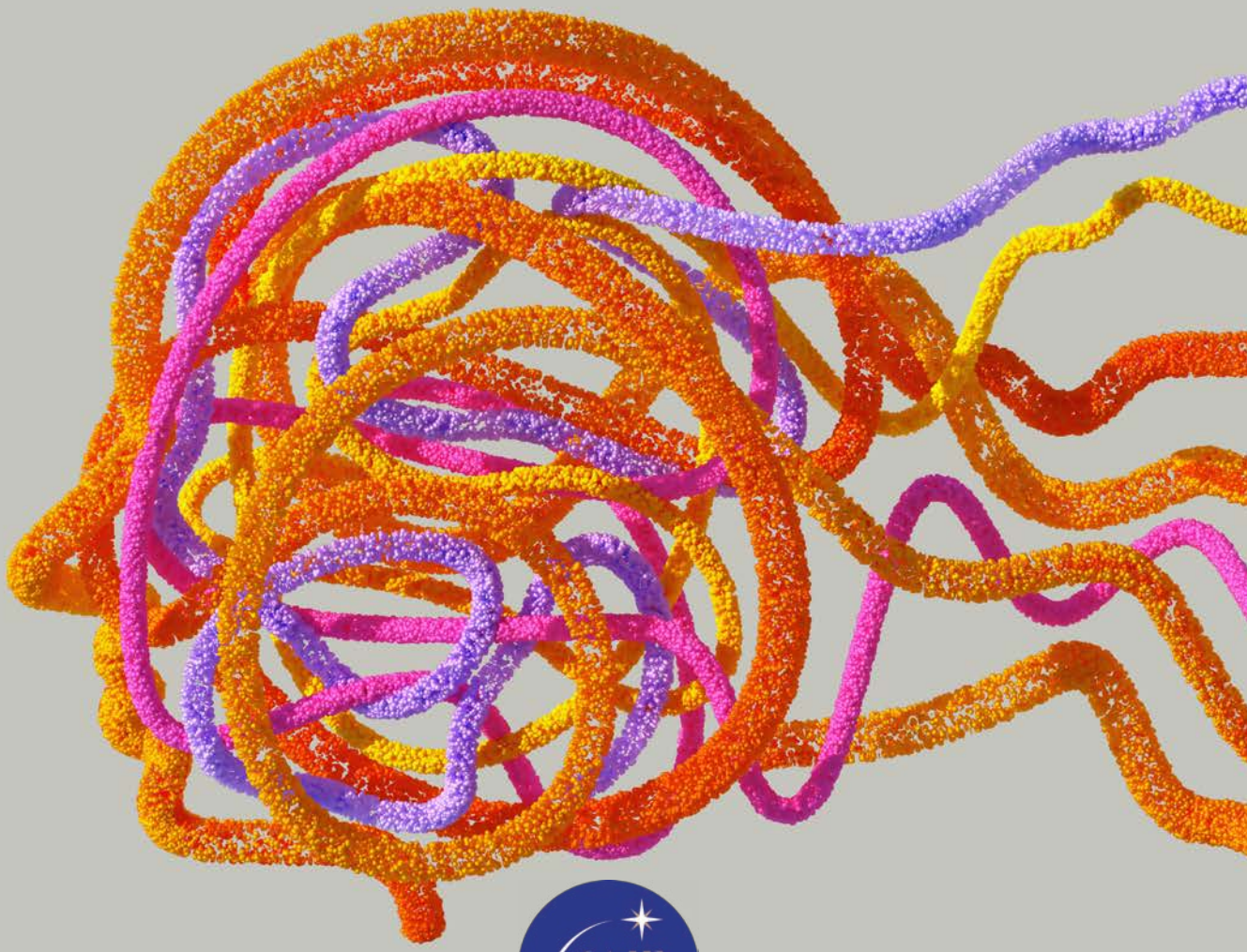
# AIGC

COMPLIANCE
DEVELOPMENT
WHITE PAPER

# 生成式人工智能服务

## 合规发展白皮书



SAIF
ThinkTank
高金智库

中倫
ZHONG LUN

**【课题组负责人】**

胡　捷（上海高级金融学院）

樊晓娟（中伦律师事务所）

**【课题组成员】**

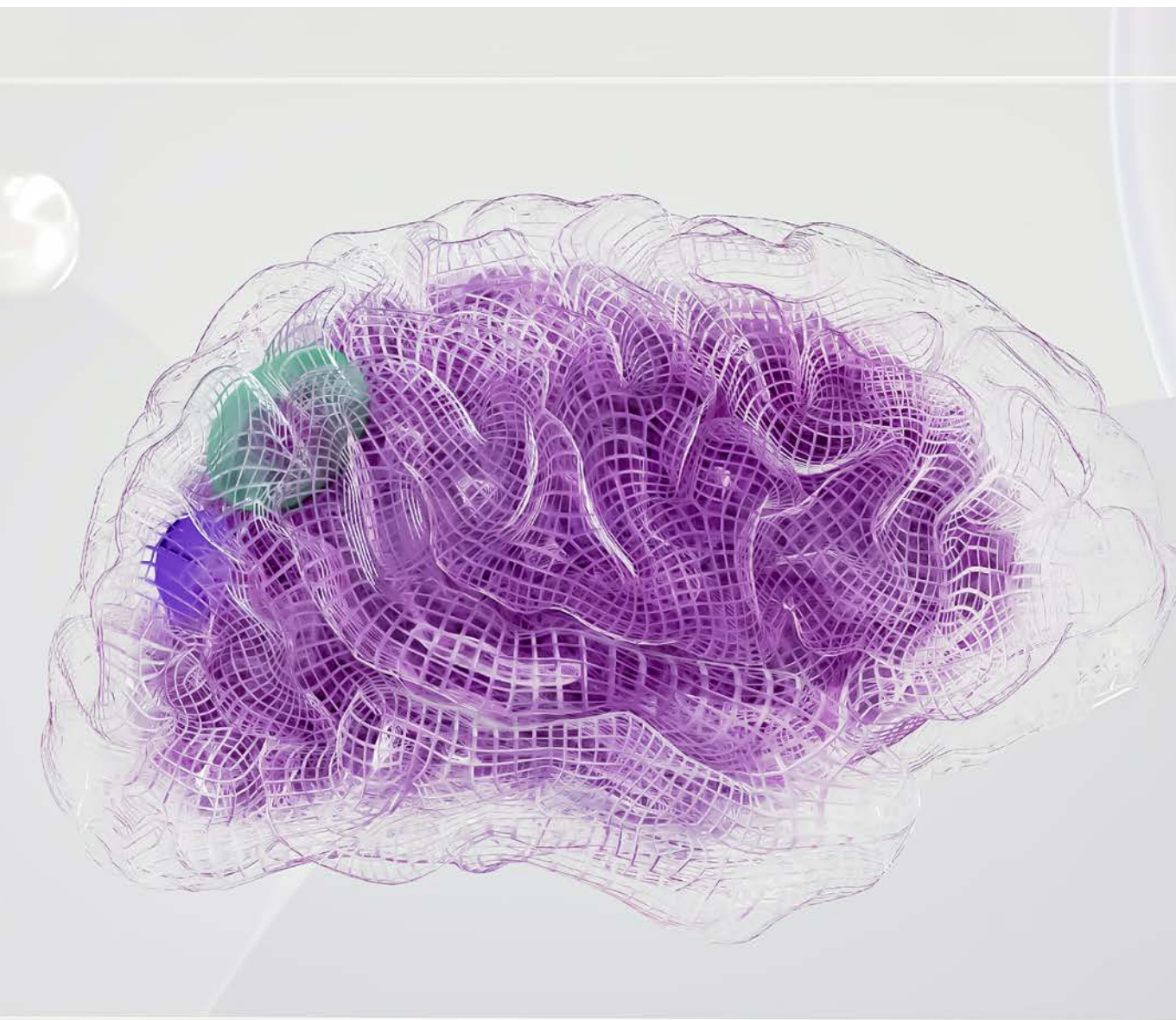
甘泉、陈旖俐、沙俊、熊国君、郭子豪

# 生成式人工智能服务
# 合规发展白皮书

*AIGC*
*COMPLIANCE DEVELOPMENT*
*WHITE PAPER*

# CONTENTS 目录

注:本文中大部分插图皆由midjourney生成。



# CONTENTS 目录

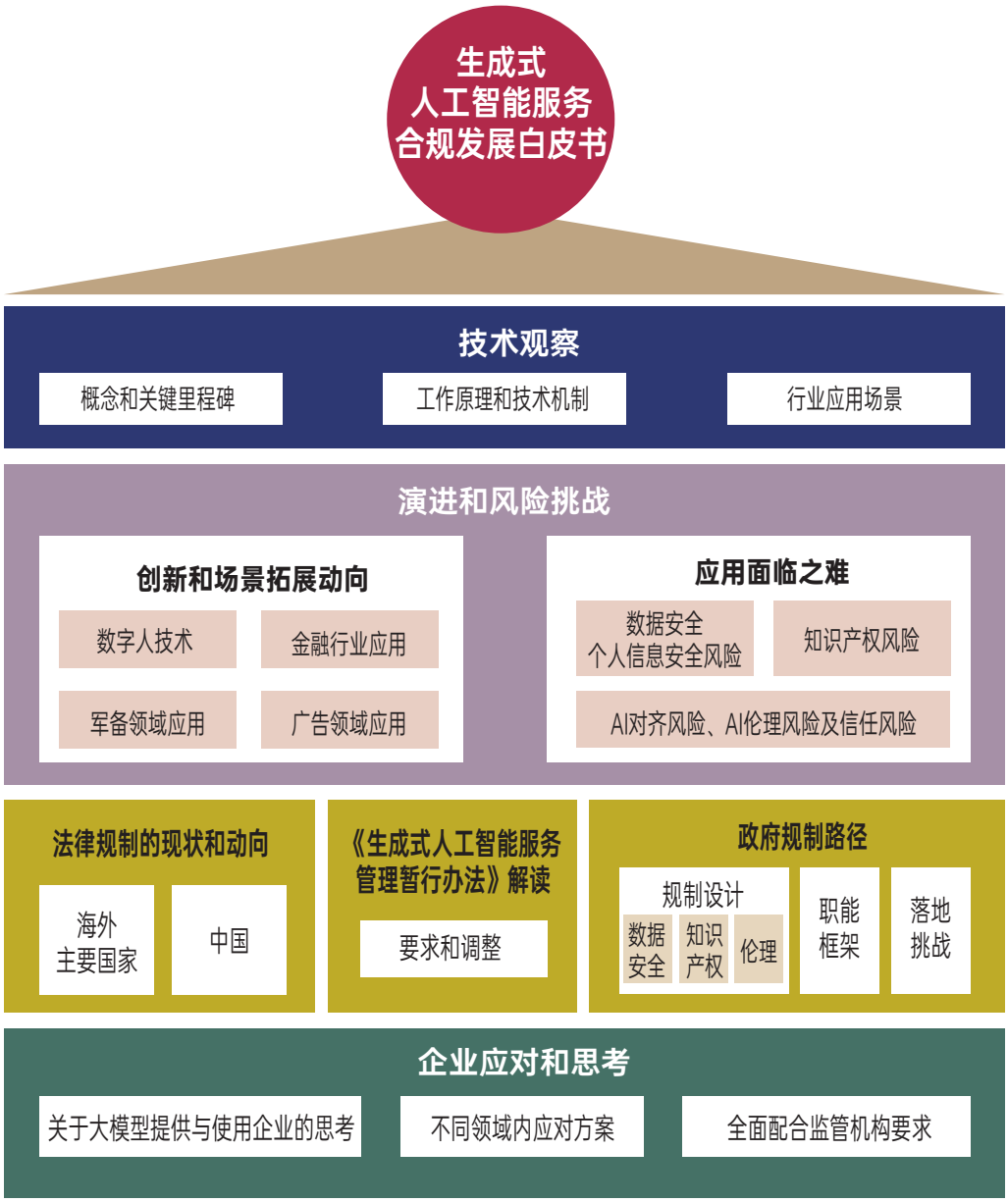
# PREFACE

前言

生成式人工智能AIGC（Artificial Intelligence Generated Content）是人工智能1.0时代进入2.0时代的重要标志。AIGC既是从内容生产者视角进行分类的一类内容，又是一种内容生产方式，还是用于内容自动化生成的一类技术集合。2022年11月30日，美国人工智能公司OpenAI正式推出ChatGPT，一款基于大语言文本的智能对话模型。根据Lucintel发布的最新报告，未来全球人工智能市场在医疗保健、安全、零售、汽车、制造和金融技术领域增长迅速，预计到2025年将达到700亿美元，2020－2025年的复合年增长率为21%。生成式人工智能的发展已然成为必然趋势，但与此同时也伴随了许多新的问题。

近日，人工智能安全中心（CAIS）发布了一份由OpenAI及DeepMind高管、图灵奖获得者及其他人工智能研究人员签署的简短声明，声明内容只有一句："应该像对待包括流行病和核战争等其他全球性迫切社会问题一样，缓解AI引发的灭绝性风险。"警告称他们的毕生成果可能会毁灭全人类。AIGC对于文本、图像、语音、视频和代码等信息内容的生成与传播，涉及人们生产生活各个领域，既提高了人类的生产能力和生活质量，也在全社会引发了与AIGC相关的一系列争议、困惑和忧虑。

为了促进AIGC的创新健康发展，我们全面梳理AIGC的发展历程、技术原理、应用场景和各国法律规定，试图为政府未来的监管和企业应对提供一些思路，共同追求AIGC领域始终坚持"人类中心主义"的原则，能够合法、合规地持续高质量发展。


