Capgemini 凯捷

生成式人工智能
道德&合规
风险白皮书

如何理解和应对生成式人工智能
引发的数据合规风险

2023年11月

目录

引言

在当今数字时代，生成式人工智能（GenAI）已经成为科技和商业界的前沿领域，为我们带来了前所未有的创新和机会。生成式人工智能技术的快速发展不仅提高了生产力，还在医疗、教育、娱乐、金融和众多其他领域中掀起了一场革命。生成式人工智能的解决方案预计在2–5年能达到全球认可的成熟度，率先采用生成式人工智能技术的企业将从重塑的业务模式和流程中获益最多。

凯捷咨询认为生成式人工智能的力量将全面重塑未来商业架构的DNA，例如生成式人工智能将改变企业和客户的沟通交流模式、使用数据和保障隐私的方式以及向潜在客户营销的方式，可以将工作流程由自我服务(Self–serving)转变为自动生成(Self–generating)，并且利用互联的情境化数据增强组织能力等。

凯捷咨询始终关注生成式人工智能的商业应用落地，专注于提供定制化解决方案。凯捷研究院（CRI）发布凯捷生成式人工智能主题系列报告：《解锁生成式人工智能的价值》。为了解企业管理层对生成式人工智能的看法以及应用情况，我们对全球来自13个国家的1000家企业进行了调研。报告显示，在全球受访的企业中，96%的企业将生成式AI列为高层级规划方向。大多数受访高管（78%）认为生成式AI可以使产品和服务设计下更高效。

尽管生成式人工智能在不同行业和领域中都有应用，但企业仍面临一些障碍。预训练模型的底层数据缺乏明确性、可能存在偏见以及缺乏包容性等问题，会给企业带来法律和声誉风险，甚至自定义的内部模型也可能出现“幻觉”和数据泄露的问题。凯捷咨询坚信应当负责任地使用生成式人工智能，遵守相关规范约束。

本白皮书旨在提供有关生成式人工智能的全面概述，以帮助组织了解并遵守相关合规要求。我们将讨论生成式人工智能的定义、应用领域、法律法规、伦理原则和最佳实践，以帮助您在生成式人工智能领域的工作中确保合法性、公平性和透明性。无论您是技术专家、法务从业者还是决策者，这份白皮书都将为您提供宝贵的参考，助您在生成式人工智能的世界中保持合规并取得成功。

1 生成式人工智能概述

1.1 生成式人工智能的发展历程

在此小节，我们将通过时间线图引导我们回顾生成式人工智能技术的关键事件，帮助理解生成式人工智能技术的演化，为合规和伦理讨论提供基础。

(资料来源：公开资料整理)

1.2 生成式人工智能的研究趋势

1.2.1 大模型对齐和幻觉

在生成式人工智能的研究中，最关键的问题之一是如何使通用人工智能与人类的价值和意图保持一致，这被称为对齐问题。大语言模型的本质是数学模型，而不是知识模型，即神经网络根据用户提示和上下文计算每个词汇符号的概率分布，逐步生成句子，但其生成的文本有时与用户的意图不符甚至完全相反。

一个常见的现象是，在用户刻意或无意的某些特定提示词下，大语言模型会在回答中参杂毫无根据或胡编乱造的“假事实”。这类毫无根据的错误回答可能会引导用户产生错误认知，甚至在极端情况下表现出对特定群体的偏见或敌意。这些幻觉现象的来源通常是模型训练数据中未被验证或恶意生成的语料、训练过程中未被准确定义的目标函数、或特定具有误导性的提示词输入。

当对齐问题于2021年提出时，Kenton等人将其描述为“我们如何设计一个能满足人类期望来行动的代理人”。然而，这个问题中缺少对于代理人的具体描述和定义。因此，在Sam Bowman后续的定义中，对齐问题变为了“如果人工智能系统拥有某些重要的能力，人类如何利用人工智能来可靠可信地完成目标”。而缺乏对人类期望定义，以及对模型对齐这一目标的追求将人们引入了提示工程这一新兴研究领域。

1.2.2 提示工程和检索增强

在与大语言模型同时兴起的提示工程研究领域中，科研人员致力于设计和优化对大语言模型的提示词以理解大语言模型的能力边界，并提升大语言模型在推理任务和其它复杂场景任务中的表现。最具代表性的提示工程方法包括少样本提示 (Few-shot Prompting)、自我一致性 (Self-consistency)、思维链 (Chain of Thoughts)、最少到最多提示 (Least-to-most Prompting)、和检索增强生成 (Retrieval-Augmented Generation) 等。

在思维链方法中，提示模型在生成回答时还输出其思考的过程，这有助于模型在回答中包含有逻辑的思考步骤，从而生成更易于理解和准确的答案。自我一致性的方法更为直观，模型会根据简单提示生成多个基于思维链方法的答案，然后选择最一致的答案作为结果。

检索增强生成是当前采用最广泛的知识增强方法之一。它通过匹配结构化和非结构化数据中的知识片段，把最符合当前提示的知识片段注入到提示词中，辅助大语言模型生成有根据的回答。思维链和最少到最多提示等提示方法在某些语言模型指标上，甚至能超过经过特别精细人工标注数据训练的模型，通过低成本的提示词优化, 达到了出色的模型性能。检索增强生成方法更是避免了对模型进行昂贵的微调和重新训练以获得有关特定领域的知识，从而显著优化了模型的幻觉现象，证明了提示工程的必要性和可用性。

1.2.3 通用人工智能和代理

通用人工智能 (Artificial General Intelligence, AGI) 是人工智能领域科研的神圣目标，旨在让人工智能系统能够自主学习并完成复杂的任务。

以ToolLLM项目为例, 研究人员训练了一个能够跨越49个领域的16000多个现实世界RESTful API的代理模型，该代理模型基于Llama基座模型，被称为ToolLLaMA，能够熟练掌握泛化的复杂任务分解和未见API调用的能力。

基于大语言模型对自然语言的理解能力, 人们开始研究如何使用自主工作或半监督的代理 (Agent) 来完成复杂的任务。代理的核心组件在于为模型接入例如计算器、API和搜索引擎的函数工具，使其拥有与世界交互的能力，通过多轮思维链和结果传递，帮助用户完成复杂的代理任务。

1.2.4 快速起步使用生成式人工智能

根据凯捷研究院的调查，在生成式人工智能快速普及的当下，全球超过95%的企业领导层正在探索利用这个强大的工具提升生产力并创造更多商业价值的可能性。

现在最便捷的大模型应用是基于非开源的大语言模型服务。例如OpenAI、PaLM、文心一言等大语言模型的文字生成能力需要通过官方提供的API接口进行访问，让开发者快速将大语言模型能力嵌入自己的应用中，避免了训练和部署模型涉及的大量储存和算力成本，并能通过服务提供商假设的高性能计算设备，快速获得强大且持续更新的文字理解和生成能力。然而大语言模型服务在费用、访问频次、隐私考虑上的限制。当开发者将大语言模型服务嵌入至高访问量的应用中时，基于文字token数量收费的潜在高成本是无法忽视的一环。而在例如金融、保险或医疗行业中涉及敏感用户数据的应用场景中，将用户数据上传至第三方的API请求服务也面临着无数的数据合规考虑。

因此，大部分企业在涉及大语言模型应用的时候，会考虑将开源的大语言模型私有化部署到能被透明化管理和运维的服务器上。

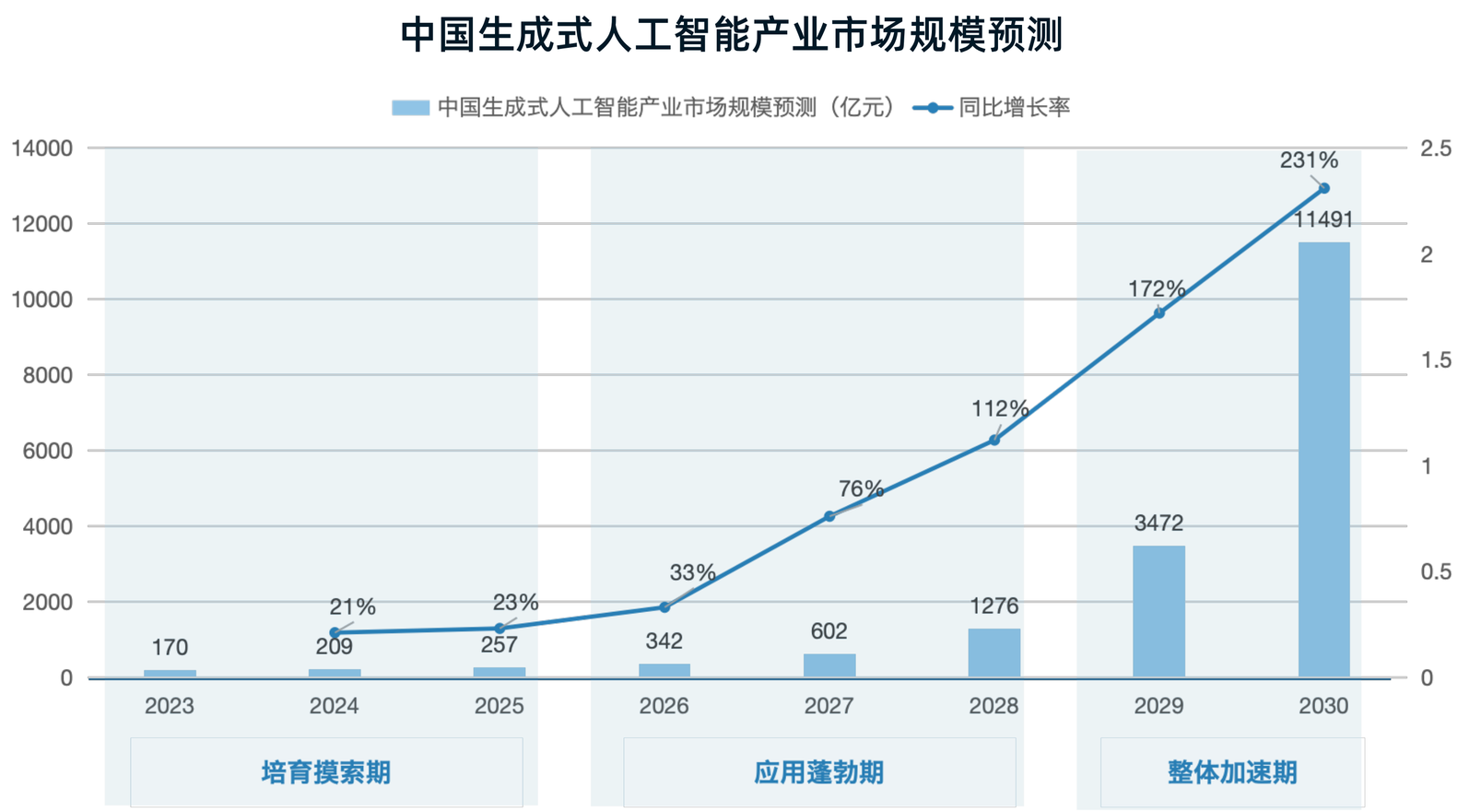
HuggingFace是目前最大的数据科学开源社区；包括微软、Meta AI等科技公司和Stability AI、BigScience、智谱AI等科研机构的开源模型参数都能在该社区上找到，而无数的开发者正在使用他们的私有数据对这些基础模型微调，并将掌握了不同垂直领域知识和能力的模型重新贡献到社区中。最知名的开源中文大语言模型之一，ChatGLM是由清华大学基于GLM (General Language Model) 训练的项目；其6B参数的版本经过约1TB的中英双语数据训练，能够完成文案写作、信息抽取、角色扮演、评论比较等中文语言任务，并且INT4量化版本的模型可以在大部分消费级显卡上运行甚至微调。

因此，对于有私有化模型需求的企业和商用场景，这类开源的大语言模型成为了首选。克服了高性能计算的成本，并在服务器上配置环境和部署模型后，企业可以完全掌握大模型运行中消耗、运算和产生的数据，确保敏感数据的隐私和安全。

1.3 生成式人工智能应用领域

1.3.1 市场规模总览

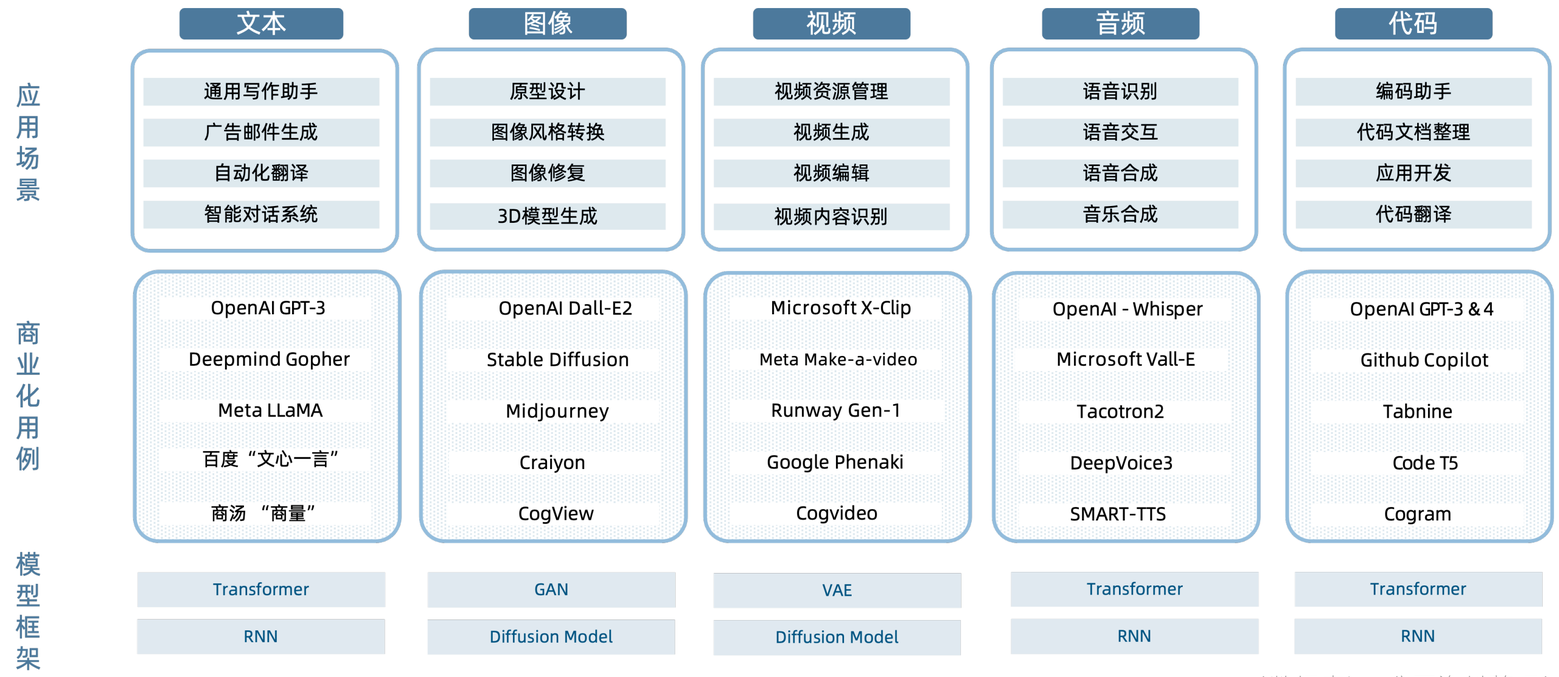
目前生成式人工智能产业正处于培育摸索期，大部分技术还未在实际生产过程中大规模使用，商业应用场景边界和商业模式还有待探索，用户体验仍需优化。随着大模型技术发展、垂类数据的积累、用户需求的识别细化和产业生态的完善，生成式人工智能的应用层走向垂直化和业务场景趋向多样化，生成式人工智能市场有望进入万亿级规模。



（来源：量子位 - 中国AIGC产业全景报告暨AIGC-P7）

1.3.2 多模态应用，赋能生产力：从数据类型划分

按照生成数据类型或者模态划分，生成式人工智能的应用涵盖了文字、图像和音频等领域。生成式人工智能技术可以用于参与数字内容创作，突破传统内容创作的数量约束，有着更为流畅和高效的人机交互模式，减少了重复性的任务负担，实现生产力解放。



（数据来源：公开资料整理）

1.3.3 聚焦个性化场景，创造业务价值：从行业划分

随着人工智能算法的迭代、算力的进步和数据的增加驱动生成式人工智能的技术变革，生成式人工智能模型的大范围连续对话能力、生成内容质量、语言理解能力和逻辑推理能力上都得到大幅提升。相比通用大模型，垂直大模型深耕特定行业和应用场景如医疗行业和金融行业，凭借其专业化和精准化的优势，更容易解决特定领域的问题，创造业务价值，实现商业变现。垂直大模型主要通过“预训练大模型+微调”的开发方式，只需针对具体任务对大模型进行二次开发，降低了企业应用的开发门槛。

