人工智能安全合规监管与应对

敏于知

自 2022 年 11 月 30 日 ChatGPT 的发布以来，人工智能大语言模型已引起全球广泛关注。人工智能的发展成为不可避免的趋势，但也引发了数据和网络安全合规、技术伦理等方面的风险隐患，成为当前法律监管讨论的焦点。为应对这一挑战，各国相继制定了相关法律法规并不断的加以完善和补充。2023 年 10 月 18 日，中国国家互联网信息办公室发布了《全球人工智能治理倡议》。该倡议系统阐述了人工智能治理中国方案，重点涵盖人工智能发展、安全、治理三个方面，致力于推动国际合作，构建全球开放、包容、透明的人工智能治理体系。2023 年 12 月 8 日，欧洲议会、欧盟成员国和欧盟委员会三方成功达成了《人工智能法案》的初步协议。该法案被认为是全球首部人工智能领域的全面监管法规，为人工智能系统的开发、市场推广和使用制定了全面而细致的监管框架。

本文旨在探讨人工智能领域的变革和发展，关注人工智能合规的推动因素以及当前的监管态势。通过提出切实有效的合规对策，并通过对生成式人工智能的安全评估和备案流程的相关规定进行详细解读，为企业提供实质性的操作建议，确保其在人工智能的创新和发展中保持持续的合规。

人工智能的变革 — 决策式 AI 和生成式 AI

人工智能（AI）是一系列技术的集合，其目标是赋予计算机系统执行各种高级功能的能力，涵盖口头和书面语言的视觉、理解和翻译，以及数据分析和提出建议等方面。

人工智能模型可分为决策式人工智能和生成式人工智能两大类。决策式人工智能基于规则和逻辑，采用预定义的规则和算法，通过海量大数据分析做出决策。与之相对，生成式人工智能是一种基于学习和理解的系统，通过学习大量数据，并试图从中提取规律，从而具备创造新数据或内容的能力。

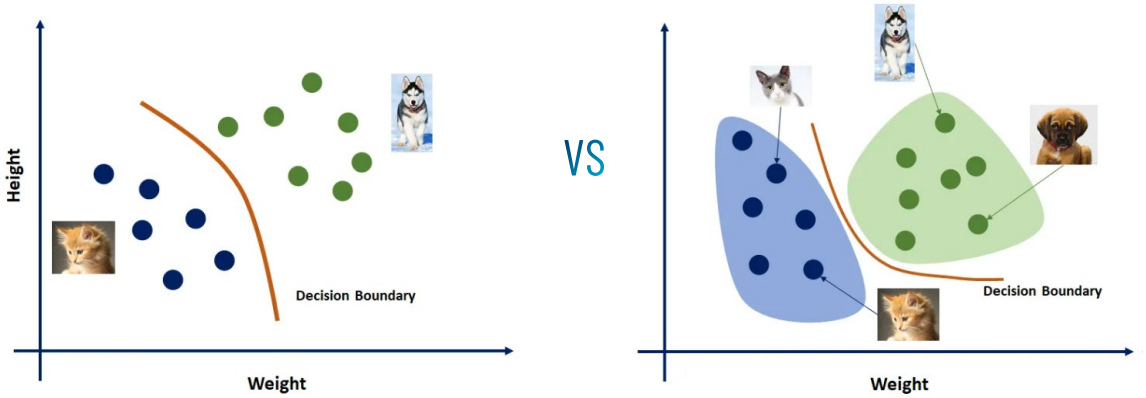
决策式 AI 和生成式 AI 原理[1]

决策式人工智能

基于预定义的规则和算法，通过大数据逻辑推理做出决策，如将图像分类为狗或猫。

生成式人工智能

通过学习和理解大量数据从中提取规律后创造新数据或内容。如生成逼真的猫或狗图像。



[1] Generative-and-discriminative-models: https://learnopencv.com/generative-and-discriminative-models/