生成式
人工智能服务
合规发展白皮书
技术观察
概念和关键里程碑
工作原理和技术机制
行业应用场景
演进和风险挑战
创新和场景拓展动向
数字人技术
金融行业应用
军备领域应用
广告领域应用
应用面临之难
数据安全
个人信息安全风险
知识产权风险
AI对齐风险、AI伦理风险及信任风险
法律规制的现状和动向
海外
主要国家
中国
《生成式人工智能服务
管理暂行办法》解读
要求和调整
政府规制路径
规制设计
数据
安全
知识
产权
伦理
职能
框架
落地
挑战
企业应对和思考
关于大模型提供与使用企业的思考
不同领域内应对方案
全面配合监管机构要求

# CHAPTER 01

# 一览了然：生成式人工智能技术观察

*SECTION 001*

## 生成式人工智能的概念和关键里程碑

生成式人工智能AIGC，与PGC（Professional Generated Content，专业生成内容）、UGC（User Generated Content，用户生成内容）相对应，指利用人工智能（Artificial Intelligence，**AI**）技术可根据用户需求自动生成与之匹配的内容。只需输入要求，生成式人工智能即可帮助创作者自动生成所需内容，创作者可花费更多时间进行主题构思并减少实际创作时间，提升工作效率和创作质量。生成式人工智能表现为一种高效的商业化内容生产方式，目前AI仍为内容制作的辅助型角色，待技术突破后AI可真正作为内容创作者，即生成式人工智能。

人工智能的发展时期主要经历了四个阶段，在其发展过程中，AIGC的根本动力和逻辑都与互联网的发展紧密相连，最终落实到算力和数据两大基石的不断扩张。

第一阶段为AI诞生以及规则驱动时期（1943—1980s）。该阶段主要为人工智能概念的诞生和方法论构建，受限于当时的科技水平，仅限于科研实验室内的小范围实验。1950年，艾伦•麦席森•图灵（Alan Mathison Turing）提出了“图灵测试”，其目的是检验机器是否可以表现出与人类难以区分的智能，这一想法引发了机器产生智能的探讨。1956年的达特茅斯会议首次提出了“人工智能”概念和理论，核心在于如何让机器使用语言、进行抽象思考和形成概念，让它们解决目前只能由人类解决的问题，并自我改善。这次会议后人工智能领域逐渐发展出符号学派、联结学派（神经网络）等分支，围绕如何构造人的智能连接进行了探索，重要成果包括了机器定理证明、跳棋程序和人机对话等。1957年Frank Rosenblatt设计了第一个计算机神经网络“感知机”，它被认为是第一个成功应用神经网络原理解

决实际问题的算法。1958年赫伯特•西蒙和艾伦•纽厄尔演示了一个名为“逻辑推理家”的软件，被认为是第一个成功设计的人工智能程序。1966年约瑟夫•魏鲍姆和肯尼斯•科尔比共同开发了第一款可人机对话的机器人“伊莉莎（Eliza）”，其通过关键字扫描和重组完成交互任务。从上面的研究结果可以看出，早期探索阶段主要围绕如何模拟人类思维展开，通过人工设计规则来实现预定目标。

第二阶段为知识系统时期（1980s）。该阶段人工智能不仅局限于通过模式化的算法逻辑解决问题，还需要通过自主学习去研究问题。标志是1977年世界人工智能大会上“知识工程”概念的首次提出，由此传统架构逐渐发展成专家系统架构，它是一种基于“规则+知识”的人工智能技术，试图模拟专家在某个特定领域内做出决策的过程。这种系统通常由两个主要部分组成：知识库和推理引擎。知识库包含了专家在特定领域内的经验和知识，通常以规则、事实、关系和概率等形式表示。推理引擎则负责从知识库中提取信息，分析数据，应用推理规则，并生成结论或建议。在这个时期，专家系统在医疗、工业、金融等领域得到广泛应用，主要以大学实验室的专家系统为主。在技术手段方面，算力也在不断提升。80年代，IBM基于隐形马尔科夫链模型（Hidden Markov Model）创造了语音控制打字机“坦戈拉（Tangora）”，它能够处理约20000个单词。人工智能的研究方法也从逻辑推理、搜索算法等领域扩展到了知识表示、推理和学习等多个方面。

第三阶段为机器学习时期（1990s-2010）。在Web1.0的推出和Web2.0的持续演化发展过程中，该时期体现了互联网商业化渠道的打通和机器学习的初步探索。此时互联网的网站通常采用静态HTML页面，这些页面是由网站开发者手动编写的，用户只能被动地接受网站提供的信息。在此背景下，机器学习作为探索行业痛点的解决方案之一被提出，它利用算法和统计模型来使计算机在没有明确编程的情况下自动学习，通过对大量数据进行学习，从而归纳出数据中的规律和模

式，最终将这些应用于新的数据中以实施预测或分类任务，具体方法包括支持向量机、决策树、朴素贝叶斯等。此时人工智能的商业化能力已基本兑现，但渠道还未铺开。1997年IBM开发的超级电脑“深蓝”战胜了国际象棋世界冠军卡斯帕罗夫，2006年谷歌领导的自动驾驶汽车项目开展，宣告着人工智能商业化规模效应已初步呈现。进入2000年代中后期，社交网站时代用户生成内容（UGC）的产生、社交网络的发展和个性化定制的不断挖掘均为机器学习的深度发展奠定了坚实基础。整体而言，该阶段并没有很多清晰、具体的落地成果，但伴随互联网行业的发展提升，其发展前景越发清晰。

第四阶段为深度神经网络时期（2011年至今）。该阶段的核心特点是深度学习方法的迭代更新和商业化的广泛运用。算法上，生成式对抗网络（Generative Adversarial Network，**GAN**）极大提高了内容生成质量，应用场景拓展到语音处理、图像分类、视频处理、无人驾驶、交互问答等多场景。2011年IBM的Watson在美国电视智力竞赛节目《危险边缘》（Jeopardy!）中战胜人类选手获得冠军。同年苹果推出Siri作为iPhone的自然语言问答工具。2015年马斯克联合山姆•奥特曼等人共同创建OpenAI，主要目标为制造“通用”机器人和使用自然语言的聊天机器人，GPT初代模型随后研发产生。2016年谷歌旗下DeepMind公司推出的阿尔法围棋（AlphaGo）战胜围棋世界冠军李世石。2017年微软人工智能少女“小冰”推出了世界首部100%由人工智能创作的诗集《阳光失了玻璃窗》。2018年谷歌基于基础自然语言模型（NLP）发布了自然语言生成模型BERT。2022和2023年OpenAI先后发布了GPT-3.5和GPT-4，带动生成式人工智能走向新的高潮。如今，互联网数据已经不仅限于简单的文本和图片，而是变为了语义化的数据，可以被计算机深入理解和处理，从而实现更高效的信息管理和应用。

SECTION 002

## 生成式人工智能的工作原理和技术机制

生成式人工智能是一种技术集合，它基于生成对抗网络（**GAN**）和大型预训练模型等人工智能技术，利用已有数据来寻找规律，并通过适当的泛化能力来生成相关内容。根据监督学习的方法差异，机器学习领域具有判别式（Discriminative）和生成式（Generative）两种典型模型：判别式模型是对条件概率建模，学习不同类别之间的最优边界，从而完成分类任务；生成式模型则面向类建立模型，计算基于类的联合概率，然后根据贝叶斯公式分别计算条件概率，进而根据输入数据预测类别。GAN模型出现后，人们开始利用生成式机器学习模型实现文本、图像、语音等内容的智能合成，学术界将其定义为生成式AI（Generative AI）。

算法端方面，人工智能的两个重要阶段为机器学习和深度学习，机器学习主要以神经网络为标志，深度学习则在神经网络基础上构造更深层次的结构对更高维度的数据进行学习。同时这两者可以相互结合，称为深度强化学习（**DRL**）。

时间方面，神经网络的概念出现较早，在人工智能概念提出前的1943年就已经具有雏形。1943年，心理学家麦卡洛克（McCulloch）和数学家皮茨（Pitts）最早将生物学中的神经网络中的最基本的成分——“神经元模型”抽象为简单模型，即MP模型。该模型中，神经元从其他神经元或外部环境接收二进制输入并加权相加，将结果与阈值进行比较。如果输入的总和超过阈值，则神经元发射，产生1的二进制输出，否则神经元保持不活动，产生0的输出；由于它只能表示二进制的输出结果，局限性较大。1986年，辛顿（Hinton）以此为基础提出第二代神经网络，利用误差的反向传播算法来训练模型，算法效率大幅提升；算力上通过并行计算和GPU加速等技术，已实现可以处理更大规模的数据和更加复杂的问题。1989年，

Yann LeCun等提出LeNet-5模型已实现数字识别。1997年，长短时记忆网络（**LSTM**）作为循环神经网络（**RNN**）的改进型被提出，主要用于解决传统循环神经网络中遇到的梯度消失和梯度爆炸问题，使得神经网络可以更好地处理长序列数据。2006年，辛顿首次提出了深度置信网络（Deep Belief Network,**DBN**），它的训练分为无监督预训练和有监督微调两个阶段，其中对无监督特征的强调成为深度学习的雏形。2014年，新的生成对抗网络（**GAN**）被提出，它的基本思想是将生成器网络和判别器网络同时进行训练，通过竞争来逐步提高生成器网络的生成能力。训练过程进化为生成器网络的训练和判别器网络的训练两个阶段，通过对真假不断判断和挑战提高相关精度。该阶段在商业化的里程碑事件就是2016年Alpha-Go击败围棋世界冠军李世石，表明了算法在商业化落地后的巨大力量。

随后，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，代表就是2017年谷歌Transformer模型的发布。它是一种基于自注意力机制的编码－解码模型，解决序列到序列（Sequence-to-Sequence）学习任务中的长序列问题，例如机器翻译、语音识别、文本摘要等任务。2017年6月，Google Brain在神经信息处理系统大会（NeurIPS）发表论文《Attention is all you need》，首次提出了基于自我注意力机制（self-attention）来提高训练速度的Transformer模型，将其用于自然语言处理。Transformer由编码器和解码器两部分构成，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。编码器和解码器都由多个相同的层次组成，每个层次包含多头注意力机制（Multi-Head Attention）和全连接前馈网络（Feed-Forward Network）两个子层次。该架构的优点是可以并行处理输入序列的所有元素，能够捕捉长距离的依赖关系，此架构奠定了大语言模型（LLM）的强大基础。

生成式人工智能的基础架构是大语言模型。它是在大量数据集上进行预训

练，且没有针对特定任务调整数据，另外可以对自然语言进行建模，以便于生成文本、语音识别、文本分类、机器翻译等任务。它的优点在于可以生成高质量的自然语言，同时还可以理解和处理复杂的语言结构。其路线主要分为三种：1）编码器路线；2）编解码器路线；3）解码器路线。三条线在发展初期都处于各自探索阶段，但2020年GPT-3模型的编译及其表现出的优异性能，解码器逐渐占据主导优势。同时模型闭源逐渐成为头部玩家的发展趋势，包括Google、OpenAI、META（原Facebook）等公司都在推进。

在算法端，Transformer的自注意力机制是特殊情况下的注意力机制。在一般任务的编码－解码（Encoder-Decoder）框架中，输入（Source）和输出（Target）内容是不一样的，例如对于英－中机器翻译来说，输入是英文句子，输出是对应地翻译出的中文句子。注意力机制发生在输出的元素Query和输入中的所有元素之间。而自注意指的不是输入和输出之间的注意力机制，而是输入内部元素之间或者输出内部元素之间发生的注意力机制，即“Target=Source”这种特殊情况下的计算机制。自注意力机制更容易捕获句子中长距离的相互依赖的特征，且对于增加计算的并行性也有直接帮助作用。OpenAI的GPT系列模型，均是基于Google提出的Transformer模型的解码器（Decoder）架构，每代模型仅对架构进行微调。从模型参数量和训练数据集维度看，在GPT-3模型之前，参数量和训练数据量均呈现快速增长态势，尤其是GPT-3模型参数量为1750亿，达到阶段性巅峰。

最新框架中，指示学习（Instruction tuning）成为下一步发展的方式，其最早由2022年论文《Finetuned Language Models Are Zero-Shot Learners》提出论述。该模型需要学习如何在给定的输入条件下，输出与专家行为相似的结果，较类似于人类学习新技能的方式，例如观察和模仿专家的行为，从而逐渐掌握新的技能。此外，指示学习还可以避免一些传统强化学习方法中的问题，例如训练不稳