行业	环节	应用
医疗卫生行业	分诊导诊	生成式人工智能有着迅速和强大的信息检索和匹配能力，全天候在线，可远程提供服务。AIGC能够更及时和详细地解答病人的就诊疑问，提供分诊导诊的指导，优化病人的就医体验。
	辅助诊断	生成式人工智能可以根据医学数据库、病历信息、过往就诊数据等信息进行学习、归类和分析，形成辅助诊疗建议，提供更为优化的临床治疗方案，以提高诊疗效率和效果。
	医学影像	生成式人工智能利用大模型的图像识别、特征提取等能力，更为精准地识别影像中微小的数据差异，辅助医生产生更为精准的诊断结果。同时利用大模型的分析预测能力， AI有望预测未来病人体征的变化，评估疾病发生的风险。
	诊后管理	生成式人工智能可24小时为病人提供病情指导、饮食指导和用药注意事项，实时服务患者，提高患者满意度的同时减少医疗人员负担。
	药品研发	生成式人工智能可以学习从蛋白质的序列到蛋白质的结构之间的映射关系，并基于其强大的算力解决复杂的高维数据映射处理问题，从而实现蛋白质结构预测。
银行、金融行业	营销	生成式人工智能依托大模型的上下文理解、逻辑推断能力，加之输入客户实时的信息和需求，可一键生成个性化营销内容，以更为精准的产品推荐文案和海报来吸引客户，触达客户的核心需求。
	风险控制	生成式人工智能有着强大的信息搜集和分析能力，可实时收集公司信息如股权变动、跟踪股价变化和舆情信息，利用深度学习网络强大的分析和预测能力，可帮助机构评估投资潜在的风险点，并对风控人员进行提示。
	虚拟营业厅	生成式人工智能结合计算机图形学，生成虚拟营业厅和数字人的3D模型，为用户提供更为直观和沉浸的对话体验，打造有温度的银行和金融服务。
	客服助手	生成式人工智能可接入企业的知识文档，利用大模型的逻辑理解能力，更为准确地理解用户的需求，同时不受时间和空间的约束，可全天候远程解答用户的问题如金融产品的信息、现有投资组合查询等等。

（数据来源：公开资料整理）

2 生成式人工智能相关法规浅析

随着我们对生成式人工智能技术进行了全面概述，现在我们将转向更深入的话题，探讨与生成式人工智能合规密切相关的法规问题。各国积极制定相关法规，目的是更好地管理生成式人工智能的使用，确保其对社会和个体产生积极、合法的影响，并且符合伦理要求。通过这一深入的法规解析，我们将更全面地了解如何在不断演变的生成式人工智能领域中维护合规性，构建可信的人工智能系统。

2.1 外国法

2.1.1 美国

美国关于生成式人工智能立法现状

2022年10月4日，美国白宫科技政策办公室发布了《自动化系统的开发、使用和部署蓝图》，又称《生成式人工智能权利法案蓝图》。不同于欧盟的生成式人工智能法案草案，该蓝图并不具有法律约束力，而是列出了五项原则，旨在最大限度地减少生成式人工智能系统的潜在危害。另外，美国国家标准与技术研究院(NIST) 于2022年8月18日发布了《生成式人工智能风险管理框架》的第二稿，目前处于征求意见阶段。该框架的初版可以追溯到2022年3月，并以2021年12月的概念文件为基础。生成式人工智能风险管理框架旨在帮助公司评估和管理与开发或部署生成式人工智能系统相关的风险。

另一方面，美国一些州已颁布立法，规范了在各类背景下使用生成式人工智能的情况，包括:

- 阿拉巴马州规定了使用面部识别技术(FRT)匹配结果来确定刑事调查或逮捕的潜在原因。
- 科罗拉多州限制了州和地方机构在没有意向通知、问责报告和对产生法律效力的决定进行有意义的人工审查的情况下使用面部识别服务(FRS)。
- 蒙大拿州限制了执法部门在特定情况下使用FRT，并禁止持续的面部监控。
- 爱达荷州已制定规定，对审前风险评估算法的使用进行了约束，要求提高透明度，并取消了审前风险评估工具的商业机密保护，以确保相关信息不受保密限制。
- 路易斯安那州和德克萨斯州已宣布，使用"深度伪造"技术来模仿未成年人是非法的。
- 康涅狄格州的法律要求州机构对所有使用生成式人工智能的系统进行年度审查和持续评估，以确保不存在非法歧视或差别影响的情况。

美国生成式人工智能道德伦理挑战和解决方案

2021年10月，美国平等就业机会委员会启动了一项倡议，以确保在招聘和其他就业决策中使用生成式人工智能和其他技术驱动工具符合联邦反歧视法。人工智能导致歧视性结果的能力，特别是以不明显或不易识别的方式导致的歧视性结果，以及相关的已知和未知后果，已导致全球各地采取措施，实施更严格的监督，以防止人工智能在就业中被滥用。如果算法识别出申请人的身体残疾、精神健康或其他不明显的临床诊断，则可能触发《美国残疾人法》。例如，雇主对显示震颤的数据的审查可被视为与残疾有关的调查，因为震颤可能显示出某些神经系统疾病，如脑瘫或中风。

尽管法律和监管领域仍处于起步阶段，许多科研组织和顶尖的科技企业已经开始了自我监管，以促进负责任的生成式人工智能开发和部署，并帮助防止生成式人工智能工具提供可能延续甚至加剧非法就业歧视的有偏见的结果。例如微软（Microsoft）这样的跨国公司开发和发布自己的生成式人工智能原则或指导方针已经成为一种常见的做法。

与此同时，一些专家学者把重点放在创新和前瞻性的非立法建议上。例如，一些人认为，企业应该借鉴金融领域企业十多年来成功实施的模型风险管理框架。该框架的支持者认为，公司和开发人员可以有效地管理与生成式人工智能相关的风险，通过使用从金融行业吸取的经验教训，并经过测试和时间的既定流程。

美国正在准备实施一个总体的立法和监管框架，激励将进一步推进生成式人工智能和相关技术能力的创新。比如雇主应当监控和审计人工智能的使用和流程，以主动识别故意滥用或潜在的歧视性结果。公司必须认识到虽然有生成式人工智能监管及合规审计等方法做事后评估，同时必须要加入公平和道德规范参与到雇佣决策过程。公司需要考虑和理解的其他考虑因素是供应商的责任以及对生成式人工智能立法和诉讼的持续态势的感知。

2.1.2英国

英国生成式人工智能立法现状

2023 年 3 月 29 日，英国政府发布了一份白皮书，概述了其支持创新的人工智能监管方法。根据目前的情况，现有的行业监管机构将被授权在各自的行业内监管人工智能，而不是制定新的法律或单独的人工智能监管机构。重点是加强现有制度以涵盖人工智能，并避免可能阻碍创新的高压立法。

白皮书中概述的拟议监管框架基于两个关键特征来定义人工智能，即适应性和自主性。白皮书认为，通过参照这些特征来定义人工智能，并设计监管框架来应对这些特征所带来的挑战，英国立法者可以使该框架在未来应对不可预期的新技术。

白皮书还列出了监管机构在应对人工智能相关风险时应遵守的五项 "注重价值观的跨部门 "原则。这些原则包括 (i)安全性、保障性和稳健性，(ii)适当的透明度和可解释性，(iii)公平性，(iv)问责制和治理，以及(v)可竞争性和补救。

白皮书发布后，英国政府将继续与企业和监管机构合作，着手建立已确定的核心职能。英国政府将在回应白皮书咨询的同时发布人工智能监管路线图。从长远来看，在白皮书发布12个月或更长时间后，英国政府计划实施所有中央职能，支持监管机构应用跨部门原则，发布人工智能风险登记簿草案，开发监管沙盒，并发布监测和评估报告以评估框架的绩效。

英国生成式人工智能技术如何应对伦理挑战和解决方案

生成式人工智能技术给个人隐私带来了两种威胁。第一种威胁类型涉及机构的意外披露：一个机构将缺乏足够隐私保护的数据集故意上传到云或境外，导致数据泄露和失控，而这些数据集往往包含有关个人的敏感信息和可识别信息。研究人员需要耗时耗力去分析这个逃逸的数据集，获取这些信息并重新识别到个体；第二种威胁类型涉及研究者偶然披露。研究人员发布基于受限的数据计算的产品(例如，训练有素的机器学习模型)。发布的产品缺乏足够的隐私保护，研究产品的外部消费者从研究人员使用的原始数据集中了解到个人或个人的敏感信息。

针对这些伦理挑战，英国于2017年通过的《数字经济法案》和其后的配套措施为研究人员获取政府数据和使用数据进行计算提供了合法途径。在保证不具体说明个人身份的情况下，可以对公共事务局所持有的与该当局职能有关的数据进行大规模算法研究。数据访问主要通过经过认证的机构的安全物理设施或与该设施的安全连接，并且政府监管部门对研究人员的活动和产出进行密切监测,任何产出在发布前都要进行检查。

从研究者的角度来看，获取数据集需要以下步骤：

- ☐ 研究员向机构提交项目提案。
- ☐ 项目经相关小组批准。
- ☐ 研究人员参与培训并可进行评估(例如，访问国家统计局持有的关联数据需要获得国家统计局安全研究服务和认证，并且可以亲自访问数据，也可以通过远程连接获得额外认证)。
- ☐ 所需数据由该机构确定，然后由相关数据中心摄取。
- ☐ 通过安全的数据服务提供去身份化数据。
- ☐ 研究人员进行分析;监测活动和产出。
- ☐ 对输出进行受试者隐私检查。
- ☐ 改进联邦数据管理方法，对其进行补充和修正。

从政府监管的角度来看，获取数据集需要做到保护公共利益：

- ☐ 与研究人员共享的任何数据都是匿名的，个人标识被删除，并进行检查以防止再次识别研究人员和拟议的研究都有严格的认证程序，以确保公众利益不受生成式人工智能导致的损害。

2.1.3 欧盟

欧盟生成式人工智能立法现状

欧盟一直走在全球生成式人工智能监管运动的最前沿，2023年6月14日，欧洲议会投票结果通过欧盟《人工智能法案》（EU AI-ACT，下文简称《法案》）草案。《法案》很可能成为世界上第一个全面管理生成式人工智能的法规，该《法案》对违规公司可以处以4000万欧元或年营收7%的罚款。随着《法案》进入采用的最后阶段，其拟议的语言为所有司法管辖区的公司在使用生成式生成式人工智能时将面临的重大合规障碍提供了宝贵的见解。欧盟的做法也将成为未来全球生成式人工智能监管的蓝图，为数据治理、透明度和安全性设定新的要求。

该法案是首次尝试为人工智能制定横向法规。拟议的法律框架重点关注人工智能系统以及生成式人工智能的具体使用和相关风险。欧盟委员会建议在欧盟法律中确立一个技术中立的人工智能系统定义并对其进行分类，根据 "基于风险的方法 "制定不同的要求和义务。该法案对人工智能系统进行风险分类，限制深度伪造，并对ChatGPT等生成式人工智能提出了更高透明度的要求。该《法案》定义了人工智能系统、供应链上涉及到的相关环节以及不同类别的生成式人工智能数据等相关要素，与GDPR中对受监管的个人数据的宽泛定义保持一致。另一方面，该《法案》明确了所有在欧盟市场投放、使用生成式人工智能系统及相关服务的国内外供应商、服务商和公共服务用户提供者，只要其生成式人工智能系统影响到欧盟及欧盟公民的，均将受到《法案》约束，从而保证了其规则的域外适用性。

《人工智能法案》对不同应用场景的生成式人工智能系统实施风险定级

级别	生成式人工智能系统应用场景	约束力度
不可接受	在平台算法推荐等具有算法偏见的场景中对人类意识和行为进行操作。	此类系统的应用和部署被绝对禁止。
	在具有算法偏见的场景中伤害儿童及残疾人等弱势群体等场景。	
高风险	关键基础设施领域：包括无人驾驶，水气热电供应等。	此类系统被严格管控，需在符合一定强制性要求的情况下才能进入欧洲市场。
	就业领域，教育领域，社会保障领域，司法领域，移民与边境管理等。	
	产品或系统的安全组件。	
有限风险	用户能够清晰意识到其与生成式人工智能系统交互，并可随时决定继续或终止交互行为的场景，包括聊天机器人，情绪识别系统，生物特征分类系统等。	此类系统需履行用户告知等透明义务，保障用户的知情，选择权。
极小风险	提供视频，游戏，邮箱等简单功能的生成式人工智能赋能工具。	未实施干预措施。

欧盟生成式人工智能法案的“长臂管辖”甚至会触及那些只生产用于欧盟市场的产品的生成式人工智能系统。因为该《法案》侧重于通过施加影响深远的义务来规范基础模型，主要体现在以下几方面：

- **风险管理：** 风险管理作为贯穿生成式人工智能模型整个生命周期的持续迭代过程，以降低风险并提高性能。这个过程包括识别和分析与该生成式人工智能系统的预期目的有关的最有可能发生的风险。
- **数据治理：** 以验证数据源和减轻偏见；根据《法案》被称为“提供者”，不应该允许生成式人工智能系统处理和使用不适合生成式人工智能训练的数据集。
- **安全性和ESG设计：** 以实现性能和网络安全，并减少能源使用。
- **技术文档（包括使用说明）：** 使下游生成式人工智能提供商能够满足某些高风险用例的透明度义务，包括生成式人工智能系统的一般描述、预期目的和预期输出等。技术文档的保存期为基础模型在欧盟市场上发布或使用后的10年。
- **质量管理：** 确保强大的上市后监控系统和持续遵守生成式人工智能法案。
- **在欧盟数据库中注册，以及其他义务。**

生成式人工智能的提供商必须采取进一步措施遵守《法案》，包括：

- **告知：** 提供商必须告知自然人，他们正在与生成式人工智能系统交互，并且内容不是由人类创建的。
- **保护：** 提供商还必须确保防止生成违反欧盟法律的内容。
- **发布：** 提供商还将提供其使用培训数据的摘要。

欧盟生成式人工智能伦理挑战及解决方案

早在2018年12月，《法案》未起草前，欧盟委员会的人工智能高级专家组(High-Level Expert Group on Artificial Intelligence, AI HLEG)就针对生成式人工智能的伦理问题和可能的解决方案发布了《可信人工智能伦理指南草案》。可信人工智能是将一般性和抽象性的伦理准则融入到生成式人工智能系统和具体应用中。

AI HLEG共提出10项要求，这10项要求均同等重要。针对不同的应用领域和行业，应根据特定环境进行评估，包括：可追责性、数据治理、普惠性设计、人工智能自主性的管控、非歧视、尊重和强化人类自治、隐私保护、健壮性、安全性、透明性。

欧盟《可信人工智能伦理指南草案》主要分为三个章节：

- 第一章通过阐述应遵循的基本权利、原则和价值观，确定生成式人工智能的伦理目标。

- 第二章为实现可信生成式人工智能提供指导，列举可信的要求，并概述可用于其实施的技术和非技术方法，同时兼顾伦理准则和技术健壮性。

- 第三章提供了评测清单以帮助组织识别和发现生成式人工智能系统的几个主要潜在问题：

  - 数据主体权利保护问题，为了维护欧洲公民的自主权，需要在生成式人工智能中合理使用监控技术。但实现可信人工智能应当区别个体识别与个体跟踪之间的差异，以及有针对性的监视和普遍监视之间的差异。

  - 隐蔽生成式人工智能系统问题，人与机器之间的边界模糊会带来如依附、影响或降低生而为人所应具有人生价值之类的恶果，因此发展人形机器人更应经过仔细的伦理评估。

  - 致命性自主武器系统（LAWS）问题，LAWS可在没有人为控制的情况下运行，但最终人类必须对所有伤亡负责。目前，众多国家和行业正在研究和开发致命自主武器系统，包括能够自主选择攻击的导弹、具有认知能力的自主杀人机器等，这都带来了基本的伦理问题。

欧盟的生成式人工智能管理框架无论在深度还是广度上都有着比较成熟的思考，围绕生成式人工智能全生命周期的流程、角色、活动等不同维度的风险识别和责任定义，使组织能够在生成式人工智能相关活动中明确企业、个人以及相关方的责任和义务。另外一方面，也是由于《法规》对相关方责任义务的充分识别以及对监管范围的放宽，这也将一定程度会制约了法规制约范围内的企业和组织在生成式人工智能领域的探索深度和商业化进程。

2.1.4 其他国家生成式人工智能法律发展简介

德国生成式人工智能法律及伦理发展

生成式人工智能是一项关键技术，在德国、欧洲乃至全世界都蕴藏着促进经济增长和提高生产力的巨大潜力。为了促进和利用这一潜力，联邦政府制定了一个行动框架，并在《人工智能战略》（AI Strategy）中采取了意义深远的措施以建立和扩大人工智能生态系统，加强人工智能的广泛应用，同时提高杰出倡议和结构的知名度。更新版还将大流行病控制、可持续发展（尤其是环境和气候保护）以及国际和欧洲网络建设作为新举措的核心。

