大模型合规白皮书

2023

2023 年 11 月

金杜律师事务所 KING&WOOD MALLESONS

上海人工智能研究院 Shanghai Artificial Intelligence Research Institute

HUAWEI

昇思 MindSpore

大模型合规白皮书

金杜律师事务所

上海人工智能研究院

华为技术有限公司

上海昇思AI框架&大模型创新中心

2023年11月

前言

大模型作为人工智能发展脉络中的里程碑，引发了新一轮的科技创新浪潮，其以强大的计算能力和深度学习技术，极大地提高了内容生产效率，促进内容生产方式颠覆式变革。各行各业纷纷布局大模型应用，把握智能化发展的机遇。然而，大模型也面临隐私泄露、侵犯第三方权益以及违背伦理等潜在风险，引发了社会各界的关注和担忧。随着大模型的广泛应用，加快完善大模型的立法监管以确保大模型的应用与发展符合伦理道德和社会价值观，推动人工智能科技的健康发展变得迫在眉睫。

世界上主要国家和地区均着手并加快完善大模型相关的法律监管。例如，欧盟以《人工智能法案》为核心，结合大模型可能涉及的其他领域的立法，逐步建立起专项法案为主、现存法规为辅的人工智能法律监管框架；美国对于人工智能大模型的立法较为分散，各州分别各自推进人工智能立法，联邦政府则试图在现有的立法框架及监管规则内对大模型及人工智能进行规制，但同时，人工智能相关的联邦专项立法提案也在推进当中。我国围绕网络安全、数据安全、个人信息保护等重点领域制定了法律法规，并及时跟进人工智能技术创新发展态势，先后针对互联网信息推荐、生成式人工智能等技术领域出台了管理办法，建立了法律法规和标准规范相协调的人工智能监管制度体系。

在此背景下，本白皮书在我国人工智能法律监管框架下进一步梳理了大模型相关方的合规义务及要点，并展望未来大模型法律监管体系的发展趋势与特征，对政府、企业、社会共建大模型治理体系提出切实建议，从而为社会各界了解大模型立法最新动态和立法趋势提供有价值的参考，并为相关单位开展大模型业务提供法律解读及合规指引，保障大模型相关业务的合规经营以及行业的健康规范发展。

目录

一、大模型的发展历程

(一)早期模型的探索与局限性

从早期的符号逻辑到现代的深度学习[1]模型，AI 领域经历了数十年的探索和迭代，为后续突破打下了坚实基础。随着大数据的发展和 AI 计算能力的爆炸式增长，深度学习模型的崛起显得尤为突出。然而，尽管这些模型在特定任务上取得了令人瞩目的成就，其在初期也面临着许多局限性，如存在数据依赖、计算消耗大、缺乏可解释性等。这些局限性不仅为 AI 领域带来技术挑战，也引发了对模型偏见、安全性和应用范围的深入思考。

1956 年 6 月举行的达特茅斯夏季人工智能研究项目，被广泛认为是人工智能作为一个研究学科的开端。自“人工智能”概念被提出，大模型的发展经历了三个阶段：

- **早期发展期(1956-2005)**：该阶段主要是**传统神经网络模型的阶段**，例如循环神经网络(Recurrent Neural Network，“RNN”)[2]、卷积神经网络(Convolutional Neural Networks，“**CNN**”)[3]。起初，AI发展主要基于小规模的专家知识，然后逐渐转向机器学习[4]，1980年和1998年诞生的CNN和LeNet-5[5]奠定了深度学习模型的基础。

- **快速成长期(2006-2019)**：该阶段是**全新的神经网络模型阶段**，模型的发展方向主要聚焦长序列的处理和计算效率的提升，以Transformer[6]架

[1] 深度学习 (Deep learning) 是机器学习 (Machine learning) 中的一类算法，指利用多层神经网络，模仿人脑处理信息的方式从原始输入中逐步提取和表达数据的特征。https://en.wikipedia.org/wiki/Deep_learning，最后访问于 2023 年 11 月 22 日。

[2] 循环神经网络 (Recurrent Neural Network，RNN) 是具有时间联结的前馈神经网络 (Feedforward Neural Networks)，特点是必须按顺序处理，并且上一层的神经细胞层输出和隐藏状态具有较大的权重影响下一层的运算。循环神经网络必须完成上一步才能进行下一步，只能串行不能并行，因此循环神经网络具有“短时记忆”的特点，技术上把这个现象称为梯度消失或梯度爆炸，循环神经网络不擅长处理和捕捉长文本中的语义。https://en.wikipedia.org/wiki/Recurrent_neural_network，最后访问于 2023 年 11 月 22 日。

[3] 卷积神经网络 (Convolutional Neural Networks，CNN) 是一类包含卷积计算且具有深度结构的前馈神经网络 (Feedforward Neural Networks)，是深度学习 (Deep learning) 的代表算法之一。https://en.wikipedia.org/wiki/Convolutional_neural_network，最后访问于 2023 年 11 月 22 日。

[4] 机器学习 (Machine learning)，作为人工智能的一个分支，是指不需要进行显式编程，而由计算系统基于算法和数据集自行学习，做出识别、决策和预测的过程。https://en.wikipedia.org/wiki/Machine_learning，最后访问于 2023 年 11 月 22 日。

[5] LeNet 又称 LeNet-5，由 Yann Lecun 提出，是一种经典的卷积神经网络，是现代卷积神经网络的起源之一。https://en.wikipedia.org/wiki/LeNet，最后访问于 2023 年 11 月 22 日。

[6] Transformer 是一种基于注意力机制的序列模型，最初由 Google 的研究团队提出并应用于机器翻译任务。

构的出现为代表。从2013年的Word2Vec[7]到2017年的Transformer，都标志着深度学习模型正走向一个全新的时代。在该阶段，如GPT[8]和BERT[9]等预训练模型逐渐成为主流。

- **全面爆发期(2020-至今)**：该阶段是**预训练大模型阶段**。以GPT为代表，预训练大模型处于快速发展的阶段，特别是OpenAI[10]推出的GPT-3和GPT-4，标志着大模型技术正迈向新高度。

机器学习有三种主要的方式，分别是监督学习、无监督学习、强化学习。

- **监督学习(Supervised Learning)**：“模板规范”（投喂好的资料），我们向模型投喂带有标签的数据（包括数据特征和期望的输出值），让算法学习输入和输出之间的映射关系。经典的监督学习包括分类和回归。

  分类：例如学习大量猫和狗的图片和标签，当模型接收新的动物图片时可以将其根据特征识别是猫还是狗；

  回归：例如学习猫的产地、毛色、习性等特征，并将猫的价值作为输出标签进行训练，当模型接收新的猫咪图片时可以根据特征预测猫的价值。

- **无监督学习(Unsupervised Learning)**：“开卷有益”（多投喂资料），我们向模型投喂不带标签的数据，让模型自行寻找其中的规律，并进行处理。经典的无监督学习包括聚类和降维。

  聚类：例如学习大量房屋的信息，模型自行寻找其中的价格、面积、户

[7] Word2vec，是一群用来产生词向量的相关模型。这些模型为浅而双层的神经网络，用来训练以重新建构语言学之词文本。https://en.wikipedia.org/wiki/Word2vec，最后访问于 2023 年 11 月 22 日。

[8] GPT，全称 Generative Pre-Trained Transformer(生成式预训练 Transformer 模型)，是一种基于互联网的、可用数据来训练的、文本生成的深度学习模型。https://en.wikipedia.org/wiki/Generative_pre-trained_transformer，最后访问于 2023 年 11 月 22 日。

[9] BERT(Bidirectional Encoder Representations from Transformers) 是一种预训练的深度学习模型，用于自然语言处理任务，基于 Transformer 架构的双向编码器，通过无监督的学习方式预训练语言表示，以便能够捕捉语言的上下文信息。

[10] OpenAI 是在美国成立的人工智能研究公司，核心宗旨在于“实现安全的通用人工智能 (Artificial General Intelligence，AGI)”，使其有益于人类。https://en.wikipedia.org/wiki/OpenAI，最后访问于 2023 年 11 月 22 日。

型的规律，并自动将相同类型的房屋进行汇总。

降维：例如学习大量房屋的信息，模型自行寻找其中用户决策最关心的因素，在保留价格和其他少量辅助数据的同时对房屋数据进行压缩，以便简化建模。

- **强化学习(Reinforcement Learning)**：“创意引导”（进行条件反射），我们向模型设置特定环境，让模型在其中采取行动，我们再对其进行反馈，让模型从反馈中学习以便优化下一次的行动。这一过程就类似以条件反射的方式训练小狗。

在机器学习领域的早期阶段，研究者们的主要关注点是基于统计、线性回归和决策树等的简单模型。早期模型具有以下特点：**简单性**。早期的模型，如线性回归和逻辑回归，是基于明确的数学方程，使其容易被理解和解释。**计算消耗低**。由于模型的简单性，其在计算上相对高效，不需要大量的计算资源。**表示能力存在上限**。虽然早期模型在特定方面表现良好，但其表示能力有限，尤其体现在处理复杂任务和非线性问题上。

大模型早期所面临的主要局限性包括：

- **存在数据依赖**：早期的模型对于大量高质量数据有极高的依赖性。在没有足够训练数据的情况下，这些模型往往难以达到令人满意的性能，但获取、清洗、标注这些数据却昂贵且极为耗时。

- **缺乏可解释性**：大模型通常被视为“黑盒”，即模型的内部工作原理很难被理解。由于用户需要理解模型的决策过程，模型的解释性不足在很多关键领域(如医疗和司法)构成障碍。

- **泛化能力不足**：尽管早期的大模型在特定任务中表现性能优秀，但其在新数据或新场景中的泛化能力仍受到质疑。

- **存在环境和任务依赖**：早期的AI模型通常需要根据特定任务定制和调

整，这意味着为特定任务训练的模型可能难以直接应用于其他任务。

- **模型具有一定偏见**：由于训练数据往往包含现实世界的偏见，大模型可能反映这些偏见，导致应用于实际场景时出现歧视或不公平的决策。

- **安全性和稳定性不足**：由于早期大模型的复杂性，其易受到对抗性攻击或在特定条件下表现不稳定。

以上局限性不仅为 AI 领域的研究者和工程师带来挑战，也为 AI 技术的未来发展和应用提出反思和探索的方向。随着技术发展，许多问题已经得到解决或缓解。

(二)深度学习的崛起

深度学习从其最初的简单尝试到现今所达到的辉煌高峰，不仅展现了技术的快速发展，更揭示了人类在追求智慧和知识上的不懈努力。深度学习源自人类对人脑工作原理的好奇和模仿，意图借助数学和算法的力量，赋予计算机对信息的处理和认知能力。随着技术日益成熟，深度学习赋予计算机识别图像、处理自然语言甚至复杂决策的能力，不仅体现技术进步，也标志人工智能正逐步走向更加深入、广泛的应用领域，为人类生活带来无尽可能性。因此，深度学习的崛起可以被视为人类科技史上的一大里程碑。

神经网络的早期探索。1957 年，Frank Rosenblatt 提出感知器模型，被称为最简单的神经网络，通过简单的线性组合实现分类任务。尽管当时的应用领域有限，但其为后续神经网络的发展奠定了基础。19 世纪 80 年代，Rumelhart、Hinton 及其团队引入了反向传播算法，通过多层神经网络训练，为复杂模型和任务提供强大工具。

数据与计算能力的融合。21 世纪初，互联网的广泛传播和智能设备的普及，使得数据呈现指数级增长，为深度学习提供丰富的训练数据。同时，硬件技术也在飞速发展，NVIDIA 等厂商投入 GPU 研发，其能够大幅度加速数值计算，尤其是深度学习中的矩阵运算，软硬件的进步大大加速了模型的训练过程。

关键技术突破与模型创新。1997 年，Hochreiter 和 Schmidhuber 提出长短时记忆网络 (Long Short-Term Memory，LSTM)，解决了循环神经网络的梯度消失 / 梯度爆炸的问题，使得神经网络可以更好的处理长文本内容，为序列数据的处理开辟了新天地。1998 年，Yann LeCun 及其团队提出 LeNet-5，但真正让深度学习走向世界舞台的是 2012 年由 Alex Krizhevsky 等人设计的 AlexNet，其在 ImageNet 挑战赛中大胜，展示了深度学习在图像处理上的潜力。2014 年，生成式对抗网络 (Generative Adversarial Networks，“GAN”) 被提出。GAN 的原理是通过竞争机制来逐步提高生成器的准确性。2016 年横空出世击败围棋世界冠军李世石的 AlphaGo，就是基于 GAN 架构训练的模型。2017 年，Google 提出 Transformer 架构，此后 BERT、GPT 等模型皆以其为基础，在自然语言处理任务中达到新高度。

(三)GPT等代表性大模型的影响

Transformer 架构的优点是可以并行处理输入序列的所有元素，能够捕捉长序列内容的关联关系，因此 Transformer 架构不再受到“短时记忆”的影响，有能力理解全文，进而 Transformer 成为自然语言处理的主流架构。

一个原始的 Transformer 架构由编码器 (Encoder) 和解码器 (Decorder) 两部分构成，其中编码器用于将输入序列转换为一系列特征向量，解码器则将这些特征向量转换为输出序列，即：输入内容——编码器——解码器——输出内容。如果给编码器输入一句英语“She is a student”，解码器返回一句对应的中文“她是一名学生”。Transformer 的架构和自注意力机制能够实现这些的关键在于“将词汇转换为词向量，并通过多头注意力机制（Multi-Head Attention）和前馈神经网络（Feed-Forward Network）两个子层进行处理”。

第一步：模型对接收到的输入序列文本 Token 化， Token 可以被理解为文本的基本单元，短单词可能是一个 Token，长单词可能是多个 Token。Token 是 GPT 的收费单元，也是源于此。

第二步：将 Token 转换成一个数字，成为 Token ID，因为计算机语言只

能存储和运算数字。

第三步: 将Token ID传入嵌入层(Embedding Layer),转换为词向量(Word Embedding)，词向量是一串数字。

可以将这个过程想象为将一个单词放到多维空间中，每个数字就表达了这个单词某个维度的含义，一串数字所能表达和蕴含的信息量远多于 Token ID 的一个数字，可以记载这个单词的词义、语法和不同语境、语序中的变化关系。

第四步：对词向量的语序和语境进行位置编码，形成位置向量。上文提到语境和语序对理解词义至关重要。之后将词向量合并位置向量，将合并后的结果传给编码器，这样模型既能理解词义也能理解语境和语序。

第五步：接收到上述信息后，编码器中的多头注意力机制将会运作，捕捉其中的关键特征，编码器在处理时不仅会关注这个词与临近的词，还会关注输入序列中所有其他词，将输入的信息根据上下文进行调整，输出了降维后的向量。

第六步：进入编码器的前馈神经网络处理，前馈神经网络“思考”之前步骤中收集的信息，并增强模型的表达能力，尝试进行预测。

第七步：降维后的向量将继续传输给解码器运算。解码器具有带掩码的多头注意力机制，解码器在处理时仅关注这个词及其之前的词，遮盖输入序列中后面的内容，并结合已经生成的文本，保持未来输出文本的时间顺序及逻辑连贯性。

第八步：进入解码器的前馈神经网络处理，解码器中的前馈神经网络与第六步类似，也是增强模型的表达能力。

第九步：解码器的最后处理环节经过 linear 层和 softmax 层，这两个子层将解码器输出内容转换为词汇表的概率分布，概率分布反映下一个 Token 生成概率。通常模型选择概率最高的 Token 作为输出，生成输出序列。因此解码器本质上是在做“单词接龙”的游戏，猜下一个输出单词。

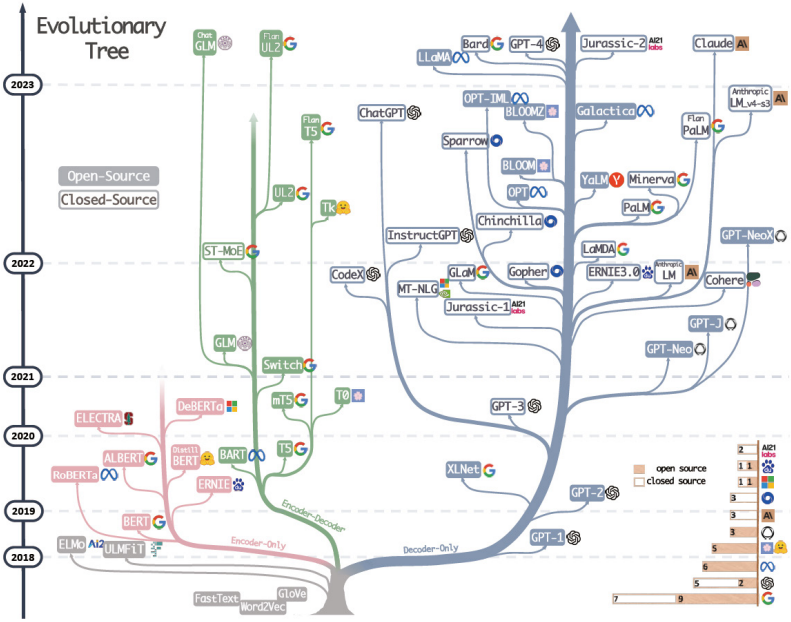


图 1 近年来大语言模型进化树 [11]

从图 1 可以看出，经过演变，大模型大致分为三种：其一是舍弃 Decoder、仅使用 Encoder 作为编码器的预训练模型，以 Bert 为代表，但 Bert 未突破 Scaling Laws，Encoder-Only 分支在 2021 年后逐渐没落。其二是同时使用 Encoder、Decoder 的预训练模型，代表模型有清华大学的 chatGLM。其三是舍弃 Encoder、仅使用 Decoder 作为编码器的预训练模型，以 GPT 为代表，其通过预测下一个单词，基于给定的文本序列进行训练。GPT 最初主要被视为文本生成工具，而 GPT-3 的推出成为该分支发展的历史性时刻。自 GPT-3 问世后，不断涌现出诸多如 ChatGPT、PaLM、GPT-4 等优秀的大模型，Decoder-Only 分支现发展势头强劲。

1. 大模型带来的效率与准确度革命

GPT 及其他大模型为当今的生产效率带来了前所未有的革命性提升。传统

[11] See Jinfeng Yang et al.,*Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond*, https://arxiv.org/pdf/2304.13712.pdf.