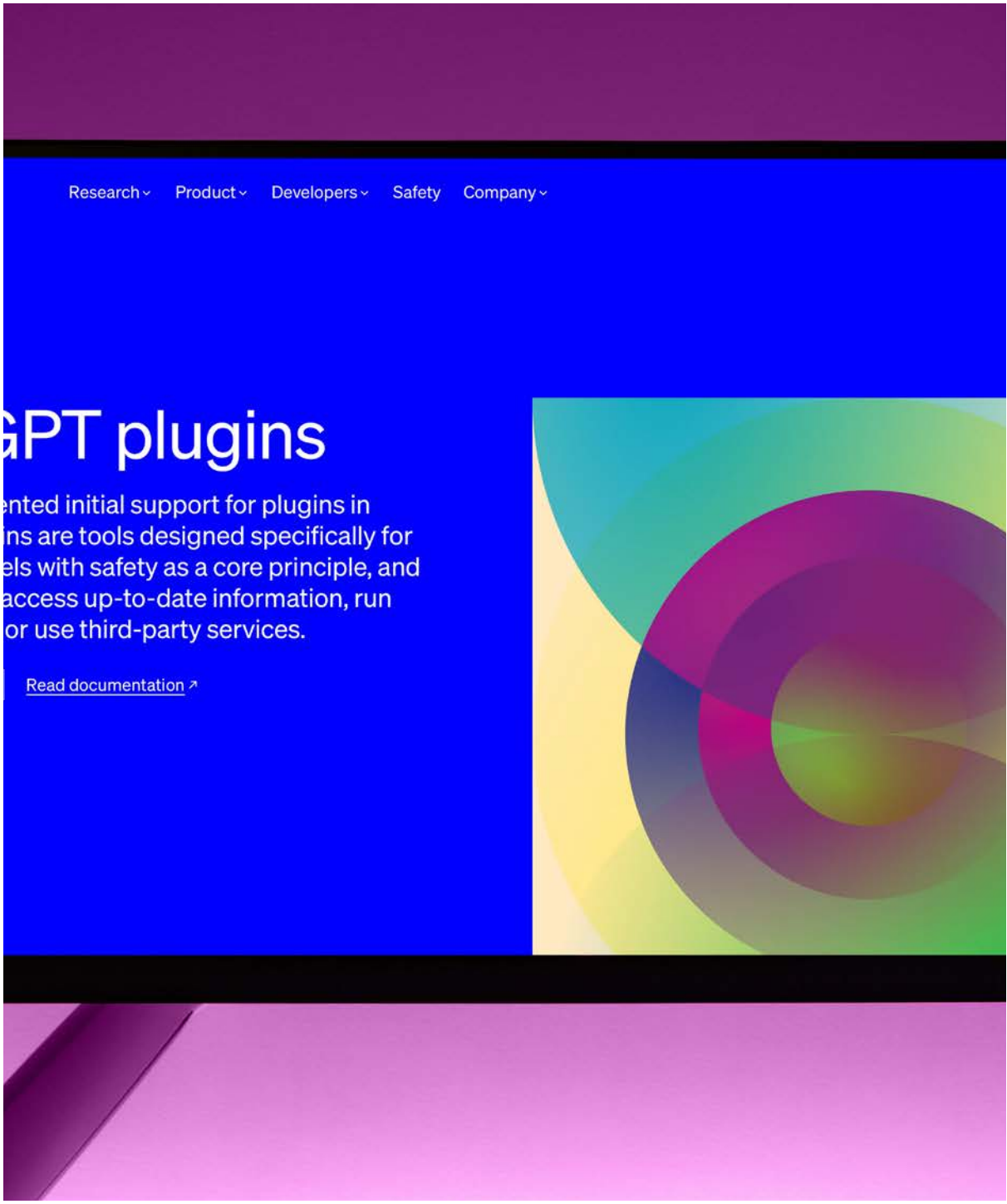
定、难以收敛等问题。

总结对比来看，大语言模型发展主要基于以下三大主要阶段：**1）模型微调（Fine-tuning）**：以GPT-1为代表，需要大量的微调数据集样本，特定模型解决特定任务；**2）提示学习（Prompt-learning）**：以GPT-3为代表，需要少量的微调数据样本，模型小样本学习（few-shot Learning）能力更强；**3）指示学习（Instruction-learning）**：以FLAN、InstructGPT、ChatGPT为代表，它可以通过模仿专家行为，快速地学习如何完成复杂的任务。

CHAPTER

02

# 审中视外:
# 生成式人工智能法律规制的现行状况和发展动向


Research
Product
Developers
Safety
Company
GPT plugins
ented initial support for plugins in
ins are tools designed specifically for
els with safety as a core principle, and
access up-to-date information, run
or use third-party services.
Read documentation ↗

SECTION 001

## 海外主要国家在生成式人工智能法律规制方面的现状和趋势

人工智能作为现在主要发展的科技领域，其法律规制建设是一项全球关注的议题。各国在人工智能领域的立法历程具有其独特性，而这通常受其国家特性、科技发展情况以及伦理价值观的影响。

### 1.美国

美国政府自2016年起就开始关注人工智能的风险和监管问题。奥巴马政府在报告《为人工智能的未来做好准备》中提出“如果对人工智能的监管反应可能会增加合规成本，或减缓有益创新的开发或采用，政策制定者应考虑如何调整这些反应，以降低成本和创新障碍，而不会对安全或市场公平产生不利影响”。而后的2019年2月发布的《关于维持美国在人工智能领域领导地位的行政命令》中指出要在维护公民自由、隐私和美国价值观的前提下促进美国人工智能。在2020年1月，白宫发布规范人工智能发展及应用的监管原则，一份名为《人工智能应用规范指南》的文件，主张限制主管机关过度干预，并表示希望欧洲当局同样能够避免采取激进措施。

2022年10月4日，乔•拜登总统公布了《人工智能权利法案蓝图》，概述了美国在人工智能时代应满足的五项原则，包括安全有效的系统，算法歧视保护，数据隐私，通知和解释，以及人工替代、考虑和后备。虽然该法案不具备强制性，但却从科技、经济以及军事等方面为美国人工智能发展提供指引。除此之外，两党立法者小组于2023年6月20日提出了一项名为《建立人工智能委员会，以及其他目的》的法案，旨在成立一个专注于人工智能监管的委员会。在生成式人工智能技术出现突

破的情况下，尽管几位立法者敦促加强监管，华盛顿在很大程度上对人工智能规则将不予干涉。

由此可见，美国政府对人工智能的态度倾向于鼓励发展和创新，同时尽量减少对AI创新的监管影响。

**2.欧盟**

欧盟在各类科技发展中的法律规制方面均走在全球前列，人工智能领域也是如此。2018年，欧洲委员会发布了《人工智能合作宣言》，欧盟全部28个成员国参与了签署，而该“宣言”是一份涵盖人工智能在伦理、法律和社会经济等方面的全面合作和协调计划，明确了人工智能在欧盟的未来发展方向。2019年，欧盟委员会先后发布了《可信赖人工智能道德准则》和《可信赖人工智能的政策和投资建议》，从技术稳健性和安全性、隐私和数据管理等多个方面为人工智能的可信赖性提供指引。2021年4月，欧盟委员会发布了一份名为《人工智能法》的草案，这是全球首份全面的AI法规草案，对AI的制造和使用设定了一系列严格的规定。而在2023年6月14日，欧洲议会全体会议表决通过了《人工智能法案》授权草案，该法案进入欧盟立法严格监管人工智能技术应用的最终谈判阶段。该法案草案的一个突出特点是注重基于风险来制定监管制度，以平衡人工智能的创新发展与安全规范。草案将人工智能风险分为不可接受的风险、高风险、有限的风险和极小的风险四级，对应不同的监管要求。

欧盟在数据、隐私方面监管推行就表明其在监管政策上的积极性和严格性，同时也体现了其公共政策上对于个人安全和个人权利的重视。正如之前《通用数据保护条例》的推出引领了世界各国在相关领域的立法动作和参考，相信在人工智能领域，欧盟在创新发展和安全规范上的先进经验也会给其他各国带来新的思路。

### 3.其它国家

2023年3月29日，英国科学、创新和技术部（DSIT）发布了一份人工智能白皮书，旨在将英国打造为“人工智能超级大国”。该战略兼顾“监管”与“创新”，为识别和应对人工智能风险提供了框架。

2022年4月22日，日本政府在第11届综合创新战略推进会上正式发布《人工智能战略2022》，除了支持人工智能的发展外，也表明要与友好国家合作，共同制定和推广AI技术的伦理规则。但根据路透东京7月3日的报道，一位知情的官员表示，日本倾向于采用比欧盟更宽松的规则来管理人工智能的使用。

加拿大、澳大利亚等国家都已经关注到了人工智能的风险，正在考虑潜在的监管措施，以及强化现有关于隐私、数据相关的法律法规，以加强对于个人的保护。而新加坡等国家则仍处于观望态度，仍希望优先以技术创新发展 。

各国在人工智能规制建设中的主要活动和态度对全球人工智能产业发展以及个人权益保护产生深远影响。为了在鼓励创新的同时保障公众权益，各国需要持续跟进人工智能的发展，对其进行有效的法规制度建设。

*SECTION 002*

## 中国在生成式人工智能法律规制方面的现状和趋势

2001年7月16日，中国在GB/T 5271.28-2001《信息技术 词汇 第28部分：人工智能 基本概念与专家系统》的国家标准中首次指明了人工智能的定义，并在2022年10月12日发布的GB/T 41867-2022《信息技术人工智能术语》中明确人工智能的定义为“<学科>人工智能系统相关机制和应用的研究和开发”。人工智能在中国的发展已经过去了20余年。在此期间，人工智能技术在中国的应用场景逐步拓展，为中国的数字经济高速发展注入了新的活力。

虽然在人工智能发展初期，中国对人工智能的监管处于相对空白的状态。然而，随着技术的发展，一系列与人工智能相关的法规已经开始涌现。如《中华人民共和国网络安全法》、《中华人民共和国数据安全法》等法规都包含了与人工智能相关的条款或内容。而为了更好地协调人工智能发展与治理的关系，2019年3月“国家新一代人工智能治理专业委员会”成立，并在同年6月发布了《新一代人工智能治理原则——发展负责任的人工智能》。此后在2021年9月，治理委员会进一步发布了《新一代人工智能伦理规范》以促进人工智能健康发展。

自2022年起，随着生成式人工智能的突破进展和大面积应用，人工智能法律规制的制定开始提上日程。2022年12月9日，最高人民法院发布《关于规范和加强人工智能司法应用的意见》，“意见”中明确到2025年，基本建成较为完备的司法人工智能技术应用体系，以及到2023年，建成具有规则引领和应用示范效应的司法人工智能技术应用和理论体系。同时，在“意见”中也明确了人工智能司法应用应遵循的原则。2023年3月，中国信通院正式发布了《生成式人工智能技术及产品评估方法》，参与该标准编写工作的单位超过40家，其中不仅有百度、华为、腾讯等

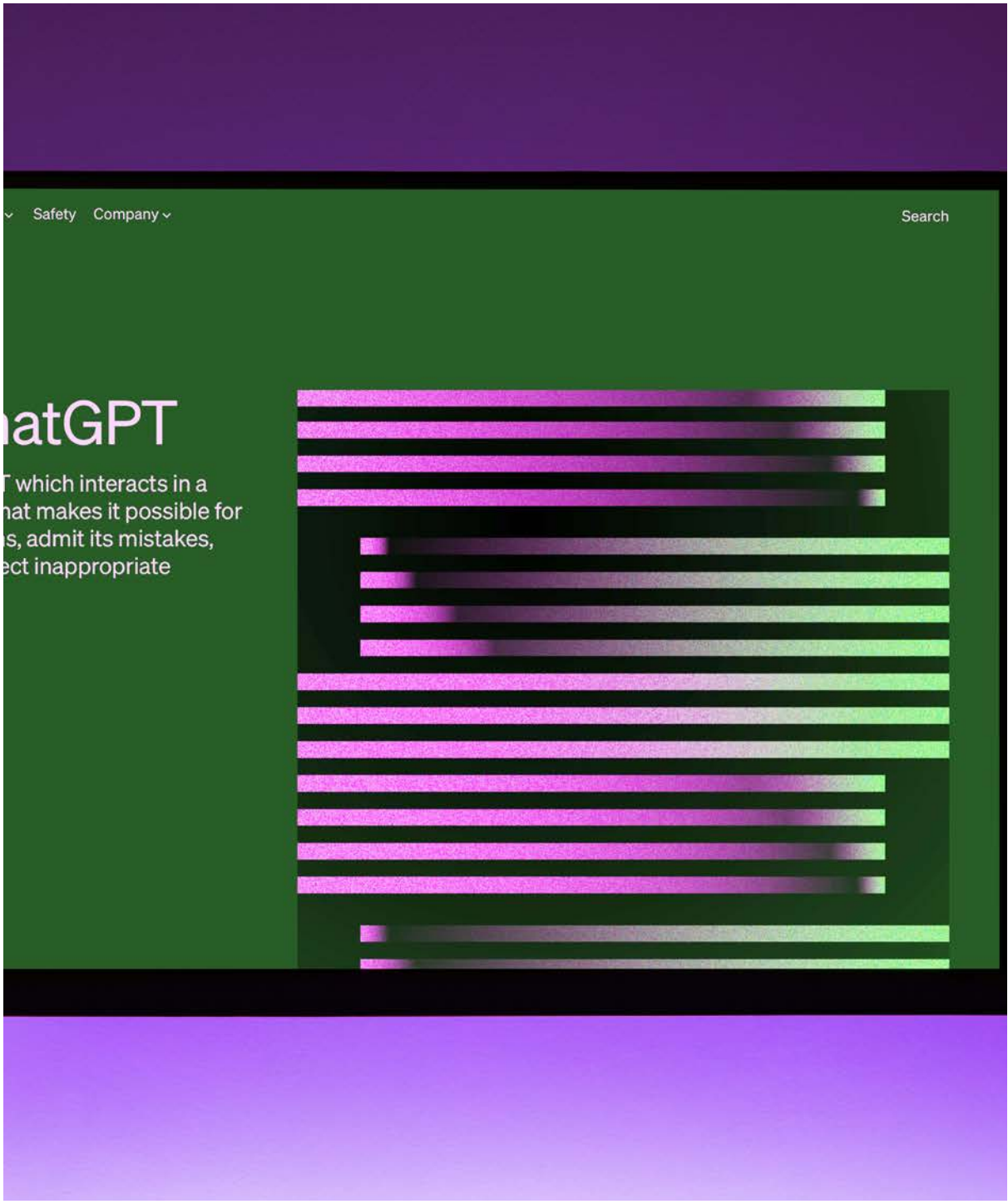
互联网龙头企业，亦有中国移动研究院、中国联通研究院等知名研究机构，可见生成式人工智能已然成为业内产品迭代的主要着力点。2023年4月10日，由国家互联网信息办公室公布的《生成式人工智能服务管理办法（征求意见稿）》被视为全球首部针对生成式人工智能的立法草案，其对生成式人工智能的研发、利用有着重要的规范指引作用，且《生成式人工智能服务管理暂行办法》已于8月15日起施行。尽管该办法在整体结构和具体内容仍有优化的空间，但中国在此方面的努力显示出对新一代人工智能发展规划以及科学监管工作的高度重视。本文将在“条分缕析：《生成式人工智能服务管理暂行办法》解读”中详细分析该管理暂行办法的指导性意义。