2019年10月10日，委员会发布《针对数据和算法的建议》，旨在回答联邦围绕数据和生成式人工智能算法提出来的系列问题并给出政策建议。围绕“数据”和“算法系统”展开，包括“一般伦理与法律原则”、“数据”、“算法系统”、“欧洲路径”四部分内容。德国数据伦理委员会认为，人格尊严、自我决策、隐私、安全、民主、正义、团结、可持续发展等应被视为德国不可或缺的数字社会行为准则，这一理念也应在“数据”和“算法系统”的监管中贯彻。

法国生成式人工智能法律及伦理发展

法国对生成式人工智能的伦理治理问题高度关注，发布多项指导生成式人工智能安全应用的指南和条例，联合工业龙头企业发布《工业人工智能宣言》，积极推动人工智能健康发展。法国国家信息与自由委员会（CNIL）作为法国的数据监管机构，围绕算法和系统安全等方面出台多项条例和安全指南。在算法安全方面，发布了《人工智能与算法伦理风险》，深入分析了生成式人工智能算法可能引发的系列伦理问题，并提出治理举措建议。在系统安全方面，发布了《人工智能系统自评估》《人工智能系统安全指南》，致力于为公众、专业机构和相关领域专家提供有关生成式人工智能系统安全性的知识、理论工具和实施指导，围绕规划设计、数据资源安全性、保护和强化学习过程、使用可靠应用程序、考虑组织战略5个方面，提出强化生成式人工智能系统安全性的操作建议。2023年5月16日，CNIL发布了一份人工智能行动计划，内容分为四个方面：了解生成式人工智能系统的运作及其对个人的影响；支持和监管尊重隐私的生成式人工智能的发展；整合和支持法国和欧洲生态系统中的创新者；审计和监控生成式人工智能系统并保护个人。通过这项关键的协作工作，CNIL希望制定明确的规则，保护欧洲公民的个人数据，以促进尊重隐私的生成式人工智能系统的发展。

日本生成式人工智能解读

日本政府于2019年3月公布了由综合创新战略促进委员会通过的《以人为中心的生成式人工智能社会原则》，体现了生成式人工智能社会的基本原则，这七项社会准则分别为：(1)以人为本，(2)教育/扫盲，(3)数据保护，(4)确保安全，(5)公平竞争，(6)公平，问责制和透明度，以及(7)创新。这一系列法律制度涵盖了当前关于生成式人工智能相关机遇和风险的政治共识。就内容而言，日本在包容性增长、可持续发展和社会福祉方面的生成式人工智能方法符合经合组织的生成式人工智能原则。

加拿大生成式人工智能解读

2022年6月16日，加拿大联邦政府提交了C-27法律草案，也被称为2022年数字宪章实施法案。该立法方案的第三部分包括《生成式人工智能和数据法案》(AIDA)，这是加拿大的第一个生成式人工智能法案。AIDA旨在规范生成式人工智能系统的国际和省际贸易，要求某些人员采取措施，减少与高性能生成式人工智能系统相关的伤害风险和偏见结果。它规定了公开报告，并授权部长下令披露与生成式人工智能系统相关的记录。该法案还禁止处理可能对个人或其利益造成严重损害的数据和生成式人工智能系统的某些做法。目前，截至2023年3月，该法案正在下议院进行二读，仍需得到参议院的批准。

2.1.5总结

凯捷观点：

由于生成式人工智能技术涉及到隐私增强技术的使用尚处于起步阶段和不确定性，隐私应主要通过数据访问策略来解决。虽然在某些情况下欧美及日本立法者会建议甚至是强制要求合规设计，但是技术处理和访问策略仍是主要的防线：通过控制谁可以访问数据来确保敏感数据集受到保护。这种处理方法的表现形式之一就是采用分层访问策略，即将更敏感的数据集放在更受限制的层中。

例如，高度限制的获取数据可能对应于个人健康数据，而最低限度限制的获取数据可能对应于测量数据。这使得访问高度受限数据的提案将面临更高的审查标准，研究人员可能一次只能访问一个受限访问数据集。这种方法反映了目前的制度，即研究人员接受特殊训练来处理某些类型的数据。

2.2 中国本土法律

2.2.1 生成式人工智能的伦理道德问题讨论

2.2.1.1我国生成式人工智能伦理问题的基本原则

人工智能系统在社会上引发了广泛的伦理问题，如就业、社交、医疗卫生、医药保险、ESG、治安、商业运营、人权等等。这些问题的核心在于生成式人工智能算法，它们有可能复制和加深现有的偏见，导致各种歧视问题，带来全新的伦理挑战。

为了解决这些挑战，中国政府采取了一系列政策举措。2021年修订的《科学技术进步法》第103条设立了国家科技伦理委员会，旨在完善科技伦理规范，推进科技伦理教育和研究，并建立审查、评估和监管体系。2019年成立了国家科技伦理委员会，下设了人工智能、生命科学和医学三个分委员会，负责制定行业规范和进行伦理审查。

2022年，中国政府发布了《关于加强科技伦理治理的意见》和《生成式人工智能服务管理暂行办法》，这两份文件是关于生成式人工智能的首批全面法律文件。从《科技伦理审查办法（试行）》的征求意见到正式发布，中国的科技伦理监管体系经过了全面的顶层设计，各相关部门，包括国家网信办、工业和信息化部、公安部、新闻出版总署等，都在各自领域内强化了对生成式人工智能服务的管理。因此，企业需要密切关注中国国家机关在生成式人工智能领域的执法案例和指导意见。

《治理意见》提出了中国政府对科学技术伦理审查的5个维度，即：（1）增进人类福祉；（2）尊重生命权利；（3）坚持公平公正；（4）合理控制风险；及（5）保持公开透明。凯捷认为《治理意见》中的前两项明确要求技术创新和应用的最终目的是增进人类福祉，科技进步的同时务必尊重生命权利和公平利益，这意味着在中国发展生成式人工智能的商业体不能为了追求科技领先而牺牲人的安全保护、身体健康、精神健康，不可以通过损害人的隐私和安宁达到盈利目标。《治理意见》同时要求科技活动申办者和组织者全过程（全生命周期）秉承公平、公正、包容地对待社会群体，防止针对不同群体的歧视和偏见，防范技术加深偏见和排挤特定人群的风险，确保用户信息安全，并且鼓励公众参与监督，保持科技应用的透明度。

值得注意的是《审查办法》将特定种类的“算法模型、应用程序及系统的研发”以及特定场景的“自动化决策系统的研发”也归入了需要开展科技伦理审查复核的科技活动中。

2.2.1.2 我国生成式人工智能在商业领域的伦理审查要求

2.2.1.2.1 生成式人工智能在医药领域内伦理审查

医药行业审查重点

医药行业的伦理审查一直是该行业常规工作内容，生命科学和医学领域的从业者对医药健康领域的伦理审查要求更加熟悉。我国的法律对医药健康领域的伦理审查要求分散在不同法规中，建立了以相关研究事项的事前审查为核心的伦理审查机制。

在我国医药行业，对于生成式人工智能的应用，总体要求是不应以科技进步为代价而牺牲人的生命安全、身体健康，以及精神和心理健康。同时，科技活动的全过程需要以公平、公正、包容的方式对待各个社会群体，以避免歧视和偏见。此外，科技公司和医药企业需要接受全社会公众的监督，以确保透明度和合规性。

我国法律对于医药健康领域的伦理审查要求散见于各个法规中，包括但不限于《生成式人工智能服务管理暂行办法》、《互联网信息服务深度合成管理规定》、《中华人民共和国民法典》、《人类遗传资源管理条例》、《药品管理法》、《生物安全法》和《医师法》，这些原则和法规确保了医药行业的伦理审查在技术进步的同时保护了人的权益和健康，并倡导了公平、公正和透明的科技发展。

医药行业中生成式人工智能伦理缺陷的对策

我国在《互联网诊疗监管细则（试行）》中明确规定，“医疗机构开展互联网诊疗活动，处方应由接诊医师本人开具，严禁使用人工智能等自动生成处方，且医师接诊前需进行实名认证，确保由本人提供诊疗服务，人工智能软件不得替代医师本人提供诊疗服务。”

生成式人工智能在协助诊疗和医生决策中的界限是当前医疗行业广泛讨论的话题。生成式人工智能应用在医疗领域，尤其是医保方面，面临着特有的伦理问题，其中之一是算法偏见。如果用于训练AI应用的数据集未能充分覆盖女性、少数族裔、老年人、农村人群等多样化群体，可能导致最终算法的建议存在偏见。此外，如果用于AI药物研发的数据集在种族、环境和文化上过于同质化，可能导致AI识别的有效活性物质仅适用于有限的群体。此外，许多“AI+医药健康”应用需要积累患者数据。对于基于大数据计算并输出结论和诊疗建议的AI应用来说，数据积累至关重要，因为缺乏足够的数据将限制其工作和发展。然而，这个过程中如何保护患者数据和隐私，是企业和医疗机构需要特别关注的问题。

生成式人工智能在医疗领域的应用引发了伦理问题，涉及算法偏见、数据多样性和患者数据隐私，这些问题需要细致考虑和合理解决，可参考以下几个方面开展相关工作：

- 根据《人工智能医用软件产品分类界定指导原则》对于AI产品的分类和管控进行企业自查自纠，对照进行分类分级。
- 伦理需要全面涵盖，包括算法数据抓取以及用于模型训练的数据。
- 确保数据训练结果的公平性。

2.2.1.2.2 其他行业的伦理审查

随着《暂行办法》的发布，其将伦理审查的范围从直接以人为研究对象的科技活动扩展到所有存在伦理风险的科技活动，弥补了医学伦理审查范围以外关于科技活动伦理审查相关规定的空白。《暂行办法》第八条从监管角度对于服务提供者的数据标注义务提出了更为具体的要求：不仅明确要求服务提供者进行数据标注要制定清晰、具体、可操作性的标注规则，而且要求对数据标注进行质量评估，抽样核验标注内容的准确性，并对标注人员进行必要培训；第十九条更是规定有关主管部门有职责开展监督检查，要求服务提供者对于训练数据来源、规模、类型、标注规则、算法机制机理等予以说明，并提供必要支持和协助，《暂行办法》的出台在维护科技活动的伦理合规性方面起到了积极作用，以确保科技的发展与伦理原则相协调。

智能汽车行业

自动驾驶及自动升级会触发的伦理问题，在自动驾驶汽车中，生成式人工智能可以用于决策制定，如何选择在危险情况下采取何种行动。这引发了道德问题，例如，在公共交通遇到事故情况下，应该优先保护乘客还是驾驶员的生命，这涉及到道德伦理的权衡，同时，自动驾驶会引发大面积失业问题，需要考虑如何帮助受影响的工作人员转换职业或获得新的技能，以减轻社会不平等。《暂行办法》第七条规定生成式人工智能服务提供者应当依法开展预训练、优化训练等训练数据处理活动，使用具有合法来源的数据和基础模型。举例来说，智能汽车的OTA升级可能引发软件质量和安全性问题，可能会侵犯消费者的知情权，OTA如何平衡消费者期望的新功能和车辆的可持续性和环境和可持续性，降低消费者被剥夺知情权的风险和减少电子垃圾都是在生成式人工智能设计过程中要加入考量的因素。

互联网行业

虚假信息传播是该行业对生成式人工智能最大的疑虑，新技术往往被用来生成虚假信息，从而威胁社会的信息生态系统。在互联网上，虚假新闻、欺诈广告和虚假评论可能通过生成式人工智能传播，损害用户的信任和影响决策，《暂行办法》第四条在算法设计、训练数据选择、模型生成和优化、提供服务等过程中，采取有效措施防止产生民族、信仰、国别、地域、性别、年龄、职业、健康等歧视，要求企业必须采取技术措施来减轻虚假信息的伦理调整，并加强监管。

互联网行业

生成式人工智能在教育中用于个性化学习，但这可能涉及到潜在的隐私侵犯，因为系统需要访问学生的个人数据来定制教育内容。同时，这也引发了道德问题，对此《暂行办法》第十条规定提供者应当明确并公开其服务的适用人群、场合、用途，指导使用者科学理性认识和依法使用生成式人工智能技术，采取有效措施防范未成年人用户过度依赖或者沉迷，并建立明确的伦理准则基础上设计游戏安全控制内容，比如严格的防沉迷控制。

其他行业

国家对利用生成式人工智能服务从事新闻出版、影视制作、文艺创作等活动另有规定。

这些伦理道德风险点在不同行业中都需要认真对待，以确保生成式人工智能的应用不仅符合法规，还遵循伦理原则，尊重用户的权益和社会价值。此外，监管和自律机制也需要不断改进，以应对不断发展的伦理挑战。

2.2.1.3 我国生成式人工智能存在的伦理缺陷导致的法律责任和社会问题

生成式人工智能在数据采集和模型训练阶段，内容输入阶段和内容生成阶段都极其容易出现数据来源的合法合规问题，如果用来训练的基础数据包含他人拥有著作权的作品，则有可能构成侵犯著作权的法律问题，与此同时缺乏伦理规制的生成式人工智能应用还可能导致严重的社会问题。

2.2.1.3.1 生成式人工智能可能引发的法律责任

数据采集和模型训练阶段生成式人工智能可能引发以下法律纠纷：版权侵权：

如果生成式人工智能使用了受版权保护的数据或文本来进行训练，而未获得合适的授权或许可，这可能构成版权侵权；侵犯公民的隐私权，数据采集可能牵涉到个人数据，如果这些数据未经合法授权或适当的隐私保护机制，就可能触犯数据隐私法律。

在内容输入阶段生成式人工智能可能引发以下法律纠纷：版权侵权：

同上文，用户提供的输入包含受版权保护的材料，生成式人工智能系统生成的内容可能包含未经授权的版权材料；侵犯 商业机密，如果内容输入涉及公司的商业机密或机密信息，使用这些信息进行内容生成可能触犯商业机密法律。

在内容生成阶段生成式人工智能可能引发以下法律纠纷：涉及《刑法》的诽谤罪：

生成式人工智能生成的虚构小说内容可能包含诽谤、侮辱或虚假陈述，可能导致名誉损害诉讼；生成式人工智能生成的艺术品极其容易构成肖像权侵权，往往存在生成内容中包含个人肖像并且未经授权的情况；侵犯商标权，如果生成的内容包含未经授权使用的商标，可能构成商标侵权。

这些都是潜在的法律问题示例，具体涉及的法律问题取决于行业性质和企业如何应用生成式人工智能，以及生成式人工智能的具体用途是否遵循《暂行办法》。因此，在开发和使用生成式人工智能技术时，必须遵守相关法律法规，确保数据采集、模型训练、内容输入和内容生成都在中国的伦理框架内进行，以避免潜在的法律问题。

2.2.1.3.2 生成式人工智能可能引发的社会问题

人工智能生成内容的广泛应用确实带来了伦理挑战，主要涉及到公平性、社交隔离、虚假信息和诈骗等方面的问题。以下是对这四个方面的讨论：

公平性：

生成式人工智能系统在生成内容时可能受到数据集偏见的影响，这可能导致内容的不公平性。如果培训数据中存在性别、种族、年龄或其他偏见，生成式人工智能可能会在生成内容时反映这些偏见，加剧社会不平等。这引发了公平性问题，需要确保生成式人工智能系统不会强化或传播社会偏见，而是产生公平、无偏见的内容。

科技时代的“种族隔离”：

生成式人工智能系统可以根据用户的兴趣和行为，推送与其偏好相关的信息，导致社交隔离。尽管这可以提高用户体验，但也可能使用户陷入信息“自动过滤”“自动推荐”中，仅接触与其立场一致的信息，而忽视其他观点。这加剧了社会和政治分歧，威胁到公共辩论和民主价值观。

虚假信息：

生成式人工智能系统可以被用来生成虚假信息，这对社会和政治稳定构成威胁。恶意用户可以滥用生成式人工智能以制造虚假新闻、欺诈性内容或虚假评论，混淆信息的真实性。这强调了需要加强监管和技术措施，以应对虚假信息的传播。

诈骗：

生成式人工智能系统可以用于诈骗活动，如欺诈电话、电子邮件诈骗和社交媒体欺诈。通过模仿真人声音或虚假身份，生成式人工智能系统可以欺骗个人或组织，导致财务损失和隐私泄露。