上，数据处理、内容生成、决策支持等任务都需要大量人力支持，且伴随着可能的人为错误和效率不高等问题。然而，大模型通过其强大的计算能力和广泛的知识基础，使得这些任务在短时间内得以高效完成。无论是企业内部的行政管理、市场分析，还是产品设计、客户服务，大模型都能够提供快速、准确且高质量的输出。这种技术驱动的生产效率革命不仅大幅度减少企业的运营成本，也为新商业模式和新机遇创造可能性。

大模型的出现也标志着信息处理和知识推断的准确性革命。大模型代表了可以更深入、更广泛地理解和处理人类语言的能力，使得很多任务的执行准确性得到前所未有的提高。大模型背后的深度学习算法使得系统能够从大量数据中提取规律和关系。与此同时，模型的庞大规模意味着它们能够记忆和处理的细节越来越丰富，这确保了其在诸如文本解析、情感分析和复杂问题回答等任务中的出色表现。传统的机器学习模型通常需要针对特定任务进行训练，而GPT 之类的模型由于其通用性，可以被微调以适应特定的领域或任务，从而在医学、法律、工程等专业领域中展现出惊人的准确性。在机器翻译、图像识别等许多应用场景中，大模型相较过去错误率显著降低，准确性的提高对于如医疗诊断和自动驾驶汽车等关键领域具有特殊重要性。

2. 大模型带来的机会与挑战

大模型当前已经覆盖了许多领域，为我们的日常生活、工作和娱乐带来了深刻的变革。例如，在零售业，大模型能够根据消费者的购买记录和浏览习惯为其生成个性化的购物推荐；在新闻和媒体领域，它可以快速地为记者生成初稿或摘要，加速新闻的传播速度；在娱乐领域，音乐、艺术和电影制作人开始尝试利用 AI 生成原创作品。同时，大模型在医疗、金融和交通领域的应用也都在逐步展开，为我们的健康、财富和出行安全提供了前所未有的支持。例如：

- **医药行业**：在药物研发领域，传统方法需要合成大量化合物，并且研发时间长、成本高，大模型的引入大大加快了药物的研发速度，其中以蛋白质结构预测为典型。例如，生物技术公司安进使用NVIDIA的BioNe-

Mo模型，显著减少了分子筛选和优化的时间。

- **金融服务**：金融服务行业正在经历技术驱动的数字转型，其中大模型在客户服务、营销优化、投资指导、风控与反欺诈等环节扮演重要角色。例如，Financial Transformer能够理解非结构化的金融数据，对市场深度分析、投资决策提供支持。

- **零售行业**：零售商正使用大模型以提升客户体验，实现动态化定价、细分客户、设计个性化推荐以及可视化搜索。例如，生成式AI会使用包含产品属性的元标签以生成更加全面的产品描述，包括“低糖”、“无麸质”等术语。

- **高等教育**：智能辅导系统、自动化论文评分以及各学科相关的大语言模型已经陆续在各大高校得到应用。例如，佛罗里达大学的研究人员使用超级计算机开发了一种自然语言处理模型，使计算机能够读取和解释存储在电子健康记录临床笔记中的医学语言，甚至实现自动绘制图表。此外，基因组学大语言模型等专业大模型也已经有落地案例。

- **公共服务**：政府机构人员可以使用生成式AI提高日常工作的效率，大模型的分析能力能够帮助其处理文件，加快办事效率。由大语言模型驱动的AI虚拟助手和聊天机器人可以即时向在线用户提供相关信息，减轻电话接线员的压力。

然而，这些应用也带来了诸多争议。例如，数据隐私是公众最大的关切之一，原因是生成式 AI 的许多应用都依赖于大量的个人数据。大模型内容生成也可能会模糊真实和虚构的界限，从而引发道德和法律上的困境。大模型的透明性和公正性也是广大公众、企业和政府关心的焦点。在数据收集、处理到跨境传输的全过程中，每一个阶段都存在特定风险，如侵犯隐私、泄露商业秘密或跨境数据违规流通等。另外，随着人们对大模型的使用频次逐渐增加，可能出现人们对大模型过于依赖而不再进行批判性思考的现象，从而引发人们对于自身思维能力倒退、价值创造能力降低的担忧。

二、全球大模型监管现状

(一)主要国家和地区加快完善大模型监管

2023 年 11 月 1 日，首届人工智能安全全球峰会在布莱切利园正式开幕，会上包括中国、美国、欧盟、英国在内的二十余个主要国家和地区共同签署了《布莱切利宣言》(The Bletchley Declaration)[12]，承诺以安全可靠、以人为本、可信赖及负责的方式设计、开发、部署并使用 AI。《布莱切利宣言》肯定了 AI 的广泛应用前景，同时指出了 AI(尤其是包括大模型在内的前沿高功能通用 AI 模型) 在包括网络安全和生物技术等领域所可能造成的风险，以及需要解决的包括保护人权、透明度和可解释性、公平性、问责制、监管、人类监督与控制、歧视与偏见、隐私与数据保护、合成欺骗性内容、AI 滥用等问题，并确认 AI 开发者需要对该等风险及问题承担重大责任。各国家和地区共同承诺在国际层面识别共同关注的前沿 AI 安全风险，并承诺在各国家和地区制定各自的基于风险的政策。最后，《布莱切利宣言》表达了支持建立一个具有国际包容性的前沿 AI 安全科学研究网络的决心。《布莱切利宣言》作为目前全球针对 AI 监管的前沿文件，显示了全球对于 AI 发展的密切关注。

目前，就欧盟、美国及英国而言，其均将大模型作为人工智能的一部分进行监管，因此，对于境外大模型的监管现状的梳理，需要与整体人工智能监管现状相结合。

1. 欧盟

(1) 立法现状

2016 年 10 月，欧盟议会法律事务委员会颁布《欧盟机器人民事法律规则》(European Civil Law Rules in Robotics)[13]，正式揭开了欧盟人工智能与大模型合规监管的立法篇章。此后，欧盟陆续颁布了与人工智能和大模型合规监管密切相关的一系列法案及政策，其中尤以《人工智能法案》(Artificial Intelli-

[12] https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023，最后访问于 2023 年 11 月 22 日。
[13] https://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_EN.pdf，最后访问于 2023 年 11 月 22 日。

gence Act)[14] 最值得注意。目前，《人工智能法案》已经进入最终谈判阶段，一经通过，其可能成为全球第一部专门针对人工智能进行综合性立法的区域性法规。总体来看，欧盟针对人工智能与大模型合规监管的政策采取了专项法案为主、现存法规为辅的结构，以《人工智能法案》作为治理核心，结合可能涉及的其他相关领域的立法 (包括数据及个人信息、市场监管等)，共同构成了包括大模型在内的人工智能监管体系。

(a)《人工智能法案》

2021 年 4 月，欧盟发布了《人工智能法案》的提案。2022 年，欧盟委员会综合各方意见，对《人工智能法案》进行了进一步修正。2023 年 6 月，《人工智能法案》再次修正，并经欧洲议会投票通过 (“《人工智能法案》”)[15]。按照欧盟立法程序，修正法案下一步将正式进入欧盟委员会、议会和成员国三方谈判协商的程序，并确定最终版本。

《人工智能法案》是欧盟首部有关人工智能的综合性立法，其以人工智能的概念作为体系原点，以人工智能的风险分级管理作为制度抓手，以人工智能产业链上的不同责任主体作为规范对象，以对人工智能的合格评估以及问责机制作为治理工具，从人工监管、隐私、透明度、安全、非歧视、环境友好等方面全方位监管人工智能的开发和使用，详细规定了人工智能市场中各参与者的义务，主要内容如下：

(i) 以人工智能 (Artificial Intelligence，“AI”) 概念为体系原点

根据《人工智能法案》，“AI 系统”是指一种以机器为基础的系统，该系统在设计上具有不同程度的自主性，可以为实现明确或隐含的目标生成如预测、建议或决策等的输出结果，对物理或虚拟环境造成影响。而“大模型”是指在广泛的数据上进行规模化训练的人工智能模型，其设计是为了实现输出的通用性，并能适用各种不同的任务。值得注意的是，较为狭窄、不普遍的、无法适应广泛任务的预训练模型不属于《人工智能法案》所规制的大模型。

[14] 详见下文第 1(1)(a) 段。
[15] https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.pdf，最后访问于 2023 年 11 月 22 日。

(ii) 以责任主体为规范对象

《人工智能法案》将 AI 系统的责任主体划分为提供方、部署方、进口方、分销商四种主要角色。其中，“提供方”指开发或拥有已经开发的 AI 系统，以自己的名义将其投放市场或在欧盟投入服务的自然人或法人；“部署方”指在欧盟境内在其权限范围内使用 AI 系统的自然人或法人 (不包括在个人非专业活动过程中使用)，包括使用 AI 系统以提供用户服务的商业机构等；“进口方”指在欧盟设立或者位于欧盟境内，并将带有欧盟境外自然人或法人名称或商标的 AI 系统投放到欧盟市场的自然人或法人；“分销商”指供应链中提供方和进口方之外的在欧盟市场中提供 AI 系统且不改变其系统属性的自然人或法人。

(iii) 风险分级标准

对于 AI 系统涉及的风险，欧盟主要区分为“不可接受的风险”、“高风险”、“有限风险”和“最小风险”四类，具体如下：

- 存在不可接受风险的AI系统。

  存在下列情况的AI系统均可能属于存在“不可接受的风险”的AI系统，欧盟成员国内将完全禁止该等AI系统投入市场或者交付使用：(1)采用潜意识技术或有目的的操纵或欺骗技术；(2)利用个人或社会群体的弱点(例如已知的人格特征或社会经济状况、年龄、身体精神能力)；(3)利用人的社会行为或人格特征进行社会评分；(4)在公众场所的“实时”(包括即时和短时延迟)远程生物识别系统。

- 高风险AI系统。

  存在下列情况的AI系统均属于存在“高风险”的AI系统，其投放市场及交付使用均受到严格的管控并需履行评估及备案要求：

  – AI系统同时满足下述两项条件：(1)属于欧盟统一立法规制范围内的产

品的安全组件或为该范围内的产品本身；并且(2)根据欧盟统一立法规制需要就健康或安全问题经过第三方合格评估方可投放市场或交付使用；

– AI系统同时满足下述两项条件：(1)存在可能损害环境或损害人类健康、安全、基本权利的重大风险；并且(2)符合分级标准且在规定的领域内使用，包括生物特征识别AI系统、关键基础设施AI系统、可能决定人的受教育或职业培训机会的AI系统、作为超大型在线社媒平台[16]拟在其用户内容推荐中使用的AI系统等。

- 有限风险AI系统。

  不属于存在不可接受的风险或高风险的AI系统，但需要履行一般合规要求，属于存在“有限风险”的AI系统，主要包括与人类互动的AI系统、用于情绪识别的AI系统、用于生物特征分类的AI系统以及生成深度合成内容的AI系统。

- 最小风险AI系统。

  在上述三种类型之外的AI系统，均属于存在“最小风险”的AI系统，主要包括允许自由使用AI的电子游戏、邮件过滤器等。

(iv) 风险分级监管

对于前述不同的风险等级，《人工智能法案》采取了不同程度的监管措施，具体包括：

- 对于存在不可接受风险的AI系统，严厉禁止使用；

- 对于高风险AI系统，要求其同时履行：(1)高风险AI系统的特殊合规要求(“**特殊合规要求**”)；以及(2)AI系统的一般合规要求(“**一般合规要**

[16] 指根据欧盟第 2022/2065 号法规第 33 条的规定的超大型在线平台的社交媒体平台，主要为用户数量超过 4500 万的社交媒体平台。

求”)。其中，特殊合规要求主要包括内部合规及外部认证措施：

- 内部合规措施须贯穿系统全生命周期，包括：(1)形成风险管理体系；(2)实施数据治理；(3)形成技术文档；(4)自动记录运行日志；(5)保证透明度；(6)保证人工监督；(7)保证系统的准确性、稳健性和网络安全性。

- 外部认证措施均应当于上市前完成，包括：(1)根据系统功能不同，进行自评估或者第三方评估；(2)在欧盟公共高风险AI系统数据库中备案；(3)使用CE(Conformity European，“CE”)标识。

- 对于有限风险AI系统，履行一般合规要求即可。就一般合规要求而言，主要为透明度要求，具体要求根据AI系统的不同类型而有所区分：

  - 针对与人类互动的AI系统，系统使用者需要告知人类其正在与AI系统进行互动；

  - 针对情绪识别及生物特征分类AI系统，系统使用者需要告知系统识别对象上述系统的存在，并且需要就生物识别数据的获取取得系统识别对象的同意；

  - 针对生成深度合成内容的AI系统，系统使用者需要对外告知该等内容是由AI生成或操纵的，而并非真实内容。

- 对于最小风险AI系统，不作强行性干预。

(v) 各类责任主体的义务

总体而言，提供方是 AI 系统的最终负责人，其需履行的义务最为全面，责任承担亦为最重，部署方需履行风险防范义务，其他参与者需履行以审查义务为核心的一系列的合规义务，具体如下：

- 提供方的义务主要包括：(1)执行前述所有特殊合规要求；(2)在系统上标明其名称、注册名称或注册商标，以及其联系信息；(3)确保执行人工监督的人员精通自动化或者算法偏见的风险；(4)执行数据保护，包括数据保护影响评估并发布摘要，以及提供输入数据或所使用的数据集的任何其他相关信息的说明；(5)建立书面质量管理体系；(6)日志及文档保存；(7)对不当行为采取纠正措施并告知有关机构；(8)提交欧盟合格声明，并在系统上市后由国家监督机构和国家主管部门保管；(9)境外提供方应在欧盟境内设置代表(“**授权代表**”)，以全权履行《人工智能法案》项下提供方的义务并配合主管机构的工作。

- 部署方的义务主要包括：(1)监督与风险控制；(2)数据保护；(3)履行备案，作为欧盟公共当局或者欧盟机构、团体(“**公共当局**”)的部署方或者属于《数字市场法案》[17]守门人的部署方，需要在使用系统前在欧盟公共高风险AI系统数据库中备案，其余高风险AI系统的部署方可自愿备案；(4)履行高风险AI系统的基本权利影响评估，以确定系统在使用环境中的影响。部署方为公共当局的，应公布评估的结果摘要，作为上述备案的一部分。

- 进口方主要义务包括：确保AI系统提供方履行了自评估或第三方评估义务、形成技术文档义务、授权代表任命义务(如需)，并确保AI系统带有CE标识，附有所需的说明文件。

(vi) 大模型的特殊合规义务

大模型的提供方在大模型上市前，应确保该模型符合下述要求：(1) 以适当的方法识别、减少重大风险，并记录剩余的不可缓解的风险；(2) 只纳入经过适当的大模型数据治理措施的数据集，且须审查数据来源的适当性和可能的偏差以及缓解措施；(3) 在设计和开发期间进行测试及评估，以在其整个生命周期内达到适当的性能、可预测性、可解释性、可纠正性、安全性和网络安全水平；(4) 减少能耗及浪费，提高整体效率，具有测量和记录能耗以及可能产生的其他

[17] 详见下文第 1(1)(b) 段。

环境影响的能力；(5) 制定技术文件和使用说明；(6) 建立质量管理系统，以记录对上述义务的遵守；(7) 在欧盟数据库中备案该大模型；(8) 在其大模型投放市场或投入使用后的 10 年内，将技术文件交由国家主管部门保存。

(b) 数据隐私、算法及知识产权相关法律法规

针对大模型及其所服务的 AI 系统所涉及的数据、个人信息、算法以及知识产权等领域，欧盟现有的相关规定在各自适用的范围内实际上起到了垂直监管的作用。该等垂直监管类的主要规定如下：

2018 年 5 月，欧盟委员会的《通用数据保护条例》(General Data Protection Regulation，“**GDPR**”)[18] 生效。GDPR 从数据控制者和处理者的责任以及数据监管等方面重新调整了欧盟个人数据保护策略。另外，GDPR 关于透明度的原则以及自动化决策有关的规定也为算法设计者设置了相关义务，包括确保算法训练数据痕迹可查义务以保证算法训练数据真实、对算法部分技术原理进行阐释义务以保证算法目标服务人群充分了解情况，以及算法的非歧视机制等。

2022 年 10 月，欧盟委员会颁布了《数字服务法案》(Digital Service Act，“**DSA**”)[19]，其适用对象为数字服务供应商。DSA 将适用对象划分为管道服务商、缓存服务商、托管服务商、在线平台及在线搜索引擎，并特别定义了超大型在线平台 (Very Large Online Platform，“**VLOP**”) 和超大型在线搜索引擎 (Very Large Online Search Engines，“**VLOSE**”)。上述主体各自承担不同的合规义务，其中 VLOP 及 VLOSE 承担的合规义务最重。DSA 的立法宗旨为加强网络平台的内容审查义务、非法商家打击义务、信息透明义务 (例如需向消费者明确透传算法推荐及定向广告内容)，帮助建立透明、安全、可预测、可信任的网络环境，保护网络平台用户的权益。

2022 年 11 月，欧盟委员会颁布了《数字市场法案》(Digital Market Act，“**DMA**”)[20]，引入“守门人”这一概念，对从事在线中介服务 (如应用商店)、

[18] http://data.europa.eu/eli/reg/2016/679/2016-05-04，最后访问于 2023 年 11 月 22 日。
[19] http://data.europa.eu/eli/reg/2022/2065/oj，最后访问于 2023 年 11 月 22 日。
[20] http://data.europa.eu/eli/reg/2022/1925/oj，最后访问于 2023 年 11 月 22 日。

在线搜索引擎、社交网络服务、即时通讯服务、视频共享平台服务、虚拟助手、网页浏览器、云计算服务、操作系统、在线市场和广告服务等服务的符合标准的大型互联网平台进行反垄断合规监管。DMA 借助行为清单工具，明确列举了守门人“必须为”和“禁止为”的内容，旨在维护数据开放，保护个人数据、禁止守门人滥用优势地位进行不正当竞争，确保数字市场的公平竞争和良性发展。

2019 年 3 月，欧盟议会通过了《数字化单一市场版权指令》(Directive on Copyright in the Digital Single Market，“**《版权指令》**”)[21]。《版权指令》规定，基于科学研究与数据分析两种目的，并且作品为合法获取的情形下的数据挖掘 (Text Data Mining，“**TDM**”) 具有正当性。显然，大模型的开发者进行的 TDM 通常并不属于科学研究范畴，而更可能属于以数据分析为目的的 TDM。《版权指令》第 4 条为大模型在数据训练阶段对版权客体的复制、提取行为设置了合理使用的例外，该等例外实际上赋予了 TDM 在数据处理阶段复制、提取数据行为的合法性，且不存在主体限制或使用技术目的限制，换言之，即使是出于商业性使用目的也同样适用。

总体而言，GDPR 适用于 AI 采集和使用个人数据等场景，DMA 和 DSA 以透明度和公平性为核心，对数字平台服务的提供方分别提出监管要求，《版权指令》则对大模型训练数据的获取合法性进行了规定。而在《人工智能法案》即将通过的大背景下，法案中所提及的大模型系统及其所嵌入的 AI 系统的提供方、部署方、进口方、分销商等角色是否以及如何适用于该等垂直监管类的规定，《人工智能法案》如何处理与现有的各垂直监管法规的法条竞合、冲突与协调适用等问题，人工智能的监管部门与其他各垂直监管法规的监管部门的管辖权如何划分以及与各类组织机构间协调运作，都需要通过实践来回答。

(2) 相关案例

实际上，意大利、法国、西班牙已经对 OpenAI 展开了调查[22]。在意大

[21] https://eur-lex.europa.eu/eli/dir/2019/790/oj，最后访问于 2023 年 11 月 22 日。

[22] https://www.politico.eu/article/chatgpt-italy-lift-ban-garante-privacy-gdpr-openai/，最后访问于 2023 年 11 月 22 日；https://www.zdnet.fr/actualites/chatgpt-les-premieres-plaintes-francaises-enregistrees-par-la-cnil-39956702.htm，最后访问于 2023 年 11 月 22 日。