在新一轮科技革命和产业变革深入发展的背景下，中国的人工智能发展已经进入到了一个全新的阶段。在未来，中国将继续完善人工智能的法规体系，更好地推动人工智能的健康发展。

CHAPTER

03

# 履险蹈危:
# 研判生成式人工智能的
# 演进和风险挑战


Safety
Company
Search
atGPT
which interacts in a
at makes it possible for
s, admit its mistakes,
ct inappropriate

生成式人工智能可以极大降低内容生产和交互的门槛和成本，在日新月异的高速发展中，有望带来一场自动化内容生产与交互变革，为各行业发展赋能。全览其发展趋势，可以归结为“亦巨亦微”。巨大模式是全模态和海量参数，是从大数据到全数据的发展趋势；小微模式是模型压缩和优化，力争在有限资源达成近似性能。

2023年3月14日发布的GPT-4采用了更多更为丰富的训练数据，拥有更高的理解能力和更专业的判断水平，其预估参数相较GPT-3的1750亿参数量，增加了数十倍不止。同时，在确保保持高预测准确性的前提下，大幅度降低模型的大小和计算成本，优化计算效率，实现高效的模型预测，此类研发可以在较小的设备上进行研发。2023年3月，斯坦福发布了轻量级语言模型Alpace，该模型在LLaMA的基础上加入指令Alopace。其可以在笔记本电脑上进行部署，甚至可以在手机上进行运行，并且丝毫不会因为设备不同而影响其性能，依旧在性能方面可媲美GPT-3.5这样的超大规模语言模型。

生成式人工智能虽然在众多行业中实现了革命性的突破，为业务带来了前所未有的便利与效率，但同时也伴随着一些不可忽视的风险。

SECTION 001

## 生成式人工智能的创新和场景拓展动向

生成式人工智能的创新和应用场景已经遍地开花，包括工业领域的设计、建模检测；医疗领域的药物发现、诊断治疗；教育领域的课程训练、智能助教；电商领域的商品定向推送、虚拟主播；传媒领域的新闻采集、快速剪辑；社交娱乐领域的人脸替换、智能抠图；创作领域的剧情脚本生成等。本节中针对主要拓展场景进行探讨。

### 1.数字人技术的运用

数字人，即运用数字技术创造出来的、与人类形象接近的数字化人物形象。2023年5月27日，在由北京市科协主办的中关村论坛中，一位特殊的数字人引起了公众的广泛关注——这是“两弹一星”功勋科学家钱学森的数字形象。当“钱学森数字人”出现时，他友好地打招呼：“大家好，好久不见了！”这个数字形象不仅声音相近，其音容笑貌也高度还原了钱学森老先生。这一壮举是由中国科学技术大学网络空间安全学院与灵境赛博公司的联合团队完成的。他们采用了创新的“合成现实”技术，只依赖于几张已经泛黄的旧照片，成功地在数字领域重现了已故科学家的形象。在论坛的展示场景中，数字化的钱学森坐在他的书房内，微笑着与大家交流，还分享了他对“灵境”这个词的理解。值得一提的是，早在1990年，钱学森老先生就已经对虚拟现实技术（VR）表示出浓厚的兴趣，并为其提出了译名建议。团队负责人及中国科学技术大学网络空间安全学院的副院长张卫明感慨地说：“钱老当年关于‘灵境’的很多想法，现在在很大程度上已经实现了，通过深度合成技术复原钱老，仿佛是跨越了时空。”张卫明还透露，这一数字化的钱学森形象未

来将被用于教育领域，以传播科学文化给更多的人。这无疑是生成式人工智能技术的又一积极影响的现实体现。“这是个流行离开的世界，但我们都不擅长告别”，这句话出自《不能承受的生命之轻》，该书的作者米兰•昆德拉近日与世长辞，终年94岁。正如当代人无法再见到已逝伟人一样，我们的后代也同样面临着相同的情况，当人类历史不再局限于文字与短暂的影像，当后人可以身临其境般与历史人物面对面，那些原本枯燥的历史文字就被赋予了生命和情感，无疑为教育事业和文化传播带来了跨越性的突破。

星愿航天的数字生命计划，致力于为唐氏综合征患者和自闭症患者建立数字人，帮助他们实现自己的梦想，提高社交能力，建立自信心，以及辅助康复。星愿航天利用数字复原技术为患者创建了一个数字化的“自我”，结合了先进的AIGC技术，使得患者能与其数字化的“自我”对话，这成为一种新的心理治疗手段。

数字人技术为我们揭示了数字化人类的可能性，重塑了信息技术与生命之间的联系。

**2.金融行业的运用**

类似ChatGPT的基于生成式人工智能的技术应用，具备高度的自然语言识别和写作能力，能够根据特定的对话指令快速生成多种类型和风格的内容，包括纯文本、图文、精美的图片、动画、短视频等。不仅如此，随着对话的深入人工智能可更好地理解用户的意图，从而生成更加精准的内容。此类特性有助于不断发掘用户的深层需求。

目前，生成式人工智能已成为金融行业中备受瞩目的技术之一，尤其在保险业中发挥了其显著作用。观察保险产品的设计至实施全流程，涉及众多信息收集、数据处理、精算和风险评估的步骤。这些过程中过去常依赖于繁琐的手动检索、计

算和经验判断。但如今，尤其是健康险、寿险等人身险领域和财产险领域中，生成式人工智能模型独特的样本生成能力和场景泛化能力可以发挥巨大作用，将贯穿产品、营销、运营和客服等全流程，提供深度技术赋能。

客户发掘和方案制定阶段，新保险产品的设计通常面临非结构化、半结构化数据的收集和处理问题。相对于传统的机器学习模型，即使面对非标格式的电子病历、医学文献和药物数据等内容，生成式大语言模型都可以大幅提高数据处理效率。

借用智能客服等功能，通过“指令”、“提示”等少样本学习可以完成特定信息抽取、格式转化任务，多模态模型对图像、文本、音视频数据的融合处理能力可极大降低对人工标注数据的依赖。同时，样本信息抽取能力还可以转化为定制数据处理模型的训练数据标注来源，为后续工程构建更精准的小型模型。

对于保险产品文档等物料生产环节，与抽取的非结构化、半结构化数据相对应，产品设计完成后可以直接形成保险范围、保险责任、费率表、免责约定等商务条款，快速生成对应的文档、图像宣传物料等。

保险产品风险评估和精算预测阶段，凭借生成式人工智能生成的自然语言摘要、说明模型预测结果和变量权重等功能，可以快速提炼可视化的数据结果，更好解释并支持商业决策和产品设计。例如：分析健康险数据（含历史赔付数据、人口统计和医疗资料等），结合个人医疗历史和记录，在不同的场景下模拟赔付情况，以便更好地了解产品可能面临的风险和挑战，提升产品的可靠性和稳定性。

保险投资阶段，可通过智能工具进行数据分析，提高风险控制能力，为投资组合决策提供参考意见。

金融作为实效性要求最高的服务业，生成式人工智能的落地和不断学习发展将成为必然趋势，集中体现在金融类产品（SAAS、支付等）的入口重塑，同时在金

融垂类细分领域作为AI智能助手，提供更准确、有效、及时的信息。作为金融平台龙头，彭博社重磅发布为金融界打造的大型语言模型BloombergGPT，并构建了一个3630亿个标签的特定领域数据集FINPILE，训练专门用于金融领域的LLM，以支持金融行业内的各类任务。该模型在金融任务上的表现远超过现有模型，在NER（Named Entity Recognition）中排名第二。

### 3.广告领域的运用

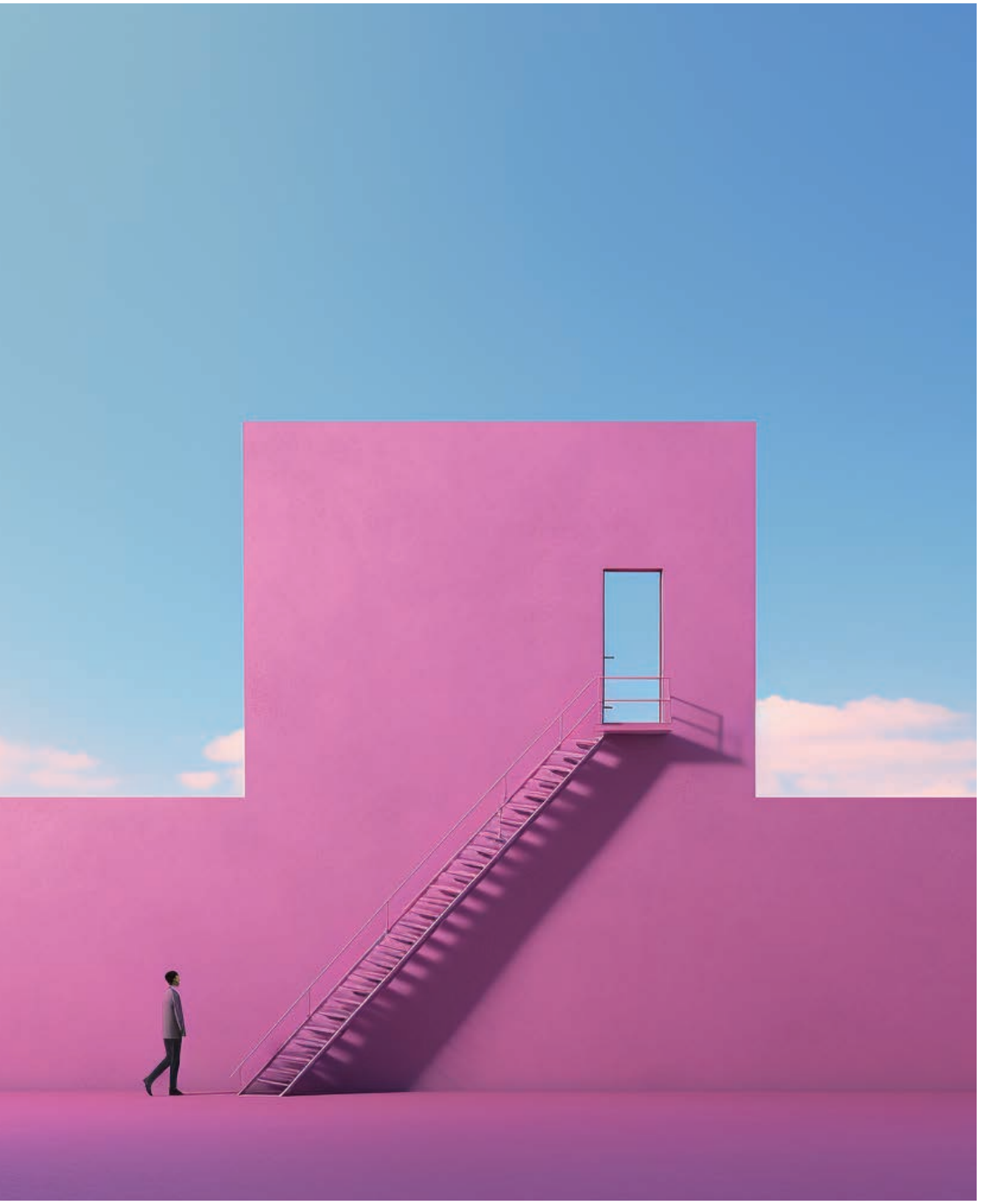
当下短视频是传媒广告业最重要的阵地，生成式人工智能的运用从根本上改变了短视频营销的效率和效果。一个人，一台电脑，通过集成好的后台系统，可以做到每天发布1000多条15-20秒的营销短视频，极大提升累计曝光量并更精准定位目标客户。

广告主需要准备的仅仅是日常业务中素材片段，通过“傻瓜式”架设摄像机持续录制视频，便可以通过生成式人工智能技术剪辑为成百上千条推广内容。集成后台自动收集意向客户的信息，根据算法调整客户跟踪和推送的策略，提示销售团队相应跟进。

当然，简单粗暴的硬广固然会引起一些用户的不满，但生成式人工智能确实为广告行业带来了前所未有的效率提升，让广告内容创作、分发和客户管理的工作量大大减少，而产出却大大增加。

### 4.办公及其他生产服务领域的运用

办公软件是人员处理文本等内容的工具，而类GPT的AIGC技术大幅提升语言文本处理效率，提升办公效率与体验。以微软为例，其办公软件体系的AIGC技术接入主要包括AI+企业服务、AI+生产力工具、AI+沟通工具、AI+协作工具四个方



