从原则到实践：
动态监管下负责任的人工智能

AI Governance and Compliance
Working Group

CSA GCR cloud security
GREATER CHINA REGION alliance ®

人工智能治理和合规工作组永久的官方网址是：

https://cloudsecurityalliance.org/research/working-groups/ai-governance-compliance

联盟简介

云安全联盟（Cloud Security Alliance,CSA）是中立、权威的全球性非营利产业组织，于2009年正式成立，致力于定义和提高业界对云计算和下一代数字技术安全最佳实践的认识，推动数字安全产业全面发展。

云安全联盟大中华区（Cloud Security Alliance Greater China Region, CSA GCR）作为CSA全球四大区之一，2016年在香港独立注册，于2021年在中国登记注册，是网络安全领域首家在中国境内注册备案的国际NGO，旨在立足中国，连接全球，推动大中华区数字安全技术标准与产业的发展及国际合作。

我们的工作

联盟会刊下载地址
了解联盟更多信息

加入我们

CSA大中华区官网
(https://c-csa.cn)

点击会员

加入联盟

填写相关申请信息

成为CSA会员

JOIN US

致谢

报告中文版支持单位

北京数安行科技有限公司（简称：数安行）是一家数据安全厂商。公司主营产品涵盖数据分类分级、数据安全沙箱、下一代数据泄露防护、数据安全检查、数据安全风险态势感知、数据安全风险监测与风险评估、数据合规与隐私保护、数据运营安全等，主要为政府、军队、企事业单位提供数据运营安全产品和服务。数安行以DataSecOps为理念，以AI人工智能技术为核心驱动，将数据安全左移，在数据处理的第一现场对数据采取安全措施，平衡业务与安全，打造以数据运营为核心的数据安全生态体系，助力数字化转型，致力于让用户的数据安全地创造价值。公司核心团队拥有20余年产研和市场服务经验，技术积累雄厚，服务于能源、电力、金融、运营商、教育、高端制造、软件与信息技术服务、互联网、医疗、政府、军队军工等各行业客户。

报告英文版编写专家

主要作者

Maria Schwenger
Louis Pinault

参与编辑

Arpitha Kaushik
Bhuvaneswari Selvadurai
Joseph Martella

审稿人

Alan Curran MSc
Udith Wickramasuriya
Piradeepan Nagarajan
Rakesh Sharma
Gaetano Bisaz
Hongtao Hao
Jan Gerst
AshishVashishtha
Gaurav Singh
Ken Huang
Frederick Hänig
Dirce Hernandez
Tolgay Kizilelma, PhD
Saurav Bhattacharya
Michael Roza
Gabriel Nwajiaku
Vani Mittal
Meghana Parwate
Desmond Foo
Lars Ruddigkeit
Madhavi Najana

CSA 全球工作人员

Ryan Gifford
Stephen Lumpe

目 录

安全声明

本文仅供参考，不构成法律建议。

本研究文件由云安全联盟编写，探讨了当前围绕人工智能的监管治理情况。虽然本文涉及各种法律和监管框架，但必须强调的是，所提供的信息不适用于任何特定情况的法律指导。

人工智能的监管环境正在迅速演变，法律法规的解释和应用会因各种因素而存在很大差异，这些因素包括：

- 管辖范围（国家或地区）
- 具体的情景（如行业、应用场景等）
- 具体的人工智能技术或应用

因此，云安全联盟和本文作者强烈建议，如果有任何与人工智能开发、部署或使用相关的法律层面的问题或疑虑，应单独寻求法律顾问的意见。

前瞻性声明和人工智能的发展前景

本文包含了一些可能具有前瞻性的陈述。为确定其适用性，我们鼓励向相关的国家监管机构和法律顾问寻求指导。需要注意的是，这些陈述是作者和云安全联盟基于当前的知识和预期所做，受固有风险、不确定性和假设的影响，部分陈述可能与实际结果存在差异。

以下是可能影响人工智能领域未来发展和相关监管环境的一些重要因素，也是可能影响本文件中前瞻性陈述准确性的因素所在：

- **技术的快速进步：**人工智能领域不断发展，新的技术和应用层出不穷，很难预测这些技术进步的确切轨迹及其对人工智能监管各方面的影响。
- **监管框架的不确定性：**对人工智能的监管方法仍在开发，不同管辖范围内对人工智能开发、部署和使用的具体规定可能存在较大差异，并可能随着时间的推移而发生变化。

- **新兴的伦理考量：**随着人工智能应用变得越来越复杂，可能会出现新的伦理考量，促使更多有关负责任的开发和使用这些技术的法规或指导原则出台。
- **经济和社会因素：**整体经济环境和对人工智能的社会态度，可能会影响新技术的开发、采用及监管环境。

这些关于前瞻性的陈述仅反映作者和云安全联盟本文件发布之日的观点，作者和云安全联盟不承担更新或修改本文档中任何前瞻性陈述以反映未来事件或情况的任何责任。请读者不要过度依赖这些陈述。

文档摘要

本文围绕人工智能和生成式人工智能（GenAI）的法律和监管环境论述。主要强调了生成式人工智能在复杂多变的环境中面临的挑战，这些挑战源于生成式人工智能自身的多样化应用、全球监管机构采取的不同监管方法，以及对现有规定的延迟适应。

本文旨在为各组织提供基本知识，帮助其从根本上了解自身的现状，并为他们在快速变化的负责任、合规地使用人工智能方面的需求提供指导。本文探讨了部分现行法规，并阐述了在地区、国家和国际层面开发和部署负责任的人工智能的注意事项和最佳实践。

本文高度概括了当前人工智能（包括生成式人工智能（GenAI））的法律和监管情况。虽然内容并非详尽无遗，但对于组织来说，这是一个了解自身现状并确定负责任和合规的使用生成式人工智能应该考虑哪些关键因素的起点。

由于技术的不断进步以及法律和政策环境的演变，提供一份完整的概述是具有挑战性的。因此，我们建议将这些信息作为了解不断演变的人工智能法规和监管机构的基础。重要的是要意识到，人工智能法规来自全球各级政府和司法管辖区。此外，尽管数据隐私和反歧视法规等法律不是专门为人工智能设计，但这些法律将决定人工智能的使用范围和方式。例如，在美国，人工智能将受到城市、州和联邦法律、政府行为、行政令、自愿行业协议甚至普通法的监管。

在准备人工智能项目时，需要考虑到人工智能法规的起源并不总是直观的，因此需要细致分析。首个具有深远影响的法律框架是欧盟《人工智能法案》，因为它保障了个人和企业的安全及基本权利。如果某些人工智能应用干扰或威胁到公民权利，则会遭到禁止。如大语言模型等高风险人工智能系统可能会对健康、安全、基本权利、环境、民主和法治造成重大损害，预计将出台相关法规加以监管。

执行摘要

人工智能正在迅速改变我们的世界，并且具有重塑社会基本结构的巨大潜力。然而，这种变革力量也带来一个严峻的挑战：当前的法律和监管环境很难跟上人工智能，尤其是生成式人工智能爆炸性增长的步伐。本文旨在提供现有法律法规及其对人工智能开发、部署和使用影响的高层次概览。我们的目标是确定立法滞后的领域，并寻求实际的方法部署负责任的人工智能。当前的环境缺乏完善的立法，在解决日益复杂的人工智能功能的潜在风险方面存在差距。这导致现有规定，如《通用数据保护条例(GDPR)》和《加州消费者隐私法案(CCPA)》/《加州隐私权法案(CPRA)》，虽然为数据隐私提供了基础保障，但并未针对人工智能开发的独特挑战提供具体的指导，而且不足以满足例外情况下的需求。随着大型科技巨头计划向人工智能投资数千亿，预计技术创新的步伐不会放缓，技术革新的快速步伐已经超出了立法适应的能力。

一个令人担忧的缺口正在出现：生成式人工智能的广泛使用，无论是个人还是专业用途，都伴随着治理缺失的问题。恶意行为者已经开始利用生成式人工智能执行复杂的攻击，公司也将生成式人工智能视为一种竞争优势，从而进一步加快了生成式人工智能的应用。

尽管这种快速应用令人兴奋，但需要伴随着负责任的人工智能开发实践，而这些实践不能抑制创新。理想的解决方案是营造一个鼓励负责任的、透明和可解释的人工智能使用的全球环境，并辅以清晰实用的指导原则做支持。为了

弥合人工智能的无限潜力与负责任开发需求之间的差距，我们需要一种三管齐下的合作方法：所有科技公司对负责任的人工智能做出承诺，政策制定者提供明确的指导，立法机构制定有效的法规。

本文以立法和法规为重点，开启了关于人工智能治理的重要论述。它为涉足人工智能领域的从业者和企业提供了对当前人工智能治理环境及其不足之处的基础理解。旨在通过强调这些不足，促进一场关于负责任人工智能开发和应用所需法律框架的公开讨论。

引 言

人工智能领域的迅速扩展，需要法律和监管环境的不断演变，以确保在保护个人和社会的同时，负责任地发展、部署和创新。

了解人工智能的道德和法律框架有助于组织实现三个关键目标：

- **建立信任和品牌声誉：**通过展示透明、负责任的人工智能实践，与利益相关者建立信任，并提升品牌声誉。
- **降低风险：**积极采用这些框架并利用基于风险的方法，有助于降低与不负责任的人工智能使用相关的潜在的法律、声誉和财务风险，从而保护组织和个人。
- **促进负责任的创新：**通过坚持最佳实践、保持透明度、问责制和建立强大的治理结构，组织可以培育一种负责任的和安全的人工智能创新文化，确保人工智能在发展的同时对社会产生积极影响。通过多样化的团队、全面的文档记录和人类监督，负责任的人工智能将通过减轻偏见、及早发现问题以及与现实世界保持一致，增强模型表现。

范围和适用性

由于人工智能，更具体地说是生成式人工智能（GenAI）本身具有多样性，如何应对复杂的法律环境就成为了一个巨大的挑战。本文深入探讨了围绕人工智能的监管环境，涵盖了诸如生成逼真文本格式（代码、脚本、文章）的深度学习模型、处理视觉内容（面部识别、深度伪造）的计算机视觉应用、稳定扩散（文本到图像模型）以及在自主系统（自动驾驶汽车、机器人）中使用的强化学习算法等多样化系统。更广泛的类别，如生成式对抗网络和大语言模型等，是众多生成式人工智能应用的基础，要求在监管中将其纳入考虑。由于现行立法在适应这一动态环境方面面临挑战，因此有必要采取细致入微的方法管理这一广泛、快速发展的系统。由于竞争压力，快速发展的技术渗透到我们的生活和商业实践中，但与此同时，法律框架却不完善且适应缓慢，造成了一种严峻的局面。本文将探讨：

- 最广泛使用的现有法规如何参与解决生成式人工智能的特定领域问题。
- 制定新法规面临的一些挑战和机遇。
- 使用可解释的人工智能技术制定负责任的人工智能原则的高级建议和最佳实践。

本文采用阶段性的方法分析人工智能治理，重点关注以下几个方面。

表1：治理领域范围

现行文件	未来考虑因素
国家最高级别的政府机构或联邦政府的立法： ●美国： ○ 行政令 （例如，维持美国在人工智能领域的领导地位，以及关于安全、可靠和值得信赖的开发与部署人工智能技术的行政令） ○ 国会法案 （例如，2023年算法责任法案（提案） ● 欧盟: ○ 欧盟委员会政策文件 （例如，《可信人工智能伦理指南》） ○ 法规 （例如，《人工智能法案》）	国家层面: ● 亚太地区的一些规定： 中国（已颁布）（科学技术部）、日本（内阁府）、韩国（科学技术信息通信部）、新加坡、印度的国家政策是“全民人工智能” ● 其他正在制定人工智能政策的国家（加拿大、英国、澳大利亚）
主要地区性法规: ● 《加州消费者隐私法案(CCPA)》，由《加州隐私权法案(CPRA)》修订 ● 《通用数据保护条例(GDPR)》	**国际组织**：正在探索框架的有： ● 经济合作与发展组织（关于人工智能的建议） ● 联合国教科文组织（关于人工智能伦理的建议）。 ● 全球人工智能合作伙伴关系（GPAI） 汇集来自科学界、产业界、民间社会、政府、国际组织和学术界的专长，以促进国际合作。 ● ISO/IEC 42001:2023（人工智能管理系统标准） ● OWASP大语言模型应用的10大安全风险

1. 生成式人工智能法律和监管的重点领域

1.1 数据隐私和安全

生成式人工智能在数据隐私和安全领域存在独特的挑战，能够从大量数据中学习，从而引发了人们对在整个人工智能开发和部署生命周期中如何收集、存储、使用、共享和传输个人信息的关注。包括《通用数据保护条例(GDPR)》、《加州消费者隐私法案(CCPA)》、《加州隐私权法案(CPRA)》和《医疗电子交换法案(HIPAA)》在内的多项现行法律法规，旨在保护个人隐私和数据安全，具体如下：

1.2 通用数据保护条例(GDPR）（欧盟）

- **适用范围：**《通用数据保护条例》适用于在欧洲经济区(EEA)内处理个人数据的任何组织。
- **主要条款:**
  - **处理的合法依据、公平性和透明度：**组织处理个人数据必须有合法依据（如用户同意、正当利益等）。它要求向个人明确提供关于数据收集和处理目的的具体的信息。
  - **数据最小化：**将个人数据的收集和保留限制在所述目的所规定的范围内。
  - **数据主体的权利：**授予个人对其个人数据的各种权利，包括访问、更正、删除和限制处理的权利。
  - **安全措施：**要求组织采取适当的技术和措施来保护个人数据免遭未经授权的访问、披露、更改或破坏。
  - **自动化个体决策制定，包括画像：**对于包括画像的自动化决策制定，必须征得数据主体的明确同意（《通用数据保护条例》第22条）。
- **生成式人工智能的《通用数据保护条例》合规性：**

欧盟《通用数据保护条例》 要求在处理个人数据（包括用于人工智能系统的数据）时必须征得数据主体同意。此外，数据保护的要求意味着系统必须遵守《通用数据保护条例》原则，如合法性、公平性、透明度、目的限制、数据最小化、准确性、存储限制、完整性和保密性。

1.2.1 合法、透明的数据收集和处理

● **对训练数据和提示词数据的限制：**《通用数据保护条例》概述了以下处理数据的主要原则：

○ 目的限制：数据的收集和使用只能用于特定、明确界定或兼容的目的。

○ 必要性：只能收集和使用实现这些目的所必需的个人数据。

○ 数据最小化：应尽量减少收集和使用的个人数据量，只收集绝对必要的数据。

○ 存储时限：必须尽可能缩短个人数据的存储时间，并且必须定期设定和审查存储期限。

训练数据（以及提示词数据，它们也可能成为“训练数据”），即只在实际需要的范围内收集和使用数据，以达到特定的训练目标。

● **知情同意：**《通用数据保护条例》要求在收集和处理用于训练生成式人工智能模型的个人数据时获得用户的明确同意。这可以确保个人了解其数据将如何被使用（例如，用于模型训练或微调），并有权拒绝。人工智能开发者必须为那些数据被人工智能/机器学习系统处理的个人提供行使这些权利的便利。