随着人工智能技术的不断演进，AI 正逐渐从传统决策式模式转向生成式模型（见下图），以更有效地应对复杂、非线性和不确定性任务，并在各个应用领域广泛展示其应用价值。

决策式人工智能（2014 年之前和现在）		生成式人工智能 - AIGC（2014 年以后及未来）
效能提升	主要功能	内容创造
注重分析和优化 根据可用数据做出决策	特 点	注重创造力和创新 生成前所未有的内容
个性化推荐（内容、服装、票务等） 内容审核系统 数据分析和模式识别 资源的有效分配和管理 风险管理体系 医疗诊断系统 精准营销测算 人脸识别技术 智能交通（自动驾驶、交通流量监测等） ……	应 用	内容生成（新闻、自动摘要等） 图像生成 代码生成 语音生成 设计 聊天机器人 AI 数字人 药物研发 虚拟现实（VR）和增强现实（AR） ……

人工智能合规的驱动因素及监管态势

人工智能合规和风险管理的驱动因素

在当前数字化时代，我们识别到了以下两个可能引发企业 AI 合规和风险管理行动的关键信号：

信号一：《人工智能法案》引发人工智能合规和风险控制

随着《人工智能法案》的制定和立法的展开，企业对人工智能合规和风险控制的关注进一步升温。法律框架的出台将深刻影响业务运营。目前，欧盟《人工智能法案》作为首部人工智能法规，已于 2023 年 12 月 8 日成功达成了初步协议，随后中国、美国以及欧盟也陆续发布了各项草案和标准，这些法规和标准成为当地商业发展的推动力，对人工智能的监管和控制起到了重要作用，并且全球越来越多的司法管辖区正在努力跟上这场竞赛，以实现创新与监管之间的最佳平衡。

各司法管辖区在推进本身框架的同时，也致力于协调和统一各异的方法，经济合作与发展组织的人工智能原则在多个法规背景下得到重申，七国集团（G7）、联合国教科文组织、国际标准化组织、非洲联盟和欧洲委员会等均在努力构建多边人工智能治理框架，致力于为全球人工智能合规体系提供更为完备的指引。

全球主要 AI 立法管辖区域图示（Global AI Legislation Tracker）[2]

[2] iapp Global AI Legislation Tracker: https://iapp.org/media/pdf/resource_center/global_ai_legislation_tracker.pdf

信号二：2024 年可信 AI（Trustworthy AI）即服务的技术趋势将推动人工智能的合规实践

2024 年，可信 AI 即服务的技术趋势对企业提出更高标准的责任和可信度要求。根据 Gartner 2024 年十大技术预测，到 2026 年，引入 AI 信任、风险和安全管理（TRiSM）概念控制于 AI 应用程序的企业有望通过消除 80%的错误和非法信息来提高决策的准确性。此外，从持续威胁暴露管理（CTEM）的角度来看，到 2026 年，根据 CTEM 计划优先考虑安全投资的组织将实现漏洞数量减少三分之二。这些技术趋势将推动企业采取更积极主动的措施，确保人工智能应用达到可信的标准。

我国人工智能最新监管态势

我国已相继制定多项法律法规，以确保人工智能技术的平衡发展和安全应用。2017 年，国务院发布的《新一代人工智能发展规划》首次将人工智能的发展纳入国家战略。在此基础上，中央于 2021 年陆续发布了《关于加强互联网信息服务算法综合治理的指导意见》和《互联网信息服务算法推荐管理规定》，标志着算法治理的开端。而后，2022 年和 2023 年发布的《互联网信息服务深度合成管理规定》和《生成式人工智能服务管理暂行办法》，从纵向上进一步完善了我国人工智能的监管框架。

