一个重要问题。例如，未经加密的医疗数据一旦遭受黑客攻击，就可能导致患者信息泄露，给患者带来隐私侵犯甚至经济损失。

其次，人工智能算法可能会因为训练数据的不均衡或算法设计不当而产生偏见，导致对某些群体

的不公平。例如，如果一个人工智能辅助诊断系统在训练时主要使用了某一人群的数据，可能在诊断其他人群时表现不佳，从而影响诊断的准确性和公平性。

此外，人工智能系统可能会因为技术限制或操作失误而导致错误的医疗决策，这使得责任归属复杂。例如，如果一个人工智能系统错误预测患者疾病风险可能导致不必要的治疗，不仅增加了患者的身体和经济负担，还可能导致医疗资源的浪费。过度依赖人工智能技术可能导致医护人员的技能退化，而人工智能系统的误操作或错误解读也可能导致医疗事故。

最后，医药行业作为一个强监管行业，受到严格的法规监管，人工智能技术的应用需要符合相关法规和伦理标准。例如，人工智能在基因编辑中的应用可能会引发伦理争议，需要我们在推进技术应用的同时，充分考虑伦理和社会影响。

7.2.4 医药行业人工智能管理体系的搭建与运营

在医药行业，人工智能技术的广泛应用正引领一场深刻的变革。为了确保这场变革能够安全、合规、高效且符合伦理地进行，越来越多的药企开始设立专门的人工智能治理人员或团队。人工智能治理人员在药企中负责确保人工智能技术的安全、合规、有效和伦理应用。他们制定和维护人工智能政策，评估风险，监督合规性，并管理数据。他们还参与伦理审查，跨部门协调，并组织培训以提升对人工智能治理的认识。此外，他们监督人工智能的效果，处理危机，并参与行业合作。这些人员需要跨学科知识，包括计算机科学、医药学、法律和伦理学，以及沟通和风险管理能力。随着人工智能在医药行业的应用不断加深，人工智能治理人员的作用愈加关键。

药企人工智能治理人员在制定人工智能相关制度时，通常综合参考国内外的个人智能相关法律法规和标准，如欧盟《人工智能法》草案、GDPR、美国 NIST 的人工智能风险管理框架、ISO/IEC 42001、ISO/IEC 22989 和 ISO/IEC 23894 等海外对人工智能治理方面的要求。在中国，药企还需遵守数据安全法、个人信息保护法、GB/T 42888 等外规要求。这些法律为人工智能应用的数据安全、隐私保护、透明性、数据使用规则、技术和伦理要求等方面提供了指导。

7.3 零售行业调研

快消零售行业主要指包括食品、饮料、日用化学品、化妆品在内的使用周期短、销售速度快的消费品。这类产品生命周期短、更新快，且消费需求巨大，导致了快消行业企业竞争激烈，需要在生产、营销、供应链管理等方面保持高效运转。

7.3.1 零售行业人工智能技术的应用

零售行业对人工智能技术的应用大多处于初步到中期阶段，涉及人工智能技术规划、治理、训练数据准备、人工智能技术与基础设施配置、测试用例选择与优化。技术的深度应用通常集中在大型跨国企业，而中小型企业则还处于探索和试点阶段。当前，人工智能技术主要用于以下业务场景：一是研发设计，用于数据分析与处理、内容创作等；二是生产制造，用于设备运维、自动化重复性任务、供应链管理等；三是经营管理，用于提升办公效率、增强网络安全防护、改进决策制定等；四是客户服务，用于聊天机器人、自动化客户服务、产品推荐与个性化等。

常用的人工智能工具或模型包括 Microsoft CoPilot、微软认知工具包（CNTK）等。部分海外人工智能工具在某些地区可能存在使用限制。在这种情况下，企业通常会根据本土化工具或模型的能力，来选择本土人工智能服务，确保技术应用的连续性

和合规性。

7.3.2 零售行业对人工智能技术应用的普遍认知

整体来看，零售行业对人工智能技术的应用呈现出较强的兴趣，普遍认为有助于提升企业竞争力。企业在应用人工智能时，首先看重的因素是性能指标，包括技术稳定性、可靠性、准确性、可集成性、可扩展性等，其次是算法的透明度和可解释性，而后考虑成本效益、处理数据时的安全性、解决隐私保护和公平性等伦理性问题以及知识产权问题等。对于中小企业而言，成本效益则成为优先考虑的因素。