● **透明度：**欧盟的个人对其个人数据享有访问、更正、删除、限制处理和数据可携性的权利。组织在人工智能/机器学习系统中使用个人数据时必须保证其目的、法律依据和数据保留期限的透明，让用户了解他们的数据是如何贡献到生成的输出结果中的。

1.2.2 数据安全与责任

数据安全：《通用数据保护条例》第25条规定，组织必须采取“隐私融入设计原则”和“默认隐私保护”的方法，并实施适当的技术和组织措施，以确保基础模型中使用的个人数据的安全，包括加密、访问控制和数据泄露通知程序。此外，由于大语言模型是整个供应链的一部分，保证其安全性需要特别关注对抗性攻击、数据中毒和模型偏差等恶意技术。

- **责任：**组织对在启用生成式人工智能的系统中使用个人数据负有责任，并必须证明其符合《通用数据保护条例》，这包括执行数据保护影响评估和维护适当的记录。
- **数据匿名化和假名化：**虽然匿名化和假名化有助于降低隐私风险，但在生成式人工智能的背景下，它们并不能每一次都完美应对，因为即使是有限的信息也可用于推理身份。
- **生成式人工智能输出的潜在危害：**虽然《通用数据保护条例》看起来似乎只影响用于训练模型的数据，但实际上该法规也适用于模型输出。这包括处理意外生成的输出和恶意使用深度伪造技术,这些都可能损害个人声誉和违反道德原则。所以制定明确的指导方针和保障措施对于确保负责任地开发和使用生成式人工智能、降低风险和保护个人免受潜在伤害至关重要。

1.2.3 个人权利和控制

- **访问和更正权：**个人有权了解和访问生成式人工智能使用的其个人数据，并在数据不准确或不完整时要求更正。这包括个人直接提供的信息和通过与生成式人工智能互动交互而生成的数据。然而，与传统数据库不同的是，纠正人工智能训练数据具有挑战性，因为其数据规模庞大且相互关联，这可能需要重新训练整个模型，并可能导致意想不到的后果。迄今为止，纠正人工智能模型训练数据中已经获取的不准确信息的可行性尚不明确。虽然

有关数据标记和隐私保护技术的研究仍在进行，但确保 "更正权 "仍然是一个开放性的挑战，应持续监督促进对这一需求的研究。

● **删除权（被遗忘权）：**个人有权要求删除其个人数据，这可能会影响人工智能/机器学习系统模型的训练和使用方式。由于个人数据在训练后可能会深嵌于模型内部复杂的表征中，因此落实这项权利对这些模型是一项独特的挑战。目前，从训练好的模型中移除特定数据点的技术可行性和伦理影响仍不明确，缺乏可靠的流程和既定的指导来处理此类请求，这就引发了如何平衡个人隐私与模型整体功能和社会利益这一关键问题。

● **反对权：**个人有权反对出于特定目的处理其个人数据，包括在生成式人工智能的背景下。然而，在生成式人工智能的背景下行使这项权利面临着独特的挑战。目前还没有可靠的标准化流程可以在模型训练完成后将个人数据从训练集中删除。

此外，反对权可能只适用于特定的数据元素或特定的目的，而不一定适用于用于训练模型的所有信息，这可能会限制个人反对权的范围。这凸显了开发透明且负责任的生成式人工智能以尊重个人隐私权的必要性。

● **合规性：**《通用数据保护条例》要求对数据处理活动执行数据隐私影响评估(DPIA)，这也适用于人工智能系统的数据处理及其对数据主体造成的风险。在用于训练大型生成式模型的大数据集中识别个人数据非常困难，目前还不清楚欧盟将如何处理生成式人工智能中的《通用数据保护条例》合规性问题。

● **自动化决策管理治理：**《通用数据保护条例》第22条规定，个人有权反对对其具有法律影响或重大影响的对个人资料分析执行的自动化决策。这意味着个人有权选择退出自动化决策或对自动决策做出的决定提出异议，尤其是当自动化决策可能引起偏见而对其生活产生重大影响的时候。因此，使用自动化决策的公司必须有人工申诉审查流程。

1.3 《加州消费者隐私法案》/《加州隐私权法案》(CCPA/CPRA)

- **适用范围：**适用于在加州开展业务且符合其他要求（如全球年收入超过2500万美元）的营利性企业。该法案赋予加州居民了解收集其个人信息目的的权利，并有权要求删除或更正信息以确保信息准确性。企业必须将收集和处理个人信息的范围限制在披露目的所必需的范围内。《加州消费者隐私法案》的适用范围还包括依赖于这些数据的人工智能/机器学习系统，要求企业确保这些系统在涉及加州居民个人信息的训练和输出生成过程中遵守隐私要求。企业在利用加州居民的个人信息开发和部署生成式人工智能模型时，必须绝对遵守《加州消费者隐私法案》的义务。
- **主要条款：**
  - **知情权：**允许消费者要求了解所收集的个人信息的类别和具体内容。
  - **删除权：**赋予消费者要求删除企业收集的个人数据的权利。
  - **退出权：**赋予消费者拒绝其个人信息被出售的权利。

注：《加州消费者隐私法案》和《加州隐私权法案》对消费者数据的定义比常用的“个人可识别信息”(PII)更为宽泛。因此，本文采用“个人信息”(PI)这一术语，以确保与《加州消费者隐私法案》的范围保持一致。个人可识别信息 通常指的是可以直接识别个人身份的具体数据，如姓名或社会保险号。然而，《加州消费者隐私法案》对个人信息的定义涵盖了更广泛的数据内容，包括浏览历史记录、IP地址或地理位置数据等，这些数据本身不足以识别个人身份，但在与其他信息结合后可用于识别个人身份。因此，“个人信息 ”更准确地反映了《加州消费者隐私法案》有关消费者数据隐私的意图。

- **生成式人工智能对《加州消费者隐私法案》/《加州隐私权法案》合规性：**

虽然《加州消费者隐私法案》/《加州隐私权法案》没有直接针对生成式人工智能提出技术要求，但《加州消费者隐私法案》/《加州隐私权法案》对个人数据权利的关注可能会给生成式人工智能带来重大的数据管理挑战，需要生成式人工智能采取谨慎的做法确保自身的合规性，并注意可能影响模型的性能和功能。重要的是，《加州消费者隐私法案》/《加州隐私权法案》只保护加州居民的个人数据。一些注意事项如下：

1.3.1 《加州消费者隐私法案》/《加州隐私权法案》下的数据收集、存储、使用和披露

《加州消费者隐私法案》/《加州隐私权法案》主要侧重于监管企业对加州居民个人信息的收集、使用和披露，包括用于训练人工智能/机器学习系统模型的数据，以及由此产生的包含个人信息的输出结果。加州居民有权根据《加州消费者隐私法案》/《加州隐私权法案》访问自己的个人信息，这项权利可能适用于用于训练模型的数据，但需要注意区分包含个人信息的输出与更通用的模型输出。加州居民有权了解出于人工智能目的收集了他们哪些个人信息、收集的目的以及与之共享这些信息的第三方类别。虽然《加州消费者隐私法案》/《加州隐私权法案》不一定要求披露具体的训练数据来源，但强调了透明度的重要性。

对数据来源的追溯对于《加州消费者隐私法案》/《加州隐私权法案》合规性至关重要，尤其是考虑到生成式人工智能通常使用庞大的数据集，复杂的数据来源追踪对于满足“访问权”和“知情权”请求变得困难。强大的数据治理实践、适当的记录以及可能使用匿名化的训练数据披露可有效应对这些挑战。

1.3.2 消费者权利

《加州消费者隐私法案》/《加州隐私权法案》 授予消费者有关其个人信息的特定权利，包括访问、删除、更正其个人信息以及拒绝个人信息被出售的权利。具体内容如下:

- **知情权：**要求披露为训练模型而收集和使用个人信息的详情，包括说明用于训练的数据类别（如文本、图像、音频或姓名、位置等），确定个人信息的来源（如用户互动、购买/第三方数据集、社交媒体、公共记录等），详细说明个人信息的用途（如模型训练、性能评估等）。
- **访问权：**用户可以要求访问在训练数据中使用到的具体数据点，训练过程可能会泄露可识别信息，需要在训练数据集中实施识别和隔离单个数据点的机制，如果采用匿名或聚合技术，这在技术上可能具有挑战性。

● **删除权：**用户有权要求删除用于训练的个人信息，这将对模型产生多方面影响：

○ **数据删除：**这可能需要用剩余数据重新训练模型，从而可能影响模型的性能和适用范围。

○ **数据修改：**根据训练过程的不同，可能需要对特定数据点执行匿名处理或编辑，这可能会影响模型的准确性和可解释性。

○ **知识移除：**如何识别千亿层深度神经网络中的已学知识，并移除特定的信息呢？实际上，这意味着需要从头开始重新训练大型语言模型，从而既不经济，也不环保。

从技术可行性角度看，在复杂的训练数据集中识别并移除单个数据点可能计算成本高、耗时长，甚至是对于高级人工智能系统（如大语言模型）来说有时根本不可能实现的。如何处理那些需要移除数据的训练模型这一问题至今仍然没有答案。

● **拒绝出售权：**如果生成的人工智能输出内容（如"深度伪造技术"）被认为符合 《加州消费者隐私法案》/《加州隐私权法案》规定的"个人信息"，用户有权拒绝将这些信息出售或披露给第三方。这涉及在《加州消费者隐私法案》框架下明确定义生成式人工智能输出并对其分类，可能需要进一步的澄清和法律解释。

1.3.3 合规与执行

遵守《加州消费者隐私法案》/《加州隐私权法案》主要涉及实施技术和程序性保障措施保护个人信息。

加州隐私保护局(CPPA)是一个相对较新的机构，成立于2020年，目前仍在包括消费者数据和隐私在内的不同领域建立监管规定。加州隐私保护局负责实施和执行《加州隐私权法案》和《加州消费者隐私法案》。虽然目前还没有发布专门针对人工智能管理的具体法规，但以上两项重要进展已经涉及人工智能和生成式人工智能。

1.3.4 自动化决策技术(ADMT)法规草案

● 该草案于2023年11月发布，重点关注负责任的使用自动化决策技术(ADMT)，其中包括用于面向消费者决策的多种形式的人工智能，其性质类似于《通用数据保护条例》第22条。

● 该草案概述了对企业使用自动化决策技术的要求，例如：

○ 使用前通知：使用自动化决策技术做出影响消费者的决策之前告知消费者。

○ 拒绝权：允许消费者选择拒绝仅由自动化决策技术做出的决定。

○ 访问和解释权：向消费者提供有关如何使用自动化决策技术做出关于他们的决策信息，并解释这些决策是如何形成的。

○ 风险评估：要求企业进行风险评估，以识别和减轻使用自动化决策技术可能带来的的潜在危害，如偏见和歧视。

虽然该规定没有明确提及“生成式人工智能”，但它适用于任何用于对消费者做出自动化决策的人工智能技术，以及企业在加州部署和使用生成式人工智能的方式。

1.3.5 加州关于生成式人工智能的行政令

● 2023年10月，加利福尼亚州州长加文-纽森发布了一项行政令，成立了一个工作组，负责探索在州政府内负责任地开发、使用和实施生成式人工智能。

● 该行政令强调了生成式人工智能的潜在益处，但也承认其潜在风险，如虚假信息的传播，强调了负责任的部署生成式人工智能的必要性。

● 该工作组的任务是就为加利福尼亚州州级机构制定建议，包括但不限于：

○ 识别部署生成式人工智能的潜在益处和风险。

○ 制定使用生成式人工智能的伦理原则。

○ 实施保障措施以预防潜在危害。

虽然该行政令并不直接监管私营企业，但它标志着加州以积极主动的方式了解并可能塑造生成式人工智能未来的发展和使用态势。

CPPA正在不断发展以适应生成式人工智能的复杂性，预计会有更多的合规要求。这强调了在促进负责任地开发和部署生成式人工智能的同时，持续努力应对不断变化的监管环境的必要性。

1.4 欧盟人工智能法案(EU AI Act/EIAA)

- 适用范围：EIAA适用于欧盟境内参与人工智能系统开发、部署和使用的提供者、部署商、进口商、经销商和其他运营商，不适用于军事、国防或国家安全目的。它针对人工智能系统的开发者和使用者提出了一系列规则和要求，重点关注四个风险等级：不可接受的风险、高风险、中风险和低风险。法案旨在确保对隐私和非歧视等基本权利的保护，确保人工智能系统的安全性和透明度，及如何负责任的使用人工智能系统。该法案适用于那些人工智能系统是在欧盟市场上提供或使用，或者使用影响到欧盟境内人们的运营商。法案涵盖了包括生物识别、自动驾驶汽车和关键基础设施在内的广泛的人工智能应用。
- 主要条款:
  - 禁止行为（第5条）：法案第5条概述了与人工智能系统有关的禁止行为。禁止这些行为是为了确保个人的权利和安全，防止不道德和有害地使用人工智能系统。欧盟将禁止被认为具有不可接受风险的人工智能系统，包括操纵人类行为的人工智能、社交评分系统以及用于执法目的在公共场所使用“实时 ”远程生物识别系统。
  - 基于风险的方法（第 9 条）：EIAA第9条引入了基于风险的方法监管欧盟内的人工智能系统，并平衡监管与创新，确保人工智能系统安全可信，同时避免不必要的合规成本。人工智能系统被划分为高风险、中风险和低风险，监管水平将根据其对个人构成的潜在危害程度而有所不同。

■ 高风险人工智能系统，如用于关键基础设施的系统，必须满足严格的要求，通过审查，并在部署前获得预先批准。此类系统的提供者必须遵守最严格的法规规定，包括透明度和可解释性、人类监督和独立验证。

■ 中风险人工智能系统的风险较低，但仍必须遵守特定要求。这些系统的提供者必须确保它们符合相关的法律义务、透明度和可追溯性等规则。

■ 低风险人工智能系统对个人几乎不构成风险。这些系统不受相同监管要求的约束，但仍应遵守适用于人工智能系统的法律框架。

○ 数据治理（第10条）：第10条旨在确保人工智能系统对数据的使用应透明、负责，并尊重个人隐私和数据保护权利。它要求用于训练和支持人工智能系统的数据必须遵守《通用数据保护条例》和其他相关数据保护法律的规定。高风险人工智能系统的提供者必须确保用于训练和支持人工智能系统的数据是相关、可靠、无偏见和无差错的，还应确保数据得到适当记录、标记和注释，以便监测和审计系统的性能。此外，数据管理必须透明，被使用数据的个人必须知情并同意。