近年来，我国人工智能主要的法律法规文件和关键内容如下：

	法律法规文件	重点内容
1	《新一代人工智能伦理规范》	人工智能活动应符合增进人类福祉、促进公平公正、保护隐私安全、确保可控可信、强化责任担当、提升伦理素养等 6 项基本伦理要求。
2	《关于加强互联网信息服务算法综合治理的指导意见》	提出健全算法安全治理机制，构建以算法监测、算法安全评估、算法备案、监管模式创新为主的安全监管体系，促进算法生态规范发展。
3	《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》	规定了有舆论属性或社会动员能力的信息服务提供者开展安全评估的内容、程序和要求。
4	《互联网信息服务算法推荐管理规定》	规范由算法推荐服务产生的算法歧视、大数据杀熟、网络诱导沉迷、未成年人保护等问题。
5	《移动互联网应用程序信息服务管理规定》	规范移动互联网应用程序应用程序提供者、应用程序分发平台的合法权益以及备案、审核等相关义务，维护国家安全和公共利益。
6	《互联网信息服务深度合成管理规定》	明确了在深度合成技术方面技术支持者、服务使用者和服务提供者三方主体的技术规范、备案，内容合法性审查，深度合成标识等法律义务。
7	《人工智能深度合成图像系统技术规范》	针对深度合成图像（含视频）系统框架，规定了系统技术要求，描述了对应的测试评价方法。
8	《生成式人工智能服务管理暂行办法》	明确了服务提供者的主体责任和义务，对数据合法性、知识产权保护等进行规制，并规定了算法的安全评估与备案制度。
9	《GB/T 42888 信息安全技术机器学习算法安全评估规范》	保障机器学习算法生存周期安全，开展机器学习算法安全评估。
10	《TC260- 生成式人工智能服务安全基本要求》（征求意见稿）	提出了生成式 AI 服务提供者需遵循的安全基本要求，可为生成式 AI 服务算法备案前的安全性评估提供一套可参考操作的标准。

目前，《国务院 2023 年度立法工作计划》已将《人工智能法草案》列为“预备提请全国人大常委会审议”的法律案，标志着中国人工智能领域的通用顶层立法《人工智能法》已经正式进入法律制定的程序。也预示着，中国在人工智能领域将秉承发展与安全并重、促进创新和依法治理相结合的原则，在治理中不断完善监管模式，将行业监管作为主要监管手段，以建立更为全面的人工智能治理监管体系框架。

人工智能安全合规要点解读

企业人工智能应用合规要点

近年的监管重点对计划提供人工智能服务的企业提出了要求，我们总结了以下值得关注的合规要点，旨在协助企业在人工智能应用中确保符合最新的法规标准和道德准则。

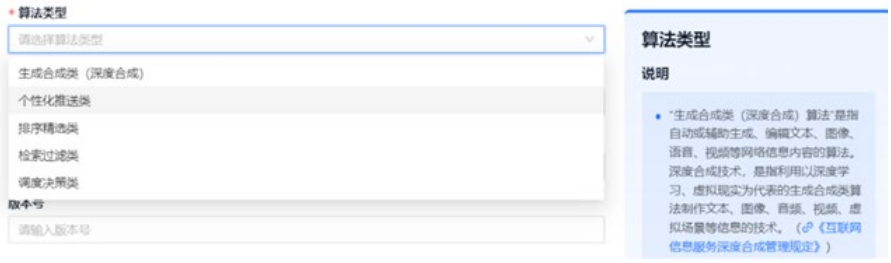
企业人工智能合规重点

收集和训练合规性	信息内容合规性	企业内部管理合规	准入合规
• 使用具有合法来源的数据（外采、自采、公开数据、爬取） • 遵循个人信息、知识产权等权益保护要求 • 采取有效措施避免歧视 • 采取措施增强训练数据的真实性、准确性、客观性、多样性	• 采取措施确保遵循伦理道德要求 • 加强内容管理，采取技术或者人工方式进行审核 • 执行知识产权、个人信息处理内容和反垄断审查 • 建立健全用于识别违法和不良信息的特征库 • 遵循特别标识要求	• 建立算法安全治理相关制度规范 • 落实内部管理机制和人员职责 • 定期展开人员培训 • 定期执行个人信息保护风险审查和合规整改 • 建立投诉举报和违法行为处置机制，协助执法	• 进行 ICP 备案 / 许可、公安联网备案、app、小程序备案等手续 • 履行安全评估要求 • 履行算法备案要求 • 取得其他必要许可（网络文化经验许可、网络出版服务许可等）