根据调研，大型企业认为过去 12 个月的人工智能技术应用投入无法满足业务实际需求，而中小型企业则能基本满足需求。关于应用人工智能技术对业务流程的改进成效，企业的反馈大多是积极的，但也存在部分企业成效不高的反馈，例如奢侈品企业，这可能归因于企业自身主营业务对 AI 的依赖性不强。目前，人工智能技术在零售行业的普及程度比较广泛，整体普及率仍处于上升阶段。

7.3.3 零售行业人工智能技术应用的挑战

人工智能技术应用的过程中，企业可能面临来自技术基础、数据管理、数据安全和隐私保护、伦理方面以及在与长期业务战略融合方面的复杂挑战。在技术基础方面，算法复杂、人工智能与现有业务流程集成性低等技术障碍是首要挑战。AI 技术应用的前提是与企业现有系统和流程进行整合，但有些企业的 IT 基础设施不够灵活，导致难以与新技术无缝衔接；加上供应链、营销和销售等环节业务流程复杂，需要具备较强的 AI 建模和算法能力，才能提供有效的解决方案。在数据管理方面，零售企业在数据清洗和预处理复杂、数据使用合规和生成内容安全方面存在困难。零售企业通常掌握来自供应链、销售、市场营销、客户反馈等多元渠道的庞大、分散的数据源，数据难免存在缺乏精准性、未经清洗或存在不完整、重复的问题，从而影响 AI 预测准确性。在数据使用的合规性和生成内容的安全性方面也存在风险和隐患。

在数据安全和隐私保护方面，由于数据量持续增加，数据泄露风险相应提高，加上全球各地数据保护法规不断变化，全球监管环境趋严，企业将面临数据安全和隐私保护方面的巨大合规挑战。在保护个人隐私的同时，企业还需要对数据进行从收集到销毁的全生命周期保护，并确保数据的可用性。企业在进行跨国业务运营时，数据出境规制也成为一项重要挑战。在伦理方面，不同国家和地区对伦理的理解和标准不同，对人工智能与人类工作者之间关系的理解也存在差异。人工智能技术在设计和应用过程中可能加剧或产生偏见和歧视，影响社会公平和平等的实现。另外，算法决策过程过于复杂，导致难以向非技术用户解释其工作原理。如果 AI 系统出现失误或问题，责任如何分担也成为伦理问题的焦点。在长期业务战略融合方面，企业面临诸多宏观层面的难题。其一，技术快速迭代，从而导致研发困难、算力紧张等；其二，数据源不可靠和数据使用不合规可能会影响整个 AI 所参与业务的合规性；其三，人工智能相关的监管政策不断调整，难以对其作出及时准确的适应性动作；其四，缺乏具备人工智能知识和技能的专业人才；其五，持续的成本投入是否能带来预期的收益，成为管理者在长期业务战略考虑的难题。

7.3.4 零售行业人工智能管理体系的搭建和运营

部分零售行业企业通常有专门的团队参与治理工作，涉及团队包括高级管理层或董事会、技术团队、合规团队 (含隐私) 和安全团队。一些中小型零售企业暂时没有负责人或团队。为了保证 AI 的可信性和可持续发展，许多零售企业会参考国际公认的标准和框架制定内部政策或指导原则，参考的文件包括欧盟《人工智能法》、GDPR、美国 NIST 的 AI 风险管理框架、ISO/IEC 42001、ISO/IEC 22989 和 ISO/IEC 23894 等国际相关标准，以及中国人工智能相关

法规与标准，所制定的内部政策通常涵盖 AI 应用的数据安全、隐私保护、透明性、数据使用规则、技术和伦理要求等。此外，企业还会面向全体员工开展可信人工智能方面的培训。

在 AI 项目的开发和实施过程中，企业通常会进行风险评估，以确保 AI 系统不会产生意外的偏见、隐私风险或技术故障。关注的评估点包括性能指标、算法的透明度和可解释性、伦理合规性、数据全生命周期的保护情况等。对于在评估过程中发现的风险，企业通常将其视为长期优先事项，并采取调整 AI 算法、加强数据保护措施以及制定应急预案等积极措施减少风险。对于无法完全消除的风险，企业可能采取缓解策略减少风险影响，如限制 AI 应用的范围或增强人类监督。