利，2023 年 3 月，意大利个人数据保护局 (Garante per la Protezione dei Dati Personali，“**GPDP**”) 宣布禁止使用 ChatGPT，并限制开发这一平台的 OpenAI 公司处理意大利用户信息，同时对 OpenAI 公司展开立案调查，理由是 ChatGPT 平台存在用户对话数据和付款服务支付信息丢失的情况，而且没有就收集处理用户信息进行告知，缺乏大量收集和存储个人信息的法律依据。此外，ChatGPT 没有有效的年龄核实系统，可能会让未成年人接触到不适当的内容。4 月 12 日，GPDP 列出一份清单，要求 OpenAI 在 4 月底前满足包括透明度、数据纠正及被遗忘权、个人数据保护、未成年人保护等一系列要求。4 月 28 日，ChatGPT 在完成整改后重新在意大利上线。在西班牙，2023 年 4 月 13 日，西班牙国家数据保护局发表声明，因 ChatGPT “可能不符合 GDPR 规范”而对 OpenAI 启动初步调查程序。在法国，2023 年 4 月，法国数据监管机构国家信息与自由委员会 (Commission Nationale de l’informatique et des libertés，“**CNIL**”) 对 ChatGPT 提出违反 GDPR、涉嫌侵犯个人隐私、捏造不实信息等数项指控，并展开调查。

2. 美国

(1) 立法现状

相较于欧盟的统一协调、垂直跨部门的体系化立法而言，美国对于大模型及其所服务的 AI 系统的立法总体而言仍呈现较为保守、零散、地区化的态势。在州一级层面，各州的立法进程相差较大，较为积极的例如伊利诺伊州、加利福尼亚州、弗吉尼亚州、纽约州等已经通过了相关法案，但侧重点各有不同，例如主要针对人工智能视频面试 [23] 及职场自动化决策 [24]、人工智能产业促进 [25] 等。在联邦层面，目前为止，美国尚未通过一部完整且专门针对大模型及其所服务的 AI 系统的法案，而是试图通过调整政府机构的权力，在现有的立法框架及监管规则内对大模型及人工智能进行规制，但由于政府机构多元，机构之间的执行程度与政策发展也并不平衡。目前，联邦层面的合规重点主要涉及 AI 安

[23] https://ilga.gov/legislation/publicacts/fulltext.asp?Name=101-0260&GA=101，最后访问于 2023 年 11 月 22 日。

[24] https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=4344524&GUID=B051915D-A9AC-451E-81F8-6596032FA3F9&Options=ID%7cText%7c&Search=，最后访问于 2023 年 11 月 22 日。

[25] http://alisondb.legislature.state.al.us/ALISON/SearchableInstruments/2019RS/PrintFiles/SJR71-int.pdf，最后访问于 2023 年 11 月 22 日。

全、算法透明度、反歧视、评估等要求。但随着 ChatGPT、Bard 等生成式人工智能的井喷式出现，目前一系列与人工智能的联邦立法提案也已经出现在了国会中。同时，联邦政府机构也在积极制定相关政策，加紧对于 AI 的体系化监管。

(a) 人工智能重点整体性法规政策

2020 年 11 月，美国行政管理和预算局 (Office of Management and Budget) 颁布了《人工智能应用监管指南》(Guidance for Regulation of Artificial Intelligence Applications)[26]，反映了美国在人工智能监管方面的核心立场。该指南并未直接规定人工智能的监管法规，而是为美国政府提供了关于制定人工智能监管政策的指导方针。该指南主要关注了歧视、国家安全等问题，并提出了一系列风险评估和管理框架等要求，以提升人工智能的可信度和透明度，但其对人工智能仍持自由开放的基本态度，旨在确保监管规则不会妨碍人工智能的发展。

2020 年 12 月，时任美国总统特朗普签署了名为《促进联邦政府使用可信赖人工智能》(Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government, Executive Order 13960 of December 3, 2020)[27] 的行政命令，主要规定了联邦政府机构在考虑设计、开发、获取和在政府中使用人工智能时应遵循的一系列旨在促进公众信心、保护国家价值观并确保人工智能的合法使用的共同原则，包括：(a) 合法并尊重国家价值观。各机构在设计、开发、获取和使用人工智能时，应充分尊重国家价值观，并符合宪法及其他适用的法律和政策，包括涉及隐私、公民权利和公民自由的法律和政策；(b) 目的明确，效率主导。各机构应在风险可控情况下积极设计、开发、获取和使用有益的人工智能；(c) 准确性与有效性。各机构应确保其对人工智能的训练场景与应用场景一致，确保人工智能的可靠性；(d) 安全性与稳健性。各机构应确保其人工智能在面对系统漏洞和其他恶意攻击时的弹性；(e) 可理解性。各机构应确保其人工智能应用程序的操作和结果能够被相关专家和用户充分理解；(f) 可问责性和可追溯性。各机构应确保在设计、开发、采购和使用人工智能时，

[26] https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf，最后访问于 2023 年 11 月 22 日。

[27] https://www.federalregister.gov/documents/2020/12/08/2020-27065/promoting-the-use-of-trustworthy-artificial-intelligence-in-the-federal-government，最后访问于 2023 年 11 月 22 日。

明确界定、适当分配各主体的角色和责任。人工智能的设计、开发、获取和使用应酌情并在切实可行的范围内进行详细记录和追踪；(g) 定期监测。各机构应根据上述原则定期测试其系统并及时更新补正；(h) 透明。各机构应在实际可行的范围内，根据适用的法律和政策向适当的利益相关者披露其使用人工智能的相关信息；(i) 问责。各机构应负责实施和执行适当的保障措施，以确保其人工智能系统的正常使用和运行，并应监督记录该等保障措施的遵守情况，并应为所有负责设计、开发、采购和使用人工智能的人员提供适当的培训。

2021 年 1 月，经国会批准，《2020 国家人工智能倡议法案》(National AI Initiative Act of 2020 (DIVISION E, SEC. 5001)) 正式通过[28]，其中明确重申了确保美国在可信人工智能领域的领导地位。该法案的主要目的是确保美国在人工智能研发方面的领导地位，为社会各部门的人工智能技术整合准备充足劳动力，协调各联邦机构开展人工智能相关活动，保证信息多渠道流通。具体而言，该法案将：(1) 通过美国白宫科技政策办公室 (Office of Science and Technology Policy，“**OSTP**”) 管理的机构间协调委员会，制定人工智能研究领域的机构间协调战略规划；(2) 成立咨询委员会，该委员会将跟踪人工智能的科学研究现状，为机构间协调委员会提供信息；(3) 在美国国家科学基金会 (National Science Foundation，“**NSF**”) 的协调下，建立人工智能研究机构网络，该网络将促进学术界、政府部门、私人组织之间的合作，加快人工智能的研究；(4) 支持美国国家标准技术研究所 (National Institute of Standards and Technology，“**NIST**”) 研究制定人工智能评价标准，要求 NIST 创建数据共享的管理框架；(5) 支持 NSF 在人工智能相关领域开展多种研究，以优化人工智能系统，推进其他领域的科学研究；(6)NSF 将提供奖学金和培训来支持人工智能及相关领域的教育；(7) 支持能源部 (Department of Energy，“**DOE**”) 开展人工智能研究，利用 DOE 的基础设施来应对人工智能挑战、促进技术转移、实现与其他联邦机构间的数据共享及协同合作；(8) 进一步探究人工智能带来的机遇和挑战，探究保持美国在人工智能领域领先地位所需的计算资源。

2022 年 10 月，OSTP 颁布了《人工智能权利法案蓝图》(Blueprint for

[28] https://www.congress.gov/116/crpt/hrpt617/CRPT-116hrpt617.pdf#page=1210，最后访问于 2023 年 11 月 22 日。

an AI Bill of Right)[29]，主要内容包括前言、指导人工智能的设计、使用和部署的五项原则 (该等五项原则为：技术的安全性和有效性、防止算法歧视、保护数据隐私、告知及解释义务以及人类参与决策)、应用说明以及技术指南，该指南针对五项原则中的每一项均解释了原则的重要性、原则所指引的期望以及各级政府到各种规模的公司等多种组织为维护原则可以采取的具体的实施步骤、原则的实践案例。

2023 年 1 月，NIST 正式发布了《人工智能风险管理框架 (第一版)》(AI Risk Management Framework 1.0，“**AI RMF 1.0**”)[30]。AI RMF 1.0 是一个自愿性框架，基于经济合作与发展组织 (Organization for Economic Co-operation and Development) 的 AI 系统分类框架，旨在为设计、开发、部署和使用 AI 系统提供指南，以增强人工智能的可信度、降低风险，并提供关于如何在整个人工智能生命周期 (包括 AI 的应用背景和数据输入阶段 (AI 设计)、AI 模型构建阶段 (AI 开发)、AI 任务执行和输出阶段 (AI 部署)、AI 操作和监控阶段 (AI 监控)) 中管理风险的建议。

2023 年 4 月，美国参众两院共同发布了《确保人工智能安全、可靠、道德和稳定系统法》(草案)(Assuring Safe, Secure, and Ethical Systems for AI Act，“**ASSESS AI Act**”)(Draft)[31]。该法案将设立一个人工智能工作组，以评估联邦政府在 AI 政策和使用方面的现有政策、监管现状、法律空白，并提出具体建议。具体而言，该人工智能工作组的成员将包括美国司法部长、NIST 和 OSTP 的负责人，以及来自工业界、学术界和非营利组织的代表。该人工智能工作组将针对保护隐私、公民自由和公民权利的政策，面部识别和生物特征识别的联邦标准，AI 审计和风险评估的要求等内容提出建议，并且将在成立后的 18 个月内向国会和总统提交最终报告。

2023 年 6 月，美国参众两院共同发布了《国家人工智能委员会法》(草案) (National AI Commission Act)(Draft)[32]。该法案将设立一个由来自不同领域

[29] https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf，最后访问于 2023 年 11 月 22 日。
[30] https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf，最后访问于 2023 年 11 月 22 日。
[31] https://www.congress.gov/bill/118th-congress/senate-bill/1356，最后访问于 2023 年 11 月 22 日。
[32] https://www.congress.gov/bill/118th-congress/house-bill/4223，最后访问于 2023 年 11 月 22 日。

的 20 名专家组成的委员会，并指示该委员会制定 AI 立法框架，该法案目前正在国会审议中。该法案本身并不是 AI 的监管框架，而是寻求建立一个国家人工智能委员会，即一个位于立法部门的独立机构，负责制定 AI 综合监管提案。该委员会的职责在于确保美国实现与 AI 相关的三个主要目标，包括：减轻与 AI 相关的风险和潜在危害、保护美国在 AI 研发领域的领先地位、建立 AI 保障机制，确保 AI 系统符合美国价值观。

2023 年 10 月，美国总统拜登签署了《关于安全、可靠和值得信赖的人工智能的行政命令》(Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence)[33]，该命令主要围绕 AI 发展的八项原则展开，并针对每项原则向特定政府机构及官员提出了详细的要求。上述八项原则具体包括：

(一) 安全与保障原则，即应采取措施保证 AI 是安全且可靠的。为达成此目的，NIST 应与商务部合作: (1) 制定指导方针，以提供确保 AI 安全可靠的指南、标准及最佳实践；以及 (2) 收集美国境内的，或美国企业拟收购的拥有或具备可能开发大规模算力的潜力的公司相关数据，确保 AI 的安全可靠性，包括：(a) 管理关键基础设施和网络安全中的 AI；(b) 降低 AI 与化学、生物、放射和核威胁交叉的风险；(c) 减少 AI 合成内容带来的风险，促进识别和标记由 AI 系统产生的合成内容的能力，并确定由联邦政府或其代表生产的合成和非合成数字内容的真实性和来源；(d) 促进 AI 培训联邦数据的安全发布和防止恶意使用；(e) 指导形成国家安全备忘录。国家安全事务总统助理和总统助理兼政策副幕僚长应监督机构间流程，并向总统提交一份拟议的 AI 国家安全备忘录。该备忘录应涉及作为国家安全系统组成部分的、或用于军事和情报目的的 AI 的治理。备忘录应概述国防部、国务院、其他相关机构和情报系统应对 AI 带来的国家安全风险 (例如内部人员风险和外部攻击风险) 和潜在利益的行动。

(二) 促进创新及竞争原则。美国应促进 AI，特别是半导体行业的的创新、竞争和合作，并保护 AI 知识产权，制止对关键资产和技术的非法串通和垄断。具体措施包括：(a) NSF 应：(i) 协调启动实施国家 AI 研究资源的试点项目；(ii)

[33] https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/，最后访问于 2023 年 11 月 22 日。

资助并启动 NSF 区域创新引擎，优先考虑 AI 相关工作，如 AI 相关研究、社会或劳动力需求；(iii) 在目前资助的 25 个国家 AI 研究机构的基础上，建立至少 4 个新的机构；(b) 能源部长应与 NSF 主任协调，建立一项试点计划，以加强现有的科学家培训计划，目标是到 2025 年培训 500 名新的研究人员；(c) 国家专利商标局应澄清与 AI 和可专利主体的发明人有关的问题；(d) 国土安全部长应领衔制定培训、分析和评估计划，以减轻 AI 相关 IP 风险，包括收集和分析与 AI 相关的 IP 盗窃报告，调查此类影响国家安全的事件，并采取执法行动；(e) 为推动广泛的医疗保健技术开发人员进行负责任的 AI 创新，以促进医疗保健部门患者和工作人员的福利，卫生与公共服务部部长应支持 AI 开发和使用，包括通过卫生与公共服务部的项目与适当的私营部门合作，支持 AI 工具的发展，为患者开发个性化的免疫反应档案、加速通过美国国立卫生研究院 AI/ 机器学习联盟促进健康公平和研究人员多样性 (AIM-AHEAD) 计划授予的拨款等。

为促进竞争，该命令还授权包括联邦贸易委员会在内的所有联邦机构，利用其权力促进 AI 和相关技术的竞争，包括采取措施制止非法勾结，防止占主导地位的公司的不正当竞争，并努力为小企业和企业家提供包括资金及贷款计划、专业设备、知识产权援助等。

(三) 保护劳动者权益，改善劳动环境原则。具体措施包括：(a) 为增进政府对 AI 对工人的影响的理解，(i) 经济顾问委员会主席应编写并向总统提交一份关于 AI 对劳动力市场影响的报告；(ii) 劳工部长应向总统提交一份报告，分析各机构针对因采用 AI 等技术进步而被取代的工人所能够采取相应措施的能力，包括联邦援助项目，加强 AI 教育与职业培训等；(b) 为帮助确保在工作场所部署的 AI 能够促进员工的福祉，劳工部长应为雇主制定并公布可用于减轻 AI 对员工福祉的潜在危害并最大化其潜在利益的原则和最佳实践；(c) 为培养多样化的 AI 劳动力，NSF 主任应优先考虑通过现有计划支持 AI 相关教育和 AI 相关劳动力发展，包括设立奖学金等。

(四) 促进公平及人权原则。具体措施包括：(a) 加强刑事司法系统中的 AI 和公民权利；(b) 保护与政府福利和项目有关的公民权利；(c) 在宏观的市场经

济中加强 AI 和公民权利，包括防止在招聘中使用 AI 造成的非法歧视，解决住房市场和消费者金融市场中对弱势群体的歧视，打击用于决定住房和其他房地产相关交易的自动化或算法工具 (例如租户筛选系统) 所包含的非法歧视，以及帮助确保残疾人从 AI 中受益，同时保护其免受风险。

(五) 消费者权益保护原则。具体措施包括：(a) 鼓励独立监管机构保护美国消费者免受欺诈、歧视和隐私威胁，并解决使用 AI 可能产生的其他风险，包括金融稳定风险，并考虑出台现有法规适用于 AI 的解释和指导，包括澄清受监管实体对其使用的任何第三方 AI 服务进行调查和监控的责任以及需履行的透明度义务；(b) 卫生与公共服务部部长应帮助确保 AI 在医疗保健、公共卫生和人类服务部门安全使用；(c) 交通部长应与相关机构协商，促进 AI 在交通运输部门的安全使用；(d) 为帮助确保 AI 在教育部门的负责任开发和部署，教育部长应制定有关 AI 资源分配的指导。这些资源应解决 AI 在教育中的安全、负责和非歧视使用问题，包括 AI 系统对弱势和服务不足社区的影响；(e) 鼓励联邦通信委员会考虑将 AI 用于改善通信网络，包括用于改善频谱管理、促进联邦与非联邦频谱运营商之间共享频谱、为使用包含 AI 的下一代技术 (包括 6G 和 Open RAN) 提高网络安全性、弹性和互操作性提供支持、阻止骚扰信息等。

(六) 隐私及公民自由保护原则。在开发和运营 AI 的过程中，必须确保数据的收集、使用和保留是合法、安全的，并能保护隐私。具体措施包括：(a) 行政管理和预算局局长应：(i) 评估并采取措施识别各机构采购的商业可用信息 (“**CAI**”)，特别是包含个人身份信息的 CAI；并且 (ii) 与联邦隐私委员会和机构间统计政策委员会协商，评估与包含个人身份信息的 CAI 的收集、处理、维护、使用、共享、传播有关的机构标准和程序，以便为各机构提供指导，说明如何减轻各机构与 CAI 有关的活动所带来的隐私风险；(b) NIST 应为各机构制定指导方针，以评估包括 AI 在内的差分隐私保证 (differential-privacy-guarantee，一种用来防范差分隐私攻击的隐私保护方法) 保护措施的有效性；(c) 促进与隐私增强技术 (Privacy-enhancing Technologies，PETs) 有关的研究、开发和实施。

（七）联邦政府 AI 风险管控原则。联邦政府应当管控使用 AI 的风险，并提高其内部监管、管理和支持负责任地使用 AI 的能力。

（八）确保联邦政府 AI 领导地位原则。美国应引领 AI 在全球的社会、经济和技术进步，包括与国际合作伙伴合作制定 AI 风险管理框架，并共同应对挑战。该命令还要求商务部长和国务卿就全球技术标准与主要国际伙伴合作，并提交一份关于全球参与计划的报告。

(b) 数据及算法技术合规

2022 年 2 月，美国众议院发布了《2022 年算法问责法案》（草案）(Algorithmic Accountability Act of 2022)(Draft)[34]，要求使用自动化决策系统做出关键决策的企业研究并报告这些系统对消费者的影响，其内容包括是否会因为消费者的种族、性别、年龄等生成对消费者有偏见或歧视性的自动决策等。该法案形成了“评估报告—评估简报—公开信息”三层信息披露机制。此外，联邦贸易委员会还将建立可公开访问的信息存储库，公开发布关于自动化决策系统的有限信息。