生成式人工智能算法备案流程解读

《互联网信息服务算法推荐管理规定》和《互联网信息服务深度合成管理规定》中规定了算法备案的具体要求，明确了备案主体和备案对象。

- 《互联网信息服务算法推荐管理规定》第 24 条：具有舆论属性或者社会动员能力的算法推荐服务提供者，应当在提供服务之日起 10 个工作日内，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，履行备案手续。
- 《互联网信息服务深度合成管理规定》第 19 条：具有舆论属性或者社会动员能力的深度合成服务提供者，应当按照《算法推荐管理规定》履行备案和变更、注销备案手续。深度合成服务技术支持者应当参照前款规定履行备案和变更、注销备案手续。



备案主体可通过网信办的互联网信息服务算法备案系统（https://beian.cac.gov.cn/）提交算法备案申请并查询相关信息，根据主体信息填报、算法信息填报、算法特征详细信息填报流程准备相关信息和文件。

企业算法备案准备流程

根据《互联网信息服务算法推荐管理规定》第二十五条，网络信息部门应在三十个工作日内完成备案。而基于甫瀚咨询的实际项目经验，由于在深度合成算法备案过程中涉及补正材料，完成备案通常需要约两个月的时间。

网信办安全评估要求解读

安全评估的监管要求始于 2017 年网信办发布的《互联网新闻信息服务新技术新应用安全评估管理规定》，然而，该法规明确规定了开展安全评估的主体责任范围仅限于“互联网新闻信息服务提供者”。

2018 年 11 月 15 日，网信办和公安部联合发布了《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》。其第 3 条规定，互联网信息服务提供者具有下列情形之一的，应当依照本规定自行开展安全评估，并通过全国互联网安全管理服务平台（http://www.beian.gov.cn/）向所在地地市级以上网信部门和公安机关提交安全评估报告：

（一） 具有舆论属性或社会动员能力的信息服务上线，或者信息服务增设相关功能的；

（二） 使用新技术新应用，使信息服务的功能属性、技术实现方式、基础资源配置等发生重大变更，导致舆论属性或者社会动员能力发生重大变化的；

（三） 用户规模显著增加，导致信息服务的舆论属性或者社会动员能力发生重大变化的；

（四） 发生违法有害信息传播扩散，表明已有安全措施难以有效防控网络安全风险的；

（五） 地市级以上网信部门或者公安机关书面通知需要进行安全评估的其他情形。

《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》和《生成式人工智能服务管理暂行办法》也对安全评估的要求进行了进一步重申。

在实际操作中，安全评估的实施主要由公安部门负责，整个安全评估过程从递交报告到通过通常在一个月内完成。国家目前尚未制定针对人工智能安全评估的详细细则。在实践中，甫瀚咨询根据丰富的实践经验，对企业组织人员、制度建设要求、数据隐私安全和运营管理等基础方面进行全面的基础安全评估，并就训练数据治理、信息审核、伦理和社会影响等方面展开专项安全评估。

达于行

甫瀚咨询在人工智能领域可提供的服务

甫瀚咨询始终将科技视为业务革新的主要推动力，以专业的技术视角准确把握行业技术趋势。我们能够实时且长期地进行技术弹性研究，深入分析可能影响客户 IT、技术战略以及业务的因素。在执行技术战略咨询时，我们专注于多个维度，包括人工智能的选择策略、供应商分析、应用框架评估和风险规划，从而及时调整战略，通过提高企业以技术为驱动的应变能力，不断为客户的技术战略和业务战略赋能。