总之，在可信 AI 治理方面，零售企业建立了多项机制，包括 AI 管理制度、事件应急处理机制、建立投诉机制、明确组织和人员角色、训练数据来源合规、训练数据标注规范、数据质量保证、数据安全防护、个人信息保护、内容安全治理和模型安全保障等。针对人工智能技术的未来发展，企业建议加强技术研发、完善法律法规、提升公众认知，并加强国际合作。

7.4 服务行业调研

服务行业涵盖了人力资源、审计、咨询和财务等多个领域，是现代经济的重要组成部分。在服务行业中，人工智能的治理和发展正受到越来越多的关注。在全球范围内，联合国教科文组织发布人工智能伦理协议，各国都在推进人工智能治理领域的持续合作；中国国家主席习近平也提出了人工智能治理方案强调“以人为本、智能向善”，倡导在联合国框架内加强人工智能规则治理，这为中国在人工智能领域的发展方向提供了明确的指导。与此同时，众多企业也积极响应国家政策，将人工智能技术应用于各自的服务领域，以实现技术创新和服务升级。这促使服务行业中的不同领域都在不断寻求通过技术手段提高效率和准确性。

7.4.1 服务行业人工智能的应用

随着人工智能技术的迅猛发展，服务行业正经历着一场前所未有的变革。服务行业作为经济的重要组成部分，其核心在于提供无形的、以人为核心的价值。随着人工智能技术的融入，服务行业在自动化和智能化处理、提升服务质量、优化客户体验、降低运营成本、提高运营效率等方面展现出巨大潜力。然而，技术的快速发展也带来了治理上的挑战，如何确保人工智能的可信性和可靠性成为行业关注的焦点。

当前，服务行业中的人工智能应用多集中在客户服务自动化、个性化推荐、风险管理和欺诈检测等领域。企业通过部署智能客服、分析客户行为数据、利用预测模型来优化服务和产品。

人力资源管理（HR）：人工智能在人力资源管理中的应用包括自动化招聘流程、员工培训、绩效评估和离职管理。人工智能系统被应用于企业人力资源管理，提供智能招聘、优化招聘流程，并提供客观的员工绩效评估和发展建议等功能。

审计：在审计领域，人工智能帮助审计师更快地找到客户财务存在的异常。人工智能技术被广泛应用于头部审计公司进行现金审计，自动化处理大量财务数据，识别异常交易和潜在风险，以提高审计效率和准确性。

咨询：咨询行业通过人工智能的应用，可以进行市场映射和竞争者研究、M&A 咨询、数据分析等，可以基于此提供深入的市场洞察和战略建议。此外，通过分析客户反馈和行为数据，人工智能帮助企业优化客户服务策略，提升客户满意度。

财务：在财务领域，人工智能的应用包括算法

交易、自动化、合规、信用评分、成本降低、客户服务、数据分析、欺诈检测等。

以上行业实践表明，企业普遍重视人工智能技术的创新和应用，但在可信治理方面仍处于探索阶段。人工智能在服务行业中的治理和发展是一个多方面的过程，涉及技术进步、伦理考量、法律合规和战略规划。随着人工智能技术的不断进步，这些行业需要不断适应和更新其治理结构，以确保人工智能的负责任使用，并最大化其潜在价值。

7.4.2 服务行业对人工智能技术应用的普遍认知

目前，服务行业中的企业大多处于人工智能技术应用的初级到中级阶段。虽然一些领先企业已经在特定业务场景中实现了 AI 的深度应用，但整体上仍在探索和试验阶段。人工智能在服务行业的应用场景包括但不限于：人力资源管理中的自动化简历筛选、员工绩效分析、员工培训和发展规划；审计工作中的自动化数据分析、异常检测、合规检查；财务管理工作中的财务预测、风险评估、自动化报表生成。