○ 透明性和可解释性（第 13 条）：本条要求高风险人工智能系统必须具有透明性和可解释性，解释其运作方式，并为用户提供获取文件的途径，使个人了解其工作方式和所做的决定。人工智能模型必须保存适当的记录和日志，确保可以对其审计。本条还规定了知情权和寻求人工干预的权利，以便对人工智能系统做出的决定提出质疑，确保人工智能系统以完整、负责和透明的方式运行。

○ 人类监督（第14条）：人类监督的目的是防止或最小化风险，可通过内置在系统中的措施或由部署者实施的措施实现。此外，人工智能系统的设计必须允许由人类操作员检查。监督系统的自然人应能了解其功能、监控其运行、解释其输出并在必要时干预。对生物特征识别系统提出了具体要求。高风险人工智能系统将受到严格的义务约束，确保人为监督，并设计为能有效的由自然人监督。

○ 独立测试和验证（第57至63条）：它要求高风险人工智能系统应通过独立测试和验证，确保安全性和可靠性。

○ 治理和认证（第64至70条）：欧盟将建立治理结构和认证框架，确保欧 盟的人工智能系统符合规定的标准和条例。该法规建立了一个治理框架协调和支持国家层面和联盟层面的法规实施。治理框架旨在协调和构建联盟层面的专业知识，利用现有资源和专业知识，并支持数字单一市场。

○ 罚则（第99条）：本条规定了对违反条例规定的制裁、措施和处罚。它 规定成员国必须制定适当的行政或司法程序执行条例的规定。通过对违规行为处以重罚来遏制违规行为，以确保条例得到有效执行。它旨在确保以负责任和合乎伦理的方式开发、部署和使用人工智能系统，保护个人的权利和自由。EIAA规定，制裁和罚款根据违法行为的严重程度分级，旨在确保处罚与违规行为造成的危害程度相称。

1.4.1 生成式人工智能的EIAA合规性

1.4.1.1 要求、义务和规定

该法规旨在改善内部市场的运作，促进以人为本、值得信赖的人工智能系统在欧盟的应用。它规定了人工智能系统投放市场、投入使用和使用的统一规则，以及对高风险人工智能系统的特定要求和义务。它还包括某些人工智能的禁止行为，并为某些人工智能系统制定了透明度规则。

此外，它还涉及市场监管、市场监督治理和执法。

高风险人工智能系统提供者的义务：确保高风险人工智能系统符合概述的要求。

● 风险管理（第9条）：提供者必须对高风险的人工智能系统执行全面的风险评估，考虑对安全性、基本权利和系统预期用途的潜在风险，必须为高风险的人工智能系统建立风险管理系统，包括识别和分析已知和可预见的风险，评估可能出现的风险，并采取风险管理措施。风险管理措施的目的应是消除或减少已识别的风险，并应对条例规定和要求的综合影响。风险管理系统应确保高风险人工智能系统的总体残余风险被认为是可接受的。

● 数据质量和管理（第10条）：提供者必须确保高风险人工智能系统接受过高质量、相关性和代表性数据集的训练。它们还必须实施适当的数据管理措施，防止出现偏差以确保数据的准确性。使用数据训练技术的高风险人工智能系统必须使用接受过高质量的训练、验证和测试的数据集。必须实施设计选择、数据收集过程、数据预处理操作等数据管理措施以解决偏差和数据缺口问题。

● 技术文档（第11条）：提供者必须为高风险人工智能系统创建并维护准确和最新的技术文档。这些文件应包括系统的设计、开发、配置和运行信息。高风险人工智能系统的技术文档必须编制并保持更新。这些文档应证明系统符合法规要求，并提供必要的信息供主管部门和通报机构评估。应为属于欧盟统一立法范围的高风险人工智能系统准备一套技术文件，委员会可通过授权法案修订技术文件要求。

● 记录保存（第12条）：高风险人工智能系统必须能够在其整个生命周期内自动记录事件（日志）。日志记录功能应能识别风险情况，便于上市后监控，并监测高风险人工智能系统的运行。

● 透明度和信息提供（第13条）：提供者必须确保高风险人工智能系统的透明度，并向用户提供关于系统功能和局限性的相关信息。高风险人工智能系统必须以透明的方式运行，使部署者能够正确理解和使用系统的输出结果。使用说明应包括提供者的相关信息、系统特点和能力、已知风险、解释输出的技术能力以及输出的规定。

● 人类监督和干预（第14条）：提供者必须在高风险人工智能系统中纳入适当的人类监督和干预机制，这包括确保系统在必要时可被人工操作覆盖或停止。高风险人工智能系统的设计必须允许自然人在系统使用期间有效监督。人类监督措施应旨在预防风险或最小化风险，并可集成到系统中或由部署者实施，被指派执行人类监督的自然人应能够了解系统的功能和局限性、检测异常、解释系统输出并在必要时干预或覆盖系统决策。

● 准确性、鲁棒性和网络安全（第15条）：提供者必须确保高风险人工智能系统准确、可靠且稳健。应尽量减少与系统性能相关的错误和风险，并采取必要措施解决准确性和鲁棒性问题。应执行安全风险评估，并结合系统的设计

实施必要的缓解措施。高风险人工智能系统必须执行全面的风险评估，并遵守网络安全标准。在使用语言模型时，它们还应达到适当的准确性、鲁棒性和网络安全水平。可针对准确性和鲁棒性的技术方面制定基准和衡量方法，应在随附的使用说明中声明准确度和相关衡量标准。

● 对某些人工智能系统的具体要求（第53和55条）：该条例确定了对特定类型的高风险人工智能系统的特定要求，如生物特征识别系统、关键基础设施中使用的系统、教育和职业培训中使用的系统、用于就业目的的系统以及执法机关使用的系统。

○ 在高风险人工智能系统应在其包装/文件上注明其名称、注册商号或注册商标以及联系地址。

○ 建立质量管理系统，确保符合法规要求。

○ 保存文档，包括技术文件、质量管理体系文件、由认证机构批准的变更、认证机构发布的决定以及欧盟符合性声明。

○ 将高风险人工智能系统生成的日志保存一段时间。

○ 在将高风险人工智能系统投放市场或投入使用之前，执行相关的合格评定程序。

○ 制定欧盟合格声明并贴上CE标志，以表明符合法规要求。

○ 遵守注册义务。

○ 采取必要的纠正措施并提供所需的信息。

○ 应国家主管部门的合理要求，证明高风险人工智能系统符合要求。

○ 确保符合无障碍要求。

进口商义务:

● 在将高风险人工智能系统投放市场之前，核实其是否符合要求。

● 确保高风险人工智能系统带有符合要求的CE标志，附有欧盟合格声明和使用说明。

● 确保高风险人工智能系统得到妥善储存和运行。

● 附有认证机构颁发的证书、使用说明和欧盟合格声明的副本。

● 按要求提供国家主管部门要求的必要信息和文件。

● 与国家主管部门合作，降低高风险人工智能系统带来的风险。

经销商义务:

● 核实高风险人工智能系统是否带有CE标志，是否附有欧盟合格声明和使用说明。

● 酌情在包装/文件上注明其名称、注册商号或注册商标以及联系地址。

● 确保储存或运行条件不会违反高风险人工智能系统的合规性。

● 检验是否具有认证机构颁发的证书、使用说明和欧盟合格声明的副本。

● 按要求提供国家主管部门要求的必要信息和文件。

● 与国家主管部门合作，降低高风险人工智能系统带来的风险。

1.4.1.2 促进创新（第57条到63条）

支持创新的措施如下:

人工智能监管沙箱:

● 成员国必须在国家层面建立人工智能监管沙箱，这有助于在产品上市前开发、测试和验证创新型人工智能系统。

● 沙箱提供一个可控的环境，促进创新并允许识别和降低风险。

● 其目标是提高法律确定性、支持最佳实践的分享、促进创新与竞争力，有助于基于证据的法规学习，并为中小和初创企业的人工智能系统进入欧盟市场提供便利。

● 国家主管部门对沙箱具有监督权，并必须确保与其他相关机构的合作。

人工智能沙箱中的个人数据处理：

● 为其他目的收集的个人数据被允许在人工智能监管沙箱中处理的情况，仅包括以公共利益为目的的开发、训练和测试个人数据的人工智能系统。

● 必须满足一定的条件以确保遵守数据保护法规，包括有效的监督机制、数据主体权利的保障措施以及保护个人数据的适当技术和组织措施。

在真实环境中测试高风险人工智能系统:

● 高风险人工智能系统的提供者或潜在提供者可以在人工智能监管沙箱之外的真实环境中测试。

● 他们必须制定并向市场监督管理局提交一份真实环境测试计划。

● 测试可以独立进行，也可以与潜在用户合作完成。

● 联盟或国家法律可能要求进行伦理审查。

指导和支持:

● 人工智能监管沙箱内的主管部门为参与者提供指导、监督和支持。

● 引导提供者获得预部署服务，如法规实施指导、标准化和认证。

● 欧洲数据保护监督员可专门为联盟机构、团体、办事处和机关建立一个人工智能监管沙箱。

管理与协调:

● 该条例建立了一个管理框架，协调和支持在国家和联盟层面实施人工智能条例。

● 人工智能办公室由成员国代表组成，负责发展联盟在人工智能方面的专业知识和能力，并支持人工智能法的实施。

● 设立委员会、科学小组和咨询论坛，为法规实施提供意见、建议和专业知识。

● 国家主管部门在委员会内开展合作，并就人工智能监管沙箱的进展和结果提交年度报告。

● 委员会为人工智能监管沙箱开发单一信息平台，并与国家主管部门协调。

市场监督与合规:

● 由成员国指定的市场监督机构负责执行该法规的要求和义务。

● 它们具有执法权，可以独立公正地履行职责，并协调联合行动和调查。

● 可通过包括降低风险、限制市场供应、撤销或召回人工智能模型等措施来强制执行合规。

数据保护机构的参与:

● 国家数据保护机构和其他具有监督职能的相关国家公共机关或机构有责任根据保护基本权利的联盟法律对人工智能系统进行监督。

● 他们可以查阅根据本条例创建的相关文件。

与金融服务机构的合作:

● 负责监督欧盟金融服务法的主管机关被指定为监督人工智能法规实施的主管机关，包括对受监管和监督的金融机构提供或使用的人工智能系统有关的市场监督活动。

● 委员会与其合作，确保义务的一致应用和执行。

促进符合道德和值得信赖的人工智能:

● 鼓励未被归类为高风险的人工智能系统的提供者制定行为准则，自愿适用于高风险人工智能系统的部分或全部强制性要求。

● 人工智能办公室可邀请所有通用型人工智能模型的提供者遵守行为守则。

透明报告和文件:

● 要求提供者建立一个上市后的监测系统，分析其人工智能系统的使用情况和风险。

● 必须向有关部门报告因使用其人工智能系统而导致的严重事件。

● 人工智能监管沙箱的技术文件和退出报告可用于证明其符合法规要求。

● 委员会和理事会可查阅相关任务的退出报告。

1.4.1.3 人工智能被禁止的行为

● 严重扭曲人类行为：禁止在市场上投放、投入使用或使用以严重扭曲人类行为为目的或效果的人工智能系统，因为这会对身体健康、心理健康或经济利益造成重大损害。这包括使用潜意识成分或其他操纵性或欺骗性技术破坏和损害个人对决策过程的自主性和自由权选择。

● 基于生物特征对敏感个人信息分类：禁止使用基于自然人的生物识别数据的分类系统推断敏感的个人信息，如政治观点、工会成员身份、宗教或哲学信仰、种族、性生活或性取向。

● 提供社会评分的人工智能系统：根据自然人的社会行为、已知的、推断出的或预测的个人特征或性格特征来评估或分类自然人的人工智能系统，可能会导致歧视性结果和对某些群体的排斥。使用此类人工智能系统执行社会评分是禁止的，因为这类人工智能系统会导致个人或与数据生成或收集的背景无关的群体受到不利或不平等对待。

● 用于执法的实时远程生物特征识别：以执法为目的在公共场所对个人执行实时远程生物特征识别被认为具有侵扰性，可能会影响个人的私生活。这种做法是被禁止的，除非是在严格必要且为了实现重大公共利益的情况下，如搜寻失踪人员、生命或人身安全受到威胁，或识别特定严重刑事犯罪的作案者或嫌疑人。

1.4.1.4.合规、违规与处罚

该条例提供了各种术语的定义，并规定了其适用范围。它强调了在人工智能系统方面保护个人数据、隐私和机密的重要性。它还包括对违规行为的处罚以及对受影响者的补救措施。

此外，它还允许未来评估和审查该法规，并将实施权力下放给欧盟委员会，规定在生效后的特定时间内适用。

合规规定:

● 通用型人工智能模型的提供者必须采取必要措施，在条例生效之日起36个月内履行条例规定的义务。

● 高风险人工智能系统的运营商，如果在其投放市场或投入使用之前（即条例生效之日起24个月之前）已经使用这些系统，则只有在需要重大修改设计的情况下，才须遵守条例的要求。

● 使用高风险人工智能系统的公共机构必须在条例生效之日起六年内遵守条例的要求。

违规处罚:

EIAA 规定违规罚款遵循分级制度:

● 对于向认证机构或国家主管部门提供不正确、不完整或误导性信息的违规行为，最高可处以750万欧元的行政罚款，如果违规者是企业，则最高可处以其上一财政年度全球年营业总额1%的行政罚款，以数额较高者为准。

● 对于未获得高风险人工智能系统认证、未遵守透明度或监督要求（如风险管理）、以及未履行提供者、授权代表、进口商、经销商或使用者义务等违规行为，拟议罚款最高达1500万欧元，或其全球年营业额的3%，以较高者为准：

○ 根据第16条规定的提供者的义务。

○ 根据第22条规定的授权代表的义务。

○ 根据第23条规定的进口商的义务。

○ 根据第24条规定的经销商的义务。

○ 根据第26条规定的部署者的义务。

○ 根据第31、33(1)、33(3)、33(4)或34条对被通知机构的要求和义务。

○ 根据第50条规定的提供者和用户的透明度义务。

● 对于使用被认为构成不可接受风险的人工智能系统，或不遵守条例第5条所列的人工智能行为实践的违规行为，拟议的行政罚款最高可达3500万欧元，或其全球年营业额的7%，以较高者为准。