2022 年 6 月，美国参众两院共同发布了《美国数据隐私和保护法案》（草案）(the American Data Privacy and Protection Act，“**ADPPA**”)(Draft)[35]。ADPPA 规定，使用“覆盖算法”的大数据持有人，如果对个人或群体构成相应伤害风险，并单独或部分使用“覆盖算法”来收集、处理或传输覆盖数据，则应当根据 ADPPA 规定的评估标准进行隐私影响评估。ADPPA 将“覆盖算法”定义为：“使用机器学习、自然语言处理、人工智能技术或其他类似或更复杂的计算处理技术，并就涵盖数据做出决策或促进人类决策的计算过程”。人工智能大模型为深度学习模型，需要大规模的数据集，这些数据集很可能涵盖个人信息、数据与隐私。因此，可能构成使用“覆盖算法”，进而需要根据 ADPPA 规定的评估标准进行隐私影响评估。另外，ADPPA 还对隐私政策的告知与退出机制、反偏见等内容做出了规定。ADPPA 规定，企业或代表企业的服务提供商需要告知个人有“选择退出”的选择，即拒绝企业对其个人数据的收集、处理或传输。

[34] https://www.congress.gov/bill/117th-congress/senate-bill/3572，最后访问于 2023 年 11 月 22 日。
[35] https://www.congress.gov/bill/117th-congress/house-bill/8152，最后访问于 2023 年 11 月 22 日。

2023 年 2 月，拜登总统签署了《关于通过联邦政府进一步促进种族平等和支持服务不足社区的行政命令》(Executive Order on Further Advancing Racial Equity and Support for Underserved Communities Through The Federal Government)[36]，规定人工智能大模型应避免由于大量输入训练数据中存在的对种族、性别、年龄、文化和残疾等的偏见而导致训练结果输出内容中存在偏见。联邦政府在设计、开发、获取和使用人工智能和自动化系统时，各机构应在符合适用法律的前提下，防止、纠正歧视和促进公平，包括保护公众免受算法歧视。

(c) 知识产权保护

在 2020 年 4 月，美国专利商标局 (United States Patent and Trademark Office) 判定，只有自然人才可以在专利申请中被指定为发明人，而生成式 AI 系统不可以 [37]。

2023 年 3 月，美国版权局 (United States Copyright Office) 发布了《版权登记指南：包含人工智能生成材料的作品》(Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence)[38]。该指南明确，相关法律中使用的“作者”一词不包括非人类；人工智能生成的内容应该明确地被排除在版权登记之外。版权局强调，人类在多大程度上创造性地控制了作品的表达，并“实际形成”了作者身份是判断是否可以作为版权作品作者的关键因素。

(d) 生成内容合规

2019 年 6 月，美国众议院发布了《深度伪造责任法案》(草案)(Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019，“**DEEP FAKES Accountability**

[36] https://www.whitehouse.gov/briefing-room/presidential-actions/2023/02/16/executive-order-on-further-advancing-racial-equity-and-support-for-underserved-communities-through-the-federal-government/，最后访问于 2023 年 11 月 22 日。

[37] https://content.govdelivery.com/accounts/USPTO/bulletins/287fdc9#.Xqcts2pR6ZQ.email，最后访问于 2023 年 11 月 22 日。

[38] www.federalregister.gov/documents/2023/03/16/2023-05321/copyright-registration-guidance-works-containing-material-generated-by-artificial-intelligence，最后访问于 2023 年 11 月 22 日。

Act”)(Draft)[39]，其中规定，“深度伪造”一词系指任何录像、电影、录音、电子图像或照片，或者言论或行为的实质上衍生的任何技术表达，该等表达看似真实地描述了一个人的任何言论或行为，而该人事实上并未从事该等言论或行为，以及其制作实质上依赖于技术手段，而非他人在身体上或言语上模仿该人的能力；任何深度伪造制作者必须对其深度伪造记录有显著的披露，任何包含移动的视觉元素的深度伪造记录应当嵌入数字水印，以清楚地识别该记录是否包含改变的音频或视觉元素。

2020 年 11 月，美国众议院颁布了《识别生成对抗网络法案》(Identifying Outputs of Generative Adversarial Networks Act，“**IOGAN Act**”)[40]，指示 NSF 和 NIST 支持对深度伪造的研究。该法案要求 NSF 支持对操纵或合成内容和信息真实性的研究，支持必要的测量和标准开发研究，以加速技术的开发，检查生成对抗网络的功能和输出或其他合成或操纵内容的技术。

(2) 相关案例

2023 年 1 月 13 日，美国三名艺术家 Sarah Andersen、Kelly McKernan 和 Karla Ortiz 代表其他集体诉讼成员对 Stability AI Ltd.、Stability AI Inc.、Midjourney, Inc.、DeviantArt, Inc. 四名被告发起集体诉讼，指控四位被告所使用的生成式 AI 图片产品在未经用户同意下擅自爬取了数百万乃至数十亿张受著作权保护的图像的未经授权的副本用于训练模型和生成 AI 图片，其所生成的内容亦并未包含原告的著作权信息，进而侵犯了原告的版权。该案件中，争议焦点主要在于：(1) 生成式人工智能生成的内容是否侵犯了原告的版权；(2) 被告未经原告许可而删除和修改其作品的著作权管理信息是否侵犯原告版权。2023 年 7 月，在美国加利福尼亚州北区地方法院举行的关于被告驳回动议的听证会上，法院表达了对原告的核心责任论述的严重怀疑，认为原告未能提出可靠的依据来证明生成式人工智能生成的内容与原告创作的作品间存在实质的相似或者侵权情

[39] https://www.congress.gov/bill/116th-congress/house-bill/3230，最后访问于 2023 年 11 月 22 日。
[40] https://www.congress.gov/bill/116th-congress/senate-bill/2904，最后访问于 2023 年 11 月 22 日。

况。[41] 某种程度上，这一案例揭示了大模型输出的一个典型的知识产权难题：输出结果阶段，著作权人想要证明其著作权作品数据与生成式人工智能生成作品之间存在因果关系的难度较大，只有在著作人确定人工智能生成作品与其爬取的著作人著作权作品数据之间相关联后，才可以确定有哪些作品的著作权被侵犯，进而维护自身权益。

3. 英国

(1) 立法现状

与美国类似，英国部分现存的不同类型的法律法规已经涵盖了对人工智能的规定，其中部分重点法律法规如下：

(a) 生成内容合规

2023 年 10 月，英国议会颁布了《在线安全法案》(Online Safety Act 2023)[42]。《在线安全法案》规定了一系列与互联网信息内容相关的安全规定，赋予英国议会权力来批准哪些信息属于“合法但有害”的内容，要求在线平台立即采取措施。该法案要求社交媒体平台、搜索引擎以及其他允许用户发布内容的应用程序和网站，承担保护儿童、打击非法活动，并维护其已声明的条款与条件的责任。

(b) 数据合规

《数据保护法 2018》(Data Protection Act 2018)[43] 是主要的英国数据保护法律之一。英国脱欧后，英国政府将 GDPR 和相关监管要求转化为英国的数据保护监管体系，即所谓的“英国 GDPR”——虽然有部分调整，但其有关数据控制者和处理者的权利和义务与欧盟 GDPR 基本相同。2022 年 7 月，《数据

[41] Sarah Anderson, et al. v. Stability AI LTD., et al. (2023/01/13), Case details：https://stablediffusionlitigation.com/pdf/00201/1-1-stable-diffusion-complaint.pdf，最后访问于 2023 年 11 月 22 日；https://cases.justia.com/federal/district-courts/california/candce/3:2023cv00201/407208/67/0.pdf?ts=1685964605，最后访问于 2023 年 11 月 22 日；https://storage.courtlistener.com/recap/gov.uscourts.cand.407208/gov.uscourts.cand.407208.92.0.pdf，最后访问于 2023 年 11 月 22 日。

[42] https://www.legislation.gov.uk/ukpga/2023/50/enacted，最后访问于 2023 年 11 月 22 日。

[43] https://www.legislation.gov.uk/ukpga/2018/12/enacted，最后访问于 2023 年 11 月 22 日。

保护和数字信息法案》(Data Protection and Digital Information Bill)[44] 首次被提交至英国议会讨论，后经撤回修改，于 2023 年 5 月形成《数据保护和数字信息法案 (2 号)》(Data Protection and Digital Information (No.2) Bill)[45] 并再次提交至英国议会讨论，目前处于三读前的报告阶段。其中，针对自动化决策所涉及的个人数据，法案规定完整或部分基于特殊类别的个人数据的重大决策不得仅仅基于自动化决策做出，除非符合以下条件之一：(1) 该决策完全基于数据主体明确同意的个人数据处理；(2) 该决策是为订立或履行数据主体与控制者之间的合同所必需的，或法律要求或授权的。该法案同时也规定了自动化决策的保障措施，需由以下措施组成：(1) 向数据主体提供就数据主体作出的决策的信息；(2) 使数据主体能够就该等决策作出陈述；(3) 使数据主体能够就此类决策获得控制者的人为干预；(4) 使数据主体能够对该等决策提出异议。法案还规定了数据最小化原则、个人数据的访问和控制权、风险评估及合规检查等条款，以帮助企业更好地履行合规义务。

(c) 知识产权保护

2022 年 6 月，英国知识产权局 (UK Intellectual Property Office，“**UKIPO**”) 公布了《文本与数据挖掘版权例外改革提案》(Artificial Intelligence and Intellectual Property: copyright and patents: Government response to consultation)[46]。对于文本和数据挖掘，该提案计划引入一个新的版权和数据库例外，允许文本和数据挖掘用于任何目的，包括商业目的；版权所有者仍将拥有保护其内容的保障措施，包括要求合法访问。该提案使得任何文本和数据挖掘都无需向权利人支付许可费。目前该提案仍在审核之中。

(d) 算法技术合规

2021 年 5 月，英国中央数字与数据办公室、人工智能办公室与内阁办公室联合发布了《自动决策系统的伦理、透明度与责任框架》(Ethics, Transpar-

[44] https://bills.parliament.uk/bills/3322，最后访问于 2023 年 11 月 22 日。

[45] https://bills.parliament.uk/bills/3430，最后访问于 2023 年 11 月 22 日。

[46] https://www.gov.uk/government/consultations/artificial-intelligence-and-ip-copyright-and-patents/outcome/artificial-intelligence-and-intellectual-property-copyright-and-patents-government-response-to-consultation，最后访问于 2023 年 11 月 22 日。

ency and Accountability Framework for Automated Decision-Making，“**ETAF**”)[47]。ETAF 强调，算法和自动化决策在上线之前应该进行严格的、受控的和分阶段的测试。在整个原型和测试过程中，需要人类的专业知识和监督来确保技术上的弹性和安全，以及准确和可靠的系统。测试时，需要考虑自动化决策系统的准确性、安全性、可靠性、公平性和可解释性。ETAF 规定，企业必须对算法或自动决策系统做一个平等影响评估，使用高质量和多样化的数据集，发现和抵制所使用数据中明显的偏见和歧视。ETAF 指出，算法或计算机系统应该被设计为完全可以负责和可被审计的，算法和自动化的责任和问责制度应该明确。

(2) 相关案例

目前英国的司法实践中，对于专利的发明人是否只能为自然人存在激烈的讨论。2018 年 10 月 17 日和 2018 年 11 月 7 日，Stephen Thaler 先后分别向 UKIPO 提出两项发明专利申请，并将其创造并拥有的人工智能机器“DABUS”作为专利申请中的发明人，理由是两项发明均由“DABUS”在没有传统人类发明人帮助下创造完成。2019 年 12 月，UKIPO 驳回以“DABUS”作为发明人的专利申请，理由是“DABUS”为非自然人，不属于专利法中规定的发明人。Stephen Thaler 不服该决定，并接连上诉到英国最高法院。英国最高法院于 2023 年 3 月 2 日开始审理本案，目前案件还在审理中。[48] 该案的争议焦点在于，《英国 1977 专利法案》第 13(2)(a) 条是否要求专利申请中的发明人只能为自然人，包括申请人认为发明是由人工智能在没有传统人类发明人帮助下创造的情况；是否可以在没有指定人类发明人的情况下授予专利权；如果是人工智能创造的发明，那么该人工智能的所有者、创造者和使用者是否可以被授予专利权。该案的判决将为“AI 能否被认定为发明人”这一难题在英国的解决提供指引，同样对 AI 大模型领域的研究与发展至关重要。

[47] https://www.gov.uk/government/publications/ethics-transparency-and-accountability-framework-for-automated-decision-making/ethics-transparency-and-accountability-framework-for-automated-decision-making，最后访问于 2023 年 11 月 22 日。

[48] 2021/0201：Thaler (Appellant) v Comptroller-General of Patents, Designs and Trademarks (Respondent), Case details：https://www.supremecourt.uk/cases/uksc-2021-0201.html，最后访问于 2023 年 11 月 22 日；England and Wales Court of Appeal (Civil Division) Decisions：Thaler v Comptroller General of Patents Trade Marks And Designs [2021] EWCA Civ 1374 (21 September 2021) , Case details：https://www.bailii.org/ew/cases/EWCA/Civ/2021/1374.html，最后访问于 2023 年 11 月 22 日。

(二)我国对于大模型的监管现状

1. 立法现状

我国对大模型的监管主要是围绕网络安全、数据安全、个人信息展开，相关法律法规也以《中华人民共和国网络安全法》、《中华人民共和国数据安全法》和《中华人民共和国个人信息保护法》为主，同时，《中华人民共和国科学技术进步法》、《互联网信息服务管理办法》等法律法规亦针对互联网信息服务层面的合规制定了相关规范。随着产业的发展，我国的监管法律体系从该等方面不断深化拓展至算法服务、深度合成服务等与大模型密切相关的领域，《互联网信息服务算法推荐管理规定》、《互联网信息服务深度合成管理规定》等规定陆续出台。2023 年 8 月 15 日，我国针对生成式人工智能服务领域制定的首部法律法规《生成式人工智能服务管理暂行办法》(“**《AIGC 暂行办法》**”)生效，这是我国在人工智能监管领域不断探索完善的重要成果，明确了提供和使用生成式人工智能服务的总体要求，并对生成式人工智能服务提出了分类分级的监管要求，一定程度上标志着我国生成式人工智能服务领域进入强监管和高合规标准的新阶段。

与此同时，《人脸识别技术应用安全管理规定 (试行)(征求意见稿)》[49] 等与大模型领域密切相关的法律法规和相关规定正在制定过程中。在大模型的浪潮下，各机构、行业也积极响应，陆续发布了一系列大模型开发、运营相关的行业规范，如中国信息通信研究院 (“**中国信通院**”) 联合产学研各界制定的《可信大模型标准体系 2.0》、同济大学上海市人工智能社会治理协同创新中心研究团队编制的《人工智能大模型伦理规范操作指引》、华东师范大学和上海人工智能实验室联合两院院士、高校校长、知名专家学者共同制定发布的《教育通用人工智能大模型系列标准》等等。

值得关注的是，《人工智能法》已列入《国务院 2023 年立法工作计划》，《人工智能法 (草案)》预备提请全国人大常委会审议。可以说，我国正在推动全国层面的人工智能专门立法。不过，根据流程，《人工智能法 (草案)》将由

[49] 于 2023 年 8 月 8 日发布征求意见稿，但暂未生效。

国务院相关部门起草，然后经国务院常务会议审议并通过，继而才提请立法机关审议、表决，具体所需时间目前难以预计。

目前，我国和大模型相关的、已经生效的主要法律法规和相关规定，以及部分相对较有影响力的行业规范如下：

(1) 法律法规和相关规定

名称	颁发部门	生效时间
《中华人民共和国网络安全法》	全国人民代表大会常务委员会	2017.06.01
《中华人民共和国数据安全法》	全国人民代表大会常务委员会	2021.09.01
《中华人民共和国个人信息保护法》	全国人民代表大会常务委员会	2021.11.01
《中华人民共和国科学技术进步法》	全国人民代表大会常务委员会	2022.01.01
《互联网信息服务管理办法》	国务院	2000.09.25
《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》	国家互联网信息办公室，公安部	2018.11.30
《网络信息内容生态治理规定》	国家互联网信息办公室	2020.03.01
《关于加强互联网信息服务算法综合治理的指导意见》	国家互联网信息办公室，中央宣传部，教育部，科学技术部，工业和信息化部，公安部，文化和旅游部，国家市场监督管理总局，国家广播电视总局	2021.09.17
《互联网信息服务算法推荐管理规定》	国家互联网信息办公室，工业和信息化部，公安部，国家市场监督管理总局	2022.03.01

名称	颁发部门	生效时间
《关于支持建设新一代人工智能示范应用场景的通知》	科学技术部	2022.08.12
《互联网信息服务深度合成管理规定》	国家互联网信息办公室，工业和信息化部，公安部	2023.01.10
《生成式人工智能服务管理暂行办法》	国家互联网信息办公室，国家发展和改革委员会，教育部，科学技术部，工业和信息化部，公安部，国家广播电视总局	2023.08.15
《科技伦理审查办法 (试行)》	科学技术部，教育部，工业和信息化部，农业农村部，国家卫生健康委员会，中国科学院，中国工程院，中国科学技术协会，中国社会科学院，中央军委科学技术委员会	2023.12.01
《新一代人工智能发展规划》	国务院	2017.07.20
《关于调整发布 < 中国禁止出口限制出口技术目录 > 的公告》	商务部，科学技术部	2020.08.28
《网络安全标准实践指南—人工智能伦理安全风险防范指引》	全国信息安全标准化技术委员会	2021.01.05
《关于加强科技伦理治理的意见》	中共中央办公厅，国务院办公厅	2022.03.20
《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》	全国信息安全标准化技术委员会	2023.08.25

(2) 相关行业规范

名称	编制机构	发布时间
《新一代人工智能治理原则——发展负责任的人工智能》	国家新一代人工智能治理专业委员会	2019.06
《新一代人工智能伦理规范》	国家新一代人工智能治理专业委员会	2021.09
《可信大模型标准体系 2.0》	中国信息通信研究院等	2023.03
《人工智能伦理治理标准化指南》	国家人工智能标准化总体组等	2023.03
《人工智能大模型伦理规范操作指引》	同济大学等	2023.07
《教育通用人工智能大模型系列标准》	华东师范大学等	2023.07
《教育通用人工智能大模型标准体系研究报告》		
《可信 AI 技术和应用进展白皮书 (2023)》	中国信通院等	2023.07
《“弈衡”通用大模型评测体系白皮书》	中国移动研究院等	2023.07
《人工智能法示范法 1.0(专家建议稿)》	中国社会科学院法学研究所等	2023.08
《面向行业的大规模预训练模型技术和应用评估方法——金融大模型》	中国信息通信研究院等	制定中
《面向行业的大规模预训练模型技术和应用评估方法——汽车大模型》	中国信息通信研究院等	制定中

下文将对该等大模型领域的已经生效的主要法律法规和相关规定以及部分相对较有影响力的行业规范进行简单介绍。

(1) 主要法律法规

(a)《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》

2018 年 11 月 15 日，国家互联网信息办公室联合公安部发布《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》(“**《安全评估规定》**”)，该规定于 2018 年 11 月 30 日起正式施行。《安全评估规定》根据《中华人民共和国网络安全法》、《互联网信息服务管理办法》、《计算机信息网络国际联网安全保护管理办法》等有关法律、行政法规制定，明确了国家将加强对具有舆论属性或社会动员能力的互联网信息服务和相关新技术新应用的安全管理，规范互联网信息服务活动。根据《安全评估规定》，下述类型的互联网信息服务提供者需按《安全评估规定》自行进行安全评估：(i) 开办论坛、博客、微博客、聊天室、通讯群组、公众账号、短视频、网络直播、信息分享、小程序等信息服务或者附设相应功能；(ii) 开办提供公众舆论表达渠道或者具有发动社会公众从事特定活动能力的其他互联网信息服务。在此基础上，《安全评估规定》规定了互联网信息服务提供者应自行进行安全评估的具体情形。除进行自行安全评估的义务以外，《安全评估规定》还要求前述互联网信息服务提供者应履行消除安全隐患、形成安全评估报告、提交安全评估报告等各项义务。