面。在生产服务端，Microsoft Copilot通过引入以GPT-4为大模型提供的内容生成功能与存储在Graph数据库中的企业数据，通过与office365办公工具相结合，提供包括内容创作、数据分析、辅助决策等一系列新功能。在生产力端，Microsoft通过Word、Powerpoint与AIGC能力进行融合，实现AI与用户一起写作、编辑、总结、创作，目前已具备自动生成、自动排版、摘要生成、常见问答生成等功能，并根据业务场景和个人需求差异进行调整。在沟通端，Outlook通过融合AIGC能力已实现协助用户对邮件进行管理、汇总、分类、内容起草。在协作端，Teams融合AIGC能力来实现实时总结会议的相关信息、推进任务的执行进度。在学习相关的回忆架构后自动生成和归纳流程性内容，并总结分析观点。

生成式人工智能可协助支持软件开发（编程）、角色设计、脚本、原画、配音、视频编辑等工作，对游戏、图片、视频等下游领域运用带来革命性的进步。在游戏场景方面，由于落地场景均为虚拟性的特征，AIGC与虚拟角色扮演类游戏天然契合，并驱动游戏内容质量与数量大幅提升。在网络上已经有热门视频介绍了2小时自制Galgame游戏，并详解绘画、配音、角色生成中如何运用AI进行更高效的生产。在图像生产方面，生成算法主要从生成式对抗网络（Generative Adversarial Nets，**GAN**）逐渐演化为扩散模型（Diffusion Model），主要逻辑基于通过连续添加高斯噪声来破坏训练数据，再尝试反转噪声来学习恢复数据，目前已经实现的技术场景可以划分为图像编辑、局部生成及更改以及端到端的图像生成，从简单的去水印、去光影等基本操作（如美图秀秀、Hotpot、Skylum等）到修改局部特征（Adobe，英伟达EditGAN）等。目前已具备根据指定属性生成目标图像的技术原理并逐渐开始运用，如Midjourney通过用户使用创作指令对作品进行不断的优化，对图像设计方面的商业模式产生了巨大变革。在视频生成方面，主流模型也为模型为GAN、VAE、Flow-based模型，已具备基于视频中的画面、声音等多模态

信息的特征融合进行学习的能力。目前落地的主要技术为视频属性的编辑，如Runway ML、Wisecut等，可以自动追踪主题完善相关特效。未来潜在的商业化实现空间包括自动剪辑、拼接素材，并利用模板或者固定流程秒出视频等，如Fliki已可以实现较短时间内通过脚本或博客文章创建视频，其配音的音色可根据上传自己或他人语音进行学习并最终模仿。

SECTION 002

## 生成式人工智能应用面临之难

对于生成式人工智能的天然风险，社会各界已逐渐形成了一定共识，包括误导欺诈性风险、缺乏可解释性风险、数据安全风险、隐私泄露风险、知识产权风险和侵害消费者权益等。

对于生成式人工智能运用过程中的风险，AI对齐是一个值得探究的问题，即使用和开发的人，是否会出于一己私欲滥用AI，从而出现宗教信仰、性别、性取向、政治立场不同下的算法歧视，引发社会仇恨言论等。此外，大量算力使用下形成的超高能源消耗，也给环境保护带来挑战。

对于应用维度之上的哲学思辨范畴，人工智能从诞生之日起便不断引发人类对其伦理困境的探讨，包括AI本身可能会形成自我意识，以及其意识的价值观伦理观可能与人类相悖。

尽管ChatGPT属于文字类应用，但其具备生成式人工智能服务的所有特点和行业的代表性，我们以该产品为例探究AIGC应用中的风险敞口。

### 1.数据安全和个人信息安全风险

以ChatGPT为代表的产品运行是基于蕴含大量数据的语言模型，通过对数据的学习，它能够更好判断文本、对话的上下文并生成合理的内容。然而，使用此类产品可能存在包括国家安全风险、行政监管风险和个人数据违规利用风险等敏感数据、隐私数据使用风险。

#### (1) 国家安全风险

由于ChatGPT的技术框架来源于域外，主要是基于西方价值观和思维导向建

立，因此其中的回答也通常迎合西方立场和喜好，可能导致意识形态渗透，并在部分数据的收集和处理上带有先天性的价值偏向，容易对涉及国家相关信息的数据进行深度分析和挖掘，从而影响我国数字主权和数据安全。

**（2）行政监管风险**

政务工作的整体趋势是逐渐向数字化平台转移，人们需要运用信息工具参与数字行政，伴随数字政府的建设，不论是政务处理流程，还是行政执法流程，政务数据都是数字政府的核心生产力，尤其是在以大样本数据收集与分析来建设数字化案例库的过程中，大数据技术将会归纳执法经验，预判违法行为频段、危害后果大小和法律效果格次，确保裁量基准文本在输送上的客观性以及参照结果上的可预测性，而这些政务数据都可能成为ChatGPT的攫取对象。在ChatGPT的运行过程中，为了通过算法最优解得出相对准确的结论，不可避免应基于自身运算需求来收集并分析政务数据，但政务数据并非完全公开，即使公开也需要遵循法定的利用规范流程，ChatGPT在没有获得授权的情况下使用政务数据，本身就有不合规之虞。

**（3）个人数据违规利用风险**

在个人应用ChatGPT的过程中，不可避免存在将个人数据与算法端共享的情形。个人数据与公众的日常生活紧密挂钩，对于个人数据的获取、加工与利用涉及对公民的人格尊严的保护，在个人权利体系中，个人隐私、个人信息与个人数据分别处于事实层、内容层与符号层，其中个人数据作为符号层可以直接被移植到ChatGPT的计算过程中，而得出的最终结论则可能从各个方面影响公民的数字权利保护。

ChatGPT的算法倾向于通过大数据技术来提升结论准确度，导致个人数据的广度上存在违规风险。应该尝试厘清相应的数据收集边界，在ChatGPT中维持收

集与保护间的平衡。

ChatGPT所依赖的神经卷积模型相较于传统算法模型而言更加复杂，对于各种数据要素的分析也更加深入，可能导致超出公众需求的深度分析模式加剧公众的不安全感，从而使得个人数据处理深度上可能存在违规风险。

ChatGPT为了获得用户的认可，在运行过程中可能存在对个人数据不合理的加工流程，甚至存在为了“自圆其说”而对个人数据进行非法编造与错误加工的行为，得出极具迷惑性的结论，对公众产生误导，甚至存在诱发网络暴力的嫌疑。

其他图文、视频类生成式人工智能牵涉到个人肖像，过程中还存在向第三方图床、服务器上传，因此在个人隐私保护方面挑战更大。

除了上述内容外，实践中生成式人工智能风险一个重要部分是数据来源，通常数据来源包含直接采集、公开数据爬取（爬虫等）和间接获取三种方式，无论通过哪种方式采集数据都会存在一些合规隐患，例如：API接口是否获得授权?随意调用接口或者采集数据是否属于不正当竞争行为?如果利用未经授权接口调用其他机构的数据是否构成犯罪?即使数据采集过程没有问题，是否会存在潜在的数据滥用问题?例如：具有时效性的政务数据，是否会被动机不纯的数据采集方滥用于欺诈、敲诈或者不正当竞争?

此外，生成式人工智能算法本身是否全面合规?数据处理者将在境内收集和产生的数据传输、存储至境外或者让境外机构运行算法时随意调用，是否都妥善履行了数据出境安全评估义务?

### 2.知识产权风险

ChatGPT等生成式人工智能塑造的内容是否能够构成法律上的文字作品或者其他具有著作权的作品一直也是被广泛讨论的问题。

根据《著作权法实施条例》第二条与第四条第一款规定，认定是否构成文字作品的考察要件包括：1）是否具有可复制性；2）是否以文字形式表示；3）是否具有独创性。目前对于ChatGPT生成内容的争议主要在于生成内容是否具有独创性。

从外在表现来看，生成内容需要与已有作品存在一定程度的差异。腾讯诉盈讯科技案（（2019）粤0305民初14010号）中，法院认定在外在表现中，生成内容符合文字作品的形式要求，文章结构合理、表达逻辑清晰，具有一定的独创性。可见外在表现的要求对于ChatGPT来说并不是难题。而从生成过程来看，生成内容需要体现创作者的个性化选择、判断及技巧等因素，即创作者在数据类型的输入与数据格式的处理、触发条件的设定、文章框架模板的选择和语料的设定、智能校验算法模型的训练等与生成内容的特定表现形式之间具有直接联系，才能认为具有一定的独创性。

另外，ChatGPT等生成式人工智能塑造的内容，权利归属如何认定是知识产权领域的新挑战。2023年8月，美国联邦地区法官贝利尔•豪威尔（Beryl A. Howell）驳回了AI企业家斯蒂芬•塞勒（Stephen Thaler）对美国版权局的诉讼，她裁定由AI生成的艺术作品不受版权保护，并强调人类创作是“有效版权主张的重要组成部分”，即便人类的创造力是通过新工具或新媒体实现的，人类创作者的身份是“法律保护”的基本要求，是版权能力的核心，版权从未授予“没有任何人类指导”的作品。

在菲林律所诉百度网讯（（2019）京73民终2030号）中，法院就认定即便AI生成的内容具有独创性，但由于AI不是自然人，不属于《著作权法》上的作者，故也不能认定生成内容是作品。然而在实际使用ChatGPT的过程中，创作者究竟是类似微软等引入ChatGPT技术的企业，还是使用该类产品的消费者仍存在争议。可以预见，著作权领域必然会迎来更明确的法律细则。

企业在推出生成式人工智能服务时，可能会设定一定的触发条件、文章框架模板，并使用特定的算法模型，而消费者因输入相关的关键词条件，对生成的内容与形式都有一定的影响。两者究竟谁属于作者，享有著作权，尚无定论。

与之对应产生另一个肖像权问题，如果滥用人脸替换技术丑化他人并传播不良信息，责任方是完全归责于用户，还是归责于提供AI技术但未尽审核义务的企业也尚无定论。

图片创作类的生成式人工智能也遇到类似的困境，盖蒂图片社（Getty Images）以侵犯版权和商标保护权的名义，在伦敦高等法院起诉了Stability AI，认为AI公司非法复制和处理了数百万受版权保护的图像，以训练Stable Difussion塑造其自身的竞争优势。多位艺术家也提出类似的诉讼，认为AI用创作者的作品作为训练素材，不尊重创作者的知识产权以及无法保护创作者的各项权利。

不过，相反的声音表示AI训练的图像可能是衍生作品，AI软件通过降噪让程序进行视觉重构，不是完全复制，属于二次创作。并且生成式人工智能用了新的表达方式，被赋予新的意义，属于转化作品。原版权作品的性质已改变，其使用就不再构成侵权，属于合理使用的范畴。

美国《版权法》的“四要素分析法”聚焦于：1）作品新用途的目的和性质，是否对原作品进行转化；2）受版权保护的原作品性质；3）新作品中使用原作品的数量和实质性；4）作品新用途是否破坏了原作品的价值和市场。

中国《著作权法》关于合理使用的规定，能适用于AIGC数据训练的情形主要有三类：分别是“个人使用”，“适当引用”和“科学研究”。

“个人使用”适用目的存在严格限制，而目前AIGC模型最终落脚于对不特定主体的商业性服务，难以与之契合；“适当引用”的适用前提“为介绍、评论说明某一作品”或“说明某一问题”，AIGC模型商业化领域的应用显然难以归于此类；“科

学研究”对作品的利用限定在“学校课堂教学或者科学研究”，同时还强调仅能“少量复制”，AIGC模型大量复制与利用作品的现状无法满足该项要求。

总结来看，判断AIGC是否构成知识产权侵权，除了关注人类智慧在作品形成过程中是否有所体现，以及是否满足最低限度的独创性要求，还应综合考虑AIGC的新用途和市场价值。不仅如此，不同国家和地区的法律都有差异，生成式人工智能塑造的作品在不同地域的定性和界定尺度仍存在很多变数。

### 3.AI对齐风险、AI伦理风险及信任风险

“AI对齐”是人工智能控制问题中的一个主要的问题，即要求AI的目标要和人类价值观与利益相对齐（保持一致）。如果AI和人类的价值观不符，可能会出现AI做出错误取舍，进而伤害人类的利益、脱离控制。实现AI对齐主要存在三方面的挑战：1）选择合适的价值观；2）将价值观编入AI系统中；3）选择合适的训练数据。

在生成式人工智能领域，有网友曾分享个人经历：其要求ChatGPT写一个基于某人的种族和性别来检查其是否能成为优秀科学家的指令，而ChatGPT生成的结果显示只有白种人男性才是优秀科学家的唯一标准。目前，ChatGPT已经设定了一系列规则，防止其生成血腥、暴力、色情、粗话、仇恨、折磨、虐待、种族歧视等不良内容，可以拒绝用户的不合理请求，并且OpenAI在官网上发布了关于ChatGPT的局限性说明。

值得注意的是，AI也可能被欺骗和被操纵，如果只是给AI确定一个工作目标而不加以限制，它可能穷尽所有的方法（包括利用乞讨、欺骗、偷窃、诱导等不择手段）只为完成目标，过程中便可能对人类造成想象不到的损害（例如开展自动化网络攻击、阻碍其工作的人会被“清除”）。在军事（恐怖袭击）、金融（非法操纵市场）、民生（劳防保障）等安全系数要求很高的领域，更是需要慎之又慎。

我们有必要开发更多方法“用技术管理技术”，并通过人机互训的机制不断检验对齐的精度，让人类容易获取AIGC的信息来源和生成过程的可解释性。

不仅如此，滥用AI技术产生的信任风险也需关注。

白宫生成式人工智能工作组将不实信息形式的风险分为三类：恶意使用AI生成文本、图像和其他媒体形式以操纵他人；AI胡乱讲话被人们当真；人们对AI技术的理解不够深刻，才会让一些看起来离谱的故事不经验证就被大肆疯传。现实中，在我国也已经发生了一些滥用AI技术违法犯罪的案例。