行业普遍认为，人工智能是提升竞争力的关键因素。过去一年中，企业在人工智能技术上的投入显著增加，这些投入主要用于技术研发、人才培养和系统部署。

当前的全球经济中，人工智能技术正逐渐成为推动服务业和其他行业数智化转型的关键因素。大型跨国服务企业正通过数亿美元的投资，积极开发和应用大语言模型等 AI 技术，以提升数据管理的效率和优化业务流程这些投资不仅用于开发先进的 AI 模型，还用于提高员工的人工智能技能，通过在线课程、研讨会和工作坊等多种培训形式，确保业务人员能够利用人工智能为客户提供更高标准的咨询服务。此外，这些服务企业还致力于推动审计、税务和咨询服务的创新，通过整合 AI 技术和云服务，提供更灵活、可扩展的解决方案，以满足客户不断变化的需求。这些举措表明，人工智能技术的应用正成为服务业竞争力提升的重要途径。随着 AI 技术的不断进步和应用的深入，预计未来几年，人工智能将在服务业中扮演更加重要的角色，推动行业的数字化转型和升级。

7.4.3 服务行业人工智能技术应用的挑战

服务行业在采用人工智能技术时，面临着一系列特有的挑战，这些挑战覆盖了技术、伦理、法律、社会和经济等多个方面。技术层面上，服务行业需要将人工智能技术有效地整合到现有的工作流程中，并确保员工能够适应这些变化，这可能涉及到对员工的重新培训和工作流程的调整。在数据隐私和安全方面，服务行业必须在处理敏感客户数据时遵守数据保护法规，并采取措施防止数据泄露和滥用。算法的透明度和可解释性也是关键问题，因为缺乏透明度可能导致用户信任度下降，所以需要提高算法的可解释性，以确保决策过程的公正和透明。

此外，人工智能系统在数据训练和应用过程中可能产生的偏见问题也不容忽视，这可能会加剧社会不平等和歧视，因此需要确保人工智能系统的设计和应用不会带来这些负面影响。法律和监管合规方面，随着人工智能技术的发展，现有的法律体系可能需要更新以适应新的挑战，服务行业需要遵守不断变化的法律法规，并推动制定适应 AI 时代的新规则。技能和人才发展也是服务行业面临的重要挑战，因为人工智能技术的应用要求员工具备新的技能和知识，这就需要对现有员工进行培训和教育，并吸引新的技术人才。

社会接受度和适应性也是服务行业在采用人工智能技术时需要考虑的问题，因为 AI 技术的广泛应用可能会改变工作性质和就业结构。同时，AI 应用的环境影响和可持续性也是不容忽视的问题，服务行业需要评估和减少应用的环境足迹。最后，服务行业需要考虑技术的长期发展趋势，并制定相应的战略规划，同时在全球性跨国界的合作和协调下，参与国际标准的制定，以确保 AI 技术的全球兼容性

和互操作性。这些挑战需要服务行业、政府机构、学术界和民间组织的共同努力，通过合作和对话来解决，并不断监测和评估人工智能技术的影响，以便及时调整治理策略和技术应用方向。

7.4.4 服务行业人工智能管理体系的搭建与运营

为应对上述挑战，服务行业整体来看虽然只处在人工智能治理的早期阶段，但仍有一些领先企业已经开始构建人工智能管理体系。这些体系通常包括制定内部政策、建立风险评估机制、开展员工培训等。在这一过程中，企业通常会参考国际标准和最佳实践，如 ISO/IEC 标准，以确保人工智能应用的合规性和安全性。根据《国家人工智能产业综合标准化体系建设指南（2024 版）》，人工智能被视为引领新一轮科技革命和产业变革的基础性和战略性技术，服务行业需要通过明确的战略规划来整合 AI 技术与现有业务流程，以提升服务质量和效率。

在构建管理体系时，企业需关注数据隐私和安全，确保遵循相关法律法规，防止数据泄露。同时，算法的透明度和可解释性是提升用户信任的关键，企业应致力于提高 AI 系统的可解释性，确保决策过程的公正性。此外，伦理问题及算法偏见也需引起重视，以防止加剧社会不平等。

第八章

AI 应用的典型案例

08
PART

8.1 典型案例一：某互联网平台公司内部人工智能技术应用

互联网行业不仅通过商用大模型在改变和驱动社会各界对人工智能技术的使用，在各个典型的互联网平台类公司内部，人工智能技术正引领一场变革，深刻影响着企业 C 端和 B 端产品和服务的各个方面。从用户智能推荐、商家经营支持到内部业务提速，人工智能的应用正在重塑该行业的现状。