EIAA 要求，任何行政罚款都应考虑具体情况的所有相关因素，这包括违法行为的性质、严重程度和持续时间及其后果、受影响人数以及他们遭受的损害，罚款数额应根据人工智能系统的目的加以评估。此外，还应考虑的因素包括是否已由其他主管部门处以行政罚款、运营商的规模、年营业额和市场份额等。其他决定因素还包括违规行为带来的经济利益或损失、与国家主管部门的合作程度、运营商的责任、违规行为是如何被发现的、运营商是否存在疏忽或故意的情况，以及采取的任何为减轻受影响者所受损害的行动。它还规定，在诉讼中应充分尊重当事人的辩护权，并赋予他们查阅相关信息的权利，但要考虑到个人或企业保护其个人数据或商业秘密的合法权益。

1.5 《医疗电子交换法案(HIPAA)》

《医疗电子交换法案(HIPAA)》是美国1996年颁布的一部联邦法律，主要以其关于医疗保健数据隐私和安全的规定而闻名。

● 适用范围：《医疗电子交换法案》适用于“承保实体”，包括处理个人受保护健康信息(PHI)的医疗服务提供者、健康计划和医疗保健结算中心。

● 主要条款：

○ 最小必要标准：要求承保实体仅使用和披露实现预期目的所需的最小量的受保护的健康信息。

○ 管理、技术和物理保障措施：要求实施适当的保障措施，以保护受保护的健康信息的机密性、完整性和可用性。

○ 患者权利：赋予个人其受保护的健康信息的访问、修改和要求说明披露情况的权利。

1.5.1 生成式人工智能的《医疗电子交换法案》合规性

1.数据隐私与安全:

● **数据保护要求：**《医疗电子交换法案》严格的数据保护标准已在整个技术领域内确立，适用于所有技术类型或目的数据使用（例如，强大的加密在整个开发和部署过程中都是强制性的，以保护受保护的健康信息）。然而，生成式人工智能领域的利益相关者必须将重点转移到理解和实施在生成式人工智能操作和处理中应用现有原则的具体细微差别上。虽然既定规则不需要重塑，但要使其适应这种新情况，就必须认真关注生成式人工智能带来的独特挑战。

● **训练数据的限制：**《医疗电子交换法案》限制访问和共享受保护的健康信息，这可能会限制用于训练生成式人工智能模型以供医疗保健应用的医疗数据量。跟踪训练数据的来源和合规性对于确保生成的输出是否会继承隐私问题至关重要。这可能会使诊断、治疗预测和个性化医疗等领域的开发和准确性变得复杂，并限制人工智能模型在医疗应用中的有效性和通用性。

● **去标识化要求：**即使是从受保护的健康信息训练的生成式人工智能生成的去标识化输出也可能通过微妙的模式、相关或高级技术重新标识，从而引发

隐私问题并可能违反 《医疗电子交换法案》。虽然匿名化和假名化可以掩盖身份，但在生成式人工智能的背景下，在模型内与其他数据源结合时，往往无法阻止数据被重新识别。这就需要采用强大的隐私保护方法（如差分隐私、联邦学习等）有效保护个人身份。

- **模型共享限制：**由于隐私问题，训练有受保护的健康信息上的生成式人工智能模型之间的共享也受到限制，这阻碍了该领域的合作与进步。
- **严格的访问控制、审计和追踪：**《医疗电子交换法案》 要求对受保护的健康信息的访问和使用进行严格的审计和追踪。这将延伸到生成式人工智能系统，需要具备强大的日志记录和监控机制，以确保整个供应链符合《医疗电子交换法案》 的规定。

2.模型训练、输出和使用:

- **对训练数据的限制：**如上所述，《医疗电子交换法案》 限制了对受保护的健康信息的访问和共享，这可能会限制生成式人工智能模型用于训练医疗保健的数据量。就模型训练而言，限制使用多样性和全面的医疗保健数据集训练模型的能力可能会导致输出结果有偏差或不准确。实施差分隐私或其他匿名化技术可能有助于保护患者隐私，同时仍能为训练提供一定程度的数据效用。
- **共享和披露限制：**共享或披露包含受保护的健康信息的生成内容受到严格限制，即使是匿名 的，也可能会限制使用生成式人工智能分享医学见解或进行研究合作的能力，因此需要谨慎设计和实施。
- **限制生成受保护的健康信息：**生成式人工智能不能直接输出任何可能被视为受保护的健康信息的数据，即使是用于生成训练或测试目的的合成医疗记录等，也要遵守这一规定。
- **下游使用限制：**根据受保护的健康信息训练的生成式人工智能模型不得用于可能暴露受保护的健康信息的下游应用，即使模型本身并不直接输出受保护的健康信息。

● **模型的透明度和可解释性：**了解生成式人工智能模型如何得出其输出对于确保它不会无意中披露受保护的健康信息至关重要，这就需要可解释的模型和清晰的推理说明。

确保人工智能生成的医疗结果的透明度和可解释性对于建立信任和遵守《医疗电子交换法案》的“解释权 ”规定至关重要。

3.《医疗电子交换法案》法规可能还要求:

● **对输出结果进行仔细审查和持续监控：**所有由受保护的健康信息或包含受保护的健康信息训练的生成式人工智能模型生成的输出结果都必须经过彻底审查，以确保它们不包含任何可识别信息或有可能重新识别个体的信息，这自然会增加开发时间和对模型输出持续监控的复杂程度。

● **患者同意和授权：**使用生成式人工智能执行诊断或治疗建议等任务需要获得患者的明确同意和授权，即使这可能会增加输入或输出工作流程的复杂程度。

● **审计与合规：**使用带有受保护的健康信息的生成式人工智能的组织必须实施强大的审计和合规措施，以确保遵守适用于所有其他受《医疗电子交换法案》监管系统的《医疗电子交换法案》法规。

● **风险评估和缓解计划：**生成式人工智能利益相关者必须优先考虑定期执行风险评估，以保护患者隐私并维持《医疗电子交换法案》 合规性。这些评估应全面评估人工智能/机器学习系统，以便识别潜在的隐私违规行为并实施有针对性的缓解策略。

《医疗电子交换法案》法规对生成式人工智能在医疗保健领域的应用提出了重大挑战。这些挑战要求对人工智能系统进行全面的了解、实施和持续监控。通过精心设计这些人工智能系统、采用强大的隐私保护技术并严格遵守法规，我们可以解锁生成式人工智能在改善医疗保健方面的潜力，同时保护患者隐私并确保负责任和合规地使用。在这个新兴领域里，平衡创新与患者隐私仍然是一个关键挑战。

围绕医疗保健领域的人工智能（包括生成式人工智能）和机器学习的动态监管环境需要利益相关者不断调整适应，以确保符合《医疗电子交换法案》和其他相关法规不断演变的解释，尤其是针对生成式人工智能系统。

2. 如何应对生成式人工智能的幻觉对数据隐私、安全和伦理的影响

幻觉是指人工智能系统根据其所训练的模式和数据生成逼真但与事实不符或捏造的输出，如图像、视频或文本。这些幻觉引起了人们对关于数据隐私和安全相关立法和法规的极大关注。

生成式人工智能幻觉影响的一个关键领域是数据隐私。当生成式人工智能模型被输入敏感数据时，有可能产生无意中泄露个人或组织的私密信息，这给一些监管框架（如《通用数据保护条例》或《加州消费者隐私法案》/《加州隐私权法案》）带来了巨大挑战，因为这些法规要求采取严格措施保护个人数据免遭未经授权的访问或披露。人工智能生成内容的出现模糊了真实信息与捏造信息之间的界限，使有效执行数据隐私法的工作变得更复杂。

生成式人工智能的幻觉还为监管环境带来了安全风险。恶意行为者可能会利用人工智能生成的内容，如捏造的图像或文本，欺骗或操纵个人。这对数据系统的完整性和安全性构成了直接威胁，要求监管机构调整现有的网络安全法规，以应对人工智能生成内容带来的独特挑战。随着生成式人工智能技术的发展和模型能力的进步，确保符合安全标准可能会变得越来越复杂，尤其是在生成输出的真实性仍然不确定的情况下。

政策制定者和监管者在努力治理生成式人工智能的同时，还必须面对幻觉带来的伦理影响。除了遵守法律之外，伦理考量对于制定生成式人工智能治理的监管框架也至关重要。围绕负责任地开发和使用生成式人工智能模型的问题，包括幻觉内容对个人权利、自主性和福祉的潜在影响，都需要仔细斟酌。监管举措必须在促进创新和维护社会价值观之间取得平衡，确保生成式人工智能治理框架优先考虑透明度、问责制和包容性等伦理原则。

要解决人工智能产生幻觉的问题，必须持续评估人工智能的输出结果，从多个可信来源核实信息，并在评估内容的准确性问题上采用人类判断。此外，提供明确的提示和使用精心收集的训练数据可以从一开始就降低出现幻觉的可能性。

生成式人工智能的幻觉对人工智能治理的现有立法和监管框架带来了挑战，特别是在数据隐私、安全和伦理等方面。应对这些挑战需要政策制定者、监管机构、行业利益相关者和伦理学家通力合作，共同制定全面的治理机制，有效管理与生成式人工智能相关的风险和机遇。

2.1 国土安全部政策声明139-07对生成式人工智能的影响

- **数据输入：**禁止将美国国土安全部(DHS)有关个人的数据（无论其是否为个人身份信息或匿名信息）、社交媒体内容或任何仅供官方使用的敏感但非机密信息（现称为“受控非密信息(CUI)”），或机密信息输入到商业生成式人工智能工具中。
- **数据保留：**在工具中选择限制数据保留的选项，并选择拒绝将输入数据用于进一步训练模型。
- **输出审查和使用：**确保使用这些工具生成或修改的所有内容在用于任何官方用途之前，特别是在与公众互动时，先由适当的领域的专家审查准确性、相关性、数据敏感性、不适当的偏见和政策合规性。
- **决策：**商业生成式人工智能工具不得用于任何福利裁定、资格认证、审查或法律或民事调查或执 法相关行动的决策过程。

2.2 联邦贸易委员会政策宣传与研究指导:

- **人工智能（及其他）公司：悄悄更改服务条款可能是不公平或具有欺骗性的**

随着数据成为科技和商业创新的驱动力，开发人工智能产品的公司越来越多的将其用户群体作为主要的数据来源。然而，这些公司必须在获取这些数据与对其保护用户隐私的承诺之间取得平衡，任何试图偷偷放松隐私政策以使用更多客户信息的行为都可能导致违法。公司不能回溯性地更改其隐私政策的条款，因为这对可能在不同条件下同意该政策的消费者构成不公平和欺骗性行为。联邦贸易委员会一直以来都具有质疑公司的欺骗性和不公平隐私做法的权利，并将继续对试图无视隐私

法规和欺骗消费者的公司采取行动。归根结底，对于希望与用户建立信任并避免法律后果的公司来说，透明度、诚实和诚信是至关重要的。

- **人工智能公司：履行对用户的隐私和保密承诺**

开发人工智能模型需要大量数据和资源，并非所有企业都有能力开发自己的模型。提供模型即服务(MaaS)的公司，通过用户界面和应用程序接口为第三方提供人工智能模型，帮助它们解决这一问题。这些公司持续需要数据改进他们的模型，这有时会与他们保护用户数据和隐私的义务冲突。联邦贸易委员会对未能保护客户数据和隐私，以及滥用客户数据的公司实施法律制裁。提供模式即服务的公司无论在哪里做出承诺，都必须遵守，并确保不欺骗客户或参与不公平竞争。在人工智能模型的训练和部署过程中，虚假陈述、重大遗漏和数据滥用都会给竞争带来风险，违反消费者隐私权或采用不公平竞争手段的提供模型即服务的公司可能受到反垄断法和消费者保护法的追究。

2.3 美国白宫管理和预算办公室（OMB）在联邦机构使用人工智能的治理、创新和风险管理的政策

副总统卡马拉-哈里斯宣布了一项政府范围内的政策，旨在降低人工智能的风险并利用其益处。这项政策是根据拜登总统人工智能行政令（请参阅下文）发布的，旨在加强人工智能的安全性和保障性，促进公平和保障公民权利，推动美国的人工智能领域的创新。新政策包含了针对联邦机构使用可能影响美国人权利或安全的人工智能的具体保障措施。其目标是消除负责任的人工智能新的障碍，扩大并提升人工智能人才队伍的技能，并加强人工智能的治理。政府正在通过这一政策促进各联邦机构利用人工智能的透明度、问责制以及对权利和安全的保护。这项旨在降低人工智能风险并利用其益处的政府政策的主要亮点如下：

- **具体的人工智能保障措施：**到2024年12月1日，联邦机构在使用可能影响美国人权利或安全的人工智能时，必须实施具体的保障措施。

这些保障措施包括评估、测试和监控人工智能对公众的影响、降低算法歧视的风险，以及提高政府使用人工智能的透明度。

- **医疗保健领域的人类监督：**在联邦医疗系统中使用人工智能支持关键诊断决策时，由人类来监督这一过程，以核实工具的结果，避免医疗服务的不平等。
- **人类监督欺诈检测：**当人工智能被用于检测政府服务中的欺诈行为时，由人类监督有影响的决策，受影响的个人也有机会寻求补救人工智能造成的伤害。
- **人工智能使用的透明度：**联邦机构被要求通过发布人工智能用例的年度扩大清单、报告敏感用例的指标、通知公众人工智能豁免及理由、发布政府拥有的人工智能代码、模型和数据，来提高其使用人工智能的公开透明度。
- **负责任的人工智能创新：**该政策旨在消除联邦机构在负责任的人工智能创新方面的障碍。它强调了人工智能在应对气候危机、促进公共卫生和保护公共安全方面的应用实例。
- **壮大人工智能人才队伍：**政策指导各机构扩大人工智能人才队伍并提高其技能。
- **加强人工智能治理：**该政策要求联邦机构指定首席人工智能官来协调各机构内部的人工智能使用，并成立人工智能管理委员会管理机构内的人工智能使用。
- **停止使用人工智能：**如果某个机构不能采用规定的保障措施，就必须停止使用人工智能系统，除非机构领导层能说明为什么采用规定的保障措施会增加整体安全或权利的风险，或者会对机构的关键业务造成不可接受的障碍。

2.4 拜登总统关于安全、可靠和负责任地开发和使用人工智能的行政令

拜登总统于2023年10月发布的关于安全、可靠和负责任地开发和使用人工智能的行政令,是一项具有里程碑意义的努力，旨在解决社会关注的如何建立负责任的人工智能实践的问题。