解决这些问题需要综合的方法，包括技术改进、监管政策、教育和社会意识，以确保生成式人工智能系统在尊重伦理原则的同时发挥其潜在优势。

2.2.1.3.3 生成式人工智能违反科技伦理的法律后果

监管机关会通过相关法律法规的颁布和更新执法案例等方式对违反《审查办法》的情形进行实现全面监管。根据《中华人民共和国科学技术进步法》第一百一十二条，如果从事违背科技伦理的科学技术研究开发和应用活动的，科学技术人员可能会被责令改正、终止或撤销获得用于科学技术进步的财政性资金或者有违法所得、由有关主管部门向社会公布其违法行为、禁止一定期限内承担或者参与财政性资金支持的科学技术活动、申请相关科学技术活动行政许可等；并对直接负责的主管人员和其他直接责任人员依法给予行政处罚和处分甚至刑事处罚。根据《涉及人的生物医学研究伦理审查办法》第四十五条，医疗卫生机构未按照规定设立伦理委员会擅自开展涉及人的生物医学研究的可能会被监管部门要求责令限期整改、予以警告、处以罚款等；并对机构主要负责人和其他责任人员，依法给予处分。《审查办法》中规定科技活动的承担单位和科技人员，作为以下行为的责任人：弄虚作假获批，伪造、篡改批准文件；对纳入清单管理的科技活动未通过审查和专家复核的；未按照规定获批擅自开展科技活动的；或者超出获批范围开展科技活动。科技伦理委员会及其委员作为以下行为的责任人：弄虚作假为单位获得审查批准提供便利的；徇私舞弊、滥用职权或者玩忽职守等的。上述违法行为，由有管辖权的机构依据法律、行政法规和相关规定给予处罚或处理；造成财产损失或者其他损害的，依法承担民事责任；构成犯罪的，依法追究刑事责任。

2.2.2 我国生成式人工智能的法律基线和合规要点

2.2.2.1 法律基线

同其他国家一样，中国在人工智能和算法治理方面一直非常积极，甚至远早于近期的生成式人工智能监管浪潮。2021年9月，多个监管部门联合发布了一份政策声明，承诺三年内建立互联网信息服务算法应用的监管框架。同年，由网信办牵头的多部委联合发布了一项基于算法的在线推荐技术的规定（即《互联网信息服务算法推荐管理规定》），该规定涵盖了向个人用户进行推送、推广和内容排序在内的广泛的在线服务。

基于此，政府又于2022年9月发布了针对深度合成技术应用管理的规定（《互联网信息服务深度合成管理规定》），以监管自动生成音频、视觉和文本内容的深度合成技术。除了监管人工智能和算法的具体规定外，《个人信息保护法》《数据安全法》和《网络安全法》这三大法律也构成了生成式人工智能监管框架的重要基础。

最近的讨论集中在2023年7月10日《生成式生成式人工智能服务管理暂行办法》(以下简称《办法》)上，该办法由中国网信办等七部委联合发布，自2023年8月15日起施行。

《办法》第三条强调，发展与安全并重，促进创新与依法监管相结合，采取有效措施鼓励生成式生成式人工智能创新发展。《办法》强调“分类分级监管”思路，《办法》遵循了与中国数据合规治理相同的治理原则，其中第三条和第十六条都提出了基于“分类和分级监管”的生成式人工智能治理。然而，中国的生成式人工智能措施尚未为生成式人工智能生成服务的分类和分级提供具体标准。

《办法》的第五条和第六条明确表示中国支持生成式人工智能领域的创新和国际合作，特别是在算法、框架、芯片和支持软件平台等基础技术方面。《办法》还强调采用安全可信的芯片、工具、数据资源等。在对生成式人工智能技术的监管方面，《办法》强调基于生成式人工智能服务的“分类”和“分级”进行监管（第三条）。然而，中国目前尚未建立覆盖不同级别人工智能的治理体系，目前尚不清楚中国是否将采取类似于欧盟人工智能监管体系中风险导向的治理方法。

当生成式人工智能服务提供者具有“发布/分享舆情”或“动员社会”的能力时，必须进行“安全评估”和“算法备案”(第十七条)。《办法》强调个人信息保护，在个人信息保护方面，《办法》明确禁止收集“不必要的”个人信息，禁止“非法”向他人提供用户输入信息(第十一条)。对服务提供者违反《办法》的行为，将依据《个人信息保护法》《网络安全法》《数据安全法》《科学技术进步法》以及其他治安管理和刑事法律进行处罚。由此可见，《办法》规定的处罚并没有超出现行法律法规所规定的范围。

此外，《办法》还规范了生成式人工智能服务的使用，要求服务提供者必须通过以下方式管理服务的使用：采取措施防止未成年用户过度依赖或沉迷于服务；引导用户科学理性认识和依法使用生成式人工智能服务；以及如发现用户违反法律法规、商业道德或社会公德，暂停或终止向其提供服务等。

总体来看，《办法》的发布被媒体描述为中国监管生成式人工智能的里程碑式的一步，但这些发布的措施对国际公司究竟有什么法律影响和相关性，还有待相关部门发布更具体的指引进一步说明。

2.2.2.2 合规要点

2.2.2.2.1 用户隐私、内容歧视和模型研发

用户隐私、内容歧视和模型研发是当前深度生成应用的三个重要法规风险点。《办法》明确，国家支持人工智能算法、框架等基础技术的自主创新、推广应用、国际合作。《办法》集中瞄准技术应用问题，从明确条件要求、划定责任主体等几个方面为行业划定底线。生成内容本身应符合公序良俗和国家法律法规，技术提供方担负内容责任，使用方则应被充分告知其责任，应充分了解智能技术的界限和风险。《办法》对隐私信息这一备受关注的伦理风险点也作出了回应，要求提供方对此做好预防和反馈响应机制。数据资源和预训练模型是生成技术的基础，对此《办法》也要求在技术服务成形的前序阶段就进行法规管制，不得含有违法和有违公序良俗的内容。

2.2.2.2.2 生成式人工智能服务内容标识

根据2023年3月8日发布的《网络安全标准实践指南-生成式人工智能服务内容标识方法（征求意见稿）》对此要求进行了阐述。《征求意见稿》，各类场景均须在显示区域下方或使用者输入信息区域下方持续显示提示文字。至少包含“由人工智能生成”或“由AI生成”等信息。

《征求意见稿》还要求由人工智能生成图片、视频时应采用在画面中添加提示文字的方式进行标识。自然服务转人工智能服务时，也应通过提示文字或提示语音的方式进行标识，至少包含“人工智能为您提供服务”或“AI为您提供服务”等信息。

但如《征求意见稿》中的规定，图片、音视频中增加隐式水印标识是相对较为容易的，可以通过修改代码实现，但针对显示内容基本等同于代码的文本类就较难实现使用隐式水印标识。在《征求意见稿》中也可以看出，目前并未强制要求文本类进行要在生成内容中增加隐式水印标识。

2.2.2.2.3 开展生成式人工智能服务需获取的相关资质

根据相关法律法规以及已经明确提出要求的平台规则对生成式人工智能相关APP产品在申请应用商店上架时的文件进行梳理。生成式人工智能相关产品资质要求如下：

类型		文件名称
一般上架资质		ICP许可证（增值电信业务经营许可证B25信息服务业务）或者ICP备案
		APP备案文件
		《计算机软件著作权证书》《App电子版权证书》或《软件著作权认证证书》（三者选一）
生成式人工智能功能的产品上架资质		互联网信息服务算法备案系统上提交/通过的算法备案
		《安全评估报告》（加盖公章）
		《安全评估报告》在全国互联网安全服务管理平台的提交结果截图
强制许可类型	音视频、直播产品	《网络文化经营许可证》
	新闻资讯类	《互联网新闻信息许可证》
	药品信息类	《互联网药品信息服务资格证书》
	医疗器械类	《互联网药品信息服务资格证书》
	宗教信息类	《互联网宗教信息服务许可证》

同时中国涉及生成式人工智能的法律关于申请许可或备案的规定的适用范围也需生成式人工智能厂商注意做好事前规划，首先，向中国境内在前述各规定下，境内主体向境外提供生成式人工智能服务的，无需适用前述各项规定进行算法备案。但需要提醒注意的是，如果是中国境内的生成式人工智能在向境外提供服务过程中，即使不需要适用《暂行办法》，仍需要注意是否会产生数据出境等合规问题。而境外主体向中国境内公众提供服务的，同样需要适用各项规定完成算法备案。其次，在《暂行办法》中特别强调了“面向公众”的服务，因此如果仅在公司内部研发或者以提高办公效率之目的，公司内部使用则无需适用。但同时需要注意，如果向企业提供生成式人工智能服务，该企业再使用该服务向公众提供服务时，仍属于向公众提供服务，该企业应作为技术支持者基于《算法推荐规定》《深度合成规定》《暂行办法》完成算法备案。最后，提供服务需要在结合前述三项规定中的“算法推荐服务”“深度合成服务”“生成式人工智能服务”后明确自身属于技术支持者还是服务提供者。

2.2.3 总结

根据我国生成式人工智能治理“三驾马车”《治理意见》，《暂行办法》和《审查办法》的总体要求，伦理审查工作将成常态，在管辖范围内的企业或组织宜为其相应的科技活动准确理解、尽快建设伦理审查委员会制度并完成相应登记。科技伦理（审查）委员会将作为一个企业内部的常设机构，该机构对企业内部的科研活动提供持续支持，企业至少要有能力提供及时的响应，提供场地、人员及经费的支持，科技伦理审查的法律法规逐渐健全，这项工作也会成为企业科研过程中常规工作。公司需要及时设立了内部合规委员会或类似的组织，负责确保公司在商业活动中遵守伦理准则、法规和法律法规。这些组织通常负责监督公司的商业实践，确保其符合道德、法律和社会责任方面的要求。

另外，新出台的《办法》确立了包容审慎和分类分级的监管原则，力求在关于人工智能服务的创新发展与防控风险之间寻求平衡。企业开展生成式人工智能相关工作时也应在推进算法、框架等基础技术的自主创新、推广应用、国际合作的同时，尊重社会公德、公序良俗、知识产权、商业道德，禁止非法获取、披露、利用个人信息和隐私、商业秘密，保证数据的真实性、准确性、客观性、多样性，实现生成式人工智能健康发展。

最后需要说明的是，目前世界大部分地区针对人工智能的立法相对成熟，而针对生成式人工智能的立法大多还处于提议、提案阶段，因此现阶段在实践中对生产式人工智能的约束性法律规则可以参考人工智能部分的相关规定。

3 生成式人工智能的数据合规浅析

3.1 生成式人工智能的数据合规要点

生成式人工智能的快速发展给数据隐私保护带来了严峻挑战。为了保护个人权益和维护公众对生成式人工智能的信任，我们需要遵守数据隐私保护原则以及合规性要求。数据在生成式人工智能中扮演着关键的角色，在数据收集、获取、存储、清洗、预处理、标注、注释、训练、验证、评估和测试等环节都需要考虑合规性。为提高数据采集和预处理的合规性，我们可以采取伦理审查、记录细节、定期审查和监督等措施。模型训练和验证的合规性是构建可信赖、可解释和符合监管要求的人工智能系统的基础，需要遵守适用的法规法律、尊重伦理准则、保障用户隐私和权益。数据评估和调整的合规性需要平衡合规性要求和数据科学、人工智能技术发展的复杂性，密切关注法律法规动态并与专业人士进行沟通和讨论。在输出结果方面，合规性要求我们遵守人权、法律法规和道德标准，处理敏感数据时需特别小心和谨慎，避免歧视性结果，同时还需考虑知识产权、商业机密、恶意信息、虚假宣传、非法活动等合规性问题。在生成式人工智能的发展中，我们需综合考虑各种合规性标准，并采取适当的措施保护和遵守这些标准，以确保数据合规性的同时，推动生成式人工智能的可持续发展和社会认可。

3.1.1 数据隐私保护原则

透明性

提供清晰明确的信息，告知个人数据的采集、存储和使用方式，并说明信息安全措施。

合法性和合理性

数据处理必须在法律和合理的目的范围内进行，且不得违背个人的合法权益。

最小化原则

仅收集和处理实现所需目的的最小量个人数据，不得采集与目的无关的额外个人数据。

数据质量

确保收集和处理的个人数据准确、完整和最新，纠正不准确或过时的个人数据。

安全性

采取适当的物理、技术和管理措施，保护个人数据免遭未经授权的访问、泄露或损害。

存储限制

个人数据不应保留超过必要的时间，安全地销毁或匿名化不再需要的个人数据。

主体权利

个人享有访问、更正、删除和反对个人数据处理等权利，组织应提供适当的机制和及时响应请求。

合规性和问责制

确保数据处理活动符合隐私法律和规定，并建立合适的问责制，如指派数据保护官员、制定数据保护政策和流程、进行数据隐私风险评估等。

跨境数据传输

采取技术和法律措施，确保跨境数据传输过程中个人数据得到充分保护。

基于风险的方法

根据数据处理活动的风险级别，采取适当的隐私保护措施，对风险较高的活动实施更严格的保护措施。

随着生成式人工智能的快速发展，数据隐私保护成为一个非常重要的议题。以下是一些重要的数据隐私保护原则，旨在保护个人的权益并维护公众对生成式人工智能在数据处理方面的信任。

透明性

提供清晰明确的信息，告知个人数据的采集、存储和使用方式，并说明信息安全措施。

合法性和合理性

数据处理必须在法律和合理的目的范围内进行，且不得违背个人的合法权益。

最小化原则

仅收集和处理实现所需目的的最小量个人数据，不得采集与目的无关的额外个人数据。

数据质量

确保收集和处理的个人数据准确、完整和最新，纠正不准确或过时的个人数据。

安全性

采取适当的物理、技术和管理措施，保护个人数据免遭未经授权的访问、泄露或损害。

存储限制

个人数据不应保留超过必要的时间，安全地销毁或匿名化不再需要的个人数据。

主体权利

个人享有访问、更正、删除和反对个人数据处理等权利，组织应提供适当的机制和及时响应请求。

合规性和问责制

确保数据处理活动符合隐私法律和规定，并建立合适的问责制，如指派数据保护官员、制定数据保护政策和流程、进行数据隐私风险评估等。

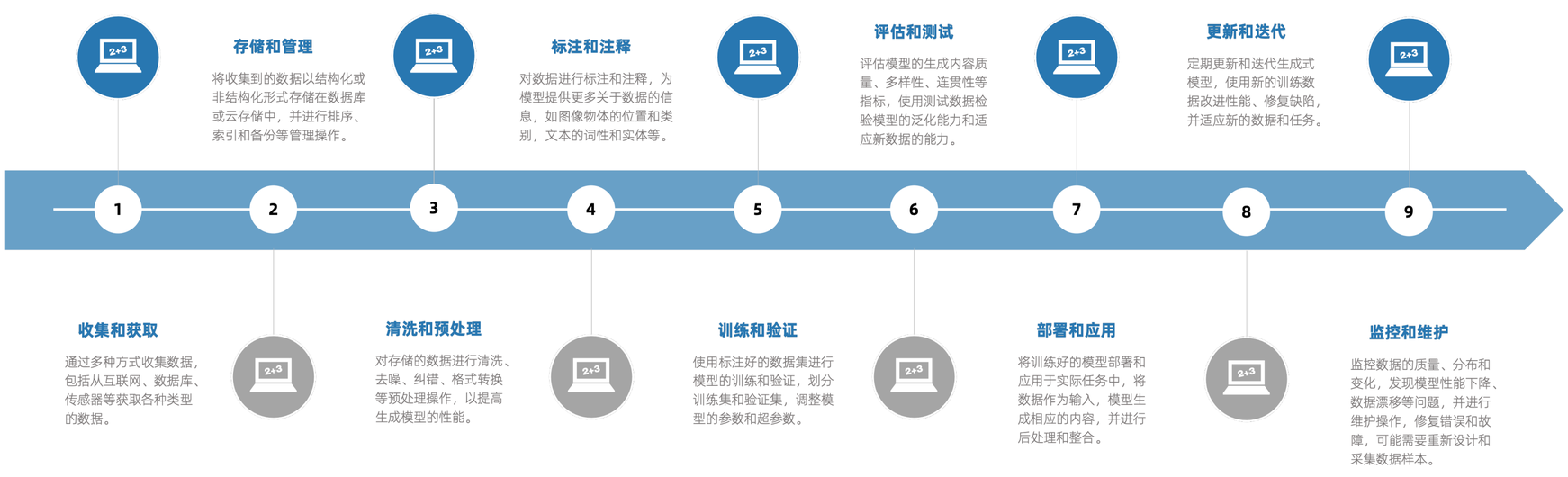
跨境数据传输

采取技术和法律措施，确保跨境数据传输过程中个人数据得到充分保护。

基于风险的方法

根据数据处理活动的风险级别，采取适当的隐私保护措施，对风险较高的活动实施更严格的保护措施。

3.1.2 数据在生成式人工智能中的角色



生成式人工智能数据生命周期流程图

生成式人工智能的应用日益广泛，它通过学习大量的训练数据，可以生成各种类型的内容，如文本、图像、音频等。在这一过程中，数据扮演着关键的角色，它经历着一个完整的生命周期，从数据的收集和获取，到存储和管理，再到清洗和预处理，标注和注释，训练和验证，评估和测试，最终到部署和应用。这一系列步骤的完成，为生成式人工智能的训练和应用奠定了基础。这一部分将从数据的生命周期出发，深入探讨数据在生成式人工智能中的使用情况以及其对系统性能的影响。