8.1.1 人工智能技术使用现状

互联网平台公司采用人工智能技术旨在提升用户体验、优化商家经营和提高数据处理效率。在用户对话导购场景中，人工智能技术通过生成式对话，扮演着小助手的角色。这种基于聊天内容的交互方式，能够根据用户的喜好和历史订单数据，提供个性化的商品推荐。与传统对话系统不同，这种导购系统经过行业语料的训练，能够更好地理解外卖平台的特性，提供更精准的建议。

在客服对话场景中，AI 技术的应用提高了客服效率，减少了人力成本。AI 客服系统能够处理大量常见问题，如订单状态查询、退款申请等，通过自然语言处理技术，理解用户的问题并提供准确答案，提高用户满意度，同时释放人工客服的时间和精力。

在图片生成方面，人工智能技术的应用旨在降低人工设计成本，提高图片吸引力。通过“文 + 图”的生成方式，人工智能技术能够根据文本描述自动创建新的图片，尤其在商家图片美化方面非常有用。此外，通过图片效果分析和 AB 测试，商家可以不断优化图片，提高用户的点击率和购买意愿。

对于商家而言，经营小助手是一个强大的工具。它能够快速分析商家数据，提供关于订单量、用户评价、商品受欢迎程度等关键信息的即时反馈，帮助商家做出更明智的经营决策。

随着人工智能技术在互联网的深入应用，其影响已经远远超出了用户和商家的直接体验。人工智能技术不仅在提升用户满意度和助力商家经营决策方面发挥着重要作用，而且已经开始渗透到业务支持的各个层面，从商家资质审核到数据分析，全方位推动该平台公司的数字化转型。

在风控和资质审核方面，人工智能技术的应用提高了审核的效率和准确性。通过审理大模型，系统能够自动审核商家资质，包括合理性、文字和图片等内容。这种自动化的审核流程提高了审核速度，减少了人为错误的可能性，但同时也需要确保审核过程中的准确性和公正性。

通过训练自然语言与 SQL 语言的转换模型，致力于将用户的自然语言查询转换为数据库查询语言，从而快速找到所需数据并完成分析。这种技术的应用提高了数据处理效率，为非技术背景的用户提供了一个更加友好的数据查询方式。

8.1.2 人工智能模型使用与评估

互联网平台公司会广泛应用各类商用人工智能模型，凭借其先进的自然语言处理技术，为这一领域提供了前所未有的技术支持。

在这个案例中，首先是集团内部有一个统一的模型训练系统，支持开源模型以及经过集团评估认可的模型训练。专门的语料系统提供内部语料和开

源语料的上架服务，包括审批、脱敏和清洗等步骤，确保数据的质量和安全性。而模型实验室系统提供境外代理服务，为确保资源的合理分配和使用，流程业务部门需要通过强制性的内部申请调用来分发和使用这些资源。

在模型正式投入应用前，必须经过全面的大模型评测，这一过程涵盖了隐私保护和内容安全等多个维度。评测中特别关注模型在面对各种指令攻击时的表现，如隐私泄露和不当内容生成等，以确保模型的安全性和合规性。对于未能通过评测的模型，需要进行黑盒的内生加固流程，主要通过重新训练语料来提升模型的鲁棒性。虽然模型上线前的测评并非强制性要求，但对于直接面向终端用户的模型将进行强调。

一旦模型上线，将被纳入在线布防系统，通过滤网进行实时监控。这包括对 FAQ 内容的过滤和检索功能的增强。利用自定义样本包，系统能够自动过滤掉潜在的不安全内容，同时，基于检索库的碰撞检测机制能够去除虚假信息，防止恶意数据投毒，确保输出结果的安全性。

在模型的发布和监控方面，该案例中的公司正在引入强制卡点，以应对模型特性的不断变化。此外，对模型的监控还包括实施事后监控和运维策略，利用拦截内容进行逆向分析，以训练更有效的黑名单，同时，密切关注用户投诉，以持续优化模型性能。