该行政令的重点是确保安全、可靠和合乎伦理地开发和使用人工智能 ，包括数据隐私、伦理、劳动力发展和国际合作等关键领域。它概述了一项制定指导方针和最佳实践的计划，指导负责任地开发和部署人工智能技术。该计划包括委派多个政府机构，如美国国家标准与技术研究所、美国国家科学基金会和美国商务部，开发与现有框架和主题相关的资源和最佳实践，例如：

- 算法的公平性和偏见
- 人工智能模型的可解释性和可解读性
- 标准化测试和评估方法

虽然具体的监管细节还未出台，但该行政令标志着政府致力于构建一个可靠的人工智能框架。虽然拜登总统的行政令并没有重新定义人工智能的法律和监管环境，但它强调了符合伦理和负责任使用的重要性，并解决了数据生命周期中的数据隐私、安全控制和网络安全等问题。

虽然还没有制定具体的法规，但安全、可靠和负责任地开发和使用人工智能的行政令为负责任的人工智能开发和使用奠定了全面的基础，通过关注数据隐私、伦理、劳动力发展和国际合作来解决社会问题。

由联邦政府层面制定和实施的人工智能相关法规和政策的缺失导致了一个复杂的局面，许多不同的州和地区正在颁布和实施各种法规。正如美国伯克利律师事务所发布的“美国各州人工智能立法快照”中所强调的，这些法规的拼凑造成了一个关键的担忧。

3. 非歧视与公平

生成式人工智能能够产生新颖的内容并影响决策，这引起了人们对歧视和公平性的严重关切，并引发了法律和监管方面的审查。让我们回顾一下反歧视法律法规如何影响生成式人工智能的设计、部署和使用。

3.1 部分现行反歧视法律法规

解决人工智能算法和决策过程中基于受保护特征的歧视问题的现行法律和拟议法律摘要：

● **美国《民权法案》（1964年）第七章：**禁止基于种族、肤色、宗教、性别和国籍的就业歧视。如果在招聘、晋升或绩效评估中使用的人工智能系统（包括生成式人工智能），长期存在对受保护群体的偏见，那么这些系统可能面临《民权法案》第七章的审查。

● **美国平等就业机会和民权法案及其执行机构：**将第七章的保护范围扩大到年龄和残疾。基于这些特征的算法偏见也是被禁止的。

美国平等就业机会委员会(EEOC)发布的技术援助文件是其“人工智能与算法公平性倡议”的一部分，该倡议旨在确保在招聘和其他就业决策中使用的软件（包括人工智能）符合 EEOC 执行的联邦民权法律。

此外，2008年颁布的《遗传信息反歧视法》是一部联邦法律，禁止在就业和医疗保险中基于遗传信息的歧视。虽然它并不直接管理算法决策（包括人工智能系统做出的决策），但它禁止基于遗传信息的就业歧视。使用生成式人工智能系统的公司仍有责任确保其系统是公平、无偏见的，且不会基于任何敏感信息（包括遗传信息）延续歧视性做法。

● **美国的《公平住房法》：**禁止基于与《第七章》相同的受保护特征的住房歧视，用于租户筛选或抵押贷款审批的人工智能驱动工具必须遵守这些保护规定。

● **美国《平等信贷机会法》：**禁止基于种族、肤色、宗教、国籍、性别、婚姻状况、年龄或残疾的信贷歧视。该法案要求必须仔细评估人工智能驱动的信用评分模型是否会产生潜在的歧视性影响。

● **多部联邦民权法（如《民权法案》第六章、《教育修正案》第九章和《康复法案》第504节）：**禁止在教育环境中基于种族、肤色、国籍、性别、残疾和年龄的歧视。学校和教育机构必须遵守这些法律，确保其实践（包括涉及机器学习和人工智能等技术的实践）不会基于上述受保护特征歧视学生。

● **欧盟《通用数据保护条例(GDPR)》：**赋予个人访问、更正和删除其个人数据的权利。这影响到生成式人工智能系统如何收集和使用个人信息，以避免出现歧视性结果。它要求数据控制者实施防止歧视性分析和自动决策的保障措施。此外，《加州消费者隐私法案》/《加州隐私权法案》还禁止组织歧视行使隐私权的消费者。

●**《算法问责法案》（美国，2019-2020年）：**旨在建立联邦偏见审计标准，评估政府机构和企业使用的人工智能算法的公平性、责任性和透明度。

● **欧盟《人工智能法案》(2024)：**对高风险的人工智能应用提出具体要求，包括解决偏见和歧视问题。

●**《纽约市算法偏见法案》（美国，2021年）：**要求对城市机构使用的人工智能算法进行审计，以确定是否存在基于受保护特征的潜在偏见。

●**《加利福尼亚州自动化决策法案》（2023年，美国）/《纽约自动化就业决策工具法案》（2023年，美国）：**两者都要求企业在使用对消费者有重大影响的自动化决策工具时，提供通知和解释。

●**《加州消费者隐私法案》/《加州隐私权法案》**禁止歧视行使隐私权的消费者。这可能会给在含有固有偏见的数据集上训练的生成式人工智能模型带来潜在挑战。在 《加州消费者隐私法案》/《加州隐私权法案》下，减少此类偏差以确保非歧视性输出变得至关重要。

●**《美国残疾人法案》：**这是一套规定为残疾人提供便利的标准和法规，与公众互动的人工智能系统需要遵守《美国残疾人法》关于无障碍的规定。

● **《公平信用报告法》：**该法对如何收集和使用消费者信息做出了规定。金融行业中使用的人工智能模型（如用于贷款决策）需要确保符合《公平信用报告法》，以避免决策中出现不公平的偏差。

最近的一些事例，如美国的诉讼，指控有偏见的人工智能算法导致的歧视性招聘行为，以及基于欧盟 《通用数据保护条例》裁决而对基于人工智能的人脸识别系统执行更严格的审查，都凸显了人们对潜在偏见、歧视性定性以及遵守反歧视法律的必要性。

这些最近的例子凸显了人工智能招聘中可能存在的偏见，用于选择候选人的工具曾面临歧视指控：

● **新闻中的执法案例：**2023年，EEOC首次针对通过人工智能进行的歧视提起诉讼：诉讼称，麦琪教育科技因年龄原因未能聘用200多名55岁以上的合格申请人。起诉方称，她用自己的真实出生日期申请，立即遭到拒绝，第二天又用更近的出生日期申请时则获得了面试机会。因此，2023年1月，平等就业机会委员会发布了《战略执法计划草案》（2023-2027），该草案表明平等就业机会委员会明确关注在整个就业生命周期中对人工智能的歧视性使用的问题，从招聘开始，包括员工绩效管理。

● **人工智能招聘中的歧视和偏见：**2023年案例：本案例研究是人工智能招聘中存在偏见的真实案例。一家银行用于筛选求职者的人工智能工具被发现出了重要的法律问题，并强调了在招聘过程中使用人工智能工具时，对潜具有歧视性。该案例提在偏见保持警惕的重要性。

● 使用人工智能监控员工信息，2024年：这篇文章重点介绍了大型企业如何利用人工智能服务监控员工信息。它不仅能分析情感，还能评估文本和图片中的“欺凌、骚扰、歧视、不合规、色情、裸露和其他行为”，甚至还能分析不同人群（如年龄组、地点）对公司举措的反应。虽然采用了数据匿名化等隐私保护技术，但这些做法引起了人们对隐私权和言论自由的担忧。

有些人认为这是对隐私的侵犯，可能会阻碍公开交流，而另一些人则认为这是发现潜在问题和加强保护公司决策的一种方式。法律前景仍不明朗，这表明这种做法可能面临监管和社会方面的障碍。

3.2 监管方面的挑战

目前的法律框架在处理生成式人工智能中的非歧视和公平问题时面临着诸多限制：

- **适用性缺口：**现有法律难以应对复杂的人工智能系统，并且在如何将“歧视”概念转化为算法和数据方面缺乏明确性。
- **难以证明偏见：**不透明的人工智能系统使得确定和证明歧视性的意图或影响变得困难，这些系统的内部因素相互关联，使得问题更加复杂化。
- **执法挑战：**有限的资源和专业知识阻碍了有效的调查和执法，再加上人工智能发展的全球性质又使之进一步复杂化。
- **创新与监管：**快速发展的人工智能能技术超越了当前的法律框架，造成了不确定性，需要在创新和伦理考量之间取得微妙的平衡。
- **定义和实现公平：**在人工智能中实现公平是多方面的。由于公平原则之间存在不同的解释和潜在的冲突，准确的定义公平是非常复杂的。实施确保公平的措施往往会带来重大的技术挑战，并需要大量资源。
- **解释的复杂性：**人工智能模型，尤其是深度学习模型，可能异常复杂。它们可能由数百万个参数组成，因此很难理解输入数据是如何转化为输出预测的。创建能准确反映这些转换的解释是一项非同小可的任务，需要大量的计算资源和时间。
- **准确性和可解释性之间的权衡：**更精确的模型，如神经网络，通常可解释性较低。另一方面，线性回归或决策树等更简单、可解释性更强，但模型在执行复杂任务中可能不如前者。如何权衡利弊，开发出既精确又可解释的模型是一个极具挑战性的过程。生成式人工智能就是用较低的可解释性换取较高准确性的最佳范例。
- **缺乏标准化技术：**尽管存在一些解释人工智能决策的技术（如局部可理解的与模型无关的解释技术【LIME】、Shap法【SHAP】等），但没有一种通用的方法。适当的技术可能会根据模型类型和特定应用而有所不同，这意味着开发可解释的人工智能通常需要定制化的解决方案。

● **验证解释：**确认由可解释人工智能技术生成的解释是否准确反映了模型的决策过程，这本身就是一项复杂的任务。这一验证过程可能既耗时又耗费计算资源。

如今，现有的法律框架还不能很好地解决快速发展的生成式人工智能领域中的非歧视和公平问题。要填补这一缺口，需要公众的理解、建立共识和制定适应性强的法规。

3.3 监管重点和技术

生成式人工智能的监管框架应解决开发和部署生命周期各阶段的偏见和公平问题。下文列出了一些监管方面的考虑事项及其对应的解决偏见和公平问题的技术。

● 数据去偏:

○ **监管重点：**可以利用数据隐私法规来确保负责任的数据收集和使用实践。特定法规可能强制要求对敏感数据使用数据去偏技术处理，或者要求在数据处理管道中提供透明度。

○ **技术：**数据清理（例如，删除有偏见的注释，识别并纠正不一致之处）、数据扩 充（例如，生成合成数据以提高代表性）、数据加权（例如，为代表性不足群体的样本分配更高的权重）。使用所谓的 “安全 ”或 “预处理”（一些专业人士更倾向于 “预处理 ”或 “去偏差”）的数据集可以作为起点，但企业应考虑到其局限性，如偏差缓解不彻底、范围有限以及潜在的信息丢失。像IBM等公司提供此类数据集，作为人工智能开发初始阶段的垫脚石，也可根据需要在网上查找参考资料（如维基百科）。

○ **适用法规：**相关法规有：《通用数据保护条例》（欧盟） 规定了数据处理的透明度和负责任的数据收集做法；《加州消费者隐私法案(CCPA)》规定个人有权访问、删除和拒绝出售其个人数据；可能管理用于训练生成式人工智能模型和人工智能输出的数据的使用；模型卡片文档框架（Hugging Face）是一个标准化文档框架，“尤其侧重于偏差、风险和限制部分”。

- 算法透明度:

  ○ 监管重点：透明度法规可要求开发人员对模型输出提供解释，特别是在影响较大的应用中，这可能涉及标准化的解释格式或获取相关数据子集以进行独立分析。

  ○ 技术：可解释的人工智能(XAI) 方法（如显著图、反事实），可解释模型的决策过程。

  ○ 适用法规：与此相关的有欧盟的《人工智能法案(EU AI Act)》，该法案要求 “高风险 ”人工智能系统必须透明且可解释，并可能强制要求使用特定的可解释人工智能技术；美国国家标准与技术研究院（NIST）的《可解释人工智能四原则》（2021 年）；以及模型卡片文档框架（谷歌研究），该框架倡导 “共同理解人工智能模型的价值”。

- 人类监督和反馈:

  ○ 监管重点：法规可能要求对关键决策或敏感领域建立特定的人类监督机制。这可能涉及对人类审查员的资格要求、规定的审查协议或已发现的偏见进行强制性报告。

  ○ 技术：人在循环系统中、具有人类反馈循环的主动学习、数据主体的明确同意和人类审查模型输出。

  ○ 适用法规：美国食品药品监督管理局（2021）提出的产品全生命周期(Total Product Life Cycle, TPLC)方法提倡人类监督，像“监控人工智能/机器学习设备，并在算法变更的开发、验证和执行过程中纳入风险管理方法以及‘决定何时为现有设备的软件变更提交510(k)’指导文件18中概述的其他方法。”

- 人工智能发展中的多样性、公平性和包容性 (DE&I):

  ○ 监管重点：平等和非歧视法律可用于确保人工智能团队内部的公平招聘和开发实践。法规可规定开发团队的多样性指标，或要求在部署前执行偏见影响评估。

  ○ 技术：在开发团队中培养多元化思维，将多元化、公平、公正和包容原则纳入设计和测试阶段，并执行偏见审计和影响评估。

○ 适用法规：虽然针对人工智能的具体DE&I法规仍在制定中，但企业必须主动采用伦理标准来确保其人工智能系统公平公正，并利用这一机会“将 DEI 嵌入公司的人工智能战略”（《哈佛商业评论》，2024年）。行业指南强调，“人工智能的方法必须合乎伦理和公平，以确保它能赋予社区权力并造福社会”（世界经济论坛，2022年），避免偏见和歧视。

● **算法的透明度和可解释性:** 确定对人工智能决策透明度和可解释性的要求（如可解释的人工智能倡议），尤其是在高风险情况下。探索要求人工智能决策可解释性的法规，特别是在高风险应用中，以及这些法规如何影响组织的做法。一些相关文件包括：

○《算法问责法》，2021-2022年：这些法案在多个州提出，旨在建立透明度，确保审计用于关键决策的人工智能系统，以减少差异影响，“应对人工智能和自动化系统已经造成的问题”。

○《算法问责法案》，2023-2024年：《算法问责法案》（2023年9月）目前处于初步阶段，旨在建立一个负责任地开发和使用人工智能系统的框架。虽然具体细节仍在制定中，以下是可能关注的一些领域：

● **透明度和可解释性：**要求开发人员解释人工智能系统是如何做出决策的，以提高公众的理解和信任。

● **数据隐私与安全：**建立保障措施，保护用于训练和部署人工智能系统的个人数据。

● **算法公平性和偏见：**通过解决数据和算法中的偏见，减少可能出现的歧视性结果。

● **风险评估与缓解：**识别并解决与人工智能相关的潜在风险，如安全、安全保障和公平问题。

随着该法案在立法程序中的推进，有关治理人工智能和生成式人工智能的具体规定将会更加清晰。

3.4 新兴监管框架、标准和指南

2023 年《人工智能权利法案》（白宫蓝图）：这套非约束性的指导方针强调了公平、无歧视地使用人工智能系统的必要性。它建议采取保障措施，防止算法歧视和有害偏见的发生。