收集和获取

通过多种方式收集数据，包括从互联网、数据库、传感器等获取各种类型的数据。

存储和管理

将收集到的数据以结构化或非结构化形式存储在数据库或云存储中，并进行排序、索引和备份等管理操作。

清洗和预处理

对存储的数据进行清洗、去噪、纠错、格式转换等预处理操作，以提高生成模型的性能。

标注和注释

对数据进行标注和注释，为模型提供更多关于数据的信息，如图像物体的位置和类别，文本的词性和实体等。

训练和验证

使用标注好的数据集进行模型的训练和验证，划分训练集和验证集，调整模型的参数和超参数。

评估和测试

评估模型的生成内容质量、多样性、连贯性等指标，使用测试数据检验模型的泛化能力和适应新数据的能力。

部署和应用

将训练好的模型部署和应用于实际任务中，将数据作为输入，模型生成相应的内容，并进行后处理和整合。

更新和迭代

定期更新和迭代生成式模型，使用新的训练数据改进性能、修复缺陷，并适应新的数据和任务。

监控和维护

监控数据的质量、分布和变化，发现模型性能下降、数据漂移等问题，并进行维护操作，修复错误和故障，可能需要重新设计和采集数据样本。

3.1.3 数据采集与预处理的合规性

生成式人工智能的模型基础源自大量数据。而模型训练的有效性来自于充足且有效的数据样本。只有在这样的基础上训练的模型才是有效的。因此，在进行数据样本采集之前，需要对模型所需的数据样本进行评估和预设，并设定数据样本的需求标准。只有按照这些标准采集的数据才是有效数据。然而，我们也需要注意制定的数据采集标准和需求与数据合规之间是否存在冲突。

在生成式人工智能中，确保数据采集和预处理的合规性尤为重要，特别是考虑到数据隐私和法规的要求。下面是一些方法和实践，可以帮助提高数据采集和预处理的合规性：

数据合规性

确保训练数据的合规性。避免使用不合法、有争议或侵犯隐私的数据。这涉及到对数据来源的审查和选择，确保获得的数据是合法和符合道德标准的。

明确数据用途

在收集数据之前，明确数据的用途和目的。只收集与项目目标相关的数据，并避免采集超出这些目的范围的数据。这确保了数据采集的合理性和合规性。

获得明确的许可

在采集个人数据时，确保获得数据主体的明确许可。这可能需要实施明示的同意机制，使数据主体了解他们的数据将被用于何种目的，并明确确认他们的同意。

数据匿名化和脱敏

在采集数据后，采用适当的方法对数据进行匿名化或脱敏，以确保个人身份无法轻易识别。这可以包括删除或替换敏感信息，如姓名、地址和手机号码。匿名化和脱敏是保护个人隐私的关键措施。

数据清洗与过滤

在生成数据之前，对原始数据进行清洗和过滤。这包括检查数据是否存在错误、缺失或选取不合规的内容。确保生成的数据是高质量和合规的。

数据安全

确保数据在采集、存储和传输过程中得到妥善保护。使用加密、访问控制、安全传输协议等安全措施，以防止数据泄露和未经授权的访问。

数据质量控制

实施数据质量控制措施，以确保采集到的数据准确无误。这可以包括数据验证、纠错和自动化检查等技术手段，以及专门的质量控制流程和团队。

数据保留和删除策略

制定数据保留和删除策略，以确保不再需要的数据会被及时删除，从而降低数据泄露的风险。根据法规和合规要求，制定合理的数据保留期限，并确保数据在到期后被安全地删除。

监管合规性

建立内部合规性团队或机构，负责监督数据采集和处理活动，确保其符合法规要求。该团队可以制定相关政策和流程，提供合规培训，并跟踪合规变化。

透明度和报告

与数据主体保持透明，向他们提供与数据采集和处理活动相关的透明度，并提供适当的报告。这包括告知数据主体数据的采集目的、数据用途和可能的风险，并根据需要提供数据主体的访问请求和修正请求。

定期审查和更新

定期审查和更新数据采集和预处理的合规性策略，以确保其与法规的变化和最佳实践的变化保持一致。随着科技和法律发展的不断变化，合规性策略也需要不断演进和更新。

3.1.4 模型训练与验证的合规性措施

在人工智能的快速发展和广泛应用背景下，确保模型训练与验证的合规性成为一项至关重要的任务。合规性涵盖了法律符合性、伦理标准、社会期望和保护用户权益等诸多方面。为了构建可信赖的人工智能系统，我们需要采取一系列措施来确保模型在运行过程中遵守适用的法规法律、充分尊重伦理准则、符合社会价值观，并保障用户的隐私和权益。这不仅是为了满足监管要求，也是为了确保技术的公正性、透明性和可解释性。因此，以合规性为中心的模型训练与验证不仅仅是一项义务，更是构建可持续发展和社会认可的人工智能系统的基石。在本章中，我们将介绍一些关键的合规性措施，以便有效管理和应对模型带来的潜在风险，并确保其符合当地法律法规和道德准则。通过统筹整合不同领域的专业知识和利益相关方的意见，我们可以迈向一个更加健康、可靠和可持续的人工智能未来。

监管和审查

在模型训练与验证的合规性措施中，监管和审查的重要性不可忽视。尽管模型的运行往往需要一定的自主性和创造性，但它必须始终在法律和道德的框架内运作。监管和审查的作用是确保模型不会产生违法或违规的结果，并且对可能出现的风险进行评估和管理。

透明度和可解释性

透明度和可解释性也是至关重要的因素。深度学习模型往往是黑箱模型，它们的决策过程通常无法被直接理解或解释。然而，透明度和可解释性是保持用户信任的关键因素。通过提供透明的工作过程和解释模型决策的文档和报告，可以帮助用户理解模型是如何生成内容的，从而增加对模型决策的接受度和可信度。

过滤和内容控制

实施内容过滤和审核机制，以防止生成的内容包含不合规的信息。这对于保护用户免受虚假、误导或有害的内容的侵害非常重要。通过使用敏感内容检测工具和算法，可以及时发现和过滤掉不良内容，保护用户免受潜在的负面影响。虽然过滤和内容控制是必要的，但过度的过滤可能会引发言论自由和创造力的抑制。因此，在制定过滤和内容控制机制时，需要权衡合规性和用户体验之间的平衡。同时，需要不断关注技术的进步，以有效应对新兴的不当内容和欺诈手段。

多方参与

多方参与在确保合规性方面发挥着重要作用。与社会活动家、隐私专家和道德哲学家等利益相关方合作，有助于引入多样的观点和价值观，避免模型仅满足特定利益或单一立场的情况，并确保生成的内容符合社会的普遍期望和原则。

数据保护

建立强大的数据保护和安全措施，以保护用户数据和生成内容的安全性。尽管数据的收集和使用对于模型的训练和改进至关重要，但数据保护必须始终是一项优先考虑的工作。这意味着在数据收集、存储和处理过程中要采取适当的加密和防护措施，以防止数据泄露或未经授权的访问。定期审查和更新数据保护政策是确保模型在保护用户数据方面持续符合最新法规和最佳实践的关键步骤。

用户教育

向用户提供清晰的隐私政策和使用条款，解释模型如何使用其数据和生成内容。同时，教育用户有关合规性问题，以提高他们的意识。这样做可以增加用户对数据隐私的了解和权益保护的重视，并使他们更加积极地参与和监督模型的合规性。不仅要依赖技术手段来确保合规性，教育用户也是建立可持续合规框架的重要一环。

3.1.5 数据评估与调整的合规性

在数据科学和人工智能的发展过程中，确保数据的合规性已成为一项重要而复杂的挑战。随着社会对数据隐私、伦理和道德规范的关注日益增加，相关的法律法规也在不断演变和加强。在这个背景下，对于数据评估与调整的合规性，需要我们以辩证思维的方式来探讨和思考。

一方面，我们要意识到合规性的重要性。合规性不仅意味着我们应当遵守法律法规，还意味着我们要对数据的质量、准确性以及所涉及到的道德和伦理问题负责。在处理数据的过程中，可能会涉及到不当的内容、偏见、歧视或其他违反规范的行为，这些都是需要严格审查和调整的。对数据的合规性进行评估和调整，有助于保护用户隐私和权益，减少不良信息的传播，并确保科学研究和智能应用的可持续发展。另一方面，也需要考虑合规性的挑战和复杂性。合规性要求不仅仅是简单地遵守规定，而是需要深入了解和解读法律法规，以及根据具体的应用场景进行权衡和调整。数据科学和人工智能技术的快速发展，使得合规性要求不断变化和更加严格。因此，需要在评估和调整数据的合规性时，密切关注最新的法律法规动态，并与专业人士和专家进行沟通和讨论，以确保我们的工作符合合规性要求，并能够适应日益变化的环境。

在实际操作中，我们可以借鉴以下措施来提高数据评估与调整的合规性：进行伦理审查，审查生成模型，记录数据处理过程的细节，定期审查和监督数据使用，并在模型训练和部署的不同阶段进行合规性测试和审核。这些措施将帮助我们建立一个合规性意识和机制，确保我们的数据评估与调整工作符合最新的法律法规要求，以及道德和伦理规范，从而为数据科学和人工智能的可持续发展提供坚实的基础。

伦理审查

进行伦理审查，评估生成的数据是否涉及不当的内容、偏见、歧视或其他潜在的不当行为。随着人工智能和自然语言处理技术的不断发展，生成的数据可能会受到更严格的监管和法律限制。因此，在进行伦理审查时，需要考虑到最新的法律法规要求，以确保数据不会违反任何相关的规定。伦理审查应该是一个全面、系统和多方参与的过程，涉及到专业领域的专家以及法律和伦理方面的专家，以确保数据的合规性。

模型审查

对生成模型进行审查，确保其不会生成不合规的内容。在最新的法律法规中，对于机器学习模型以及其生成的数据，同样存在着合规性的要求。由于机器学习模型可以根据已有的数据生成新的内容，而这些内容可能具有潜在的问题，因此需要审查模型并确保其不会生成任何非法或违反法律法规的内容。为了满足合规性的要求，可以使用预训练的语言模型进行过滤，以减少不当内容的生成，或者采用其他技术手段来确保模型生成的数据的合规性。

透明度与文档记录

记录数据评估过程的细节，包括数据处理、过滤和伦理审查等步骤。透明度成为了关注的焦点之一。为了确保数据的合规性，应该记录数据评估过程的细节，包括数据的处理、过滤以及伦理审查的步骤。这样的文档记录有助于保持透明度，并为相关的监管机构提供必要的证据。此外，透明度还可以帮助我们更好地追踪数据的来源和处理过程，以确保数据的合法性和可信度。

审查与监督

定期审查和监督生成数据的使用，确保合规性得到持续维护。对于数据的合规性进行定期的审查和监督变得更加重要。应该建立起一个长期有效的机制，对生成数据的使用进行审查和监督，以确保合规性得到持续的维护。这可以包括定期的内部审查、外部审核机制以及持续的监测和监控措施。必要时，还可以进行审计以确认合规性，并及时采取措施进行改进和调整。

合规性测试

在模型训练和部署的不同阶段进行合规性测试和审核。在模型训练和部署的不同阶段进行应当合规性测试和审核。因此应该在模型的训练阶段进行数据合规性的测试和审核，以确保训练数据的合规性。同时，在模型部署和使用的过程中，也需要进行合规性的测试和审核，以确保使用的数据和模型都符合法律法规的要求。通过合规性的测试和审核，可以及时发现和解决潜在的合规问题，保证数据的合法性和合规性。

3.1.6 输出结果的合规性

当涉及到生成式人工智能模型的输出结果时，强调其合规性要求已成为一项至关重要的任务。合规性要求我们必须遵循人权、法律法规和道德标准，以确保输出结果的可接受性和公正性。特别需要注意的是处理敏感数据和避免歧视等合规性问题。

处理敏感数据方面，我们需要特别小心和谨慎。敏感数据包括个人隐私信息、医疗记录、财务数据等，泄露或不当使用可能对个人造成严重的伤害。因此，在生成式人工智能模型的输出结果中处理敏感数据时，应遵循相关的法律法规和隐私保护准则。这意味着需要确保数据的安全存储和传输，采用匿名化或脱敏技术，以最大限度地减少个人身份的暴露风险。此外，还应提供适当的访问控制和权限管理，确保只有经过授权的人员才能访问和处理敏感数据。

避免歧视是另一个重要的合规性问题。生成式人工智能模型的输出结果不应基于种族、性别、年龄、宗教、性取向等个人特征或背景，因为这样的歧视性结果可能对个体和社会产生负面影响。为确保合规性，需要审查和评估模型的训练数据，确保其中不存在歧视性倾向。此外，进行模型审查和监督也能帮助我们及时发现和纠正歧视性输出结果的问题。

除了处理敏感数据和避免歧视，还有其他合规性问题需要考虑。例如，确保输出结果不侵犯他人的知识产权和版权，遵守与商业机密相关的法律法规；确保输出结果不鼓励或传播恶意信息、虚假宣传或非法活动；确保输出结果不违反社交媒体平台或其他在线平台的使用政策和规定。总之，输出结果的合规性要求在生成式人工智能模型的设计、训练和部署过程中综合考虑各种法律、道德和伦理标准，并通过适当的措施来遵守和保护这些标准。

为了确保生成式人工智能模型的输出结果的合规性，可以采取以下方法：

语言模型微调

对已经训练好的语言模型进行微调，以适应特定的合规性要求。通过引入合规性约束和示例，可以指导生成过程，确保生成的内容符合合规性标准。

过滤器和审核系统

引入文本过滤器和审核系统，对生成的输出进行实时监控。利用自然语言处理技术和机器学习模型，自动识别和过滤不合规内容，确保输出结果的合规性。

用户控制与反馈

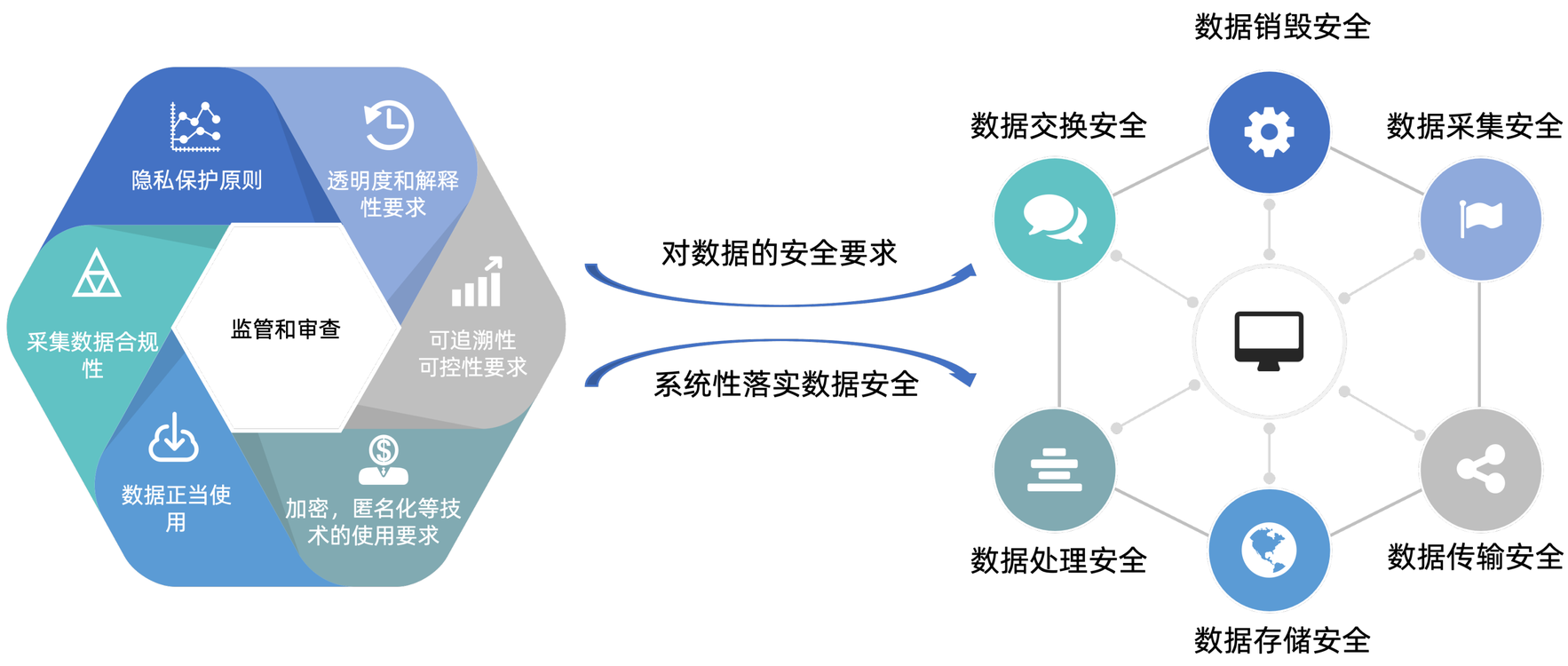
提供用户工具，使用户能够自定义人工智能生成的内容，满足其合规性偏好。同时鼓励用户报告不合规的内容，以改善系统，并及时纠正违规行为。