我们专注于提供全方位的人工智能咨询服务，包括从 AI 产品开发到合规和安全的指导与支持。在 AI 合规方面，我们设计详实的合规框架，协助企业执行备案流程，通过评估提供全维度的合规体系，降低风险。此外，我们提供专业的 AI 安全服务，包括最佳安全开发实践咨询和算法安全指引，旨在构建安全可靠的 AI 系统，提升业务稳健性，助力客户在数字化时代取得竞争优势。

AI 相关服务

AI 产品开发咨询

- AI 安全设计基线
- LLM 使用和应用咨询
- AI 风险评估和识别
- 决策式 AI 开发咨询
- 算法生命周期管理咨询

AI 合规

- AI 合规设计 (人工智能即服务 – AIaaS)
- 算法备案 & 安全评估
- AI 隐私合规设计
- AI 隐私保护评估

AI 安全

- 最佳安全开发实践咨询
- 算法安全咨询
- 算法安全开发指引

可实现价值

可信任 AI

- 确保安全的 AI 应用体
- 确保客户的可信的 AI 应用和服务，为我们的客户赋能
- 确保安全的数据科学体系，优化 AI 参数治理
- 确保 AI 环境安全，建议安全的 AI 云计算环境

数字化持续优化

- 确保 AI 参数和输入的隐私保护以及数据安全
- 确保不同 AI 角色的持续合规
- 确保 AI 模型和架构的安全和合规
- 确保基于 AI 的能力提供可靠的数字化深度自动化
- 确保企业 AI 输出的可信以及安全

预判风险和优化服务

- 提供工业自动化 AI 决策的优化生产
- 量体裁衣 AI 决策算法的可能性
- 提供安全 AI 开发流程
- 提供必要的供应链支持（开发工业决策式AI的硬件和软件安全和合规基线）

关于甫瀚咨询

甫瀚咨询是一家全球性的咨询机构，为企业带来领先的专业知识、客观的见解、量身定制的方案和卓越的合作体验，协助企业领导者们充满信心地面对未来。透过甫瀚咨询网络和遍布全球超过 25 个国家的逾 90 家分支机构和成员公司，我们为客户提供财务、信息技术、运营、数据、数字化、环境、社会及管治、治理、风险管理以及内部审计领域的咨询解决方案。

甫瀚咨询荣膺 2024 年《财富》杂志年度最佳雇主百强，我们为超过 80% 的财富 100 强及近 80% 的财富 500 强企业提供咨询服务，亦与政府机构和成长型中小企业开展合作，其中包括计划上市的企业。甫瀚咨询是 Robert Half（纽约证券交易所代码：RHI）的全资子公司。Robert Half（于 1948 年成立，为标准普尔 500 指数的成员公司。

联系我们

张国昌
董事总经理
David.Cheung@protiviti.com.cn

王志芳
项目总监
Frida.Wang@protiviti.com.cn

徐晨曦
项目总监
Melissa.Xu@protiviti.com.cn

赵欣
项目总监
Xin.Zhao@protiviti.com.cn

公司地址

北京
朝阳区建国门外大街 1 号
国贸写字楼 1 座 718 室
电话：(86.10) 8515 1233

上海
徐汇区陕西南路 288 号
环贸广场二期 1915-16 室
电话：(86.21) 5153 6900

深圳
福田区中心四路 1 号
嘉里建设广场 1 座 1404 室
电话：(86.755) 2598 2086

香港
中环 干诺道中 41 号
盈置大厦 9 楼
电话：(852) 2238 0499

protiviti 甫瀚

让每位员工享有平等的发展机会
甫瀚咨询并非一间注册会计师事务所，故并不就财务报表发表意见或提供鉴证服务。

关注甫瀚咨询
获取更多资讯