《联合国关于人工智能的全球决议》：联合国关于支持安全、可信和以人为本的人工智能的决议呼吁成员国促进开发和使用安全、可信、以人为本和透明的人工智能。决议还强调，必须确保使用人工智能是以尊重人权和基本自由为前提，且不带偏见和歧视。此外，决议还鼓励成员国共同努力，为开发和使用人工智能制定国际规范和标准。联合国关于人工智能决议的一些关键点包括：

- 鼓励成员国促进开发和使用安全、可信和以人为本的人工智能。
- 强调在使用人工智能时必须尊重人权和基本自由，同时做到透明，不带偏见和歧视。
- 呼吁各成员国相互合作，制定开发和使用人工智能的国际规范和标准。
- 鼓励成员国分享开发和使用人工智能的最佳实践和经验，以帮助确保人工智能造福整个社会。
- 呼吁与包括政府、民间社会和行业在内的各部门利益相关者继续对话和互动，以引导社会以负责任和合乎伦理的方式开发和使用人工智能。

美国国家标准与技术研究院（NIST）人工智能风险管理框架：该框架旨在帮助组织识别、管理和减轻与人工智能系统相关的风险，包括与偏见和歧视相关的风险。它鼓励在人工智能开发中纳入多样性、平等性和包容性等考虑因素。有关该框架的更多细节，请参阅《人工智能风险管理框架》（人工智能 RMF 1.0）。

有效的人工智能法规应促进标准化、问责制和国际合作三个关键方面：

- **标准化：**包括建立检测、预防和减少偏见的通用方法，例如采用ACM图书馆2019年论文中提出的“模型卡片”的标准化格式。
- **问责制：**需要明确的责任和问责框架来激励负责任地开发和部署人工智能。
- **国际合作：**通过国际合作实现跨边界的协调一致和有效地人工智能监管方法。

现有的框架、指南和资源可用于鼓励以合乎伦理、透明和可信的方式设计、开发、部署和运行人工智能。例如：

● 国际内部审计师协会（IIA）的人工智能审计框架提供了一种全面的方法评估人工智能系统的可信度。它侧重于四个关键领域：治理、伦理、控制和人的因素。更多关于三个总体组成部分（人工智能战略、治理和人的因素）以及七个要素（网络韧性、人工智能能力、数据质量、数据架构和基础设施、绩效衡量、伦理、黑盒）的详细信息可以在框架文档中找到。

● IBM 的“可信赖的人工智能”道德规范提供了确保人工智能的设计、开发、部署和运营符合伦理规范并具有透明度的指导方针。

● 微软的“负责任的人工智能实践”是值得信赖的人工智能开发和使用的指南和原则。

● AWS 的“负责任的人工智能核心要素”是安全、负责任地开发人工智能的指导方针和原则，采取以人为本的方法，优先考虑教育、科学和用户。

● 谷歌的“负责任的人工智能实践和原则”旨在采用以人为本的设计方法，以负责任的方式指导人工智能的开发和使用。

● 艾伦-图灵研究所的“理解人工智能伦理与安全”指南是有关人工智能伦理、潜在益处、挑战和案例研究的入门资源，突出强调了人工智能伦理问题的案例研究。

● 人工智能伙伴关系的人工智能事故数据库是一个收集人工智能系统造成意外伤害的真实案例的资料库，而由电气和电子工程师协会（IEEE）制定的“伦理协调设计”准则为设计符合伦理、透明和可信的人工智能系统提供了建议和框架。

这些资源为促进人工智能合乎伦理的使用提供了建议，同时对用户保持透明。这些资源还涵盖了人工智能的潜在益处、实施人工智能所面临的挑战以及与人工智能系统造成意外伤害的伦理问题和事件相关的案例研究等主题。

OWASP 大型语言模型应用程序10大漏洞项目是一项旨在教育开发人员、设计人员、架构师、管理人员和组织有关部署和管理大型语言模型时潜在安全风险的倡议。该项目提供了大型语言模型应用程序中经常出现的10大最关键漏洞的综合列表，突出强调了这些漏洞的潜在影响、利用的难易程度以及在实际应用中的普遍性。漏洞包括提示注入、敏感信息泄露（数据泄漏）、不安全的插件设计以及未经授权的

代码执行/模式窃取等。该项目的最终目标是提高人们对这些漏洞的认识，提出补救策略，改善大型语言模型应用程序的安全状况。

同样，OWASP 机器学习十大风险项目（目前正在起草）全面概述了机器学习系统的十大安全问题。该项目旨在让开发人员、设计人员、架构师、管理人员和组织了解在开发和部署机器学习系统时可能存在的安全风险。该项目提供了一份机器学习系统中经常出现的关键漏洞的综合清单，突出强调了这些漏洞的潜在影响、利用的难易程度以及在实际应用中的普遍性。这些漏洞包括对抗性攻击、数据中毒和模型窃取等。该项目的最终目标是提高人们对这些漏洞的认识，提出补救策略，改善机器学习系统的安全状况。

为支持负责任的人工智能实践，ISO专门制定了几项关键标准。下面是一些例子：

- ISO/IEC 42001:2023 是一项为人工智能系统提供管理体系框架的标准。该标准概述了管理人工智能系统生命周期（包括其开发、部署和维护）的系统方法。它有助于组织建立和实施一个运作良好的将风险、伦理、社会和法律因素考虑在内的人工智能系统管理系统。该标准强调了开发透明和负责任的人工智能系统的要考虑到各利益相关方的需求。它还鼓励组织实施负责任的人工智能能实践和治理，坚持伦理原则，包括尊重人权和隐私。

- ISO/IEC 23053:2022 是一项为使用机器学习开发、部署和管理人工智能系统提供框架的标准，该标准制定了一个流程模型，概述了开发和部署人工智能系统的数据收集和处理、模型训练和验证、系统部署以及持续监控和维护等关键活动。该标准强调了以符合伦理和负责任的方式开发和部署人工智能的重要性。它为风险评估和风险管理提供了指导，包括识别潜在风险和降低风险。该标准还涉及与人工智能系统的信任、透明度和问责制有关的问题，强调人工智能输出结果需要具有可解释性和可解读性。

如需进一步了解特定行业的人工智能治理与合规情况，请参阅云安全联盟的《人工智能韧性：一项革命性的人工智能安全基准模型》文件。

3.5 安全、责任和问责

生成式人工智能发展迅速，能够自主生成从创意文本到非常逼真的图像和视频等输出结果，这无疑开创了一个技术奇迹的新时代。然而，这一进步也迫使我们解决有关安全、责任和问责的关键问题。最近的一些例子，如Gemini生成带有偏见的视觉效果（谷歌博客，2024年2月）和加拿大航空公司机器人提供错误的退款信息（纽约邮报，2024年2月），凸显了人工智能不当行为的真实后果。这不得不引起我们的关注：当事情出错时，谁该负责，谁将首当其冲地承担生成式人工智能导致的不当甚至危险结果？我们目前是否有必要的立法管理和有效框架保证负责任地使用这一强大的技术？政策制定者和行业领导者如何合作制定负责任地使用生成式人工智能的国际标准？我们可以采取哪些技术保障措施限制恶意使用生成式人工智能的可能性？

要降低与生成式人工智能相关的潜在风险，就必须采取多管齐下的方法，包括：

- **行业标准：**为生成式人工智能的开发、部署和使用制定明确、全面的指导方针。这些指导方针必须优先考虑公平性、减少偏见和负责任的数据处理。
- **法律框架：**仔细考虑如何平衡责任归属与促进负责任的创新，制定能解决人工智能生成内容造成伤害时的复杂责任归属问题法律框架。
- **组织风险管理战略：**为组织配备工具和知识，实施强有力的保障措施和负责任的使用策略，使其能够有效地评估和管理使用生成式人工智能带来的风险。

3.5.1 生成式人工智能的责任、风险和安全考虑因素

尽管生成式人工智能有潜在的好处，但也存在固有的风险。以下是几个主要的关键领域。

3.5.1.1 生成式人工智能故障带来的法律责任风险

- **偏见和歧视：**用有偏见的数据训练的生成式人工智能模型可能会在生成的内容中延续有害的刻板印象，从而导致歧视性结果。此类法律问题可能涉及不公平的住房、就业/招聘行为、产品推荐或贷款申请/批准等方面。
- **隐私侵犯：**生成式人工智能系统通常需要访问大量数据，这引发了对用户隐私和敏感信息滥用的担忧。它们可能会无意间泄漏训练数据中使用的敏感信息，从而导致隐私泄露和法律后果。
- **信息安全和人身安全问题：**在医疗保健或自动驾驶汽车等关键领域，生成式人工智能的故障可能导致安全隐患、事故责任归属问题，甚至人身伤害。
- **虚假信息和恶意使用：**生成式人工智能可被用于生成深度伪造内容、操纵内容、生成假新闻、传播虚假信息等，从而对公众信任和民主讨论构成威胁。这可能会引发诽谤和欺诈性的法律问题。

3.3.1.2 责任分配的法律框架

确定和分配人工智能系统（尤其是生成式人工智能）所造成损害的责任是一项复杂的法律挑战。现有的法律框架往往难以应对人工智能的独特性，从而导致不确定性。虽然传统的法律原则，如产品责任法、过失法和数据隐私法可能适用于某些情况和管辖范围，但人工智能技术的动态性质要求制定新的法律框架。

如专门针对人工智能立法算法透明度的新法律，正在不同地区开始出现。这些框架旨在应对人工智能系统带来的独特挑战，重点关注与偏见、问责制、透明度和公平性相关的问题。然而，这些法规的实施和范围在不同的管辖范围，甚至不同的使用案例之间都可能存在很大差异。

国际倡议，如经济合作与发展组织（OECD）的人工智能原则，为全球范围内促进负责任的人工智能发展和部署提供了指导。这些原则倡导人工智能系统内的透明度、问责制和包容性等基本价值观，为合乎伦理的和可持续的人工智能创新奠定

了基础。虽然这些原则不具有约束力，但它们构成了制定未来人工智能政策和法规的基础框架。

尽管做出了这些努力，但驾驭围绕人工智能能责任的法律环境仍然十分复杂。法律解释和适用性高度依赖于具体情况，需要对每个案例和其所属的管辖范围进行透彻分析。因此，在确保合规性和降低人工智能相关风险方面，寻求人工智能法律专家的专业指导至关重要。

建立清晰、可预测的法律框架对于促进创新，同时确保用户安全和社会福祉至关重要。

3.5.1.3 保险

可以通过专门的人工智能责任保险政策，减轻人工智能带来的风险，分担人工智能系统造成的潜在伤害的经济负担。

3.5.2 对生成式人工智能幻觉的保险

幻觉保险是随着生成式人工智能日益融入我们生活和业务的各个方面而出现的一个新概念。顾名思义，该保险旨在减轻 “幻觉”--即生成式人工智能系统输出中的错误信息、偏见或事实性错误所造成的经济和声誉损失。

该保险试图为生成式人工智能幻觉可能带来的后果提供财务保护，包括：

- **经济损失：**包括纠正错误的成本、法律费用、声誉损失，以及因不准确或误导性输出而造成的商机损失。
- **监管罚款费用：**当人工智能生成的输出违反法规或伦理准则时，保险有可能帮助支付监管机构施加的罚款或处罚。
- **网络安全漏洞：**如果生成式人工智能系统遭受攻击或暴露敏感信息，保险可协助进行补救并承担潜在的法律后果。

以下因素促成了幻觉保险的出现：

● **对生成式人工智能的依赖日益增加：**随着各行各业越来越多地使用生成式人工智能，对风险管理策略的需求就变得更加迫切。

● **潜在的代价高昂的后果：**人工智能的幻觉有可能造成重大的经济和声誉损失，因此保险成为风险管理的重要工具。

● **不断变化的监管环境：**随着有关人工智能使用的法规不断发展，保险可以确保合规性并降低法律风险。

虽然幻觉保险仍处于早期阶段，但预计其功能将与其他类型的保险运作类似。企业或个人将支付保费以换取针对特定风险的保障，具体的覆盖风险以及侧重于财务补偿还是风险管理策略会根据生成式人工智能的应用和投保人的需求而有所不同。

虽然幻觉保险的具体形式和结构仍在确定之中，但随着生成式人工智能应用的不断增加，预计这种保险类型将成为保险业的一个新亮点。一些专家认为，幻觉保险有可能成为类似于其他形式的责任保险或网络保险一样标准的企业必需品，尤其是对于那些严重依赖生成式人工智能的公司而言。

从技术角度来看，幻觉保险并非灵丹妙药。负责任地开发和部署生成式人工智能系统，以及提高用户意识和批判性思维、人类监督仍然是最小化风险的关键因素。尽管如此，这种新颖的保险产品仍有可能为涉足人工智能领域的企业提供急需的保护，促进信任并降低这一强大技术带来的潜在风险。

3.6 知识产权

生成式人工智能引发了有关所有权、版权和责任归属的复杂知识产权讨论，而目前这些讨论缺乏明确的法律框架。面临包括所有权不明确、训练数据可能导致版权侵权以及输出责任不明确等挑战。

机遇在于制定未来法规、促进创新和探索新的知识产权模式。随时了解立法更新和法院裁决对于驾驭这一快速演变的环境并做出明智决策至关重要。2020年联合国人工智能活动报告明确概述了“人工智能技术对知识产权的大量需求”，其依据是

世界知识产权组织（WIPO）研究（2019年）中分析的“自20世纪50年代以来，超过34万项与人工智能相关专利申请和160万篇科学论文"等相关内容。

下文将介绍当前的知识产权框架如何尝试处理人工智能生成的模型、算法和数据问题，并强调了许可和保护方面的注意事项。

3.6.1 著作权、发明权和所有权

现有的知识产权框架，包括专利、版权和商业秘密，都是以人类创造者为中心建立的。例如，美国版权局拒绝为完全由人工智能创作的作品授予版权。不过，如果存在实质性的人类参与的人工智能辅助创作的作品可以被授予版权。这一概念存在灰色地带，因为需要多少人类创造力才能获得版权保护仍然不明确，这可能会在法庭案件中争论不休。

从版权局可以看出，美国的重点仍然是人类的贡献。

法院和立法者很可能需要寻找“足够的人类作者身份”的证据，比如在训练数据、提示、设计选择或创造性元素的选择等方面。这种不断发展的环境需要新的框架认可，特别是在人类和人工智能合作创造的共同发明者身份方面。