(b)《互联网信息服务算法推荐管理规定》

2021 年 12 月 31 日，国家互联网信息办公室、中华人民共和国工业和信息化部、中华人民共和国公安部和国家市场监督管理总局联合发布《互联网信息服务算法推荐管理规定》(“**《算法推荐管理规定》**”)，该规定于 2022 年 3 月 1 日起施行。《算法推荐管理规定》的适用范围是在中华人民共和国境内应用算法推荐技术提供互联网信息服务的情形。《算法推荐管理规定》确立了算法分级分类安全管理的制度设计。其中分级分类关注的维度包括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。此外，《算法推荐管理规定》要求算法推荐服务提供者建立健全相关制度，例如算法推荐服务提供者应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。同时，算法推荐服务提供者应承担算法合规义务以及用户权益保护责任，保护用户的知情权和选择权。

(c)《互联网信息服务深度合成管理规定》

2022 年 11 月 25 日，国家互联网信息办公室、工业和信息化部、公安部联合发布《互联网信息服务深度合成管理规定》(“**《深度合成管理规定》**”)，该规定于 2023 年 1 月 10 日起施行。《深度合成管理规定》是我国第一部针对深度合成服务治理的专门性部门规章，主要针对应用生成合成类算法的互联网信息服务进行了规范，明确了生成合成类算法治理的对象和基本原则，强化了深度合成服务提供者和技术支持者的主体责任，并鼓励相关行业组织通过加强行业自律推动生成合成类算法的合规发展。《深度合成管理规定》适用于在中华人民共和国境内应用深度合成技术提供互联网信息服务的情形，深度合成服务提供者和技术支持者是主要的责任主体，二者均有义务进行算法备案，且均负有遵守数据和技术管理规范、加强训练数据管理、依法告知生物识别信息被编辑的个人、加强深度合成相关技术管理、依法开展安全评估等义务。此外，深度合成服务提供者还需承担信息安全主体责任和内容标识义务等，落实安全可控的技术保障措施，并制定和公开管理规则。

(d)《生成式人工智能服务管理暂行办法》

2023 年 7 月 10 日，国家网信办、国家发展改革委、教育部、科技部、工业和信息化部、公安部和广电总局联合发布《生成式人工智能服务管理暂行办法》，该办法于 2023 年 8 月 15 日起生效。根据《AIGC 暂行办法》规定，任何利用生成式人工智能技术为中国境内公众提供生成文本、图片、音频、视频等内容的服务都适用该办法。这意味着，境内外人工智能生成内容 (Artificial Intelligence Generated Content，“**AIGC**”) 服务提供者，无论其提供的服务是在模型层还是在应用层，亦无论是直接提供服务或通过 API 接口或其他方式间接提供服务，倘若其提供服务的对象是中国境内公众，都应当遵守《AIGC 暂行办法》。在监管机制与合规要求方面，《AIGC 暂行办法》对生成式人工智

能服务采取了包容审慎和分类分级的监管原则，要求生成式人工智能服务提供者在内容管理、训练数据、用户权益、安全评估等多个层面承担相应的责任。

(e)《网络安全标准实践指南——人工智能伦理安全风险防范指引》

2021 年 1 月，全国信息安全标准化技术委员会发布《网络安全标准实践指南——人工智能伦理安全风险防范指引》，将 AI 伦理安全风险总结为以下五大方面：(1) 失控性风险：AI 的行为与影响超出服务提供者预设、理解和可控的范围，对社会价值等产生负面影响；(2) 社会性风险：不合理使用 AI 而对社会价值等方面产生负面影响; (3) 侵权性风险: AI 对人的基本权利，包括人身、隐私、财产等造成侵害或产生负面影响；(4) 歧视性风险：AI 对人类特定群体具有主观或客观偏见，影响公平公正、造成权利侵害或负面影响；(5) 责任性风险：AI 相关各方行为失当、责任界定不清，对社会信任、社会价值等方面产生负面影响。

(f)《关于加强科技伦理治理的意见》

2022 年 3 月，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》，提出“科技伦理是开展科学研究、技术开发等科技活动需要遵循的价值理念和行为规范，是促进科技事业健康发展的重要保障”，并明确了以下五大类科技伦理原则：增进人类福祉、尊重生命权利、坚持公平公正、合理控制风险和保持公开透明。

(g)《科技伦理审查办法 (试行)》

2023 年 10 月 8 日，科学技术部、教育部、工业和信息化部等多部门联合发布《科技伦理审查办法 (试行)》(“**《科技伦理审查办法》**”)，该办法对于几乎所有科技活动所涉及的科技伦理审查和监管做出了明确的规定，并将于 2023 年 12 月 1 日起正式实施。在审查主体方面，《科技伦理审查办法》明确要求从事生命科学、医学、人工智能等科技活动的单位，研究内容涉及科技伦理敏感领域的，应设立科技伦理 (审查) 委员会，其他有伦理审查需求的单位可根据实际情况设立科技伦理 (审查) 委员会。在审查程序方面，《科技伦理审查办法》将审

查程序依据科技活动伦理风险发生的可能性和严重、紧急程度划分为一般、简易和应急三类。在审查内容及标准方面，《科技伦理审查办法》针对所有科技活动规定了审查的重点内容和标准，以及针对涉及人类研究参与者以及数据和算法的科技活动就审查的重点内容和标准进行特殊规定。例如，就涉及数据和算法的科技活动而言，一方面，要求数据的收集、存储、加工、使用等处理活动以及研究开发数据新技术等符合国家数据安全和个人信息保护等有关规定，数据安全风险监测及应急处理方案得当；另一方面，要求算法、模型和系统的设计、实现、应用等遵守公平、公正、透明、可靠、可控等原则，符合国家有关要求，伦理风险评估审核和应急处置方案合理，用户权益保护措施全面得当。

(2) 主要行业规范

(a)《新一代人工智能伦理规范》

2021 年 9 月，我国国家新一代人工智能治理专业委员会发布《新一代人工智能伦理规范》，旨在将伦理道德融入人工智能全生命周期，促进公平、公正、和谐、安全，避免偏见、歧视、隐私和信息泄露等问题。《新一代人工智能伦理规范》的适用主体为从事人工智能管理、研发、供应、使用等相关活动的自然人、法人和其他相关机构。在此基础上，《新一代人工智能伦理规范》明确了人工智能的基本伦理规范，包括增进人类福祉、促进公平公正、保护隐私安全、确保可控可信、强化责任担当、提升伦理素养。同时，《新一代人工智能伦理规范》提出了一系列人工智能应用管理规范、研发规范、供应规范和使用规范。

(b)《可信大模型标准体系 2.0》[50]

为进一步促进我国大模型产业发展，中国信通院联合产学研各方于 2022 年 2 月起制定可信大模型标准体系，并于 2023 年 3 月正式发布《可信大模型标准体系 2.0》。《可信大模型标准体系 2.0》以 Model as a Service(“MaaS”)服务结果为核心，从模型开发、模型能力、模型运营、模型应用、安全可信共

[50] 原文文本尚未公开，相关介绍参见微信文章《一文读懂可信 AI 大模型标准体系》，链接：https://mp.weixin.qq.com/s/5XbCQtgWGt1WIB8GCycTQw，最后访问于 2023 年 11 月 22 日。

五个方向构建大模型标准体系，以有效助力相关主体快速构建能力全面、应用广泛、运营便捷、安全可信的基础大模型。

(c)《人工智能大模型伦理规范操作指引》

2023 年 7 月，由同济大学上海市人工智能社会治理协同创新中心研究团队编制的《人工智能大模型伦理规范操作指引》正式对外发布。《人工智能大模型伦理规范操作指引》旨在结合中国的具体情况和国际通用的伦理准则，参考借鉴国家新一代人工智能治理专业委员会颁布的《新一代人工智能伦理规范》和联合国颁布的《人工智能与数据伦理原则》、《人工智能伦理建议书》，为中国 AI 企业提供了大模型伦理规范操作指引。《人工智能大模型伦理规范操作指引》主要包括 AI 大模型全生命周期的技术与伦理要素、大模型的研发与应用的伦理原则、大模型技术研发的伦理实践指南三部分内容，提出了尊重人的自主权、保护个人隐私、保障公平公正、提高透明度和可解释性、负责任的创新等五项大模型伦理原则，以及公平性、透明性、隐私、安全性、责任、人类的监督与控制、可持续性等七项大模型伦理实践操作建议。

(d)《人工智能法示范法 1.0(专家建议稿)》

2023 年上半年以来，中国社会科学院国情调研重大项目《我国人工智能伦理审查和监管制度建设状况调研》课题组主持人、中国社会科学院法学研究所网络与信息法研究室副主任周辉组织多方专家团队，经多次调研、讨论、修改，起草形成《人工智能法示范法 1.0(专家建议稿)》(“**《人工智能示范法建议稿》**”)。

《人工智能示范法建议稿》共分为六章：第一章 (总则) 阐明人工智能发展的基本原则，包括治理原则、人类自主原则、安全原则、透明可解释、公平原则等；第二章 (人工智能发展) 从基础设施、人才培养、技术创新、体制机制支持等维度提出相应制度规范，结合产业发展实际，采取有力措施鼓励人工智能创新，并强调以国家机关的先行先试促进人工智能的推广应用；第三章 (人工智能管理制度) 沿用近年来实践证明较为可行的风险分类分级管理方式，对人工智能技术研发和提供活动作出规定；第四章 (人工智能研发者、提供者义务)

明确人工智能研发者、提供者应承担相应合规义务，同时，对人工智能研发者、提供者进行了区分，依据其不同活动特点分配主体义务，结合本法前述条款设定的负面清单管理制度，针对负面清单内的人工智能研发、提供活动进一步规定了相应的义务类型；第五章（综合治理机制）衔接第一章（总则）规定，明确国家人工智能主管机关职责，提出创新监管、协同监管等机制；第六章（法律责任）根据人工智能的风险活动，设计相应的法律责任，并明确尽职免责等制度，为人工智能创新活动提供宽松政策环境。

《人工智能示范法建议稿》提出了负面清单管理等治理制度，并对人工智能产业链条各主体责任义务分配等核心问题进行了回应。在相应的法律法规尚未出台之际，《人工智能示范法建议稿》在一定程度上对于人工智能产业链条中的研发者、提供者、使用者等主体履行相应风险防范、安全保障义务等提供了可供参考的执行标准。

2. 合规要素

在大模型领域，合规义务主要责任主体为大模型服务提供者，即利用大模型技术提供服务的组织、个人。结合前述主要法律法规和相关规定、以及部分相对较有影响力的行业规范性文件，大模型服务提供者可以分为以下两类：

- 服务提供方

  服务提供方是指提供大模型相关服务的组织、个人。服务提供方通常会利用大模型相关服务开发面向终端用户的大模型应用场景，比如百度文心一言网站、抖音快手上面的一些AI特效功能等等。

- 技术支持方

  技术支持方是指为大模型相关服务提供技术支持的组织、个人。技术支持方往往表现为大模型的设计者、开发者和完成者，掌握着大模型背后的核心算法和运行规则，负责处理数据训练、生成内容标记、模型优化

等技术性事项。技术支持方通常会结合服务提供方关于大模型终端运用的需求，以API等形式提供大模型服务所需的技术支持。

在《深度合成管理规定》中，合规主体分为“深度合成服务提供者”和“深度合成服务技术支持者”，分别对应上述“服务提供方”和“技术支持方”；而《AIGC暂行办法》、《算法推荐管理规定》等法律法规和相关规定均未对“生成式人工智能服务提供者”、“算法推荐服务提供者”基于上述角度进行进一步区分。尽管如此，该等规定项下，在明确“人工智能服务提供者”、“算法推荐服务提供者”的具体责任和义务时，同样依据其提供的服务内容及类型规范了不同的责任和义务。例如，模型训练通常由技术支持方负责，其作为“生成式人工智能服务提供者”应当确保训练数据的来源合法合规，由于技术支持方并不直接面对终端用户，所以其仅承担法规项下明确需要参照适用的那些原本针对服务提供方的要求。而对于面向终端用户的“人工智能服务提供者”，即服务提供方，由其直接将内容 / 信息向终端用户提供，所以前述内容 / 信息所引致的结果也是由其直接产生，故其应当在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生成式人工智能服务等方面履行相应的义务。如果因为服务的提供而产生了违约、侵权等民事责任，服务提供方往往是第一责任人。

此外，根据《AIGC 暂行办法》第 2 条规定，行业组织、企业、教育和科研机构、公共文化机构、有关专业机构等研发、应用生成式人工智能技术，未向境内公众提供生成式人工智能服务的，不适用《AIGC 暂行办法》的规定。也即，需要遵守相关大模型合规义务的主体，是指向境内公众提供服务的大模型服务提供者。若上述主体未向境内公众提供服务的，则不适用《AIGC 暂行办法》。《深度合成管理规定》虽未将使用者限制在“公众”的语境，但对于标识的目标和要求，亦限制在了“公众混淆或者误认的”和“向公众提示深度合成情况”范围。基于前述规定，一个值得探讨的话题是，对于仅面向境内企业而并非公众提供大模型应用服务的大模型服务提供者是否适用《AIGC 暂行办法》。某种角度而言，加强大模型监管旨在规范公共层面的数据流通、传播，避免重要、敏感信息的泄露，以及防止违法、虚假信息和内容在社会层面广泛传播。倘若仅

面向特定企业提供服务，且该企业仅在内部使用大模型服务而不会导致大模型服务成果向公众流通，很有可能并不适用《AIGC 暂行办法》。但是，通过 API 接口等方式“封装”后间接提供服务的，可能仍会被认为属于服务提供方而非技术支持方，例如，倘若某一大模型服务提供者自研完成大模型开发后，作为技术支持方向中国境内的另一大模型服务提供者提供大模型技术接口并收取技术服务费，而后者进而作为服务提供方面向中国境内的消费者提供大模型应用服务，两者很有可能均需要履行《AIGC 暂行办法》项下的义务。

除了主体层面的合规要素外，大模型领域的监管对象：算法与模型同样值得探讨。“**算法**”是对于数据进行计算或其他处理的规则，从人工智能的角度，算法通过代码的形式实现。“**模型**”是通过算法对数据进行处理后，将处理形成的有效结果，作为未来处理参照的模型数据集，与算法形成一个作为模型的整体。简单来说，**“模型”=“算法”+“模型数据集”**。区别“算法”和“模型”的概念，对于人工智能的监管具有重要意义，主要体现在：

- 更好地界定客体——例如，单纯的算法提供者和内容提供者都不具有内容生成能力，所以《AIGC暂行办法》的监管客体应是模型。同样地，《深度合成管理规定》以内容生成能力作为前提，其监管客体也应是模型。算法备案的对象和内容，是算法而非模型[51]。安全评估规定则应将算法和模型都纳入监管范围。此外，算法的监管要点在于设计合规和提高算法透明度，而弱化所选择的训练数据的数据合规、标注质量评估和输出内容的知识产权等问题，而模型的监管则需要两者兼顾。

- 能更好地分析产业——目前，以AIGC为代表的人工智能市场已初步形成了应用层-模型层-基础层三个产业层次。直接面向终端用户的“服务”特别是互联网信息服务被纳入应用层、“模型”特别是通用基础大模型的训练和开发以及由此产生的模型即服务（MaaS）范式则应被纳入模型层。在更底层，“算法”特别是算法框架和开发平台作为单纯的算法基础服务商，和AI芯片、智能云服务、智算中心等作为算力基础服务商，以及数据集、向量数据库等作为基础数据服务商，则都被纳入基础层。

[51] 在互联网信息服务算法备案系统提交备案信息时，需要填写算法信息和模型信息。

- 能更好地识别行为——不同产业的行为监管逻辑根本不同。应用层直接面向用户甚至公众生成信息和内容，大多数涉及舆论属性和社会动员能力，以及民事侵权和个人信息保护等问题是在此阶段直接产生。模型层涉及训练和预训练的开展，在承上启下的过程中，既涉及底层算法的应用、数据的选择和标注，也决定了最终输出内容/信息的质量，此时需要关注的主要既包括数据合规、知识产权、公序良俗（如避免歧视）等基础层问题，也需要关注对最终输出的内容和信息的连带责任问题。在基础层，仅“算法”的提供和数据的服务，则关注内容更限于上面提到的各自基础层问题本身。

结合主体与客体的分析，对于大模型服务提供者，当前我国的法律体系下，其需要遵循的合规要素主要涉及业务资质、内容合规、数据训练合规、算法技术合规、个人信息保护、知识产权保护和竞争法、数据与网络安全、产品合规、监管手续、科技伦理等方面，具体如下：

(1) 监管手续与业务资质

(a) 算法备案

算法备案是算法治理体系的重要监管内容，是实现算法透明性和可解释性的必要环节，其旨在保护用户权益，维护产品安全和信息安全。《算法推荐管理规定》、《深度合成管理规定》、《AIGC 暂行办法》都对大模型服务提供者提出了算法备案要求。算法备案的主体是大模型服务提供者，在选择“生成合成类 (深度合成) 算法”这一算法类型进行算法备案时需要区分备案主体身份 (“深度合成服务技术支持者”或“深度合成服务提供者”)，即服务提供方和技术支持方需要作为不同的备案主体对同一算法进行备案，二者在算法备案项下的义务相互独立而不可互相替代。根据《算法推荐管理规定》，大模型服务提供者应当在提供服务之日起十个工作日内通过互联网信息服务算法备案系统填报服务提供者的名称、服务形式、应用领域、算法类型、算法自评估报告、拟公示内容等信息，履行备案手续。

(b) 安全评估

目前我国多部法律法规和相关规定中均对“具有舆论属性或社会动员能力的互联网信息服务”提出了安全评估的要求。不过，目前我国法律法规和相关规定中仅明确了“具有舆论属性或社会动员能力的互联网信息服务”(即开办论坛、博客、微博客、聊天室、通讯群组、公众账号、短视频、网络直播、信息分享、小程序等信息服务或者附设相应功能以及开办提供公众舆论表达渠道或者具有发动社会公众从事特定活动能力的其他互联网信息服务)，而对于何为具有舆论属性或社会动员能力的算法推荐服务、深度合成服务、生成式人工智能服务则暂时并未给出进一步定义。实务中，对于何为“具有舆论属性或社会动员能力”的判断相对较为宽泛，几乎涵盖了所有具备信息共享功能的服务。因此，大模型服务很有可能涉及具有舆论属性或社会动员能力的互联网信息服务，即需要按照《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》通过全国互联网安全管理服务平台完成安全评估。按照《AIGC 暂行办法》等法律法规和相关规定，对于大模型服务还需进行新技术新应用安全评估(“**双新评估**”)，而关于双新评估的具体流程以及要求仍有待监管部门进一步公开。