透明度和解释性

增加人工智能系统的透明度，使用户能够了解生成结果的原因和依据。利用解释性人工智能技术，使人工智能的决策过程更具可解释性，并帮助用户理解和验证生成结果的合规性。

3.2 生成式人工智能的数据合规技术手段

呈上，我们一方面需要识别数据在生成式人工智能中扮演着关键的角色以及在不同阶段需要考量的合规要点，我们另一方面还需要利用合理手段切实履行这些合规要求。在第二章节法律法规浅析部分，我们已经阐述了法律法规层面对企业组织、角色和流程上的相关要求，接下来，我们将着重介绍如何从技术方面满足生成式人工智能的合规要求。



安全作为生成式人工智能功能实现不可忽略的部分，企业和组织应当根据网络安全，个人信息保护，数据安全，道德伦理等方面结合自身业务需要，合理设计、开发、使用生成式人工智能技术。

通过建立系统的安全管理体系开展对生成式人工智能业务的安全保护，如建立生成式人工智能治理框架，建立对应的安全保护团队、职责分配和有效的运转、沟通协作机制；制定适合组织的制度和流程，有明确的生存周期关键控制节点授权审批流程，规范相关流程制度的制定、发布、修订；通过技术手段和产品工具在全生存周期过程中的利用和自动化技术工具的支持，对生成式人工智能安全制度流程固化执行，落实安全要求，实现安全工作；对人员进行安全能力要求的评估，开展意识和相关专业能力的培训，建设团队安全人员能力等。从网络安全，数据安全，道德伦理使用风险和安全要求，全方位的考虑制定安全政策、策略，进行安全，道德评估，并通过技术手段落实安全保护。

3.2.1 网络安全

网络安全是我们提供生成式人工智能服务不可忽视的方面。在网络安全保护方面，我们应该遵守中华人民共和国网络安全法，根据网络安全等级保护要求对生成式人工智能系统和基础设施平台进行安全保护。

安全管理	措施
安全物理环境	基础设施要求访问管理、防火防盗等。
安全通信网络	划分不同的网络区域在重要网络区域与其他网络区域间采取技术隔离手段、进行可信验证。
安全区域边界	保证跨越边界的访问和数据流通过设备提供受控通信， RBAC模型访问控制、入侵防范防火墙、IDS、IPS等技术手段，恶意代码防范，安全审计。
安全计算环境	如单点登录、多因素身份认证、限制非法登录次数、启用安全审计功能、关闭默认共享和高危端口、及时修复漏洞、数据备份、仅采集和保存必要的用户个人信息等。
安全管理中心安全	加强对系统管理员账号的安全管控，如通过特定命令或界面进行安全审计操作等。
安全管理制度	总体方针和策略，阐明工作目标、范围、原则、框架，建立健全管理制度和健全规程等。
安全管理机构	岗位职位职能职责以及分工明确、针对系统变更、重要操作等建立执行审批机制。
安全管理人员	在人员录用、离岗进行必要审查、对人员进行必要的意识教育和培训。
安全建设管理	进行定级备案、安全方案设计、对开发或采购的软件进行必要的安全验证，如密码产品符合国家密码管理主管部门的要求，制度软件测试验收方案并形成报告。
安全运维管理	对环境管理，如机房的访问；资产管理，如资产清单及重要程度；漏洞风险管理，采取必要的措施识别安全漏洞，评估隐患进行修改；应急预案管理等等。

等以上方面进行安全建设，对组织的，单位信息化现状、保护对象列表、保护对象的概述、边界、设备部署、业务应用、及其他内容进行等级认定，进行网络安全等级保护实施建设和维护。

3.2.2 数据全生命周期合规

数据作为生成式人工智能应用的主要输入和输出信息载体，在网络安全的基础上，还必须得到充分的保护和管理，在数据安全保护方面，除了遵循整体人工智能的通用管理控制措施以外，还应当关注数据全生命周期（采集，传输，存储，使用，交换和销毁）的安全建设，不断提高组织对数据保护的安全成熟度。以下我们就数据全生命周期保护进行系统性的介绍。

生成式人工智能数据安全要点2-1

数据采集安全

数据分级分类

建立数据分级分类制度，并提供对应的级别的保护要求，应用打标或数据资产管理工具进行自动识别、标识结果发布、审核，对敏感数据进行识别标注，并持续完善改进分级分类制度。

数据采集安全

制定相应流程和方式，通过威胁建模技术，进行风险管理，使用统一工具，跟踪和记录数据采集获取过程，支持可追溯性。确保数据源的真实，有效，和最小必要原则，从而保证数据采集的合规性、正当性、一致性。

数据源鉴别及记录

明确数据源管理制度，定义追溯策略要求、数据格式、追溯数据安全存储与使用的管理制度，使用工具对产生数据的数据源进行身份鉴别和记录，对追溯数据进行安全保护，从而实现对组织内部各类数据源的统计和分析。

数据质量管理

建立数据质量管理体系，通过技术工具对关键数据进行质量管理和监控，保证对数据采集过程中对收集/产生的数据的准确性、一致性和完整性的持续改进保证。

数据传输安全

数据传输加密

明确数据传输安全管理规范，在数据分级分类的基础上，明确对不同类型、级别的数据的加密传输要求，在通道两端进行主题身份和认证的技术方案和工具，应采用TLS/SSL方式数据传输安全和密钥管理机制。保证传输通道、传输节点和传输数据的安全，防止传输过程中的数据泄露。

数据源鉴别及记录

通过网络基础设施及网络层数据防泄漏设备的备份建设，制定组织的网络可用性管理指标，对关键的网络传输链路、网络设备节点实行冗余建设，实现网络的高可用性，从而保证数据传输过程的稳定性。

数据存储安全

存储媒体安全

组织内对数据存储进行访问和使用的场景，提供有效的技术和管理手段，防止对媒体的不当使用而可能引发数据泄露风险。使用技术工具对存储媒体性能进行监控，对超过安全阈值的存储媒体进行预警。并对存储媒体访问和使用行为进行记录和审计，在媒体使用和传递过程得到严密跟踪。

逻辑存储安全

基于组织内部的业务特性和数据存储安全要求，建立针对数据逻辑存储、存储容器等的有效安全控制。明确各数据逻辑存储系统的安全管理员及职责。确保数据存储符合组织的安全要求。采用可伸缩数据存储架构，以满足数据量持续增长、数据分级分类存储等需求。

数据备份和恢复

制定数据备份与恢复管理制度，通过执行定期的数据备份和恢复，对数据备份的场景、数量、频率进行定期的统计，了解数据备份工作的开展情况。依据数据生存周期和业务规范，建立数据生存周期各阶段数据归档的操作流程，归档数据的压缩或加密要求。以实现保护数据的可用性，一致性。

生成式人工智能数据安全要点2-2

数据处理安全

数据脱敏

制定数据脱敏规范，包括静态和动态脱敏方案，使用数据脱敏技术如泛化、抑制、假名化，对脱敏效果进行验证评估。

数据分析安全

保护数据挖掘和分析过程中的隐私和安全，加强算法设计和数据匿名化，防止泄露敏感信息和危害利益。

数据正当使用

建立数据使用评估制度，满足法律合规要求，保证数据安全和隐私保护。通过技术工具有效识别、监控和预警数据滥用行为。

数据处理环境安全

制定安全控制措施，定期安全审计，确保数据处理环境的安全。参考网络安全要求和实施，实施具体控制措施。

数据导入导出安全

建立数据导入导出安全策略，执行有效的访问控制和身份验证，监控接口流量过载，及时删除通道缓存数据，降低数据泄露风险。

数据交换安全

数据共享安全

确保数据共享安全合规：脱敏、加密、安全通道、统一审核、监控审计。

数据发布安全

确保数据发布安全合规：审核制度、细则、审计流程，统一发布系统，安全验证，风险提示，实现安全可控数据发布。

数据接口安全

确保数据接口安全：限制、控制、鉴别、审计、加密，防范接口安全风险。

数据销毁安全

数据销毁处置

建立数据销毁机制，确保安全销毁和评估效果，避免数据泄漏风险。

存储媒体销毁处置

确保存储媒体安全销毁，防止数据泄漏风险。

3.2.2.1 数据采集安全：

数据分级分类

从数据分级分类的原则，框架，影响因素，流程方法以及重新等级分类几个方面建立数据分级分类制度，并提供对应的级别的保护要求，应用打标或数据资产管理工具进行自动识别、标识结果发布、审核，对敏感数据进行识别标注，并持续完善改进分级分类制度。

数据采集安全管理

明确并确认采集数据的目的和用途，采集范围，数量和频率，明确采集渠道、规范数据格式及相关的流程和方式，尽可能使用统一的采集工具，跟踪和记录数据采集和获取过程，支持可追溯性，并通过针对采集阶段的风险评估流程进行风险评估，并对结果进行确认，确保采集过程中的风险得到控制抑制，不被泄露。确保数据源的真实，有效，和最小必要原则，从而保证数据采集的合规性、正当性、一致性。

数据源鉴别及记录

明确数据源管理制度，定义追溯策略要求，追溯数据格式、追溯数据安全存储与使用的管理制度，明确关键的数据管理系统上对数据源类型的标记，使用工具对产生数据的数据源进行身份鉴别和记录，防止数据仿冒和数据伪造，对关键追溯数据进行备份，并采取技术手段对追溯数据进行安全保护，提供标记数据的数据源类型功能，从而实现对组织内部各类数据源的统计和分析。

数据质量管理

建立数据质量管理体系，采用工具从数据格式要求，完整性要求，数据源质量评价，采集过程中的质量监控规则，监控范围及监控方式以及数据质量分级标准有明确规定，对数据采集、清洗、转换和加载等操作有相关安全管理规范和流程质量要求。并通过技术工具对关键数据进行质量管理和监控，实现异常数据及时告警或更正，并应定期对数据质量进行分析、预判和盘点，明确数据质量问题定位和修复时间要求。从而保证对数据采集过程中对收集/产生的数据的准确性、一致性和完整性的持续改进保证。

3.2.2.2 数据传输安全：

数据传输加密

组织内部和外部的数据传输应采用适当的加密保护措施，明确数据传输安全管理规范，在数据分级分类的基础上，明确对不同类型、级别的数据的加密传输要求，包含对数据加密算法和密钥管理的要求，如在通道两端进行主题身份和认证的技术方案和工具，如应采用TLS/SSL方式数据传输安全和密钥管理机制，在传输链路上的节点部署部署独立密钥对和数字证书，以保证节点有效的身份鉴别，综合量化敏感数据加密和数据传输通道加密的实现效果和成本，定期评估新技术对安全方案的影响，以应对最新的安全风险，审核并调整数据加密的实现方案。保证传输通道、传输节点和传输数据的安全，防止传输过程中的数据泄露。

网络可用性管理

通过网络基础设施及网络层数据防泄漏设备的备份建设，制定组织的网络可用性管理指标，包括但不限于可用性的概率数值、故障时间/频率/统计业务单元等；基于可用性管理指标，建立网络服务配置方案和宕机替代方案等。对关键的网络传输链路、网络设备节点实行冗余建设，实现网络的高可用性，从而保证数据传输过程的稳定性。

3.2.2.3 数据存储安全：

存储媒体安全

组织内对数据存储进行访问和使用的场景，提供有效的技术和管理手段，防止对媒体的不当使用而可能引发数据泄露风险。明确存储媒体访问和使用的安全管理规范，建立存储媒体使用的审批和记录流程，购买或获取存储媒体的流程，通过可信渠道购买或获取媒体，并针对各类存储媒体尽力格式化规程，对存储媒体资产进行标识，明确存储媒体存储的数据，对存储媒体进行常规和随机检查，确保存储媒体的使用符合机构公布的关于存储媒体的使用制度，使用技术工具对存储媒体性能进行监控，包括存储媒体的使用历史、性能指标、错误或损坏情况，对超过安全阈值的存储媒体进行预警。并对存储媒体访问和使用行为进行记录和审计，在媒体使用和传递过程得到严密跟踪。

逻辑存储安全

基于组织内部的业务特性和数据存储安全要求，建立针对数据逻辑存储、存储容器等的有效安全控制。根据数据分级分类要求，对数据逻辑存储管理安全规范和配置规范，如使用分层的逻辑存储，实现授权管理规则和授权操作要求，具备对数据逻辑存储结构的分层和分级保护。明确数据分片和分层式式存储安全规则，如数据存储完整性规则、多副本一致性管理规则、存储转移安全规则，以满足分布式存储下分片数据完整性、一致性和保密性保护要求。各数据逻辑存储系统的安全管理员及职责，账号权限管理、访问控制、日志管理、加密管理、版本升级等方面。在内

部数据存储系统上线前对遵循统一的配置要求进行有效的安全配置，对使用外部数据存储系统进行有效的安全配置。明确数据逻辑存储隔离授权与操作要求，确保举报多用户数据存储安全隔离能力。通过技术工具实现对安全配置情况的实现对安全配置情况的统一管理和控制，提供数据存储系统配置扫描工具，定期对主要数据存储系统的安全配置进行扫描，以保证符合安全基线要求。监测逻辑存储系统的数据使用规范性，确保数据存储符合组织的安全要求。采用可伸缩数据存储架构，以满足数据量持续增长、数据分级分类存储等需求。

应采用应用层、数据层、操作系统层、数据存储层数据层次加密架构，以满足不同类型数据如个人信息、重要数据等敏感数据的加密存储能力的系统需求。

数据备份和恢复

制定数据备份与恢复管理制度，明确对数据备份和恢复定期检查和更新工作程序，包括范围、工具、过程、日志记录、数据保存时长，数据副本的更新频率等，通过执行定期的数据备份和恢复，实现对存储数据的冗余管理，明确数据冗余强一致性、弱一致性等控制要求，以满足不同一致性水平需求的数据副本多样性和多边形存储管理要求。对数据备份的场景、数量、频率进行定期的统计，了解数据备份工作的开展情况。依据数据生存周期和业务规范，建立数据生存周期各阶段数据归档的操作流程，归档数据的压缩或加密要求。以实现保护数据的可用性，一致性。满足业务所需的RTO, 在日常备份中使用数据全备，增量备份或差异备份等机制。

3.2.2.4 数据处理安全：

数据脱敏

根据相关法律法规、标准及业务需求，结合业务数据脱敏的具体场景制定数据脱敏的规范，规则方法和使用限制，结合分级分类数据保护要求对数据脱敏进行流程、方案设计，如静态脱敏方案，动态脱敏方案，利用技术工具实现如泛化、抑制、假名化等数据脱敏技术。对数据脱敏处理过程相应的操作进行记录，以满足数据脱敏处理安全审计要求。数据脱敏后对效果进行验证评估。

数据分析安全

通过在数据分析过程采取适当的安全控制措施，防止数据挖掘、分析过程中有价值信息和个人隐私泄露的安全风险。确立明确的数据处理与分析过程的安全规范，覆盖构建数据仓库、建模、分析、挖掘、展现等方面的安全要求，明确个人信息保护、数据获取方式、访问接口、授权机制、分析逻辑安全、分析结果安全等内容；数据分析安全审核流程，对数据分析的数据源、数据分析需求、分析逻辑进行审核，以确保数据分析目的、分析操作等当面的正当性；采取必要的监控审计措施，确保实际进行的分析操作与分析结果使用与其声明的一致，整体保证数据分析的预期不会超过相关分析团队对数据的权限范围；对数据分析结果输出和使用的安全审核、合规评估和授权流程，防止数据分析

结果数据造成安全风险。在数据分析中，组织应采用多种技术手段和工具以降低数据分析过程中的隐私泄露风险，如差分隐私保护、K匿名等；记录并保存数据处理与分析过程中对个人信息、重要数据等敏感数据的操作行为；提供统一的数据处理与分析系统，并能够呈现数据处理前后数据间的映射关系；通过技术手段降低数据分析过程中的安全风险，如加强机器学习重要数据自动识别、数据安全分析算法设计等；避免输出的数据分析结果包含可恢复的个人信息、重要数据和结构标识，以防止数据分析结果危害个人隐私、公司商业价值、社会公共利益和国家安全。