3.6.1.1保护生成式人工智能组件

- **算法与模型：**这些通常被视为商业机密，只要保密并且提供了竞争优势就可以受到保护。保护独特的算法是一种选择，但随着模型的复杂性增加，特别是其内部决策过程和数据依赖性的增加，保密工作变得具有挑战性。一些关于神经网络逆向工程和模型盗用的出版物讨论了这些挑战，强调了复杂模型保密的难度。在为模型申请专利方面，大语言模型处理和生成文本或其他输出的统计方法的核心概念和基本数学原理的属于抽象概念或自然现象而不能申请专利。但是，如果大语言模型如独特的架构或训练算法的具体实现方式符合新颖性和非显而易见性的标准，则可以申请专利。此外，将大语言模型与特定应用结合（如为医疗诊断量身定制的大语言模型）也可以申请专利，因为这种结合本身就创造了一种独特的创造性解决方

案。简而言之，虽然为整个模型申请专利可能比较困难，但如果符合新颖性和非显而易见性的标准，其中的具体技术特征和特定的实现方式是可以申请专利的。

● **数据:** 所有权取决于来源和用途。公共数据可以自由使用，而授权数据则需要遵守特定条款。人工智能训练数据的所有权可能很复杂，尤其是在数据是从多方获取的情况下。2022至2023年期间，使用全球互联网数据训练模型不考虑知识产权问题严重，导致2023年出现了多起相关诉讼，所以适用数据隐私法规和适当的许可协议是至关重要。

3.4.2 版权保护

现行版权法保护原创性表达。人工智能生成的作品提出了作者身份和原创性问题。现行法律保护人类创作的作品，这对算法生成的作品提出了挑战。截至本文发表之日，美国版权局拒绝为纯人工智能生成的作品提供保护，这引发了国际辩论。一个悬而未决的问题依然存在：人工智能生成的艺术作品（如诗歌或音乐）能否达到原创性保护标准，尤其是在严重依赖受版权保护的训练数据的情况下？当今的立法强调人类在数据选择、提示、输出内容的编辑以及训练数据的合理使用原则中的作用。

合理使用原则（美国版权局，2023年）允许在未经版权所有人许可的情况下有限度地使用受版权保护的材料，这种使用方式是对版权材料的转化性使用，即以版权材料原本未被预期的方式使用。例如，谷歌成功地辩称，转化性使用允许其从书籍中抓取文本以创建其搜索引擎，并且目前，这一决定仍然是有先例的。生成式人工智能系统可能在其创造过程中使用网络抓取数据，也属于同一类别。

3.6.3 专利保护

专利保护新颖和非显而易见的发明。如果人工智能模型中的算法符合这些标准，就可以申请专利。与版权类似，专利也需要有一个人类发明者。虽然人工智能辅助发明已经存在，但将发明权归属于算法的问题仍未解决。2024年，美国专利和商标局提供了一些指导，指出识别人工智能生成发明的非显性和发明权的归属认定可能

具有挑战性。考虑的重点应放在技术方面，突出生成式人工智能提供的进步和解决方案，而不仅仅是产出。其中一个关键要素是在专利申请中适当（透明）地披露人工智能功能，以避免未来出现异议。

3.6.4 商业秘密

商业机密是提供竞争优势的机密信息。企业必须以最谨慎的态度保护其商业机密，以免未经授权的实体获取此类商业机密。虽然根据现行法律人工智能/机器学习的算法、模型和训练数据可能符合商业机密的标准，但所有管辖范围的法院尚未对此做出明确裁决。商业秘密保护的要求可能各不相同，因此寻求具体的法律建议是必要的。

对于复杂的人工智能/机器学习系统来说，保密可能很具挑战性，尤其是在开源开发的情况下，重要的是，商业秘密只能提供针对未经授权获取的有限的保护，而不能防止类似技术的独立开发。所有利益相关者都必须采取强有力的措施保护其人工智能模型和训练数据的机密性。

3.6.5 许可和保护策略

- **开源模型/共享资源许可：**公开共享人工智能模型可以加速开发，但这是一种“复杂的方法”（Semantic Scholar.org，2021 年），可能会引发对基础训练数据的滥用和侵权潜在版权的担忧。虽然“创作共享”许可可用于人工智能生成的作品，但慎重选择适当的许可类型至关重要，因为有些许可比其他许可允许更广泛的商业使用。建议咨询熟悉开源许可的法律顾问。
- **商业许可：**开发和部署生成式人工智能模型的公司需要精心设计许可证保护其知识产权，同时允许商业使用。与合作者和用户签订的合同协议应明确界定所有权、使用权和责任。
- 数据许可：获取用于训练和微调人工智能模型的数据的适当许可是避免侵犯版权和违反隐私的关键。值得注意的是，根据使用的具体数据类型，还可能引发

其他法律问题，例如商业秘密的不当使用或特定于某些数据类型的隐私法律（例如，健康数据）。

3.6.6 商标

商标是指联邦注册的符号、单词或短语，用于识别和区分商品或服务的来源。与商业秘密不同，商标为如徽标、口号或特定的产品设计等标志性的品牌元素提供法律保护。这在人工智能生成图像领域尤为重要，因为独特的视觉输出可以成为宝贵的品牌资产。

虽然版权可以保护用于生成人工智能图像的特定代码或流程，但生成的图像本身可能根据其独特性和商业用途而受到商标保护。围绕人工智能生成商标的法律环境仍在不断演变，以下是几个关键的考虑因素：

- **独特性：**为了获得商标保护，人工智能生成的图像必须具有其本身的独特性，即不能仅仅是描述性或通用性的。如果人工智能生成的图像通用或与现有商标相似，就很难获得保护。
- **作者身份：**当前的商标法律环境通常要求有人类作者，这就让了人工智能的发展因创意角色演变而更加复杂。
- **品牌使用：**图像的使用方式必须能够识别产品或服务的来源。例如，在包装或营销材料中持续使用人工智能生成的视觉效果可以加强商标声明。

商标保护不是自动的，需要采取积极主动的措施强制执行。组织有责任注册并监控未经授权的使用。这可能涉及：

- **商标注册：**在相关商标局注册人工智能生成的商标可以在出现侵权时加强法律地位。
- **积极监控：**定期检查在线市场和竞争对手的活动是否存在潜在的商标侵权行为。
- **执法行动：**如果发现侵权行为，可能需要咨询法律顾问，并对未经授权的用户采取适当行动。

企业应采取积极措施，将人工智能生成的图像作为商标加以保护，确保消费者将独特的视觉效果与其品牌联系起来从而维护其竞争优势。

商标法非常复杂，不同管辖范围的具体规定可能有所不同。建议咨询专业的知识产权律师，以获得通过商标保护人工智能生成图像的具体指导。

3.6.7 变化的形势：

● **国际差异：**目前，各国有关生成式人工智能的知识产权法律不尽相同，这给全球企业带来了挑战，需要进行协调才能实现人工智能的国际合作和商业化。虽然在欧盟还没有明确的指令，但一些国家，如英国，现有的版权立法（1988年《版权设计和专利法》）已经涵盖了“计算机生成的作品”。该文件第9(3)条规定“在文学、戏剧、音乐或艺术作品由计算机生成的情况下，作者应被视为为创作该作品而操作计算机的自然人"。

● **持续的政策讨论：**关于调整知识产权框架以更好地应对生成式人工智能的独特挑战的讨论仍在进行中。为人工智能生成的作品提供新的特殊保护（如 “训练人工智能模型的特殊权利”），以及对版权和专利等现有类别进行修订等事宜也正在进行。

展望未来，政策制定者和法律专家正在积极探索应对人工智能产生的知识产权问题的解决方案，我们可以期待持续的辩论和即将到来的立法变革。标准化的许可模式和更清晰的所有权归属对于促进负责任和可持续的人工智能发展是迫切需要的。此外，还需要制定更多针对伦理考量的法规，例如训练数据中的偏见和人工智能生成内容的潜在滥用问题，这些问题需要立法关注。

3.6.8 相关立法

● 拜登总统在2023年10月发布的关于人工智能的行政令专注于确保人工智能的安全、可靠和可信发展与使用，并涉及数据隐私、伦理、劳动力发展和国际合作等方面。虽然该命令没有为人工智能生成的成果建立新的知识产权权利，但它承认了围绕人工智能和知识产权的复杂性，并指出“现有的知识产权法律框架可能并不完全适合解决人工智能所带来的独特挑战”。该行政令认识到需要“清晰且一致的指导”，

并指示包括美国专利商标局（USPTO, 2024）和版权局在内的几个机构，在一年内制定解决与人工智能相关的知识产权问题的推荐方案。

● 世界知识产权组织举办了一个“多方利益相关者论坛，以促进对整个经济和社 会发展中人工智能应用所涉及的知识产权问题及其对经济和文化产品与服务的创造、生产和分配的重大影响的理解”。世界知识产权组织大会的几届会议已经审议了人工智能对知识产权政策的影响。

4. 负责任人工智能的技术战略、标准和最佳实践

本节总结了我们已经讨论过的一些实施负责任人工智能技术的标准和最佳实践，并提供了一个简短的案例研究展示成功的实施方法。

组织经常面临的一个共同问题是，如何利用成熟的技术标准，在使用人工智能的过程中展示透明度、问责制和伦理实践。技术标准可以有多种分类方式。我们采用了简化的分类方法，以便将来根据需要扩展。

4.1 公平与透明度

● 数据公平性:

○ **数据集的多样性：**积极策划具有代表性和多样性的数据集，最大限度地减少输出中的意外偏差。

○ **数据集审计：**定期审计生成式人工智能模型的训练数据集，识别潜在的偏差和不足。采用数据扩充或合成数据生成等技术提高多样性和包容性。

○ **数据透明：**公布用于训练生成式人工智能模型的数据集信息，包括其组成、来源和必要的预处理步骤。这样可以执行外部审查，有助于识别数据中潜在的偏差或差异。

○ **定期评估偏差：**主动实施工具和流程来识别并减轻数据集和生成式人工智能模型中的偏差。定期测试和验证，检查是否存在歧视性输出。

○ **减少偏见：**在开发和部署阶段，积极使用和开发公平性指标和偏差缓解技术，检测和解决生成式人工智能模型中的偏差。

- 算法透明度:
  - **文档：**详细记录生成式人工智能模型的设计、架构和决策过程并以可访问的格式向利益相关者分享这些信息，以促进理解和审查。
  - **模型的可解释性：**采用可解释人工智能技术如局部可解释模型无关解释(LIME)或夏普利加性解释(SHAP)，通过深入了解生成式人工智能模型是如何得出特定结果，来识别潜在的偏差。
  - **模型卡片：**创建“模型卡片”，以透明的方式概述模型的预期用例、训练数据、性能指标、局限性和潜在偏差。模型卡片作为机器学习模型的透明文档，应该详细说明模型的训练数据、局限性和预期用途。为了促进责任的人工智能，公司可以利用Hugging Face, TensorFlow Model Garden或Papers With Code等展示的模型卡片，其中包括数据来源、组成和预处理步骤等信息。这样既能增进信任，又能让用户了解人工智能系统的潜在偏差和局限性。
  - **开源模型：**尽可能地贡献到开源生成式人工智能模型中，以便于更广泛的审查和协作改进。
- 可解释性:
  - **可解释的模型：**尽可能优先使用具有内在可解释性的生成式人工智能模型，提供对决策过程的深入了解。
  - **可解释的人工智能：**结合使用技术来解释生成式人工智能模型是如何做出决策或产生输出的。利用可解释人工智能技术生成解释，即使对于黑盒模型也是如此，突出影响输出的因素。这为用户和利益相关者提供了推理过程的见解，并促进了对模型功能的理解。
  - **可解释的界面：**在用户界面提供清晰的解释和理由来支持生成式人工智能的输出，从而培养信任和理解。

4.2 安全与隐私

- 数据安全：
  - **加密：**对静态、传输和使用中的敏感数据加密。

○ **验证协议：**采用强大的身份验证协议，如多因素身份验证和零信任安全模式，确保只有经过身份验证和授权的用户才能访问敏感信息和人工智能功能，从而降低未经授权的访问风险。

○ **定期审计：**定期执行安全审计和漏洞评估，识别和降低潜在的安全风险。

● 隐私保护技术：

○ 将融入设计：将隐私原则（如数据最小化、同意、安全性）直接纳入生成式人工智能系统的开发和实施中。

○ 隐私增强技术：探索保护敏感用户数据的技术：

■ 差分隐私：可以向数据集中添加计算噪声来匿名化信息，同时保持统计属性，使得在保护个人隐私的同时能够分析。

■ 联邦学习：在多个设备或服务器的分散数据上训练生成式人工智能模型，避免将敏感数据聚集在一个中心位置。

■ 同态加密：在不解密的情况下对加密数据执行计算。这样就可以在不泄露基础数据的情况下安全地分析敏感信息。

● 防御对抗性攻击：

○ **对抗性鲁棒性工具集(ART)：**使用对抗性示例训练生成式人工智能模型，以提高其对故意操纵的抵御能力。虽然ART在某些情况下被证明是有效的，但一些研究人员基于计算成本和易受训练分布之外的对抗性攻击等考虑因素，对ART的实际局限性提出了担忧。

○ **安全测试：**定期执行对抗性攻击模拟，识别漏洞并改进模型的防御能力。

4.3 鲁棒性、可控性和合乎伦理的人工智能实践

● 安全性和可靠性：

○ **风险评估：**进行彻底的风险评估，评估生成式人工智能系统的潜在危害和意外后果，并实施缓解措施和保障措施。

○ **测试与验证：**在各种场景和边缘情况下严格测试生成式人工智能模型，以确保其在不同情况下的可靠性和鲁棒性。

○ **最小化伤害：**设计具有保障措施的生成式人工智能系统，最大限度地减少潜在伤害。这可能涉及到"安全开关"或根据应用和所涉及的风险设计限制。

● **人类监督：**

○ **人类干预：**在关键决策过程中保持必要的人类参与，特别是对于高风险应用。允许人类干预，以便在必要时覆盖或调整生成式人工智能的输出。

○ **故障安全机制：**建立明确的升级路径和故障安全机制，以应对意外或有害的模型行为，特别是当这些行为被外部用户报告时。

● 问责制：

○ **所有权和责任：**为人工智能系统的开发、部署和监控指定明确的角色和责任，确保个人对技术的影响负责，从而高效地解决问题和改进工作。

○ **审计跟踪：**保存完整的模型开发、训练和使用日志。这些审计跟踪对于调查意外行为或伦理问题非常有价值。

○ **报告机制：**创建开放渠道供内部和外部的利益相关者报告关于生成式人工智能系统的担忧或潜在问题。这将促进积极反馈并允许迅速采取纠正措施。

○ **事件响应：**建立明确的事件响应计划和报告机制，以防出现意外结果或与人工智能相关的伤害。

○ **伦理审查委员会：**建立伦理委员会或审查委员会评估生成式人工智能应用的潜在影响，并确保其符合公司价值观。

○ **偏见和公平性审计：**定期执行审计，以识别并减少生成式人工智能系统的数据集、算法和结果中可能存在的偏差。

4.4 组织如何利用这些标准

有效采用这些技术标准并不仅仅是简单的理解。组织必须将伦理标准转化为实际行动，将最佳实践嵌入开发流程，从而切实确保人工智能得到负责任和合乎伦理的应用。例如:

- **制定明确的内部政策：**将这些标准纳入内部开发指南和组织政策。在从开发到生产的所有阶段，对负责任的人工智能提出明确的期望。
- **文档记录和报告：**定期发布有关数据使用、模型性能、偏差评估以及所采取的任何纠正措施的报告。这有助于提高对外部利益相关者的透明度。
- **伙伴关系与合作：**你并不是孤军奋战！与行业团体和合乎伦理的人工智能研究社区合作，为制定最佳实践做出贡献，并积极引导有关负责任生成式人工智能的讨论。

需要注意的是，技术标准只是起点，并不是普遍使用的--其实施应根据具体组织的需求和用例量身定制。采用合乎伦理的人工智能实践是一个持续的过程（而不是一次性的解决方案），组织需要随着技术、法规和社会期望的发展不断调整和更新其流程。

4.5 负责任的生成式人工智能的技术保障（数据管理）

表2概述了一些构建符合最常见数据管理规定的人工智能系统的关键技术和最佳实践。

表 2：构建 “负责任的 ” 人工智能系统的一些关键技术和最佳实践

数据处理	技术	说明
数据预处理	数据匿名化或伪名化	涉及从训练数据中移除或替换个人可识别信息（PII）以最小化隐私风险。 如果训练中使用了个人可识别信息 ，则需要仔细清洗输出。
	数据过滤	选择和筛选与生成模型特定目的相关的训练数据，避免不必要的数据收集或数据扩充。
	数据选择	谨慎选择训练数据，以符合预期目的并避免偏差。这包括过滤掉不相关或有害的信息。

数据整理	数据扩充	添加噪音或生成合成数据等技术可以增加训练数据的多样性，从而建立更稳健、更少偏差的模型。
模型设计、训练和优化	联邦学习	在分散的数据集上训练模型，将数据保存在单个设备上，而不传输到集中的服务器上。
	差分隐私	在训练数据中引入随机噪音 这种噪声有助于保护个人隐私，因为确切的数据被掩盖了。不过，如果数据集足够大，在不识别任何单个数据点的情况下，仍然可以观察到实际趋势和模式，因为噪音会在大量人群中产生平均效应，从而加强隐私保护。
	模型的可解释性	开发模型，使人们了解模型是如何产生结果的，从而更容易识别和减少潜在的偏差或错误
	定期监测和再训练	定期监测模型的性能，并利用更新或整理的数据对其进行再训练，以解决潜在的问题，如偏离偏差或生成不准确的输出。
	超参数调优	微调模型的超参数（控制其学习过程）可以影响输出结果，并有可能减轻意外后果。
持续监测和评估	定期审计和评估模型使用的数据	确保数据符合预期目的，不保留不必要的数据
	监控模型的输出结果，防止出现潜在偏差或意外后果	实施保障措施，解决任何发现的问题。
人在回路技术	人类监督	将人类监督纳入流程，在部署或使用前由人工审核和验证人工智能的输出结果。
	交互式生成	设计交互式系统，用户可以指导人工智能的生成过程，以实现预期结果。

可解释性和透明度	可解释的人工智能技术	采用 LIME、SHAP、Mimic 或 Permutation Feature Importance 等技术了解模型的推理，并识别潜在的偏差或局限性。
	开发和部署的透明度	对与人工智能/ 机器学习及其应用相关的局限性、偏差和潜在风险保持透明。

4.6 案例研究--在实践中展示透明度和问责制

本案例研究展示了企业将人工智能伦理原则转化为具体开发实践的实用方法。案例中，这家公司实施了生成式人工智能模型来生成图像。本案例展示了如何将透明、负责任的人工智能的具体策略和技术标准直接纳入其开发和业务流程。其中包括几个主要步骤：

- **公布模型卡片，概述模型的训练数据集、局限性和预期用途：**收集的训练数据来源清楚，并就具体用途和意图征得适当同意。数据集具有代表性和多样性--结合了获取的客户数据、公开数据、合成数据以及数据增强--所有这些都是为了最大限度地减少生成的输出结果出现意外偏差的可能。公司定期审核训练数据集以及时发现潜在的偏差或代表性不足。该公司公布了用于训练生成式人工智能模型的数据集的相关信息，包括数据集的组成、来源以及内部使用的预处理步骤。

实用方法：与Hugging Face上托管的模型卡片类似，该公司提供了基于TensorFlow Modern Garden的详细模型卡片，概述了生成式人工智能模型的训练数据。该模型卡片包括有关数据来源（如客户数据、公开可用的数据集）、组成（如文本、图像）和预处理步骤的信息。这种透明度使用户能够了解模型的潜在偏差和局限性。

- **采用可解释人工智能技术，**在生成图像的同时提供人类可理解的解释，阐明影响输出的因素。这些解释突出了促成图像输出的关键因素，使用户能够：

- **理解生成图像背后的原理：**通过使生成式人工智能模型的决策过程透明化，让用户深入了解模型生成特定图像的原因。这将促进信任，并在理解模型推理的基础上做出明智的决策。
- **识别潜在的偏见：**可解释人工智能技术解释可暴露训练数据或模型本身的潜在偏差。这使用 户能够客观评估输出结果，并确定是否存在任何歧视或不公平因素。
- **调试和改进模型：**通过分析解释并了解特定因素如何影响输出，开发人员可以找出模型的潜在缺陷，并努力提高其准确性和公平性。
- **建立人类审核流程和执行偏差验证是每个测试周期的一部分：**该流程专为敏感/高风险用例设计，有几个关键方面需要考虑：
- **标记标准：**制定清晰明确的标准，结合人类审查。这包括模型行为中被视为意外（可能有害）变化的特定输出，或模型可信度低于特定阈值的情况。
- **审核团队的组成：**组建一个由业务利益相关者与技术所有者和数据科学家密切合作组成的多元化、高素质的人类审核团队。该团队拥有必要的专业知识，能够理解模型的目的、潜在偏差及其输出结果的伦理影响。
- **审查程序：**为审查标记输出结果制定明确的标准化程序。这包括评估潜在的偏差，确保与伦理准则保持一致，并确定适当的行动，如模型重新校准或数据清理。
- **将偏差验证纳入测试周期：**偏差验证不应是一次性事件，而是贯穿 生成式人工智能模型开发和部署生命周期的持续过程。在公司层面采用了以下策略：
  - **采用不同的数据集进行测试，**以帮助诊断训练数据中存在的潜在风险。
  - **利用公平性指标：**在整个开发过程中实施并监控公平性指标，这有助于识别和量化模型输出中的潜在偏差。
- **定期对模型输出执行偏见审核，以发现并减少潜在的歧视行为：**这些审核涉及人类专家、数据科学家和业务利益相关者的合作，分析模型的输出是否存在意外偏差，例如在图像生成时偏向特定的人口统计或延续有害的刻板印象。一旦发现问题，就会实施适当的缓解策略（例如，使用增强数据集重新训练模型、调整模型算法等）。

组织应该通过借鉴这些行业标准和最佳实践，采取积极措施确保以合乎伦理和负责任的方式使用其生成式人工智能模型，并与消费者建立信任。

5. 持续监测与合规

随着生成式人工智能日益融入我们的生活和商业实践，确保安全和合乎伦理地使用生成式人工智能至关重要。持续监测和合规成为生成式人工智能有效治理的关键环节，使我们能够持续评估潜在风险并坚持负责任地使用生成式人工智能。

合规不仅仅是跟上不断发展的法律步伐。要确保负责任地使用生成式人工智能，需要对其生命周期的每个阶段进行仔细评估，同时积极规划持续合规性。这可能是一项复杂的工作，通常需要双管齐下：

1.建立健全的监测流程：

这包括持续监测生成的内容和整个开发过程。这包括检测数据、模型和输出中的偏差，同时验证数据隐私法规和道德处理的遵守情况。这种积极主动的方法可以促进公平性和包容性，同时防止生成的内容被滥用，如 深度伪造和有害内容。

持续监测有助于解决两个关键问题：它可以主动识别和消除训练数据和算法中的偏差，促进公平性和包容性；其次，它有助于识别生成内容的潜在滥用，使公司能够遵守伦理准则，防止错误信息的传播。

2.制定全面的合规计划：

该计划应概述识别和降低与生成式人工智能活动相关的潜在合规风险的程序。主要考虑因素包括：

- **数据安全和隐私：**根据适用法规，实施强有力的保障措施，在数据收集、存储和处理过程中保护敏感信息至关重要。
- **偏见与公平：**定期评估并减少训练数据和模型输出中的潜在偏差，以确保公平性和非歧视性。
- **透明度和可解释性：**确保用户了解生成式人工智能工具如何工作及其输出结果背后的原理，对于建立信任和问责至关重要。

通过积极监测合规性并实施适当的保障措施，组织可确保负责任和合乎伦理地使用生成式人工智能，从而增强用户和利益相关者的信任。

6. 管理生成式人工智能的法律与伦理考量

要有效管理生成式人工智能，就必须驾驭法律和伦理因素之间复杂的相互作用。合法性侧重于遵守既定的法律和法规，而这些法律和法规往往落后于生成式人工智能的快速技术进步。这就造成了灰色地带，为伦理框架提供了指导其发展和部署的空间。

在法律上，重点是遵守知识产权、数据隐私和非歧视等现行法律。这就需要建立相关框架，确保负责任的开发、透明的数据使用，以及对生成式人工智能产出可能造成的伤害承担明确责任。如前所述，人工智能生成的内容可能会侵犯版权，而数据隐私法规则涉及如何使用用户信息训练和运行这些系统。此外，确保公平和减少人工智能产出中的偏见对于避免社会不平等现象的长期存在也至关重要。

伦理方面的考虑不仅仅是遵守法律。它们包括更广泛的社会价值观和原则，以确保负责任地使用生成式人工智能并使其受益。围绕偏见、透明度、问责制和技术的潜在滥用等关键问题都属于伦理范畴。要解决这些问题，需要开发人员、决策者和公众之间不断对话与合作，为生成式人工智能融入生活的各个方面制定伦理准则和最佳实践。

随着生成式人工智能的日益普及，出现了几个热门话题。人们担心人工智能自动化会导致工作岗位流失、深度伪造可能会操纵公共言论，以及使用生成式人工智能创建有偏见的内容，这些都是需要认真关注的领域。要解决这些问题，需要政策制定者、开发者、企业利益相关者和公众共同努力，制定一个全面的治理框架，平衡创新与社会福祉。

7. 结论：填补人工智能治理空白，实现负责任的未来

人工智能治理的现状揭示了一个复杂的格局，其中有几个关键的挑战，需要全世界的政策制定者和监管机构立即予以关注。一方面，虽然现有法规间接涉及人工

智能，但缺乏必要的针对性，无法有效应对这一不断发展的技术所带来的独特挑战。相反，生成式人工智能 技术的迅速扩散，以及它们与日常生活各个方面的融合，凸显了对全面立法的迫切需求。这一差距要求制定新的法规，为负责任地开发、部署和使用人工智能系统（包括生成式人工智能）确立明确的指导方针。

此外，人工智能治理领域缺乏国际合作，导致法律环境碎片化，这可能会阻碍创新，并引发对未来差异和跨管辖范围的潜在不一致性的担忧。这种协调不足可能造成漏洞，并对追究人工智能相关伤害行为者的责任带来挑战。

随着生成式人工智能的迅速普及及其日益融入我们的日常生活，应对这些挑战的紧迫性也随之增强。公司将人工智能视为一种竞争优势，即使在缺乏健全法规的情况下，也推动了人工智能的快速应用。生成式人工智能在各行各业的应用日益广泛，使其有可能成为创新和颠覆的有力工具。与生成式人工智能造成的损害有关的诉讼的出现清楚地提醒我们，解决监管漏洞和防范潜在负面影响迫在眉睫。

为了迈向负责任的未来，我们应该采取多层面的方法：

1.加快制定人工智能法律法规：立法者必须优先考虑制定全面、适应性强的人工智能法规，同时考虑到与生成式人工智能相关的具体需求和潜在风险。这需要政府、行业专家和民间组织合作，建立有效的伦理框架。

2.国际合作与协调：促进人工智能治理方面的国际合作对于解决各管辖区之间的分散和不一致问题至关重要。建立国际框架和标准，同时尊重国家和地区的特殊性，将促进负责任的创新，并确保有效的跨管辖范围问责制。

3.技术标准和负责任的发展：制定和实施强有力的技术标准和最佳做法，对于在所有部门进行负责任的人工智能开发和管理至关重要。这些全面的指导方针将使公司、开发者和决策者能够建立符合伦理考量的人工智能系统，优先考虑公平性和透明度，并最终为社会做出积极贡献。

尽管缺乏全面的法规，但如今企业在人工智能系统设计方面面临着越来越严格的审查。人们越来越需要明确的指导，以正确地将人工智能集成到产品和服务中，如“内置”或“设计”。本文以实用的方法强调了正确应用技术标准的重要性，并对这些标准如何在当前立法环境下支持负责任的人工智能发展提供了初步或有限的理解。

展望未来，实现生成式人工智能的有效治理需要迅速采取行动。政策制定者必须优先考虑制定和实施兼顾创新与维护社会利益的法规。国际合作对于建立统一标准和防止相互冲突的监管框架至关重要。适当的立法将减轻公司的负担，因为它们面临越来越严格的审查，以确保其人工智能产品严格减少偏见和歧视，并遵守安全部署的最佳实践。这强调了所有利益相关者日益认识到合乎伦理和负责任的人工智能发展的重要性，突出了监管支持在指导行业实践中的关键作用。

CSA GCR cloud security alliance
GREATER CHINA REGION

Cloud Security Alliance Greater China Region

扫码获取更多报告

云安全联盟大中华区
官网：https://c-csa.cn 电话：0755-86548359
地址：上海市浦东新区东育路255弄5号3FB30（前滩国际经济组织集聚区）/
广东省深圳市南山区粤海街道北京航空航天大厦1号楼905