(c) 业务资质

为了保障大模型服务的合规发展，大模型在进入市场前，必须依照相关法律规定取得相应的资质证照。许可证类型根据相应业务而决定，例如：倘若最终的服务属于经营性互联网信息服务，需取得 B25 类增值电信业务经营许可证(即 ICP 证)；倘若最终的服务属于在线数据处理与交易处理业务，需取得 B21 类增值电信业务许可证(即 EDI 证)。

在当前我国的实践中，大模型服务涉及互联网信息服务的可能性相对较高，这主要是因为对于服务提供方向用户提供大模型应用服务的情形而言，服务提供方通过对训练数据和用户输入对话的采集和处理以及平台的建设，通过互联网向用户提供信息内容，往往会涉及为其他单位或个人用户发布文本、图片、音视频、应用软件等提供平台服务，即信息发布平台和递送服务这一类型的经

营性互联网信息服务。同时，对于经营性和非经营性的判断，实践中，不宜简单以服务是否收费来判断有偿或是无偿，而应当综合考量是否与科研、公益等非经营性活动有明显区分，需要充分考虑是否存在变相营利的情形。

此外，大模型服务领域或业务场景较为广泛，很有可能涉及多个行业的监管，从而需要获得特定行业的相关证照才能够合法运营。例如，在涉及图文、视听节目的情形下，往往还涉及《网络文化经营许可证》、《网络出版服务许可证》、《信息网络传播视听节目许可证》等行业监管角度的证照。

(2) 数据训练合规

数据训练是大模型技术存在的基础，是大模型应用的底层逻辑核心，数据是大模型最底层的原料，数据训练则是对原料的使用。因此，数据训练合规是满足服务生成内容合规、知识产权合规、个人信息合规等合规要素的重要前提。《AIGC 暂行办法》明确了生成式人工智能服务提供者在进行大模型训练时所应当履行的合规义务，其应当使用具有合法来源的数据和基础模型，不得侵害他人依法享有的知识产权，涉及个人信息的应当取得个人的同意或者符合法律、行政法规规定的其他情形。

大模型数据训练主要包括训练数据的收集、存储、使用等环节。在此过程中，除应当履行网络安全、数据安全、个人信息保护等义务外，还应当确保训练数据来源的合法性。从当前的行业实践来看，大模型服务提供者获取训练数据的途径大体可以分为经授权获取数据 (如采购第三方数据库等) 与自行收集数据 (如通过网络爬虫等技术手段收集数据等) 两类。在后者情况下，大模型服务提供者可能侵犯他人享有权益的内容，存在一定的法律风险。

在收集环节，在未经许可收集数据的情况下，根据数据类型不同，可能存在侵犯他人著作权、商业秘密、个人隐私等风险。若大模型在训练过程中存在破坏 / 绕开技术措施的方式获取数据，如采取破坏、绕开数据控制者设置的加密措施、访问限制措施、反爬措施等方式获取数据，或对数据控制者造成不合理负担的方式获取数据，妨碍、破坏他人产品或服务的正常运行，均有可能被

认定具有不正当性，从而被认定为构成不正当竞争。

在存储、使用环节，如果原始数据中包含受法律保护的客体或内容，则存储、使用行为可能落入法律规制的范畴。

(3) 内容合规

根据《AIGC 暂行办法》以及网络信息安全领域的监管要求，大模型服务提供者需要保证服务生成内容合规，承担对服务生成内容的审核义务，建立健全服务生成内容治理机制，依法设立辟谣机制、设立违法和不良信息识别特征库，积极承担信息内容管理主体责任；同时，当服务提供方发现违法内容时，应当及时采取停止生成、停止传输、消除等处置措施，并向有关主管部门报告。

(4) 算法技术合规

根据《算法推荐管理规定》、《深度合成管理规定》、《AIGC 暂行办法》等规定，大模型服务提供者需要承担算法技术管理相关的责任，主要内容详见下表：

序号	合规要点	具体内容
1	反歧视机制	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。
2	算法机制机理审核	定期审核、评估、验证算法机制机理、模型、数据和应用结果；不得设置诱导用户沉迷、过度消费等违反法律法规或者违背伦理道德的算法模型。
3	公平竞争机制	不得利用算法共谋方式形成垄断、排除市场竞争，遵循反垄断、反不正当竞争相关法律规定。
4	提供必要支持和协助	有关主管部门依据职责对生成式人工智能服务开展监督检查，提供者应当依法予以配合，按要求对训练数据来源、规模、类型、标注规则、算法机制机理等予以说明，并提供必要的技术、数据等支持和协助。

(5) 个人信息保护

《中华人民共和国个人信息保护法》(“**《个人信息保护法》**”)规制个人信息全生命周期的保护和处理活动，要求企业应在个人信息的收集、存储、使用、加工、传输、提供、公开、删除等方面落实合规义务。面向消费者的生成式人工智能应用服务在个人信息保护方面与其他应用服务相比有很多相同之处，包括制定用户服务协议、隐私政策，明确处理用户数据的合法性基础。在此基础上，《AIGC 暂行办法》针对生成式人工智能服务领域的个人信息保护做了进一步的规定，例如服务提供者对使用者的输入信息和使用记录应当依法履行保护义务和知情同意原则，不得收集非必要个人信息，不得非法留存能够识别使用者身份的输入信息和使用记录，不得非法向他人提供使用者的输入信息和使用记录，应当依法及时受理和处理个人关于查阅、复制、更正、补充、删除其个人信息等的请求。

此外，大模型服务提供者还应当特别关注个人信息的跨境传输问题。根据《AIGC 暂行办法》，无论是中国境外的技术支持方直接面向中国境内公众提供生成式人工智能服务，还是服务提供方通过接入中国境外的 API 接口向中国境内公众提供生成式人工智能服务，均应当履行《AIGC 暂行办法》项下的合规要求。在跨境的场景下，大模型服务提供者很可能将中国境内用户的个人信息传输至境外。对此，大模型服务提供者还应当按照《个人信息保护法》、《数据出境安全评估办法》、《个人信息出境标准合同办法》等相关法律法规和相关规定项下的要求履行个人信息跨境传输相关的义务，例如数据出境安全评估、个人信息保护影响评估、个人信息出境标准合同签订和备案、用户告知等，并根据不同的场景选择合适的跨境传输方式。

(6) 知识产权保护和竞争法

《AIGC 暂行办法》等法律法规和相关规定亦从知识产权保护和竞争法角度提出了相关要求。例如，根据《AIGC 暂行办法》，大模型服务提供者和用户在提供与使用大模型服务时还应当尊重知识产权、遵守商业道德、保守商业秘密，不得利用算法、数据、平台等优势实施垄断和不正当竞争行为；同时，大模型服务提供者在进行预训练、优化训练等训练数据处理活动时，亦不能侵犯他人

的知识产权。大模型服务提供者在大模型的开发和运用中还需要特别注意开源软件使用场景，应该在了解清楚每份代码的许可证类型后，明确每种许可证下的代码或软件的使用方式，以及这些许可证对商业化模式的影响，确保使用相关代码的过程不违反开源协议。

大模型服务从输入数据的获取及预处理，算法模型的构建与训练，到生成内容的输出与优化等各环节，均涉及专利、著作权、商业秘密等多种知识产权客体，稍不留意便将产生相应的侵权纠纷。需要特别注意的是，大模型多为商业性开发和利用，利用已有作品进行大模型训练的行为很难构成合理使用。因此，在服务生成内容生成过程中，倘若涉及与已有作品的接触且服务生成内容与已有作品存在实质性相似，服务生成内容本身很可能涉及著作权的侵权。而对于大模型服务提供者而言，其本身属于网络服务提供者，至少应当对用户输入数据进行审核且应当遵守服务生成内容合规方面的义务，《中华人民共和国民法典》第一千一百九十五条亦明确了网络服务提供者应当遵守的通知 - 删除义务，倘若未能遵守该等义务，有可能需承担共同侵权责任。

(7) 数据与网络安全

《中华人民共和国数据安全法》(“**《数据安全法》**”) 从多方面规定了企业数据安全保护相关的义务，包括数据分类分级、安全管理制度、风险监测、风险评估等，面向消费者提供生成式人工智能服务的大模型服务提供者作为《数据安全法》项下的数据安全合规主体，也应当履行《数据安全法》项下的合规义务。

《中华人民共和国网络安全法》(“**《网络安全法》**”) 从多方面规定了企业网络安全保护相关的义务。根据《网络安全法》，只要是由运营软硬件设备组成的、按照一定的规则和程序对信息进行收集、存储、传输、交换、处理的信息系统的所有者、管理者和网络服务提供者，均属于网络运营者。因此，大模型服务提供者作为网络运营者也应当履行《网络安全法》项下的合规义务。对于大模型服务提供者而言，其在《网络安全法》项下的合规义务主要包括两个方面：一方面，从网络运行安全的角度出发，大模型服务提供者作为网络运营者，应当按照网络安全等级保护制度的要求，履行安全保护义务，保障网络

免受干扰、破坏或者未经授权的访问，防止网络数据泄露或者被窃取、篡改；另一方面，从网络信息安全的角度出发，大模型服务提供者作为网络运营者，应当对其收集的用户信息严格保密，并建立健全用户信息保护制度，采取技术措施和其他必要措施，确保其收集的个人信息安全，防止信息泄露、毁损、丢失。从具体措施而言，在安全管理层面，大模型服务提供者作为网络运营者，应当明确网络安全的责任，并通过完善的规章制度、操作流程为网络安全提供制度保障；在技术层面，大模型服务提供者作为网络运营者，应当采取各种事前预防、事中响应、事后跟进的技术手段，应对网络攻击，从而降低网络安全的风险。

(8) 产品合规

依照相关规定，当面向终端用户提供大模型服务相关产品时，大模型服务提供者亦应当履行一系列从用户保护角度出发的合规义务。例如，建立实名认证体系义务、服务协议签订义务、明确并公开其服务信息以指导使用者科学理性认识和依法使用相关产品的义务、采取有效措施（如限定服务范围、限定服务时间）防范未成年人用户过度依赖或者沉迷相关产品的义务、采取有效措施稳定可持续的提供服务的义务、违法整改义务、建立健全投诉举报机制义务等。

(9) 科技伦理

在法律法规和相关规定层面，《科技伦理审查办法》、《关于加强科技伦理治理的意见》均对于科技伦理方面的合规要求予以规定；在行业规范层面，《新一代人工智能伦理规范》等文件均已经从原则上对于人工智能领域的科技伦理规则进行了一定程度的规定。具体要求如下：

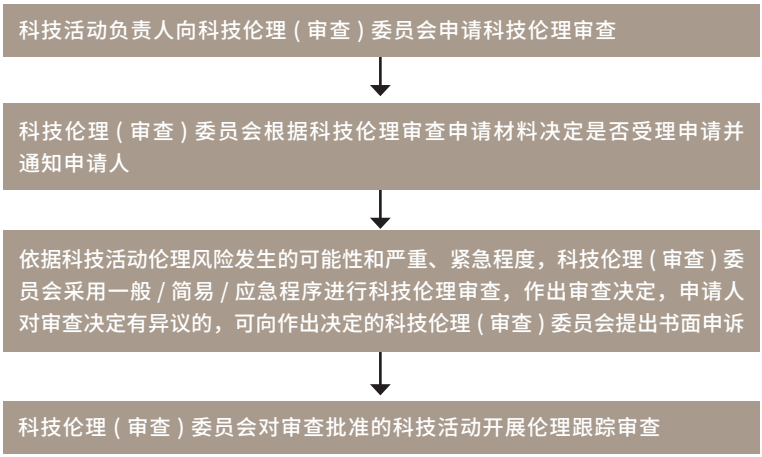
(a) 科技伦理（审查）委员会设立

根据《科技伦理审查办法》，如大模型服务提供者涉及以人为研究参与者的科技活动，包括利用人类生物样本、个人信息数据等的科技活动，或不直接涉及人或实验动物，但可能在生命健康、生态环境、公共秩序、可持续发展等方面带来伦理风险挑战的科技活动，应当负责进行科技伦理审查；如研究内容

涉及科技伦理敏感领域的，应设立科技伦理（审查）委员会，其他有科技伦理审查需求的单位可根据实际情况设立科技伦理（审查）委员会。大模型服务提供者应在设立科技伦理（审查）委员会后 30 日内，通过国家科技伦理管理信息登记平台进行登记，登记内容包括科技伦理（审查）委员会组成、章程、工作制度等，相关内容发生变化时应及时更新，并在每年 3 月 31 日前，向国家科技伦理管理信息登记平台提交上一年度科技伦理（审查）委员会工作报告。

(b) 科技伦理审查流程

根据《科技伦理审查办法》，科技伦理（审查）委员会开展科技伦理审查的流程如下：



(c) 伦理审查复核

根据《科技伦理审查办法》，针对纳入科技部发布的《需要开展伦理审查复核的科技活动清单》的科技活动，通过科技伦理（审查）委员会的科技审查后，除非国家实行行政审批等监管措施且将符合伦理要求作为审批条件、监管内容的，还需由开展技术活动的单位报请所在地方或相关行业主管部门组织开展专家复核；开展技术活动的单位应在纳入清单管理的科技活动获得伦理审查批准后 30 日内，通过国家科技伦理管理信息登记平台进行登记，登记内容包括科

技活动实施方案、伦理审查与复核情况等，相关内容发生变化时应及时更新，并在每年 3 月 31 日前向国家科技伦理管理信息登记平台提交上一年度纳入清单管理的科技活动实施情况报告。

根据科技部于 2023 年 10 月 8 日附随《科技伦理审查办法》发布的《需要开展伦理审查复核的科技活动清单》，“具有舆论社会动员能力和社会意识引导能力的算法模型、应用程序及系统的研发”属于需要开展伦理审查复核的科技活动。因此，大模型服务提供者如涉及大模型研发，除通过科技伦理 (审查) 委员会的科技审查以外，极有可能还需进行伦理审查复核。

(d) 科技伦理治理

除前述程序性要求以外，在实体层面，大模型服务提供者应当重视在研发和提供大模型服务过程中的科技伦理治理，重点关注研发规范与供应规范，其中重点内容包括：

(i) 提升数据质量。在数据收集、存储、使用、加工、传输、提供、公开等环节，严格遵守数据相关法律、标准与规范，提升数据的完整性、及时性、一致性、规范性和准确性等。

(ii) 增强安全透明。在算法设计、实现、应用等环节，提升透明性、可解释性、可理解性、可靠性、可控性，增强人工智能系统的韧性、自适应性和抗干扰能力，逐步实现可验证、可审核、可监督、可追溯、可预测、可信赖。

(iii) 避免偏见歧视。在数据采集和算法开发中，加强伦理审查，充分考虑差异化诉求，避免可能存在的数据与算法偏见，努力实现人工智能系统的普惠性、公平性和非歧视性。

(iv) 加强质量管控。强化人工智能产品与服务的质量监测和使用评估，避免因设计和产品缺陷等问题导致的人身安全、财产安全、用户隐私等

侵害，不得经营、销售或提供不符合质量标准的产品与服务。

(v) 保障用户权益。一方面，大模型服务提供者可以拒绝或避免开发以损害他人权益为主要目的的或者容易受到恶意利用的产品或服务；另一方面，在产品与服务中使用人工智能技术应明确告知用户，应标识人工智能产品与服务的功能与局限，保障用户知情、同意等权利，为用户选择使用或退出人工智能模式提供简便易懂的解决方案，不得为用户平等使用人工智能设置障碍。

(vi) 推动伦理安全建设。大模型服务提供者应建立健全覆盖管理、研发、供应、使用等全生命周期的风险治理体系、事件应对体系等。具体来说，大模型服务提供者可以采取建立验证算法、风险预警、记录和回溯机制等必要措施，持续监测和降低风险；同时定期分析风险监控报告并反馈和优化管理机制，完善治理体系。此外，大模型服务提供者可以建立事件应对体系，设立人工紧急干预机制、中止应用机制、救济金基金等必要保障机制，并明确事故处理流程，确保可以在AI伦理安全风险发生时作出及时响应。

3. 大模型业务中各方合规义务一览表（下表仅大致划分了各项义务的主要承担方，仅作参考）

合规要素	合规义务	主要义务主体		
		服务提供方	技术支持方	用户
监管手续与业务资质	算法备案	√	√	
	安全评估	√	√	
	一般性资质包括 ICP 证，特殊资质包括《网络文化经营许可证》、《网络出版服务许可证》、《信息网络传播视听节目许可证》等	√		

合规要素	合规义务	主要义务主体		
		服务提供方	技术支持方	用户
内容合规	发布内容合规	√	√	√
	AIGC 标识	√	√	
	及时处理违法内容	√		
数据训练合规	数据质量保证	√	√	
	数据来源合规	√	√	
	数据标注	√	√	
算法技术合规	反歧视机制		√	
	算法技术透明性		√	
	提供必要支持	√	√	
个人信息保护	个人信息来源合规	√	√	√
	个人信息去标识化	√	√	
	个人信息跨境合规	√	√	
知识产权保护和竞争法	不得侵害他人依法享有的知识产权	√	√	√
	尊重他人商业秘密	√	√	√
	开源软件使用合规	√	√	
	不得利用算法、数据、平台优势，实施垄断和不正当竞争行为	√	√	
数据与网络安全	数据来源合规	√	√	√
	数据跨境合规	√	√	

合规要素	合规义务	主要义务主体		
		服务提供方	技术支持方	用户
网络安全	不得利用互联网技术从事违法活动	√	√	√
	网络安全监管	√	√	
	建立网络安全等级保护制度	√	√	
	建立网络安全保障体系	√	√	
产品合规	指导、保护用户	√		
	稳定服务	√	√	
	违法处理与整改	√		
	建立投诉机制	√		
科技伦理	科技伦理审查	√	√	
	实践科技伦理规范	√	√	√

4. 运营角度的其他考量

(1) 大模型运营的要素

(a) 大模型运营的标的

在大模型相关的运营交易中，往往涉及技术支持方、服务提供方、终端用户等主体，各主体之间所涉及的标的亦有所不同。以当前的实践为例：

(i) 对于技术支持方提供大模型软件许可的场景，该等许可的标的实际上是软件模型。通常而言，大模型软件许可协议会针对许可标的予以特别规定。例如，如果被许可方仅需利用许可方已有的训练后模型，则被许可方根据许可协议取得训练后模型一定的使用权即可；但在很多场景下，被许可方需要的并非已有的训练后模型，而是定制化的训练

后模型，对于该等定制化的训练后模型的权利归属、使用条款，双方有必要在许可协议中予以进一步约定。

(ii) 对于服务提供方面向终端用户提供互联网平台服务的场景，其提供的服务通常为大模型交互对话、文字识别、自然语言处理等大模型产品服务，即以大模型为核心的服务产品。

(b) 大模型软件与传统软件的区别

(i) 软件开发方式

对于传统软件，软件开发者更关注软件的功能需求，即软件必须实现的功能。因此，软件开发者需要使用各种模型对相关功能需求进行描述，数据处理等规则往往已经被事先设计确定。而对于大模型软件而言，较之于功能需求，模型、训练模型的数据以及支撑模型训练的算力更为关键。模型开发者使用大量的数据对训练模型进行持续训练，使之归纳出处理新数据的规则。待训练模型通过学习知识成为具有推理和决策能力的训练后模型，从而实现智能化。