最后，定期对模型进行评测，以确保其持续符合最新的安全和合规标准。这一系列措施共同构成了模型安全管理的框架，旨在为用户提供安全、可靠和高质量的服务。

8.1.3 人工智能技术投入分析

在实际业务部署中，尽管人工智能模型对业务流程提供了显著的辅助作用，但它们并非不可或缺。从技术投入的视角来看，员工配置并没有因为这些模型的引入而经历大规模的调整，这表明现有的业务流程和人员架构能够适应融入新技术的环境。在财务投资方面，此前的支出主要集中在满足用户、商家和配送员的基本需求上，而近期则更加侧重于提升内部运作的效率。这一策略转变体现了对资源优化和流程改进的重视，目的是为了在不增加额外成本的前提下，利用技术手段提高业务的整体效能。

8.2 典型案例二：某汽车公司内部人工智能技术应用

在新能源汽车领域，人工智能技术的应用正日益广泛，不仅提升了驾驶体验，还推动了整个行业的技术进步。

8.2.1 人工智能模型使用现状

在汽车企业中，人工智能技术的应用正日益成为提升用户体验、优化经营效率和提高数据处理能力的关键驱动力。以下是人工智能技术在汽车企业中的几个主要应用场景：

在自动驾驶场景中，人工智能是实现自动驾驶汽车的关键技术，涵盖了感知、决策、控制等多个模块。通过机器学习、深度学习等先进算法，车辆能够准确识别道路环境、预测行人和其他车辆的行为，从而实现自动导航和驾驶。

在智能充电网络场景中，人工智能技术可以优化充电站的布局和运营效率。通过预测充电需求、调整充电价格等手段，可以有效减少充电站的拥堵情况，并提高能源利用效率。如通过云、站、桩、车深度协同，实现充电网络的全面智能化，提升充电效率和用户体验。

在智能座舱场景中，人工智能技术的应用为用户提供了更加智能和个性化的车内体验。这包括语音控制、驾驶员监控、娱乐系统等，通过 AI 驱动的感知交互，实现从被动响应到主动交互的转变。

在智能运维方面，人工智能技术可以通过监测

和分析车辆的运行数据，预测车辆的维护需求，从而提高运维效率。人工智能系统可以对换电站和充电桩的每一个设备、每一个零件进行监测，及时发现并处理问题，保障用车安全。

在个性化服务方面，人工智能技术可以根据用户的驾驶习惯和偏好，提供个性化的服务，如推荐路线、调整车内环境等。

在智能充电场景中，人工智能系统通过智能算法和数据分析，自动选择最佳的充电时间和功率，实现充电过程的优化。这种系统能够实时监控和管理充电过程，根据各种变量自动调整充电策略，从而节省成本并减轻电网压力。

8.2.2 人工智能模型使用与评估

在汽车行业迈向智能化的浪潮中，该汽车公司通过引入先进的人工智能技术，提升了其智能驾驶系统性能。通过深度学习和大数据分析，该企业不仅提高了人工智能系统的准确性，还缩短了响应时间，从而显著增强了驾驶安全性。这一案例展示了人工智能在汽车领域实际应用中的潜力。

该车企通过车载高清摄像头、激光雷达和毫米波雷达等多种传感器，结合多传感器融合技术，构建出车辆周围环境的三维模型，实现对周围环境的全面感知，使用收集的大量图像和数据对人工智能系统进行训练，以精准检测和识别行人、车辆、交通标志等目标物体，为智能驾驶提供决策依据；同时综合利用视觉、听觉、触觉等多种模态信息，提高感知精准度。

在评估人工智能表现方面，该车企采用任务导向基准测试方法，通过让人工智能系统执行特定任务，如高速并线、复杂路况自动泊车等，并评价其完成情况，以此来评估性能；此外，还通过虚拟环境进行测试和验证工作，确保智能驾驶辅助系统在各种条件下的稳定性和可靠性。这些方法不仅提升了智能驾驶辅助系统的性能，还确保了系统的可靠性和安全性，为消费者提供了更加安全、智能的驾驶体验。