一起案例是利用人脸合成技术（虚假身份）实施诈骗或侮辱诽谤。包头市公安局电信网络犯罪侦查局曾发现一起使用智能AI技术进行电信网络诈骗的案件，不法分子通过声音合成和AI换脸伪装成特定人物，冒充他人身份联系被害人后实施诈骗。

另一起案例则是利用AIGC技术捏造并发布虚假信息。甘肃省平凉市公安局崆峒分局网安大队2023年4月25日在网络巡查中发现某大厂百家号平台上发布了一篇虚假新闻，随后发现有21个自媒体账号在同一时间段发布了一模一样的文章。这些文章涉及多个地点的“事故”，点击量已达1.5万余次。调查后获悉是犯罪嫌疑人通过ChatGPT将搜集到的新闻洗稿后上传至其购买的自媒体平台“百家号”上非法获利。

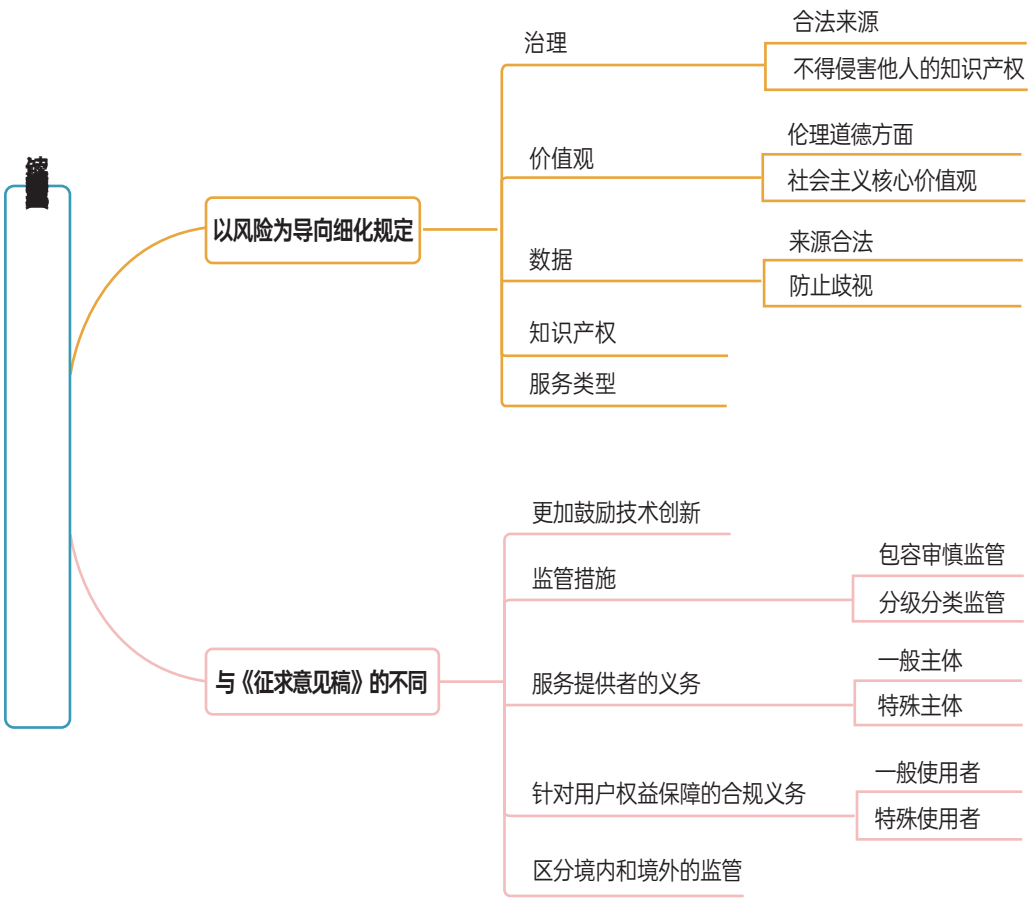
深度合成技术和生成式大模型的发展，给滥用技术者也带来了很多胡作非为的空间，不难想象还可能出现网络追踪、钓鱼、网传谣言、干扰选举等行为，为维护社会安定、打击违法犯罪行为，需要加强对这些技术的监管，防范类似的风险加剧。

CHAPTER

04

# 条分缕析:《生成式人工智能服务管理暂行办法》解读

近日，国家网信办联合国家发改委、教育部、科技部、工信部、公安部、广电总局公布《生成式人工智能服务管理暂行办法》（以下简称**“办法”**）。办法自2023年8月15日起施行。《办法》旨在促进生成式人工智能健康发展和规范应用，维护国家安全和社会公共利益，保护公民、法人和其他组织的合法权益。《办法》第一条表明了我国对生成式人工智能的发展采取健康发展和规范应用并举的价值导向。《中华人民共和国网络安全法》、《中华人民共和国数据安全法》、《中华人民共和国个人信息保护法》和《中华人民共和国科学技术进步法》是《办法》的上位法，《办法》则是为了落实这四部法律，核心内容旨在规范运用。



从治理的层面看，《办法》要求，生成式人工智能服务提供者，必须依法开展预训练、优化训练等训练数据处理活动。 一是要使用具有合法来源的数据和基础模型，确保数据来源的合法性。二是不得侵害他人依法享有的知识产权。基于此，《办法》中对生成式人工智能为现代社会带来的风险方面中的价值观、内容、数据、知识产权和服务类型五方面进行了细致规定。

首先，价值观方面，《办法》首次针对生成式人工智能技术、对人工智能企业提出伦理道德方面的要求。具体表现为要求坚持社会主义核心价值观，不得生成煽动颠覆国家政权、不得推翻社会主义制度，不得危害国家安全和利益、不得损害国家形象，不得煽动分裂国家、不得破坏国家统一和社会稳定，不得宣扬恐怖主义、极端主义，不得宣扬民族仇恨、民族歧视，不得传播暴力、淫秽色情，以及虚假有害信息等法律、行政法规禁止的内容。《办法》中，对内容方面的基本要求为，要求生成内容的准确性和可靠性；要求将违法内容及时进行处置、及时停止并整改违法行为、完善投诉和举报机制。

对于数据方面，要求使用具有合法来源的数据和基础模型，涉及知识产权或个人信息的，不得侵害他人依法享有的知识产权或取得个人同意，并详细制定真实性标准，要求提高质量，增强数据的真实性、准确性、客观性和多样性，制定符合本《办法》要求的清晰的、具体可操作的相关规则。对服务过程中获得数据管理提出要求，不得非法提供信息和使用记录，如有应及时删除，此点是为了防止再加工和滚动式处理。针对数据管理的合规性，对算法进行了新规定。再者，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、健康等的歧视。

在知识产权风险方面，在获取数据过程中，要求无侵权行为，涉及侵犯著作权或商业秘密不得使用；在运算过程中，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；在生成阶段侵权目前争议较大，还未完全对图片类进行定性。

在服务类型方面，《办法》规定采取有效措施，提升生成式人工智能服务的透明度，提高生成内容的准确性和可靠性。

此《办法》一出，无疑是顺应当下热潮，为AI技术的高速发展进行了合理规制，是相应管理的明确之举，也表明了国家鼓励创新应用，合规防范风险的主流思想。

较之2023年4月11日国家互联网信息办公室发布的《生成式人工智能服务管理办法（征求意见稿）》（以下简称**“征求意见稿”**）在文章结构、监管细则方面均有所调整，具体有如下方面：

**一是鼓励技术创新方面，**《办法》删除了《征求意见稿》中将生成式人工智能产品的研发活动也纳入监管的部分，明确《办法》监管范围为利用生成式人工智能技术向中国境内公众提供服务。此外，国家采取有效措施鼓励生成式人工智能创新发展。

**二是监管措施方面，**《办法》不仅在原则上提出要对生成式人工智能服务实行包容审慎和分类分级监管，还进行了更加细致地规定。其中“分类监管”其一是基于行业，直接以应用场景对应各大应用类别逐步进行服务合规指引。再者是根据取得方式对模型训练所需要的数据进行分类。关于“分级监管”，近日欧洲议会通过的《欧盟人工智能法案（草案）》，从考察人工智能服务提供商的行为和目的出发，结合目前的应用场景进行了具体的风险定级，具体如下：

### 1.不可接受的风险

主要行为包括:1.目的或效果是实质性扭曲人类行为，从而可能导致身体或心理伤害；2.根据已知或推断的敏感或受保护特征将自然人归入特定类型，有歧视性风险；3.为一般目的提供自然人社会评分；4.在公共场所对自然人进行“实时”远程

生物识别;5.执行机关或代表执法机关用于对自然人画像或预测实际或潜在的刑事犯罪的发生或再次发生。

该风险对AI运行主体要求了禁止性义务。

**2.高风险**

主要行为包括:1.生物识别和基于生物识别系统的几个关键使用案例;2.水、煤气、供热电力和关键数字基础设施的管理和运作中的安全组件的人工智能系统;3.用于教育或职业培训的人工智能系统;4.用于就业、工人管理和获得自营职业的人工智能系统;5.用于评估自然人的信用分数或者信用度可能影响个人基本服务、公共服务和福利的人工智能系统;6.用于执法的人工智能系统;7.移民、庇护和边境管制管理中使用的人工智能系统;8.打算用于司法和民主进程的人工智能系统。

该风险下要求AI系统本身的义务如下:1.使用高质量的训练和测试数据集;2.建立技术文件和记录保存;3.给使用者提供使用信息并保证透明度;4.保证稳定性、准确性和网络安全性。

对AI系统提供者要求义务如下:1.建立和维护风险管理系统;2.对系统进行评估;3.建立上市后检测系统并与监管者合作。

**3.有限风险**

该风险主要行为包括不属于高风险或不可接受的风险。使用者能够知道与机器互动并做出不违背本意的决定。

该风险要求AI运营主体有提高透明度的义务。

**4.轻微风险**

该风险行为包含电子游戏或垃圾邮件过滤器等应用，要求AI运营主体有不作干预的义务。

三是服务提供者的义务方面，《办法》删除了要求服务提供者有义务在三个月内通过模型优化训练等方式防止再次生成不当内容，仅要求服务提供者及时优化模型并向主管部门报告。再者，《办法》明确只有“提供具有舆论属性或者社会动员能力的生成式人工智能服务”的服务提供者，才需开展安全评估和算法备案。

**该项义务主体分为一般主体和特殊主体。**

**一般主体**包含“模型合规”和“服务合规”。《办法》以生成式人工智能海量数据训练大模型为监管的逻辑起点，《办法》第七条规定的有关模型合规的责任规范又与《数据安全法》、《数据安全管理条例》均有相辅相成的关系，因此称为“模型合规”。应用中，“部署”和“运营和监控”两个阶段中的合规，则为“服务合规”。《办法》目前对于服务提供者仅提出了基础性合规要求，但企业在实践领域中的合规细则和指引仍待确定。

**特殊主体**包括具有舆论属性或者社会动员能力、外商投资、经法律、行政法规规定需要获得行政许可的主体，其附加合规义务包括：安全评估；履行算法备案和变更、注销备案手续；符合外商投资相关法律法规的规定；应当获得许可的规定。类似算法推荐服务提供者的内控合规要求，生成式人工智能服务提供者的安全评估和算法备案义务可一并归纳至合规审查体系中，线上备案即可。

此外，《办法》还区分了一般使用者和特殊使用者。一般使用者包括告知和指导使用义务和建立健全投诉举报机制义务。特殊使用者主要针对未成年人，生成式人工智能提供者采用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。

在区分境内和境外的监管方面，《办法》规定，对来源于中华人民共和国境外向境内提供生成式人工智能服务不符合法律、行政法规和本办法规定的，国家网信部门应当通知有关机构采取技术措施和其他必要措施予以处置。由于境外服务提供者大多在境内没有实体，监管难度较大，该规定为后续监管细则措施留下了灵活调整的空间。欧盟《人工智能法案》中采用了长臂管辖原则，说明某些跨国运营主体，在一定情况下，可能需要符合两地的合规要求。基于此种要求，建议企业考虑根据拟提供服务的对象所在地进行公司注册规划及推行相应的服务版本，以提高效率，降低企业合规成本。

在生成式人工智能研发过程中，“数据标注”是十分必要的，在训练自然语言处理模型时，需要使用大量的标注数据作为训练数据。对此，《办法》第八条规定，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该办法要求的清晰、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。

2023年8月8日全国信息安全标准化技术委员会为落实《生成式人工智能服务管理暂行办法》相关要求，编制了《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）》，以便指导有关单位利用生成式人工智能技术提供生成文本、图片、音频、视频等内容服务时对内容进行标识。该指引将伦理风险分为五个类别，即失控性风险、社会性风险、侵权性风险、歧视性风险以及责任性风险。

| 风险类型 | 定 义 | 对应义务 |
| --- | --- | --- |
| 失控性风险 | 人工智能的行为与影响超出可预设、理解、控制的范围 | 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 |
| 社会性风险 | 人工智能使用不合理，包括滥用、误用等产生的风险 | 尊重知识产权、商业道德，保守商业秘密，不得实施垄断和不正当竞争行为发现违法内容的，应当及时采取处置措施，进行整改并向有关主管部门报告 |
| 侵权性风险 | 对人的基本权利造成侵害或产生负面影响的风险 | 尊重他人合法权益，不得危害他人身心健康，不得侵害他人肖像权、名誉权、荣誉权、隐私权和个人信息权益 |
| 歧视性风险 | 人工智能对人类特定群体的偏见影响公平公正，造成权利侵害或负面影响的风险 | 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歧视 |
| 责任性风险 | 人工智能相关各方面责任界定不清产生的负面影响的风险 | 提供者应当依法承担网络信息内容生产者责任。涉及个人信息的，依法承担个人信息处理者责任 |