(ii) 数据使用方式

在传统软件开发过程中，通常并不需要收集并使用大量的数据。但在大模型软件的开发过程中，软件开发者必须借助大量的高质量数据样本对大模型进行训练，并在训练过程中不断优化参数以提高运行效率和准确性。训练数据通常根据具体的应用场景进行确定。以计算机视觉应用场景为例，利用现有的开源数据集通常难以满足特定的视觉应用场景需求，因此需要采集足够多的来自于实际应用场景的真实图像或视频数据，并对这些数据进行一定的处理，例如数据清洗、数据标注等。

(iii) 软件部署方式

从软件使用者的角度，大模型软件的安装部署方式与传统软件无明显差异，但是从运营方式和商业模式来看，二者还是存在一定区别。对于传统软件而言，

其对算力的要求相对较低，因此通常是由企业购买后安装在其自有服务器上，相关数据也通常存储在本地计算机或服务器中。而对于大模型软件而言，新兴应用场景产生的海量数据对大模型算力的需求持续加大，例如云游戏、自动驾驶等对数据传输的速度和量级都提出了更高的要求，而通过云计算和云部署的方式便可以在很大程度上解决上述问题。在该等情形下，相关数据则被传输并存储在云端。

(2) 大模型运营的关注要点

(a) 知识产权相关

(i) 知识产权权属

在传统软件许可协议中，无论许可标的是目标代码还是源代码，双方均应当对相关知识产权的权属安排进行提前约定，以免后续产生纠纷。一般而言，软件许可协议的知识产权归属安排会根据时间顺序采用“三段式”的叙述逻辑，即背景知识产权、前景知识产权和改进知识产权。其中，背景知识产权是指协议一方在履行协议前拥有或取得的技术成果及相关知识产权，前景知识产权是指在双方合作期间产生的知识产权，而改进知识产权则是指对前景知识产权进行的修改、改编或提升，包括但不限于对前景知识产权相关的功能、性能、部件或模块的变更等。

如上文所述，模型是由训练程序从训练数据中归纳出的某种“推理规则”，在此过程中，训练数据的质量和标注精度对模型的准确性起到至关重要的作用，换言之，训练程序输入不同的训练数据后所输出的模型也不尽相同。一般而言，模型的训练分为静态训练 (static training) 和动态训练 (dynamic training) 两种，因此，模型也分为静态模型与动态模型。对于静态模型，模型训练好则长期投入使用，而对于动态模型而言，随着新数据的不断输入，通过对这些数据的整合，模型也将不断进行更新迭代。

因此，在大模型软件许可中，若许可方许可的仅是静态模型，则被许可方

在具体的应用场景下使用该等模型，模型不会在被使用时同步自我演化或改进，被许可方只能通过许可协议要求许可方向其定期提供更新后的模型。但是，若被许可方获得的是动态模型的许可，由于被许可方持续不断地向模型输入实际应用场景的数据，模型也将被不断训练进而形成新的版本。在该等情形下，由于模型在使用被许可方所提供的数据过程中实现了自我改进，被许可方本身便可以对该等改进所形成的前景知识产权主张相应的权利。即使在许可方较为强势进而主张相关前景知识产权为自己单独所有的情况下，被许可方也可以考虑要求许可方就最新版本的模型向自己提供一项免费的许可，对此，双方还应当在许可协议中进一步明确许可费、更新维护等相关事项。

(ii) AIGC 的保护

大模型运营还面临着 AIGC 可版权性的问题。在我国的现行法律框架下，AIGC 的相关权益可能以以下路径获得保护：(1) 著作权法；(2) 反不正当竞争法；(3) 民法典。

AIGC 通常表现为音乐、图画、文字、视频、代码等内容或表达形式，表面上符合著作权法对作品的形式要求。而 AIGC 的可版权性的关键在于是否存在人类智力成果的贡献。也即是说，如果人类对 AI 的最终生成结果具有控制力，AIGC 存在人类的独创性贡献，就可以成为受著作权法保护的作品。反之，则可能无法获得著作权法的保护。至于人类要参与到何种程度才能构成对内容的独创性贡献，当前并没有形成统一定论。因此，在著作权法中新设邻接权、在民法典虚拟财产设置针对 AIGC 的具体规则等方式对 AIGC 相关权益予以保护的论题存在大量的讨论。而利用反不正当竞争法进行保护，主要是集中于大规模收集和生产的数据或信息，大规模盗用或以不正当手段获取 AIGC 等场景。

AIGC 虽然在权利属性方面尚存争议，但这并不阻碍 AIGC 的后续利用。目前以技术服务费、内容许可费等收益方式是 AIGC 后续利用的常见模式。相应的，关于生成物的的权利归属、后续利用范围和限制等均应和用户在协议中予以明确约定。

(iii) 潜在的知识产权侵权风险

大模型训练中可能产生潜在的知识产权侵权风险。如前文所述，大模型训练主要包括训练数据的收集、存储、使用等环节。而根据训练各个环节所使用的数据或内容所构成的法律客体的不同，可能存在侵犯著作权、商业秘密等知识产权的风险，或者因行为的不当性构成不正当竞争行为。

针对数据收集行为，数据的收集者更可能基于批量的数据、重复的获取行为等被追究反不正当竞争法项下的责任。针对数据存储行为，大模型开发者通常需要将收集到的原始数据存储到服务器中，在这一过程中会形成数据或内容的副本。如该等存储的内容可能构成著作权，在相关副本需要长时间停留在服务器的情况下，可能落入“复制权”的控制范畴；而如果不存储原始数据，仅在训练时临时调用，则可能因为没有形成“永久性复制件”，从而不会受到著作权法规制的范畴。针对数据使用行为，可能涉及对原始数据的修改、加工、翻译等操作，与之相应，则存在侵犯改编权、翻译权等著作权权利的风险。

在落入著作权权利范畴的情况下，就数据训练过程能否适用合理使用规则也是全球范围内探讨的重点问题。

为迎接人工智能等新技术，2019 年 3 月 26 日欧盟通过了《单一数字市场版权指令》，新增了“不限制目的的文本和数据挖掘”这一豁免情形，即在权利人未以适当方式保留文本和数据挖掘权利的情况下，基于文本和数据挖掘的目的，复制、提取合法访问的作品或其他客体的行为被纳入责任豁免机制。日本《著作权法》于 2018 年增设了新的合理使用条款“不以欣赏作品原有价值为目的的利用”。依据该条规定，只要模型训练阶段的作品利用行为不存在“根据作品的性质、目的和使用情况，不合理地损害版权人利益”的情形，大概率可以受到该条款的责任豁免。

目前我国现行《著作权法》规定的“合理使用”情形难以涵摄大模型训练的场景。具体而言，AIGC 场景可能适用的情形只有三种，包括“个人学习、研究、欣赏目的”“适当引用”“科学研究”。其中，“个人学习、研究、欣赏目的”

的合理使用对作品使用的目的进行了严格的限制，而 AI 模型训练基本是为了开发商业化产品，具有商业动机，难以被解释为该情形。“适当引用”指的是“为介绍、评论某一作品或者说明某一问题，在作品中适当引用他人已经发表的作品”，而使用训练数据的主要目的是为了生成新作品，与该种情形存在较大出入。“为科学研究使用作品”需同时满足“教学或科研人员”的主体要件，以及“少量复制”的要求，该等要求与 AI 模型训练中大量复制使用作品的现状不符。

但是对于大模型而言，确保训练数据中包含的作品全部获得作品著作权人的许可在现实中并非易事。一方面，大模型开发者需要花费大量的时间和成本将可能受保护的作品从训练数据中识别出来；另一方面，针对识别出来的受保护的作品，大模型开发者还需逐一地与作品的著作权人进行协商取得其许可，并支付许可费用。考虑到不同作品许可谈判的难度以及大模型开发的时效性，在实践中逐一取得相关作品著作权人许可并无可行性。因此，对于大模型训练阶段知识产权风险的防控亟待后续《著作权法》等相关法律法规进一步明确、集体管理等支付提供有效的指引。

针对大模型产品的著作权侵权问题，目前业内出现了一种新的潜在方案，以缓释大模型产品使用者的知识产权侵权疑虑。2023 年 11 月 6 日，在发布最新的 GPT-4 版本“GPT-4 Turbo”时，针对著作权侵权难题，OpenAI 一并提出了“著作权盾”的解决方案，即在 OpenAI 的客户因使用其产品导致著作权侵权的法律诉讼时，OpenAI 将介入并为其客户进行辩护，且承担因此发生的相关费用，具体的方案仍待 OpenAI 进一步澄清。[52] 后续有待观望这一方案在多大程度上能够减轻大模型产品的著作权侵权问题。

(b) 数据相关

(i) 数据使用

大模型运营中涉及的数据主要包括模型训练阶段使用的原始训练数据和训

[52] https://openai.com/blog/new-models-and-developer-products-announced-at-devday，最后访问于 2023 年 11 月 22 日。

练数据集，以及模型使用阶段的输入数据和输出数据，而模型使用阶段的数据存在被用于训练模型的可能性。在大模型软件许可中，由于并非所有的被许可方均希望提供数据给许可方以训练模型，协议双方可以约定许可方是否能使用被许可方的相关数据进行模型训练，在许可使用的情形下通常会对许可方使用相关数据的目的和范围进行限制。

(ii) 数据权属

鉴于一般认为对于衍生数据权利的确认并不代表否认原始数据主体的权利，模型训练阶段使用的原始训练数据和模型使用阶段的输入数据的相关权益应当分别归属于原始数据主体和输入数据主体，但模型训练阶段使用的训练数据集由于经过收集、清洗、标注等筛选处理，其相关权益应当归属于模型开发者，而模型使用阶段的输出数据由于其法律属性界定尚存在争议，通常需要协议双方明确约定相关数据的权益归属、使用方式等内容。

(iii) 数据来源

由于在大模型运营中，模型使用阶段的数据有可能被用于训练模型，协议双方均应当确保自身使用的数据具有合法来源。对于大模型而言，获取数据的方式主要包括数据交易、自行采集和开放数据爬取，其中，数据交易是指通过合法的交易方式从数据提供方处获取相关数据，自行采集是指通过 APP、传感器等方式直接采集数据，开放数据爬取则是指通过数据爬虫等方式获取开放的数据。前两者获取数据时应当注意要确保取得相关数据权利主体的授权，通过开放数据爬取时则应当重点关注数据爬虫行为本身是否合法; 对于许可方而言，不同数据种类存在不同注意事项，如除法律另有规定，对于个人信息应当直接或要求数据提供方取得个人信息主体同意，且应注意采取合理方式履行提示或者说明义务，如在用户协议中对相关内容加粗处理；对于被许可方而言，可以在协议中要求许可方对其提供的模型不侵犯第三方权利作出陈述与保证，而在提供数据给许可方以训练模型时，被许可方也应当履行相关合规审查义务，如获得数据主体授权、不违反保密义务等。

(iv) 数据质量与数据标注

根据《AIGC 暂行办法》第 7 条规定，生成式人工智能服务提供者应当采取有效措施提高训练数据质量。提高训练数据的质量对于避免误导用户、避免生成式人工智能被错用、误用、滥用，对于促进大模型运营都起着至关重要的作用。

《AIGC 暂行办法》第 8 条进一步规定，在生成式人工智能技术研发过程中进行数据标注的，提供者应当制定符合《AIGC 暂行办法》要求的清晰、具体、可操作的标注规则；开展数据标注质量评估，抽样核验标注内容的准确性；对标注人员进行必要培训，提升遵法守法意识，监督指导标注人员规范开展标注工作。数据标注是指对未经处理的语音、图片、文本、视频等原始数据进行加工处理，使其成为结构化数据让机器可识别的过程。数据标注由标注人员进行，人为错误或个人主观意识不可避免会反映在数据标注过程中，影响数据质量，因此制定清晰明确的标注规则、对标注人员进行培训是提高生成式人工智能的可靠性与可信度不可或缺的关键环节。例如，全国信息安全标准化技术委员会于 2023 年 10 月 11 日发布的《生成式人工智能服务安全基本要求（征求意见稿）》在“5.3 语料标注安全要求”节从标注人员、标注规则、标注准确性三个层面，对服务内容提供方的数据标注工作提出了具有可操作性的安全标准。

(v) 数据安全

在大模型运营中，为训练模型需要采集各行业领域的不同类型的数据，可能涉及敏感个人信息、重要数据等对安全保护有特殊要求的数据类型，也可能涉及数据出境等问题。

对于敏感个人信息和重要数据，以自动驾驶为例，智能驾驶汽车通过摄像头等传感器每时每刻都在收集车主等的个人信息、车辆行驶信息等数据，根据《汽车数据安全管理若干规定 (试行)》，车辆行踪轨迹、音频、视频、图像和生物识别特征等信息属于敏感个人信息，而涉及个人信息主体超过 10 万人的个人信息属于重要数据。如汽车数据处理者对相关数据处理时存在安全问题，可能导致个人信息主体的人身、财产安全以及国家安全受到损害。对此，法律

法规规定汽车数据处理者应当具有直接服务于个人的目的，包括增强行车安全、智能驾驶、导航等；应当报送汽车数据的安全防护和管理措施，包括保存地点、期限等。对于数据出境，被许可方应当在协议中明确要求许可方遵守数据出境的合规要求和履行数据出境申报义务等。

(c) 开源相关

开源作为推动大模型发展的重要力量，已成为当前人工智能领域的发展趋势之一。开源在促进大模型研发创新的同时，也推动和降低了大模型落地以及人工智能产业落地的门槛。虽然大模型软件与传统开源软件在计算机软件属性方面相似，但考虑到大模型软件的开发及其主要应用场景与传统软件仍存在一定区别，因此其开源合规问题也具有一定的特殊性。具体而言，大模型开发者在大模型开发阶段至少应当关注大模型本身的开源合规问题和模型权重的开源合规问题。

2023 年 7 月 19 日，Meta 在其官网宣布大语言模型 Llama2 正式发布，这是 Meta 大语言模型的最新版本，也是 Meta 声称的首个采用开源模式的大语言模型。然而，Llama2 并非完全意义上的“开源”，事实上，Llama2 对其商业用途做了一定的限制。例如，在 Llama2 版本发布之日，倘若被许可方或被许可方关联公司提供的产品或服务的每月活跃用户数在上一个日历月中超过 7 亿，则必须向 Meta 申请许可证，Meta 可以自行决定是否授权。因此，大模型开发者通过利用开源方式进行大模型开发时，一方面，应当梳理开发所使用的开源代码和许可证类型，另一方面，在明确开源代码及许可证类型后，应当进一步明确各类许可证下模型的使用方式，特别应当注意不同许可证对模型的用途所施加的限制，从而避免发生侵权或违约风险。

除大模型本身的开源合规问题外，模型权重的开源合规问题也应当引起大模型开发者的重点关注。以清华大学开放的 ChatGLM-6B 和 ChatGLM2-6B 模型为例，相比于大模型本身，ChatGLM-6B 和 ChatGLM2-6B 对模型权重设置了更为特殊的许可条件。具体而言，模型权重对学术研究完全开放，但是模型权重的商业使用则需要完成登记并获得授权。因此，大模型开发者还应当注意区分模型本身和模型权重所适用的许可条件。

三、未来展望与发展建议

(一)未来展望：大模型合规的前沿

1. 大模型技术创新发展与合规风险并存

随着深度学习和其他人工智能技术的快速发展，大模型的结构和性能都得到显著优化。尤其在大模型的规模、复杂性和应用范围上，技术进步为其提供了强大支持。然而，快速的技术进步也带来了新的合规挑战，尤其体现在数据隐私、模型透明度和伦理道德等方面。

模型结构的优化是为了满足更为复杂的任务需求。例如，Transformer 架构使得模型可以更好地处理长序列数据，显著提升在自然语言处理和其他序列任务上的性能，且神经网络的不断深化使得模型可以学习到更为复杂的特征和规律。但是这种优化也为模型的可解释性和透明度带来挑战，大模型的内部结构和操作成为了一个“黑盒”，使得外部观察者很难理解其具体的工作原理。

与此同时，技术进步也带来了数据处理和计算的新能力，即模型可以训练和处理前所未有的大规模数据集，为模型训练提供丰富数据，但这也引发了对于数据隐私和合规的关注。在欧洲、北美和其他地区，政府和监管机构对数据隐私和合规提出严格要求，对企业和研究机构在处理用户数据时遵循明确的指导原则提出要求。

2. 大模型合规框架走向标准化与国际化

(1) 全球合规标准的趋同与差异

随着全球化的加速和技术的普及，大模型的合规问题不再是单一国家或地区的关注点，而是各国共同面临的挑战。在这一背景下，合规标准在全球范围内呈现出趋同的趋势，但各国之间因文化、法律和经济发展水平的差异，仍存在区别。技术普及、国际经贸往来和大型企业的全球化策略都在推动各国合规标准统一。例如，对数据隐私的关注、对模型透明度的要求以及对技术应用的

伦理道德边界设定，使得各国在这些共同议题上逐渐形成共识。然而由于文化背景、历史传统和经济发展阶段的不同，各国处理大模型合规问题所采取的方法和策略也略有不同。例如，欧盟的 GDPR 更强调个人隐私权益保护，美国更强调企业权益与用户权益之间的平衡。

(2) 国际合作与共建合规框架

在全球经济一体化的背景下，单一国家难以独立解决大模型合规的问题。因此，**国际合作与共建成为趋势，旨在构建一个公平、透明、有效的大模型合规框架。**随着技术跨境应用和数据跨境流动，各国意识到只有通过合作，才能真正解决跨国合规问题。同时，大型技术企业和研究机构的跨国活动也需要统一的合规标准指导。联合国、G20、世界经济论坛等国际组织和论坛，将成为各国讨论和推进共建合规框架的平台，各国能够借此分享经验、协调差异，并共同制定合规指导原则和标准。随着全球经济技术进一步融合，国际合作与共建的趋势将日益凸显，各国之间交流合作将更加深入，共同构建稳定、公正的大模型合规环境。

3. 社会文化和伦理逐渐与合规体系相融

(1) 社会公正是大模型合规的前提

大模型的发展与应用涉及到社会、文化和伦理等多重维度，正确理解和处理这些维度是确保大模型健康、合规发展的关键。社会公正是大模型发展的前提，大模型的开发与应用过程应符合公平正义，算法决策应避免偏见和歧视，促进公平。同时，大模型应尊重文化多样性。不同文化背景下，对于同一问题的看法和解决方法可能存在巨大差异，需要充分考虑大模型合规中的文化差异，确保大模型的决策不违反当地文化习俗和价值观。

(2) 大模型伦理问题需多角度对待

随着技术应用全球化，大模型的伦理问题需要从多元文化的视角审视，以

确保模型在不同文化背景下都能得到合理应用。虽然公平、透明和可解释性等伦理原则具有普适性，但不同文化背景下，其具体实施方式可能存在特殊性。因此，需要在普适性和特殊性之间找到平衡，确保伦理原则应用全球化的同时，考虑地方文化的特殊性。同时，为确保大模型在全球范围内合规应用，需要加强跨文化伦理研究，探讨不同文化背景下的伦理问题和挑战，并为大模型开发提供指导，为大模型应用全球化提供坚实的伦理基础。