数据正当使用

基于国家相关法律法规对数据分析和利用的要求，建立数据使用的评估制度，对个人信息，数据的使用前进行安全影响评估，满足国家合规要求，公司数据分级分类保护要求后允许使用数据。并应避免精确定位到特定个人，避免评价信用、资产和健康等敏感个人数据，保护国家秘密、商业秘密和个人隐私，防止数据资源被用于不正当目的，保证数据使用在声明的目的和范围内。限制用户可访问数据范围，建立相应强度或粒度的访问控制机制；具备违约责任、缔约过失责任、侵权责任等数据使用风险和处理能力。通过技术工具实现对数据滥用行为的有效识别、监控和预警。

数据处理环境安全

数据处理环境系统的设计、开发和运维阶段制定相应的安全控制措施，以及通过开展定期的安全审计，实现对数据处理环境安全的满足。具体实施细节及控制可参考网络安全要求及实施。

数据导入导出安全

依据数据分级分类要求建立符合业务规则的数据导入导出安全策略，如授权策略，流程控制策略、不一致处理策略等。通过对数据导入导出过程中对数据的安全性进行管理，防止数据导入导出过程中可能对数据自身的可用性和完整性构成的危害，明确导入导出的安全评估授权审批流程，评估导出的安全风险，对大量或敏感数据可导出进行授权审批；如采用存储媒体导出数据，应建立针对导出存储媒体的标识规范，明确存储媒体的命名规则，标识属性等重要信息，定期验证导出数据的完整性和可用性；制定导入导出审计策略和日志管理规程，并保存导入导出过程中的出错数据处理记录。记录并定期审计数据导入导出行为，确保未超出数据授权使用范围；对数据导入导出终端设备、用户或服务组件执行有效的访问控制，如多因素身份验证，用户访问管理，实现对其身份的真实性和合法性的保证；在导入导出通道提供冗余备份和对接口进行流量过载监控，在完成导入导出后对通道缓存的数据进行删除，以降低可能存在的数据泄露风险。

3.2.2.5 数据交换安全：

数据共享安全

通过业务系统、产品对外部组织提供数据时，根据共享业务需求且没有超出共享使用授权范围，通过合作的方式与合作伙伴交换数据时执行共享数据的安全风险控制，明确共享内容范围和数据共享的管控措施，及数据共享涉及机构或部门相关用户职责和权限；明确数据提供者与共享数据使用者的数据安全责任和安全防护能力，并对数据共享进行审计规程和审计日志管理的要求，明确审计记录要求，为数据共享安全事件的处置、应急响应和事后调查提供帮助；在使用外部的软件开发包/组件/代码前进行安全评估，对获取的数据应符合组织的数据安全要求。采取技术措施确保个人信息在委托处理、共享、转让等对外提供场景的安全合规，如数据脱敏、数据加密、安全通道、共享交换区域；同时建立统一的数据共享交换系统，提示数据共享交换的安全风险并进行在线审核，对共享数据及过程进行监控审计，以降低数据共享场景下的安全风险。

数据发布安全

在对外部组织进行数据发布的过程中，对安全公开发布制定审核制度,细则和审计流程，严格审核数据发布合规要求，通过对发布数据的内容、适用范围、规范、格式、发布者与使用者权利和义务执行的必要控制，定期审查公开发布的数据中是否含有非公开信，并采取相关措施满足数据发布的合规性，建立数据公开事件应急处理流程，细化各类数据发布场景的审核流程，从审核的有效性和审核的效率层面充分，考虑流程节点的制定，采用统一的发布系统，实现公开数据登记、用户注册等发布数据和发布组件的验证机制，并提示数据发布安全风险并进行在线审核，以实现数据发布过程中数据的安全可控与合规。

数据接口安全

通过建立组织的对外书接口的安全管理机制，制定数据接口安全控制策略，明确适用数据接口的安全限制和安全控制措施，如身份鉴别、访问控制、授权策略、签名、时间戳、安全协议。明确数据接口安全的要求，如接口名称、接口参数，与数据接口调用方签署合作协议，明确数据的适用目的、供应方式、保密约定、数据安全责任；对数据接口调用进行必要的自动化监控和处理，对不安全输入参数进行限制或过滤能力，为接口提供异常处理能力，对数据接口访问进行审计，并提供可配置的数据服务接口。对跨安全域间的数据接口调用采用安全通道、加密传输、事件戳等安全措施，防范组织在接口调用过程中的安全风险。

3.2.2.6 数据销毁安全:

数据销毁处置

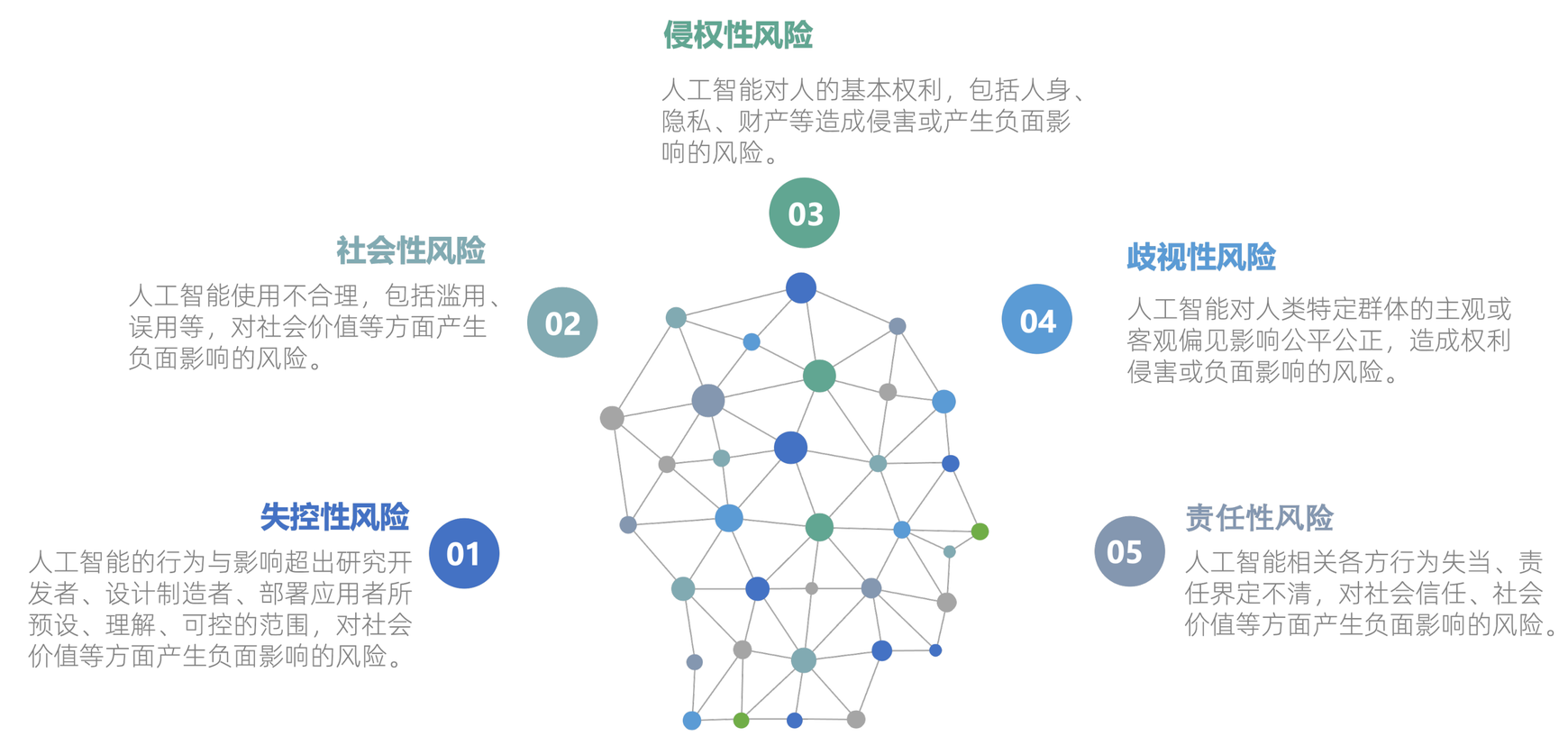
建立针对数据的删除、净化机制，实现对数据的有效销毁，依照数据分级分类建立数据销毁策略和管理制度，明确数据销毁的场景、销毁对象、销毁方式和销毁要求，建立规范的数据销毁流程和审批机制，对销毁操作工程进行监督审批，对销毁过程进行记录控制。对销毁效果建立评估机制，定期对数据销毁效果进行抽样认定。明确对已共享或已被其他用户适用的数据销毁管控措施。组织的数据资产管理系统应能够对数据的销毁需求进行明确的标识，并可通过该系统提醒数据管理者及时发起对数据的销毁。并通过技术手段避免对数据的误销毁。针对网络存储数据，建立硬销毁和软销毁的数据销毁方法和技术，如基于安全测率、基于分布式杂凑算法等网络数据分布式存储的销毁策略和机制；配置必要的数据销毁技术手段与管控措施，确保以不可逆方式销毁敏感数据及其副本内容。防止因对存储媒体中的书进行恢复而导致的数据泄漏风险。

存储媒体销毁处置

建立对存储媒体安全销毁的规程，如明确存储媒体销毁处理策略、管理制度和机制，明确销毁对象和流程。依据存储媒体存储内容的重要性，明确磁媒体、光媒体和半导体媒体等不同类存储媒体的销毁方法。通过对销毁的监控机制，确保对销毁存储媒体的登记、审批、交接等存储媒体销毁过程进行监控。并实施存储媒体销毁效果评估机制，定期对存储媒体销毁进行抽样认定，定期对销毁记录进行检查。提供统一的存储媒体销毁工具，包括但不限于物理销毁、消磁设备等工具，能够实现对各类媒体的有效销毁。对闪存盘、硬盘、光盘、磁带等存储媒体建立硬销毁和软销毁的数据销毁方法和技术。如需要时，应由经过认证的机构或设备对存储媒体进行物理销毁，或联系经认证的销毁服务商进行存储媒体销毁工作。防止因存储媒体丢失、被窃或未授权的访问而导致存储媒体中的数据泄漏的安全风险。

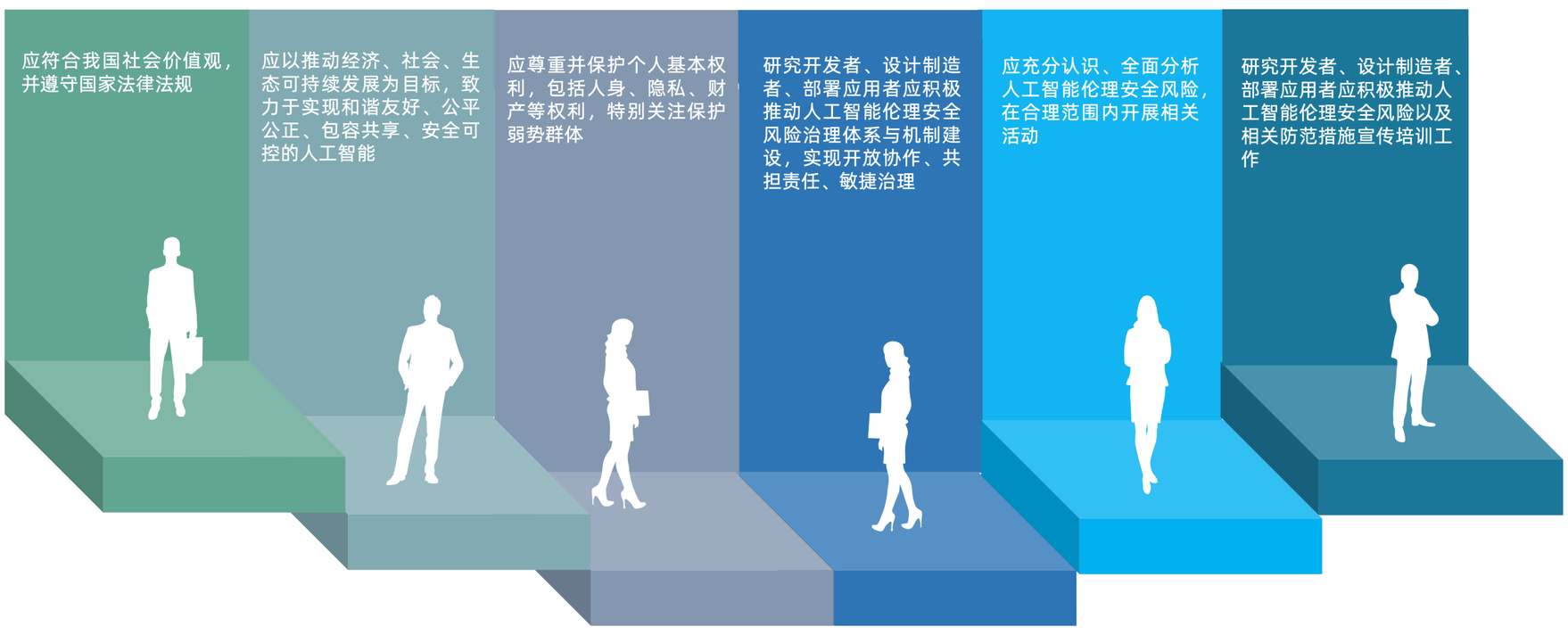
3.2.3 生成式人工智能引发的伦理道德风险和应对措施

在使用生成式人工智能，我们还需要关注伦理道德的风险，我们常见的伦理道德风险一般有以下几点：



生成式人工智能面临的道德伦理风险

在伦理安全风险防范的基本要求包括：



生成式人工智能伦理安全风险防范的基本要求

应符合我国社会价值观，并遵守国家法律法规。

应以推动经济、社会、生态可持续发展为目标，致力于实现和谐友好、公平公正、包容共享、安全可控的人工智能。

应尊重并保护个人基本权利，包括人身、隐私、财产等权利，特别关注保护弱势群体。 注：弱势群体是指生存状况、就业情况、发声途径或争取合法权益保障能力等方面处于弱势的群体。

应充分认识、全面分析人工智能伦理安全风险，在合理范围内开展相关活动。 注：合理范围是指以保障个人权利、提升社会价值为目标，具备明确边界以及清楚预期的范围。

研究开发者、设计制造者、部署应用者应积极推动人工智能伦理安全风险治理体系与机制建设，实现开放协作、共担责任、敏捷治理； 注：敏捷治理是指持续发现和降低风险、优化管理机制、完善治理体系，并推动治理体系与机制覆盖人工智能系统、产品和服务全生命周期的理念。

研究开发者、设计制造者、部署应用者应积极推动人工智能伦理安全风险以及相关防范措施宣传培训工作。

3.2.4 生成式人工智能的全生命周期合规

- 不应研究开发以损害人的基本权利为目的的人工智能技术；
- 应避免研究开发可能被恶意利用进而损害人的基本权利的人工智能技术；
- 应谨慎开展具有自我复制或自我改进能力的自主性人工智能的研究开发，持续评估可能出现的失控性风险； 注：自主性人工智能指可以感知环境并在没有人为干涉的情况下独立作出决策的人工智能；
- 应不断提升人工智能的可解释性、可控性；
- 应对研究开发关键决策进行记录并建立回溯机制，对人工智能伦理安全风险相关事项，进行必要的预警、沟通、回应；注：研究开发关键决策是指对研究开发结果可能产生重大影响的决策，如数据集的选择、算法的选取等；
- 应推动研究开发合作、互信，促进良性竞争与多元化技术发展。

- 应以良好目的使用人工智能、充分体现人工智能的积极作用，不应以有损社会价值、个人权利等目的恶意使用人工智能；
- 应主动了解人工智能伦理安全风险，积极向研究开发者、设计制造者、部署应用者反馈人工智能伦理安全风险相关信息。

- 不应设计制造损害公共利益或个人权利的人工智能系统、产品或服务；
- 应不断提升人工智能系统、产品和服务的可解释性、可控性；
- 应及时、准确、完整、清晰、无歧义地向部署应用者说明人工智能系统、产品或服务的功能、局限、安全风险和可能的影响；
- 应在系统、产品或服务中设置事故应急处置机制，包括人工紧急干预机制等；应明确事故处理流程，确保在人工智能伦理安全风险发生时作出及时响应，如停止问题产品生产、召回问题产品等；
- 应设置事故信息回溯机制； 示例：通过黑匣子实现无人驾驶的事故信息回溯；
- 应对人工智能伦理安全风险建立必要的保障机制，对引起的损失提供救济。注：可通过购买保险等手段为必要救济提供保障。