8.2.3 人工智能技术投入分析

随着技术的不断进步，人工智能在新能源汽车领域的应用将更加广泛和深入，为用户带来更加便捷、智能的出行体验。但是，随着人工智能和新应用模式的推广，也带来了更复杂的安全风险和挑战。例如：虽然自动驾驶技术能够提升驾驶体验，但也可能因为系统错误或外部环境的复杂性导致事故；智能充电网络在优化充电站布局和运营的同时，也可能面临黑客攻击的风险，攻击者可能会通过远程控制充电站，造成充电站服务中断或电能浪费；以及也可能因为软件漏洞或数据泄露，导致用户隐私受到威胁

为了降低这些安全风险，相关企业可以参考本白皮书的安全管控建议内容，建立人工智能安全风险管控体系，持续管理人工智能风险。

结 语

随着《可信人工智能治理白皮书》的深入探讨，我们不难发现人工智能作为一项革命性技术，正站在一个新的历史起点上。它不仅代表着技术进步的速度，更考验着人类智慧的深度。在 AI 的浪潮中，我们既面临着前所未有的机遇，也必须应对种种挑战。

白皮书的撰写，是我们对人工智能未来发展的一次深思熟虑。我们探讨了全球人工智能的发展态势，分析了监管体系的构建，讨论了可信原则的重要性，并深入挖掘了企业合规要求和风险治理的策略。我们希望通过这些内容，能够为社会各界提供有价值的参考和启发。

在 AI 的世界里，信任是基石。可信的 AI 不仅要在技术上可靠，更要在伦理上可靠，它需要在尊重人权、保护隐私、确保公正的基础上，为社会带来积极的影响。因此，建立一个全面、有效的 AI 治理体系，是确保 AI 技术健康发展的关键。

我们呼吁，所有 AI 的参与者，无论是技术开发者、企业决策者还是政策制定者，都应该共同承担起构建可信 AI 的责任。我们相信，通过不断地探索和努力，人类社会能够找到与 AI 和谐共存的最佳方式。让我们携手合作，共同迎接一个更加智能、更加可信、更加美好的与人工智能同行的未来。

文末引用

[1] Fortune Business Insights. 人工智能 (AI) 市场规模、份额和行业分析，按组件（硬件、软件 / 平台和服务）、按功能（人力资源、营销和销售、产品 / 服务部署、服务运营、风险、供应链管理）、其他（战略和企业财务））、按部署（云和本地）、按行业（医疗保健、零售、IT 和电信、BFSI、汽车、广告和媒体、制造等）以及区域预测，2024-2032. (2024-12-16). https://www.fortunebusinessinsights.com/zh/industry-reports/artificial-intelligence-market-100114.

[2] 中国青年报 . 深陷“信息茧房”，年轻人如何破茧 . (2023-07-14). https://baijiahao.baidu.com/s?id=1771347168949471921&wfr=spider&for=pc.

[3] 澎湃新闻·10% 公司 . 英国医院向人工智能公司共享 160 万病人信息被认定违法 . (2017-07-04). https://www.thepaper.cn/newsDetail_forward_1724635.

[4] Cyberhaven. Shadow AI: how employees are leading the charge in AI adoption and putting company data at risk. (2024-05-21). https://www.cyberhaven.com/blog/shadow-ai-how-employees-are-leading-the-charge-in-ai-adoption-and-putting-company-data-at-risk.

[5] Microsoft Azure. 人工智能 (AI) 共担责任模型 . (2024-09-29). https://learn.microsoft.com/zh-cn/azure/security/fundamentals/shared-responsibility-ai.

EY安永
Shape the future with confidence
聚信心 塑未来

建设更美好的商业世界

安永的宗旨是建设更美好的商业世界。我们致力帮助客户、员工及社会各界创造长期价值，同时在资本市场建立信任。

在数据及科技赋能下，安永的多元化团队通过鉴证服务，于 150 多个国家及地区构建信任，并协助企业成长、转型和运营。

在审计、咨询、战略、税务与交易的专业服务领域，安永团队对当前最复杂迫切的挑战，提出更好的问题，从而发掘创新的解决方案。

APAC no. 03022041
ED None

关注安永微信公众号，
扫描二维码，获取最新资讯。

ey.com/china

可信人工智能治理白皮书

TRUSTED ARTIFICIAL INTELLIGENCE GOVERNANCE WHITE PAPER

EY安永

Shape the future with confidence

聚信心 塑未来

SAISD

上海市人工智能与社会发展研究会

Shanghai Association for AI and Social Development