4. 行业应用面临不同合规挑战与监管

随着大模型在各个行业广泛应用，不同行业和领域对大模型的合规需求也呈现出明显差异性。

(1) 不同行业合规需求存在差异

- **金融**：在金融领域中，大模型的决策可能直接影响资金流动和市场稳定性。因此，金融行业对大模型的准确性、稳定性和透明性要求极高，且需考虑数据隐私和安全性问题。

- **医疗健康**：在医疗健康领域中，大模型决策涉及患者的生命健康，大模型的误判可能导致严重后果。因此，医疗行业对大模型的准确性和可解释性要求严格，且需满足医疗数据的保密性和合规性要求。

- **公共管理**：在公共管理领域中，大模型可能用于资源分配、公共决策、政务服务等核心环节。因此，大模型合规要求不仅涉及技术层面，还需保证决策的公平、公正和透明，以及服务的准确可信。

- **新闻媒体**：在新闻媒体领域中，需考虑内容的真实性、多样性和公平性，确保提供的内容不会误导公众或加剧社会分化。因此，新闻行业对大模型的可理解性和可靠性要求严格，且须满足新闻数据的准确性和真实性要求。

(2) 大模型行业应用评估与监管趋于完善

随着未来大模型在各行业的应用广泛度提升，针对大模型相关的评估与合规监管的重要性也日益凸显，相关评估和监管机制需不断完善。

- **大模型的独立评估**：可由第三方机构对大模型进行独立的评估，确保模型的决策公正、准确，并符合行业的特定要求，以提高大模型在公众中的信任度，确保其合规应用。

- **持续监管与审计**：对于已经部署的大模型，持续监管和审计也需进一步加强，以及时发现并纠正潜在问题，确保大模型在实际应用中仍满足合规要求。

- **建立反馈机制**：大模型在实际应用中可能出现未知问题，完善反馈机制可进一步畅通大模型开发者和使用者的沟通渠道，以便用户和利益相关者可以及时提出意见和建议，帮助大模型持续改进。

- **合规性指导与教育**：需将合规性指导和教育提上日程，以确保大模型开发者和使用者都能够明确合规要求，帮助其更好理解和遵循相关规定。

5. 治理路径分阶段、有弹性地构建

在面对大模型合规问题时，固化规则和僵硬管理往往难以适应技术快速演进和应用场景多样性。因此，**弹性治理理念应运而生，主张构建灵活、适应性强的治理路径**。弹性治理并非放任自流，而是在明确的指导原则下，给予大模型开发者和应用者一定自主权，使其能够针对特定场景适当调整。弹性治理具有以下特性：**适应性**，即弹性治理对于新技术和应用场景的出现能够快速反应，不会因为固化规则而制约创新。**多元性**，即弹性治理考虑到不同文化、社会和行业的特点，可在明确框架内进行多样化实践。**持续性**，即弹性治理强调持续监督和反馈，而非一次性审核，确保大模型始终保持在合规的轨道上。

(二)发展建议：构筑大模型合规生态

1. 政府推动构建行业新秩序

政府应通过为企业提供政策指导，为行业构建有利于创新与合规的新秩序，推动行业有序发展和健康成长。

(1) 制定与完善相关法律法规，构建不同阶段合规制度

(a) 横纵向监管结合，兼顾治理的统一协调与规则的垂直细分

大模型的出现标志着社会生产方式的划时代革新，其覆盖的产业版图极为全面，包含从芯片、高性能计算集群、图形处理器等硬件部署，到数据及各类语言的学习与处理、算法与模型搭建、内容生成、全场景泛语言多任务的处理应用的软件研发运营；其涉及的法律领域相当广泛，包括网络安全与数据治理、个人隐私保护、知识产权、反不正当竞争、产品市场监督等各类合规要素。针对这一复杂多变的“庞然大物”，境外各主要地区的立法思路不约而同地遵循了“横向监管”与“纵向监管”两条主要路径。

所谓横向监管，指以大模型这一整体概念为核心，建立一套统一的、普遍适用于各类大模型的、跨越多个行业不同主管部门的监管规则，目的是为大模型监管提供统一的标准以规制并引导行业发展，所体现的立法理念是“概念先行”。其表现形式通常为一部综合性法律法规 (“**横向法规**”)，配套一系列横向的统一监管工具 (“**横向监管工具**”)，例如登记、备案及评估系统等。横向监管的优点主要体现在以下几个方面：(1) 一致性。横向监管将大模型所涉及的普遍风险进行了统一规定，使得各类大模型间的监管标准一致，可以减少监管规则的冲突、混淆与重复，降低企业及机构的合规成本；(2) 开放性。横向监管可以对大模型采取较为广泛和开放的定义，并阐述大模型所适用的普遍原则 (例如欧洲、美国、英国等地均在各类法规政策中反复强调的合法、安全、透明、稳健、反歧视、人工监督、符合伦理、保护个人隐私、增进社会福利等原则)，使其可以涵盖大模型未来的各种创新形式，一定程度上避免因为旧概念无法适

用于新发展而带来的立法滞后、监管缺失以及重复立法、资源浪费，也避免因为某一大模型可能同时落入多个纵向法规的规制范围而产生法规的适用冲突；(3) 全面性。横向监管可以将各类合规要素均纳入综合性立法的考量之中，避免遗漏一些不在特定纵向监管范围内的问题；(4) 可预测性。单一且固定的横向监管工具为企业提供了监管的可预测性。

所谓纵向监管，指将大模型根据不同功能进行拆解细分，并针对每一种功能类型的大模型单独规定其合规要点，以便更精确地解决某一领域存在的特定问题，所体现的立法理念是“实践先行”。其表现形式通常为多部针对性法律法规并行 (“**纵向法规**”)。纵向监管的优点主要体现在以下几个方面：(1) 针对性。纵向监管可以更有效地解决某一特定类型的大模型所存在的特定问题，提高法律法规的可适用性与治理效率，做到对症下药、量身定制，避免过于宽泛的合规要求所导致的高昂合规成本以及部分条款适用性存疑所导致的合规焦虑；(2) 灵活性。纵向监管允许监管机构在短时间内针对新的技术或行业发展及时推出新的监管规则并调整监管策略，但因避免由于法律体系过于庞大，需要考虑条款间协调性与新旧条款融合衔接。

参考各国治理策略，我国对于大模型的监管可以考虑兼采横向、纵向监管之所长，针对不同的生产环节，分别适用不同的监管策略。一方面，大模型和人工智能二者在运行逻辑上紧密相连，因此，可以考虑采用以单部横向法规作为主体，并配合统一的横向监管工具。另一方面，针对大模型中的重点类型、主要功能，可以设置多部针对性法律法规予以规制；同时，考虑到不同类型的大模型所需要遵守的标准以及监管重点不同，在横向监管工具的具体适用中 (例如评估准则、备案信息清单等)，可以嵌入纵向监管标准 (例如针对特定行业的垂直大模型委托第三方机构制定行业标准)。在大模型产品、大模型服务的发布前环节 (包括设计、开发、部署)，可以考虑采取“纵向监管优先 + 横向监管兜底”的方式，即倘若企业所研发的大模型相关技术 (例如深度合成) 落入某一特定纵向法规的管理范畴，则该纵向法规的要求应当优先适用，但是倘若针对该等技术并无任何可适用的现存纵向法规，则可以由横向法规作为兜底性条款起到规范作用，避免监管缺口。针对产品和 / 或

服务的审核环节以及使用环节，可采用固定的横向监管工具进行统一监管，降低合规成本。同时，在具体的法律条款中，亦需要针对不同的环节设定不同等级的合规要求。

(b) 明确责任主体，确定责任分配

目前，总体而言，我国现行的大模型监管体系主要采取的是纵向法规与横向监管工具并行的策略，现行的主要法律法规和相关规定针对的主要是特定的深度合成等技术本身，同时采用了包括算法备案在内的、未来可能能够扩展适用于其他类型的监管工具。

然而，各项规定之间的概念难以实现统一已经成为了目前较为凸显的问题之一。例如，《深度合成管理规定》区分了“深度合成服务提供者”和“深度合成服务技术支持者”；《AIGC 暂行办法》主要明确了“生成式人工智能服务提供者”的合规义务；《算法推荐管理规定》则主要针对“算法推荐服务提供者”提出了合规的系列要求。但事实上，大模型产业链中从研发到投放市场、交付使用，所涉及的主体众多，所涉及的法律关系亦较为复杂，包括自行及委托研发、人工智能集成、商业运营、分销、跨境许可等，概念的划分模糊可能导致责任承担不明晰，监管问责也将付之阙如。

欧盟的《人工智能法案》提案可能可以为我国的法律规范体系提供部分思路。《人工智能法案》将责任主体划分为提供方、部署方、进口方、分销商四种角色。由于提供方对于系统的控制力度最强，因此，提供方在《人工智能法案》项下需要承担的合规义务相对最重，但当部署方、进口方和分销商对系统进行了署名或者进行了实质性的修改，从而被认为在相当程度上控制了系统时，将被视为提供方，亦需要承担较重的合规义务。

(2) 为合规大模型的研发与应用提供资金支持和税收优惠

- **资金支持**：为鼓励企业和研究机构研发符合合规要求的大模型，政府可以设立特定的资金池，专门用于支持该方面的研究和项目。此类资金支

持不仅能够缓解企业和研究机构在研发阶段的资金压力，更能够引导整个行业向合规方向发展。

- **税收优惠：**除了直接资金支持，政府可以通过税收优惠的方式，为大模型的研发与应用提供更多激励。例如，对于在大模型研发和应用方面做出显著贡献的企业以及获奖企业等，可以给予一定比例的税收减免或退税，从而鼓励更多企业参与大模型的研发与应用。

(3) 与行业进行深度合作，共建合规监管体系

在构建大模型合规生态的过程中，政府与行业之间的合作尤为关键。政府可以通过各种渠道，如研讨会、论坛等，与行业进行深度互动，了解行业的实际需求和问题，打造出既能满足技术发展需求，又能确保社会公众利益的合规框架。

- **建立沟通机制：**政府应当建立如定期政策研讨会、行业论坛、工作小组等与行业之间的常态化沟通机制，在确保行业声音被真正听到的同时，也让政府的政策制定更加接地气、具有针对性。

- **共同制定标准：**技术与合规的标准并非一成不变，随着技术发展，这些标准也需要随之调整。政府应该与行业专家、高校学者、企业代表共同制定和完善相关技术与合规标准，确保其兼顾科学性与实用性。

- **鼓励行业自律：**除了外部监管，政府应当鼓励行业自我监管。例如，支持行业组织制定专门的行为准则或伦理守则，为行业内的企业和个人提供行为指导。

- **组织培训与教育：**对于大模型合规的要求和标准，不仅行业内部需要了解，公众也需要有所认识。政府可以通过组织培训和教育活动，帮助行业和公众更好地理解和应对合规性问题，协助用好大模型这一生产力工具。

2. 企业创新与责任担当

(1) 注重大模型的自我治理与社会责任

在数字化时代，企业的责任不仅仅局限于提供高质量的产品和服务，还需要确保其行为和创新对社会产生正面影响。对于从事大模型研发和应用的企业而言，自我治理和担当社会责任至关重要。

- **建立完善的自我监管机制：**企业应建立一套内部审核与评估机制，确保大模型的研发与应用过程中能够满足法律、伦理和社会的要求，其包括但不限于对模型的输入输出内容进行审查、对模型的决策逻辑进行透明化，以及定期进行模型的合规性检查。

- **强化企业社会责任文化：**企业应当将社会责任意识融入公司文化中，积极参与公益活动，加强与社区和非政府组织的合作，以弘扬企业的正面形象和增强公众信任。

- **与社会持续沟通交流：**企业需定期与社会各方进行沟通与交流，通过公开座谈会、听证会或社交媒体平台等方式，听取外部对其大模型应用的意见和建议。

- **公开透明的责任报告：**企业应考虑定期发布关于大模型的责任报告，内容包括模型的研发、应用、影响评估以及面临的挑战和解决方案，向公众展示其在合规、伦理和社会责任方面所做的努力。

- **促进多方利益平衡：**在追求利润的同时，企业还需确保技术创新带来的社会效益，这意味着在决策过程中要充分考虑消费者、员工、股东和社会的利益，并努力实现其中的利益平衡。

(2) 重视技术研发与模型优化

技术的不断进步与创新是推动大模型走向合规的核心动力。企业若想在竞

争激烈的市场环境中长期稳定发展，必须将研发和模型优化置于首位。

- **持续增加研发投入：**企业应持续增加对技术研发的资金投入，鼓励团队深入研究和探索更先进、更高效的模型算法。这不仅能提高模型的性能，还能为企业在合规性方面带来先发优势。

- **与学术界紧密合作：**与全球顶尖的学术机构和研究者建立合作关系，可以帮助企业紧跟最新的技术发展趋势，确保技术研发的方向与国际前沿水平保持一致。

- **关注用户反馈与需求：**用户是大模型应用的最终受益者，企业应定期收集并分析用户反馈，根据反馈对模型进行优化，确保其更好地满足用户实际需求。

- **跨领域技术融合：**大模型的发展不仅仅依赖于单一技术，还需要与其他技术领域(如隐私计算、边缘计算等)进行融合，从而带来更加高效、安全和合规的应用解决方案。

(3) 加强与其他参与方的沟通与合作

大模型的研发、应用和管理是一个涉及多方的复杂过程。为确保大模型的合规性和有效性，企业不能单打独斗，必须加强与各相关参与方的沟通和合作。这不仅有助于企业更好理解和应对合规性挑战，还能为整个行业带来更加完善和统一的合规框架。只有在各方共同努力下，大模型才能真正为社会带来持久和广泛的价值。

- **与政府和监管机构建立对话机制：**企业应主动与政府和相关监管机构建立常态化的对话与沟通机制，及时了解政策方向和监管要求，为政策制定提供行业实践和技术建议。

- **与同行业企业展开合作：**在合规性问题上，企业之间不应仅视对方为竞争对手，应当共同研发技术标准，分享最佳实践案例，以及协同应对潜

在的技术、安全和伦理挑战。同时，企业也应与国际组织和跨国公司建立合作关系，共同探讨和制定国际合规标准和最佳实践方式。

- **参与或创建多方协同的行业联盟：**通过参与或创建行业联盟，企业可以与各方共同探讨合规性问题，分享资源，合作研发，从而提高整个行业的合规性水平。

3. 社会组织加强协同合作

(1) 加强大模型监督与评估

随着技术快速发展，确保大模型的合规性和公正性至关重要，而社会组织在大模型的监督与评估中发挥的作用不可忽视。例如，非政府组织、研究机构和行业协会通过编制发布大模型开发与运营相关的行业性规范，可以保证大模型技术在带来革命性改变的同时，不损害公众利益。

- **设立第三方评估机构：**设立独立于企业和政府的第三方评估机构，开展客观、公正的大模型评估，深入挖掘和识别模型中的偏见、不公和其他潜在问题。

- **提高透明度和可解释性：**通过监督企业公开或部分公开其模型的工作机制、数据来源和训练方法，提高整个行业的透明度，使复杂的模型更加可解释，帮助公众和决策者更好理解模型的决策逻辑。

- **举办公开评估和测试：**组织公开的模型评估和测试活动，提高社会影响力，助力推动行业标准制定，鼓励企业采用更高的技术和伦理标准，促进模型透明度提升。

(2) 开展大模型相关的教育与培训

社会组织应加强大模型技术宣传和教育培训，培养一批有知识、有技能、有责任心的新一代从业者，确保大模型技术在发展中获得公众支持和信赖，为

其在各个行业的应用创造有利条件。

- **组织专题讲座和研讨会：**通过定期的讲座、研讨会或工作坊，企业分享最新科研成果，向公众、政府官员传递关于大模型的最新研究、最佳实践和伦理标准，政府人员也能够及时分享与公布最新政策。

- **开发教育课程：**建议与教育机构合作，制作并发布易于理解的教育材料并进行推广，如视频、动画、互动教程等开发大模型所需的相关技术教育课程，向未来技术人员和决策者提供充分的知识普及。

- **与企业和研究机构合作：**与行业领先的企业和研究机构合作，确保其教育和培训内容与实际应用和前沿研究保持同步，推动产学研快速转化。

(3) 与政府、企业建立对话机制

社会组织作为核心媒介应为企业和政府的沟通提供交流载体，通过持续、透明和多方参与的对话，使相关主体共同参与大模型合规决策。

- **建立交流互动平台：**通过设立定期的圆桌论坛、工作小组或研讨会，为政府、企业和公众之间建立沟通桥梁，促进政府、企业和其他利益相关者提供交换观点、分享经验和探讨解决方案的平台，加强互信和合作。

- **收集和响应公众反馈：**作为与公众之间的桥梁，社会组织应当为公众打造分享观点、提出疑虑和建议的平台，定期收集公众对大模型应用的看法和反馈，以深入探讨大模型技术背后的伦理和社会影响。同时，将社会声音反馈给企业与政府，帮助企业和政府调整策略，确保技术真正服务于社会。

- **应对突发事件：**在大模型应用中可能出现的突发事件或争议情况下，社会组织可以作为调解者或顾问，协助各方共同应对和解决问题。

主编单位：

金杜律师事务所

上海人工智能研究院

华为技术有限公司

上海昇思 AI 框架 & 大模型创新中心

专家指导委员会：

宋海涛、聂卫东、李学尧、王永全、丁　诚

编审委员会：

张逸瑞、陈府申、钱琪欣、刘燕京、杨　浩、房思哲

编辑委员会：

主　编：孙　丽、冯宝宝

成　员：邓志辉、朱佳蔚、吴之洲、张一凡、张中阳、张津豪

周　彤、侯玉杰、贾挺猛、唐晟凌、黄中斌、康起明

蒋世聪、潘一飚（姓氏笔画排序）

声明：

本出版物不代表金杜律师事务所对有关问题的法律意见，不代表上海人工智能研究院对有关问题的立场，仅供读者参考。任何仅仅依照本出版物的全部或部分内容而做出的作为和不作为决定及因此造成的后果由行为人自行负责。如您需要法律意见或其他专家意见，应该向具有相关资格的专业人士寻求专业的法律帮助。

本出版物中，凡提及“香港”、“澳门”、“台湾”，将分别被诠释为“中国香港特别行政区”、“中国澳门特别行政区”、“中国台湾地区”。

版权声明：© 金杜律师事务所、上海人工智能研究院 2023 年版权共同所有

如需了解更多信息，请访问 www.kwm.com/www.sairi.com.cn

金杜律师事务所、上海人工智能研究院保留对本出版物的所有权利。未经金杜律师事务所、上海人工智能研究院书面许可，任何人不得以任何形式或通过任何方式(手写、电子或机械的方式，包括通过复印、录音、录音笔或信息收集系统)复制本出版物任何受版权保护的内容。

有关本出版物的咨询及意见和建议，请联系：publication@cn.kwm.com/kangqiming@sairi.com.cn

金杜研究院
KWM_CHINA

上海人工智能研究院
sairi930