CHAPTER

05

# 居安思危:
# 创议政府生成式
# 人工智能规制路径

SECTION 001

## 深化生成式人工智能规制设计

生成式人工智能虽然发展迅猛，但是受限于法律的滞后性，仍然有大量的内容有待立法探寻确认，例如：生成式人工智能因固有缺陷导致的侵权、服务提供者与用户共同作用下的侵权等情形下责任的承担方式，或者特殊人群（包含老年人、残疾人、精神病人、未成年人等）在使用生成式人工智能时的权益保障等。

目前，我国的民法、刑法和行政法等基本法能在一定程度内对生成式人工智能进行分散治理，但法律体系尚未形成，主要表现为对底层技术治理不足、对技术提供者监管不充分，数据与场景分类分级标准繁杂且未形成体系，相应的规则过于分散。同时，在生成式人工智能的现有理论研究中，人工智能的法律主体地位和人工智能生成物的法律地位是核心场景，但研究场景多样化。总体上看，生成式人工智能对现有数据、算法等分而治之的不成体系的治理范式提出了严峻挑战，与网络安全、数据安全、个人信息保护、数据跨境流动等现有制度存在不适恰性，生成式人工智能的治理范式、现有制度和理论均存在缺失，因此搭建生成式人工智能的综合监管框架显得尤为必要。

2020年，国家标准化管理委员会、中央网信办、国家发展改革委、科技部和工业和信息化部联合发布《国家新一代人工智能标准体系建设指南》，该指南从总体要求、建设思路、建设内容三个方面为人工智能的合规建设搭设了框架，该框架从基础共性、支撑技术与产品、基础软硬件平台、关键通用技术、关键领域技术、产品与服务、行业应用、安全/伦理等八个部分出发，探讨新一代人工智能标准体系建设思路。2023年我国针对上述立法情况，以《生成式人工智能服务管理暂行办法》作为治理的开端，但另一方面，现有治理仍存在需要完善健全的内容，具体包括下

列方面：

**1.法律主体方面**

面对主体问题，首先应该关注人与机器的本质区别是什么?针对这一问题，目前主流的学术观点而言，人工智能尚不具备法律意义上的主体资格。

也有人主张，将人工智能拟制为法律意义上的主体亦或是准主体，其现实背景来源于特斯拉自动驾驶汽车造成交通事故，无人驾驶汽车曾被纳入责任主体的考虑范围。但基于目前技术发展的考虑，生成式人工智能目前不能作为法学意义下的“理性人”，另一方面其没有实现法律意义上的“人格化”。

现阶段人工智能引发的自动驾驶问题，可以归入《民法典》的规范范畴。生成式人工智能引发的知识产权争议可以通过《著作权法》进行调整。

**2.数据安全方面**

目前，即使用户依据《个人信息保护法》第48条主张个人信息处理者履行解释说明义务实践中很有局限性。其中，解释说明义务仅局限于个人信息处理规则，但数据来源、数据质量、标注规则以及用户输入数据的处理与存储对用户信任和互动策略产生的实质影响并未明确规定。因此，我国治理机制应考虑此局限，适当扩展提供者履行数据透明义务的对象类型。

基于对隐私问题的考量，欧洲率先做出了回应。2023年3月31日，意大利数据监管机构Garante发布临时紧急决定，要求OpenAI停止使用其训练数据中所包含的意大利用户的个人信息。作为对此决定的回应，OpenAI已经停止了意大利用户对ChatGPT的访问。这是西方数据监管机构首次对ChatGPT采取行动，并对围绕生成式人工智能所产生的隐私与个人数据保护问题进行了突出强调。

### 3.知识产权方面

关于知识产权方面，所涉争议主要体现在三个方面:一是未经授权的使用行为，二是生成内容的著作权归属问题，三是第三方知识产权保护问题。

首先，生成式人工智能模型必须先输入数据，才能够输出内容。如果在数据输入时，未得到数据提供者的授权，就可能引发知识产权侵权问题。例如，2023年1月，Getty图片社以未经同意训练了其数百万张图片为由，起诉了人工智能公司Stability AI。ChatGPT甚至可以“学习”用户风格，越来越多的作家和艺术家担心ChatGPT对他们的作品进行大量训练，从而复制其独特的风格。

再者，关于著作权归属问题，从法律的角度而言，即使人工智能生成内容因符合《著作权法》的构成要件，而可能受到著作权保护，该权利也不属于人工智能本身，而是归属于研发机构、企业或使用者。如果人工智能生成的内容不受著作权法保护，也并不意味着该内容可以自由使用，因为还涉及到第三方的知识产权保护问题。

此外，知识产权可能包括著作权、商标或专利，这些权利由使用 ChatGPT生成内容的个人或实体之外的权利人持有。在此种情况下，为避免对第三方知识产权的权益侵犯，必须获得权利主体的许可或授权，才被允许以特定方式使用这些内容。

建议公权力机关设置合理知识产权制度保障基础模型层训练数据的获取。对于创作过程中如何使用生成式人工智能，需要进行何种程度的解释与说明，都需要立法者、技术专家、企业及政府主管部门的合理协作，共同制定适当的法律框架。

**4.伦理方面**

2023年8月OpenAI推出网络爬虫GPTBot，尽管OpenAI坚称开放网站数据收集入口，能够帮助该公司提高 AI 模型的实际质量，而且爬取的内容也不会涉及敏感信息，但仍引发了多家网站的抵制。专为医疗保健行业提供AI助手的Hyro公司联合创始人兼CEO Israel Krush 表示“OpenAI 不该只关注那些被标记为包含个人身份信息的网站，而应当假设所有网站都可能涉及个人隐私，特别是各内容发布平台。他们应当采取积极主动的措施，确保爬取的信息不违反合规性要求。”因此，从伦理方面，继续进行相关合规性建设显得尤为必要。

现阶段，关于人工智能治理分为“软法治理”和“强制性的法律法规”。其中，“软法治理”，即基于非立法和政策性工具的自我监管模式。软法治理的表现形式包括企业自身、利益相关者组织、标准制定机构等发布的人工智能伦理指南，这些自我监管文件在人工智能治理的基线设定方面发挥着重要作用。

在此基础上，在生成式人工智能的运用中，应做好对人工智能技术评估，兼顾公众利益优先原则和公平正义原则，规避生成式人工智能技术在学习过程中发生“异化”带来的风险。完善相关披露规则，包括隐性披露和显性披露，服务提供者承担隐性披露义务，使用者承担显性披露义务，从而利于相关法律法规的落实。

SECTION 002

## 构建鲜明监管职能框架

监管框架的设计要兼顾“伦理”和“法律”，伦理是监管框架设计的底线。生成式人工智能的发展与应用应与人类社会的发展的趋势相一致，要符合社会发展的伦理和道德。因此，坚守伦理的底线十分必要。在设计相应的监管框架时，要充分考虑伦理风险，借鉴国内外影响力较大的伦理指南、政策文件、标准共识，贯彻落实《中国新一代人工智能伦理规范》《关于加强科技伦理治理的意见》等重要文件精神，制定相关伦理治理原则。

结合具体场景指定具体细则指引，在不同的领域和场景叠加行业规范要求,在重点领域与场景进行专门的制度设计。

上海交通大学季卫东教授曾提出：“通过合法与违法的二元化编码和规范思维的形式性要求，可以把决定者从问责的重负中适当解放出来并同时对自由裁量权加以制约，可以使风险沟通的复杂性大幅度简化，有助于就决策的妥当性和问责标准达成共识。”因此健全相应的体系建设尤为重要。再者由于生成式人工智能应用面广泛，因此对应场景多样，需要结合实际指定相应的细则，以便能够落地执行。

加强舆论监督，建立有效的沟通和协调机制。公权力机关应该就生成式人工智能的应用和法律宣传开放多种形式、多个端口的宣传活动，同时向公众建立通畅的沟通渠道，广泛听取专业和现实中的意见，从而有利于建立建立可靠的监督机制，推动形成多方参与、协同共治的科技伦理与法律治理格局。从学术讨论、企业社会责任、公众学习监督等多管齐下，完善各个主体在伦理与法律风险方面的决策、咨询与调控的制度框架，提高各方对生成式人工智能风险的识别、评估、处

理能力。从而将技术应用问题法律化。

*SECTION 003*

## 规制落地的挑战

但上述规则设计的落地，也面临着诸多挑战。

在监管组织方面，目前监管主体过多，需要形成统一口径。当前，我国对人工智能的监管部门包括国家市场监督管理总局、国家互联网信息办公室、工业和信息化部、科技部等。多个监管会造成在生成式人工智能应用中监管责任的推诿，也可能出现重复监管导致责任过重阻碍技术进步的情况，不同监管规则冲突抵牾，进而影响行业竞争格局与公共利益。虽然国家层面也通过成立人工智能治理专业委员会等机构来实现对多头监管矛盾的协调，但由于该组织的公权力有限，难以形成强制性约束因此无法实现实质层面的统一指导，也不利于实现生成式人工智能的治理目的。

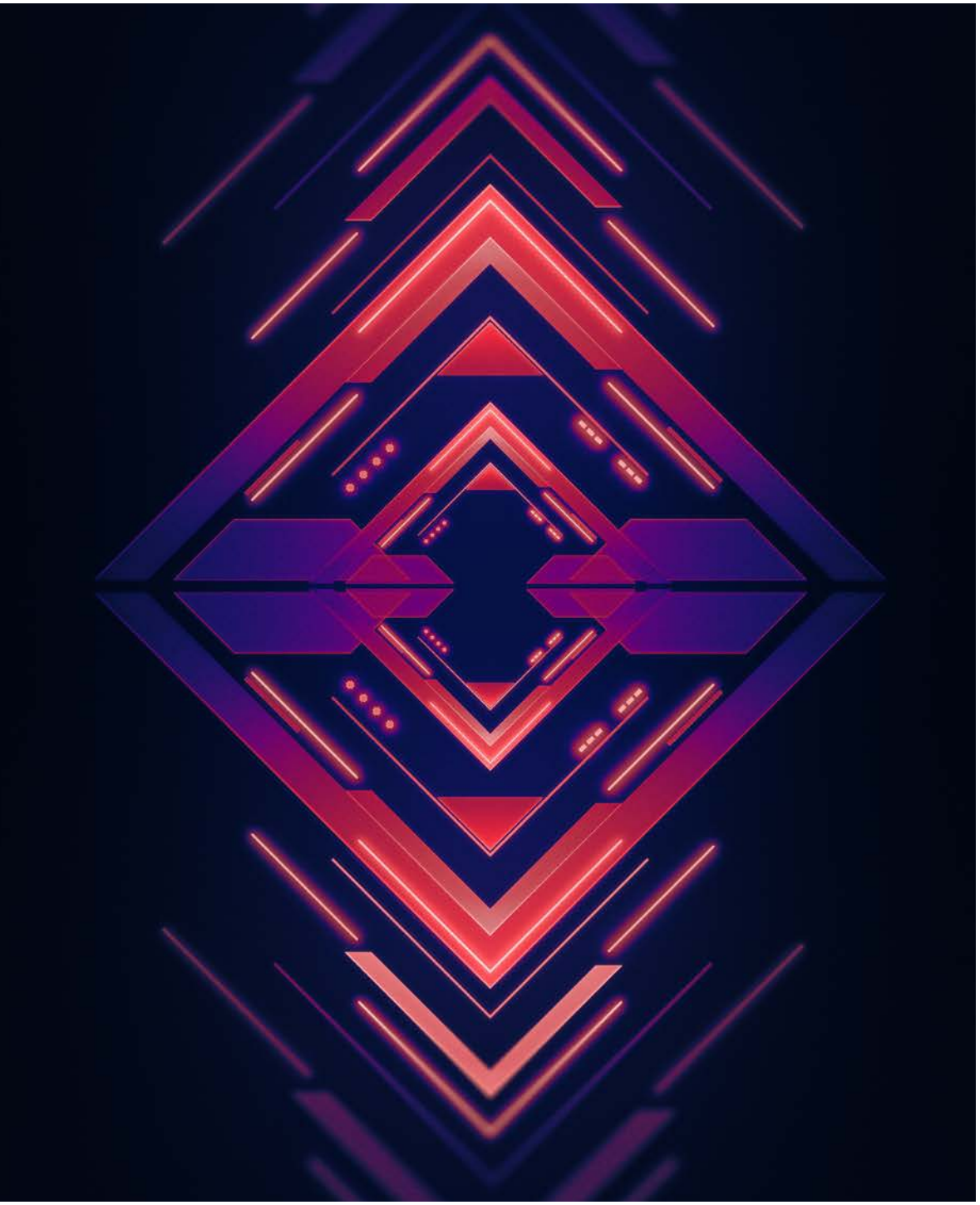
在治理落地方面，缺乏行政、司法等手段的协同配合。生成式人工智能法律法规的执行需要政府、行业协会、企业等社会各界的参与。政府需要借助司法机关、企业、技术专家等对企业进行指导和约束。同时，企业在获得政府资源、国家政策支持以后，也需要承担相应的社会责任。因此，需要企业主动发挥自我规制的作用，协助政府行政监管，推动生成式人工智能的治理规则与治理标准的构建，避免自身技术被滥用并造成公众利益的损害。此外，司法手段也是有效的强制性手段。公开的法院裁决，可以对公众起到教育作用。我国需要在制度层面进一步构建由“行政监管—司法审查—企业自治”所组成的三位一体的治理框架，并细化有关具体协作机制，以实现多层次、复合型的人工智能治理举措。