- 使用人工智能作为直接决策依据并影响个人权利时，应具有清晰、明确、可查的法律法规等依据；
- 在公共服务、金融服务、健康卫生、福利教育等领域，进行重要决策时如使用不可解释的人工智能，应仅作为辅助决策手段，不作为直接决策依据；
- 应向用户及时、准确、完整、清晰、无歧义地说明人工智能相关系统、产品或服务的功能、局限、风险以及可能的影响，并解释相关应用过程及应用结果；
- 提供用户拒绝、停止使用人工智能系统的机制和替代选择方案；
- 建立事故应急处置机制，包括紧急干预和中止应用；
- 提供清晰、易操作的投诉、质疑和反馈机制，并进行处理和补偿；
- 主动识别和改进人工智能伦理安全风险。

3.2.4.1 研究开发

- 不应研究开发以损害人的基本权利为目的的人工智能技术。
- 应避免研究开发可能被恶意利用进而损害人的基本权利的人工智能技术。
- 应谨慎开展具有自我复制或自我改进能力的自主性人工智能的研究开发，持续评估可能出现的失控性风险； 注：自主性人工智能指可以感知环境并在没有人为干涉的情况下独立作出决策的人工智能。
- 应不断提升人工智能的可解释性、可控性。

- 应对研究开发关键决策进行记录并建立回溯机制，对人工智能伦理安全风险相关事项，进行必要的预警、沟通、回应；注：研究开发关键决策是指对研究开发结果可能产生重大影响的决策，如数据集的选择、算法的选取等。

- 应推动研究开发合作、互信，促进良性竞争与多元化技术发展。

3.2.4.2 设计制造

- 不应设计制造损害公共利益或个人权利的人工智能系统、产品或服务。

- 应不断提升人工智能系统、产品和服务的可解释性、可控性。

- 应及时、准确、完整、清晰、无歧义地向部署应用者说明人工智能系统、产品或服务的功能、局限、安全风险和可能的影响。

- 应在系统、产品或服务中设置事故应急处置机制，包括人工紧急干预机制等；应明确事故处理流程，确保在人工智能伦理安全风险发生时作出及时响应，如停止问题产品生产、召回问题产品等。

- 应设置事故信息回溯机制； 示例：通过黑匣子实现无人驾驶的事故信息回溯。

- 应对人工智能伦理安全风险建立必要的保障机制，对引起的损失提供救济。 注：可通过购买保险等手段为必要救济提供保障。

3.2.4.3 部署应用

- 使用人工智能作为直接决策依据并影响个人权利时，应具有清晰、明确、可查的法律法规等依据。

- 在公共服务、金融服务、健康卫生、福利教育等领域，进行重要决策时如使用不可解释的人工智能，应仅作为辅助决策手段，不作为直接决策依据。

- 应向用户及时、准确、完整、清晰、无歧义地说明人工智能相关系统、产品或服务的功能、局限、风险以及可能的影响，并解释相关应用过程及应用结果。

- 应以清楚明确且便于操作的方式向用户提供拒绝、干预及停止使用人工智能相关系统、产品或服务的机制。

- 在用户拒绝或停止使用后，应尽可能为用户提供非人工智能的替代选择方案；注：用户停止使用包括因主观原因停止使用，以及因客观条件，如生理缺陷等，无法继续使用的情况。

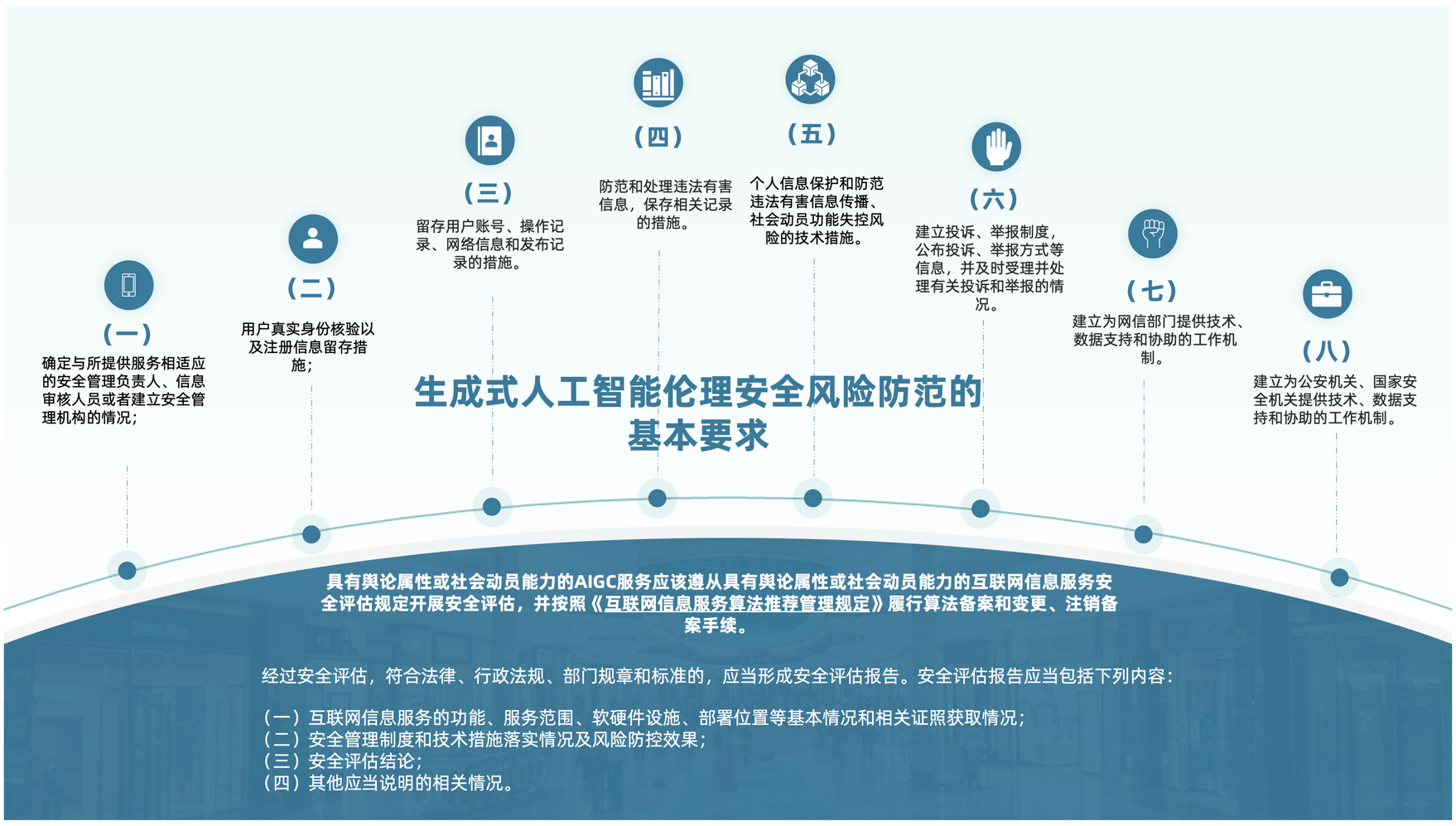
- 应设置事故应急处置机制，包括人工紧急干预机制、中止应用机制等，明确事故处理流程，确保在人工智能伦理安全风险发生时作出及时响应。
- 应向用户提供清楚明确且便于操作的投诉、质疑与反馈机制，并提供包含人工服务在内的响应机制，进行处理和必要补偿。
- 应主动识别发现人工智能伦理安全风险，并持续改进部署应用过程。

3.2.4.4 服务应用

- 应以良好目的使用人工智能、充分体现人工智能的积极作用，不应以有损社会价值、个人权利等目的恶意使用人工智能。
- 应主动了解人工智能伦理安全风险，积极向研究开发者、设计制造者、部署应用者反馈人工智能伦理安全风险相关信息。

3.2.5 生成式人工智能安全评估和算法管理

从技术使用上来说，应当按照国家有关规定开展安全评估，并按照《互联网信息服务算法推荐管理规定》履行算法备案和变更、注销备案手续。



安全评估可以包括：

- 确定与所提供服务相适应的安全管理负责人、信息审核人员或者建立安全管理机构的情况。
- 用户真实身份核验以及注册信息留存措施。
- 对用户的账号、操作时间、操作类型、网络源地址和目标地址、网络源端口、客户端硬件特征等日志信息，以及用户发布信息记录的留存措施。
- 对用户账号和通讯群组名称、昵称、简介、备注、标识，信息发布、转发、评论和通讯群组等服务功能中违法有害信息的防范处置和有关记录保存措施。
- 个人信息保护以及防范违法有害信息传播扩散、社会动员功能失控风险的技术措施。
- 建立投诉、举报制度，公布投诉、举报方式等信息，及时受理并处理有关投诉和举报的情况。
- 建立为网信部门依法履行互联网信息服务监督管理职责提供技术、数据支持和协助的工作机制的情况。
- 建立为公安机关、国家安全机关依法维护国家安全和查处违法犯罪提供技术、数据支持和协助的工作机制的情况。

经过安全评估，符合法律、行政法规、部门规章和标准的，应当形成安全评估报告。
安全评估报告应当包括下列内容：

- 互联网信息服务的功能、服务范围、软硬件设施、部署位置等基本情况和相关证照获取情况。
- 安全管理制度和技术措施落实情况及风险防控效果。
- 安全评估结论。
- 其他应当说明的相关情况。

算法推荐相关义务和责任：

算法推荐服务提供者应当落实算法安全主体责任，建立健全算法机制机理审核、科技伦理审查、用户注册、信息发布审核、数据安全和个人信息保护、反电信网络诈骗、安全评估监测、安全事件应急处置等管理制度和技术措施，制定并公开算法推荐服务相关规则，配备与算法推荐服务规模相适应的专业人员和技术支撑。 建立健全用于识别违法和不良信息的特征库，完善入库标准、规则和程序。发现未作显著标识的算法生成合成信息的，应当作出显著标识后，方可继续传输。 发现违法信息的，应当立即停止传输，采取消除等处置措施，防止信息扩散，保存有关记录，并向网信部门和有关部门报告。发现不良信息的，应当按照网络信息内容生态治理有关规定予以处置。 网信部门会同电信、公安、市场监管等有关部门建立算法分级分类安全管理制度，根据算法推荐服务的舆论属性或者社会动员能力、内容类别、用户规模、算法推荐技术处理的数据重要程度、对用户行为的干预程度等对算法推荐服务提供者实施分级分类管理。

4 凯捷提供的服务

基于生成式人工智能这个新兴领域还在不停发展和变化，保证合规和遵守伦理道德是每个正在或者计划使用这项科技的企业和组织都应该关注的核心问题。

凯捷能够在以下几个方面提供专业的咨询服务：

4.1 法律法规追踪：

法律检索

凯捷数据合规团队通过查询各国数据成文法和行政法规、规章、公共文件、期刊论文以及法律新闻等内容做为参考样本，确保信息准确、来源合规。

专业解读

团队整合法律法规、政策解读、行政处罚、判决文书等内容均来自官方机构发布，专家专栏、评论文章、热点话题等热点评论内容由国内顶尖律所、知名企业、法院、政府的专家意见组成。追踪前沿法律热点，提供统一、专业的定制化咨询服务报告。

专家问答

凯捷客户可以提出多个场景的个性化问题，凯捷数据合规团队以最便捷的方式，为客户提供专业的定制化解答和数据合规培训。

4.2 合规管理体系建设

用户隐私管理体系

包括数据采集与隐私保护、数据访问和控制、用户教育和透明性。

生成式人工智能全生命周期管理体系

包括数据采集与预处理、模型训练与验证、数据评估与调整、输出结果四个阶段。

网络安全评估

包括防护措施、访问控制、安全更新和安全培训四方面。

数据全生命周期合规

包括数据采集、数据传输、存储安全、使用安全、交换安全、销毁安全。

4.3 合规和伦理道德培训

面向企业高管：

针对生成式人工智能面临的角色和职责培训。

面向员工：

针对生成式人工智能的合规意识、伦理道德意识培训。

5 引用材料

1. Alignment of Language Agents. Kenton et al. 2021.03

2. Harnessing the Value of Gen AI. Capgemini Research Institute. 2023.07

3. Teach Language Models to Reason. Google Deepmind. 2023.09

4.《中国AIGC产业全景报告暨AIGC》，量子位，2023.03

5.《年增长率60%、市场规模已超百亿的AIGC，正一头扎进医疗》，动脉网，2023.05

6.《关于AIGC技术在金融业应用的思考与建议》，中国金融电脑，2023.08

7. 国家互联网信息办公室有关负责人就《生成式人工智能服务管理暂行办法》答记者问
http://www.cac.gov.cn/2023-07/13/c_1690898326863363.htm

8. https://www.nfx.com/post/generative-ai-tech-market-map

9. Artificial intelligence act (europa.eu)

10. Artificial Intelligence Strategy of the German Federal Government
(ki-strategie-deutschland.de)

11. https://baijiahao.baidu.com/s?id=1768733406669188370&wfr=spider&for=pc

12. https://www.nist.gov/news-events/news/2023/01/nist-risk-management-
frameworkaims- improve-trustworthiness-artificial

13. https://baijiahao.baidu.com/s?id=1768733406669188370&wfr=spider&for=pc

14. https://digitalstrategy.ec.europa.eu/en/policies/regulatory-framework-ai

15.《Enabling trustworthy innovation to thrive in the UK》
https://cdei.blog.gov.uk/2021/09/10/enabling-trustworthy-innovation-to-thrive-in-the-uk/"

16.《生成式人工智能服务管理暂行办法》
https://www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm

17.《互联网信息服务深度合成管理规定》
http://www.cac.gov.cn/2022-12/11/c_1672221949354811.htm

18. https://baijiahao.baidu.com/s?id=1762920981206832947&wfr=spider&for=pc

19. 中共中央办公厅 国务院办公厅印发《关于加强科技伦理治理的意见》
https://www.gov.cn/zhengce/2022-03/20/content_5680105.htm

20. 关于印发《科技伦理审查办法（试行）》的通知
https://www.gov.cn/zhengce/zhengceku/202310/content_6908045.htm

21. https://baijiahao.baidu.com/s?id=1778006763546306151&wfr=spider&for=pc

22.《中华人民共和国网络安全法》
http://www.cac.gov.cn/2016-11/07/c_1119867116.htm

23.《中华人民共和国数据安全法》
http://www.npc.gov.cn/npc/c2/c30834/202106/t20210610_311888.html

24.《中华人民共和国个人信息保护法》
http://www.npc.gov.cn/npc/c2/c30834/202108/t20210820_313088.html

25.《新一代人工智能伦理规范》
https://www.most.gov.cn/kjbgz/202109/t20210926_177063.html

26.《互联网信息服务算法推荐管理规定》
http://www.cac.gov.cn/2022-01/04/c_1642894606364259.htm

27.《数据安全成熟度模型》
http://c.gb688.cn/bzgk/gb/showGb?type=online&hcno=3CFD5E5A14C24D303EA1E139E6EB75C8

28.《网络安全等级保护基本要求》
http://c.gb688.cn/bzgk/gb/showGb?type=online&hcno=BAFB47E8874764186BDB7865E8344DAF

29.《网络安全标准实践指南—人工智能伦理安全风险防范指引》
https://www.tc260.org.cn/front/postDetail.html?id=20210105115207

30.《 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》
https://www.gov.cn/zhengce/zhengceku/2018-11/30/content_5457763.htm

6 关于作者

主要编写人员

王菲
凯捷中国数字化管理咨询事业部
负责人

fei.c.wang@capgemini.com

马宁
凯捷中国数字化管理咨询事业部
合规板块负责人

ning.a.ma@capgemini.com

李想
凯捷中国数字化管理咨询事业部
合规板块负责人

xiang.b.li@capgemini.com

肖倩
凯捷中国数字化管理咨询事业部
合规板块高级咨询顾问

qian.xiao@capgemini.com

参与编写人员

张秉阳　凯捷中国创新加速器负责人
bryan.zhang@capgemini.com

陈维铠　凯捷中国创新加速器算法工程师
weikai.chen@capgemini.com

陈佳敏　凯捷中国创新加速器商业分析师
jiamin.chen@capgemini.com

王东頔　凯捷中国数字化管理咨询事业部合规板块咨询顾问
dongdi.wang@capgemini.com

杨君武　凯捷中国数字化管理咨询事业部合规板块咨询顾问
junwu.yang@capgemini.com

关于凯捷集团

凯捷(Capgemini） 是全球领先的企业合作伙伴，利用技术的力量改造和管理企业业务。其宗旨是通过技术释放人类能量，创造一个包容和可持续的未来。凯捷是一个负责任的多元化组织，集团成立于1967年，总部位于法国巴黎，在50多个国家拥有近35万名团队成员。

凭借其50余年的悠久历史和深厚的行业专业知识，在快速发展的云、数据、人工智能、互联连接、软件、数字工程和平台的创新世界推动下，凯捷深受客户信任，能够满足客户从战路、设计到运营的全方位业务需求。集团2022年全球收入为220亿欧元。

Get the Future You Want | www.capgemini.com