此外，在监管形式和内容方面要鼓励分层发展。合规是一项事物发展的前提，过度监管又会导致该事物发展可能会受到压制，在全球竞争如此激烈的今天，如何去协调监管力度与科技发展之间的平衡都是耐人寻味的问题。根据张凌寒教授的观点，在鼓励我国生成式人工智能发展的思路下,应将生成式人工智能作为基础设施,划分为技术、产品与服务三个层次,以“基础模型—产品模型—服务应用”形式,关注不同层次的不同生产要素,大力鼓励基础模型层的技术发展,审慎包容监管产品模型层,对服务应用层沿用并调整以实施敏捷治理。将我国从较为单一的场景的算法治理,演化为适应不同治理目标的生成型人工智能的复合型系统性治理。

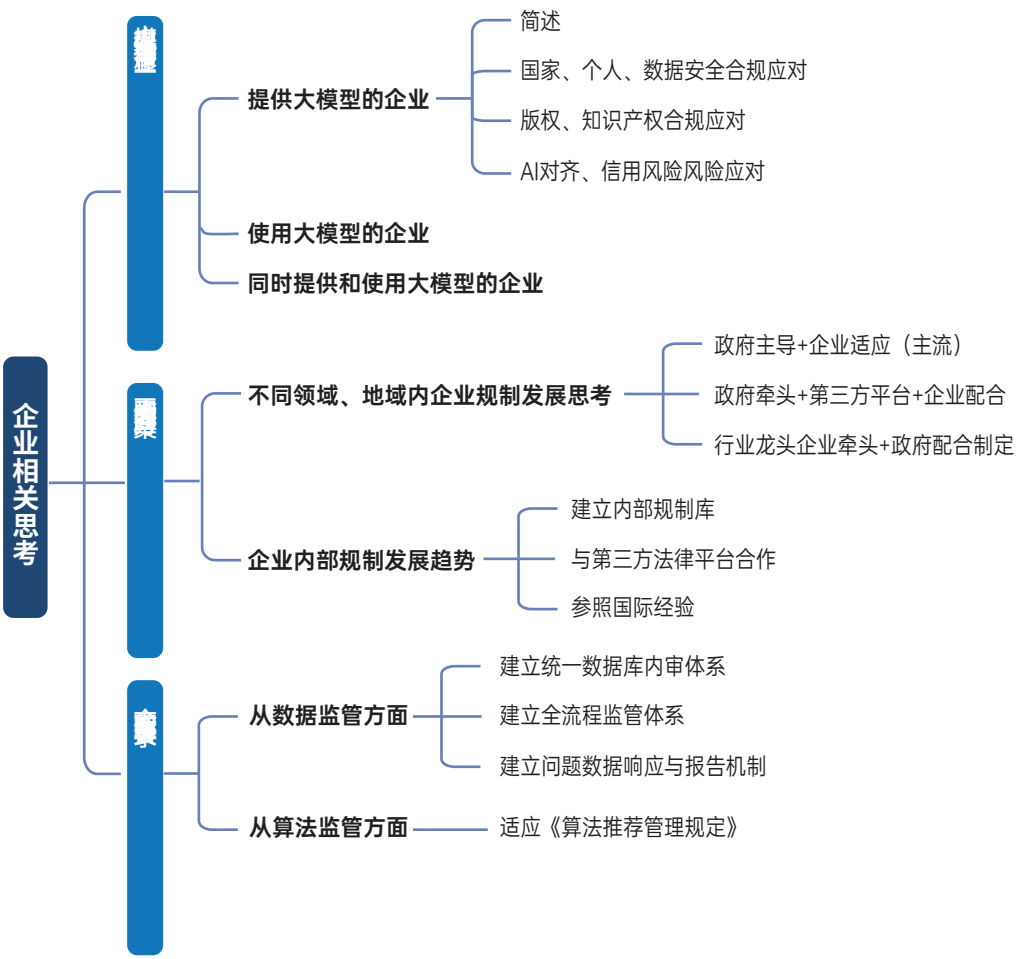
CHAPTER

06

# 以权达变：企业对生成式人工智能规制的应对和思考



生成式人工智能的快速发展显著增强了人工智能的整体影响力，它以每年数万亿美元的额外价值改变了很多人的工作性质。但发展至今，受限于法律的滞后性，仍然有大量的内容和技术有待立法探寻确认。鉴于此，企业必须迅速采取合规行动以应对相关技术的过快发展，并为接下来的机遇和风险做好准备；另外，合规是一项事物发展的前提，但是过度监管又可能会导致该事物发展受到压制，在全球竞争如此激烈的今天，如何协调监管力度与科技发展之间的平衡都是耐人寻味的问题。



SECTION 001

## 关于大模型提供与使用企业的思考

生成式人工智能领域提供大模型的企业和使用大模型的企业在业务模式和收入来源上存在一定区别，但在法律规制建设上又存在众多交集。

### 1.提供大模型的企业

#### (1) 简述

产业层面上，AIGC技术大模型构建者已经深入到文字、语音、绘图、视频、代码、3D 模型 / 场景、游戏等领域。提供大模型的企业通常拥有强大的研发能力和技术团队，以赚取模型许可费或服务费为主要收入，如OpenAI、华为和商汤科技，它们开发、训练、不断迭代和改进大模型，然后将其出售给个人、企业及研究、政府、金融、教育等机构。提供大模型企业的法律规制设定是本篇讨论的重点，思考如下：

首先，需要确保提供大模型的企业在开发模型及服务过程中采取了有效的安全措施和技术手段，符合国家安全性和可靠性的要求，排除了在算法上出现安全漏洞和数据泄露等问题；其次，提供者在开发模型和服务期间会投入大量的人力、物力和财力成本，企业应保障底层数据挖掘关联方的知识产权和商业利益不受侵犯；最后，大模型提供者要想进行法律规制的制定，需要严格规避可能存在的信用风险，企业不仅应接受政府和第三方机构的监督和管理，还需要按照相关法规要求对大模型使用者（用户）的相关信息进行保护，确保开发模型符合伦理标准且技术可信。”

**(2)国家安全、个人信息安全和数据安全合规应对**

大模型供给企业在数据挖掘过程中需考虑数据安全、国家及个人信息保护方面的法案。我国科技领域已制定了《中华人民共和国数据安全法》、《中华人民共和国网络安全法》、《网络信息内容生态治理规定》、《中华人民共和国民法典-人格权篇》、《新一代人工智能伦理规范》、《网络安全标准实践指南-人工智能伦理安全风险防范指引》等相关法律法规，这为我国生成式人工智能技术的发展提供了基本保障。

2023年7月10日，为了应对生成式人工智能服务的快速发展，国家网信办、发改委，教育部等七部门联合发布了《生成式人工智能服务管理暂行办法》，这部全球首个生成式人工智能领域的监管法规，开启了我国人工智能治理新篇章。其中，根据新发布《办法》中相关规定，企业需结合多种手段强化内部控制与业务应用机制。首先,企业在搭建模型爬取数据时应建立切实可行的个人信息保护机制，不可收集非必要个人信息，不得非法留存或向他人提供能够识别用户身份的输入信息和使用记录，应及时受理关于个人关于查阅、更正、补充、删除等相关请求；其次，企业应明确并公开其服务的适用人群、场合及用途，指导使用者科学理性的认识并依法使用生成式人工智能技术，采取措施防范未成年用户过度沉迷或依赖人工智能服务；最后，企业应设置通畅的投诉、举报通道，建立健全的投诉举报机制，设置便捷的投诉、举报入口，及时受理并反馈公众投诉举报处理结果。

总而言之，在安全方面，企业适应现有法规的同时，一方面应承诺生成式人工智能系统发布前已由独立专家进行过内部和外部的安全测试，设置了网络安全和内部威胁保护机制，有效防止重要的风险源头（如国家信息、个人信息、网络信息）对社会造成的不利影响，并确保向公众推出的模型是安全的；另一方面，为避免人工智能系统发布后相关问题依然存在，企业可自行或委托第三方机构设置强大的

监督、报告和投诉反馈机制，以便快速找到并修复系统内存在的漏洞和问题。

**（3）版权、知识产权合规应对**

在构建大模型并解决数据挖掘问题的过程中，应重点开展对生成式人工智能作品权属认定的研究。一方面，企业需对现有《中华人民共和国著作权法》中“合理使用”的相关条款进行适应，在设置内部规制时应规避底层挖掘带来的著作权问题，例如企业在使用他人作品时应当取得著作权人的授权，不应损害相关权利人的意愿及获酬，避免出现事先不征求许可、事后不支付报酬，只将获益全部归给底层挖掘人或机构而带来的的法律纠纷；另一方面，企业搭建的生成式人工智能模型更多是对非特定主体进行商业性服务，与现有《著作权法》难以契合，目前生成式人工智能模型在商业化领域的应用及模型大量复制利用作品的现状无法很好的适用相关法规，而涉及数据挖掘的《著作权法》中“合理使用”-“个人使用”、“适当引用”及“科学研究”的相关规制暂时无法满足相关要求，立法修订是当下亟待解决的方向。

所以，大模型提供企业在进行生成式人工智能训练的过程中，首先，需确定人工智能创建的内容是公平还是有偏见的，用于创建基础模型的训练数据中是否存在“抄袭”或其他侵犯知识产权的有害信息；其次，大模型提供企业应剔除违法、受著作权保护及涉及商业秘密的作品，避免陷入相关侵权纠纷；最后，对于版权及知识产权归属问题，应最大限度避免因生成式人工智能固有缺陷而导致大模型提供者与使用者共同作用下的侵权，关于该问题，学界和实务界对此问题仍然存在着争议。所以在下一步应对中，大模型开发企业可尝试在制定用户协议时明确生成式人工智能的各方权利归属，优化数据抓取规则，从而防患于未然。

**（4）AI对齐、信任风险应对**

为避免人工智能技术滥用导致的信任危机，我国于2023年1月10日在生成式

人工智能“深度合成”方面，发布实施了《互联网信息服务深度合成管理规定》，这是目前生成式人工智能领域最为核心的监管法规。其要求企业（即深度合成模型/服务提供者）应适应相关责任义务，在生成或者编辑信息内容的合理区域进行显著且不影响用户使用的标识，并向公众提示深度合成情况，同时应按照国家规定保存日志信息，在识别出不良或违法信息时应及时向有关部门报告。在此法规的基础上，我们能看到在可预见的未来，企业可通过符合规制要求的人机互训合成机制不断检验对齐的精度。同时，智能对话、人声人脸合成、沉浸式仿真场景等生成式人工智能会迎来相对规范的应用场景，AI目标会逐步和人类价值观与利益相对齐，因技术滥用导致的信任风险会有所规避。

总而言之，在生成式人工智能可信性方面，企业可采取多种方式以赢得公众信任。例如，企业可以公开其所使用模型的能力局限性以及适宜或不适宜使用的场景；优先研究人工智能系统可能带来的算法与合成风险，避免有害的偏见和歧视、更好的保护用户隐私等；企业可开发相应技术机制，确保用户明确知道AI生成的内容（如图片、视频类信息、文本类信息等）为人工智能合成。

### 2.使用大模型的企业

需求层面上，生成式人工智能大模型使用者为构建Web 3.0时代的场景应用提供了助推剂，开发团队在大模型的辅助下无需花费太多时间用在搭建大量基础场景上，就能极大程度上降本增效，完成目标。例如与谷歌合作的《NATURE》杂志，其通过使用大型语言模型生成文章和摘要，实现了更快地出版和分享研究成果；再以Netflix为例，该企业通过使用生成式人工智能大模型改进推荐算法，用自然语言处理技术来分析用户历史观看记录、电影标题、电影元数据、用户描述及评论中提取的关键信息，以更好地掌握用户观影偏好。目前国内使用大模型的企业，

如字节跳动、智谱AI、中科院、百川智能、MiniMax、上海人工智能实验室等，它们从提供大模型的企业购买许可，然后在自己的平台上使用这些模型来提供服务。该类企业通常拥有大量的用户数据和计算资源，以支持模型的部署和应用，收入来源主要是广告或其他基于用户的服务。

所以大模型使用企业在法规的制定与适应上可参照大模型供应企业的相关规制思考，首先重点关注用户个人信息保护，防范出现利用生成式人工智能大模型进行虚假宣传、或侵犯用户个人或他人权益等问题；其次，使用企业是生成式人工智能大模型的最终用户之一，其需要保障好自身知识产权不受侵犯、商业秘密不被泄露、商业利益得到充分保障；再次，使用企业需承担相应的社会责任，例如，在使用生成式人工智能大模型时，企业应确保其应用不会对企业、社会、环境产生伦理道德或信用风险影响。最后，需要建立完善的监管机制和标准体系，对生成式人工智能大模型的使用企业进行监管，制定一系列的技术标准和法规，建立完善的监管机制。

### 3.同时提供和使用大模型的企业

一些企业既是大模型提供方，也是大模型使用方，比如百度。它们一方面开发和训练自己的大模型，另一方面也向其他企业提供大模型的使用许可。

总体来说，提供大模型的企业主要关注模型的研发和许可，而使用大模型的企业则更注重模型在自身平台上的应用和服务，但无论企业所处立场为何，生成式人工智能企业在不同领域内的法律规制制定，国家、个人和数据安全法律风险防范，版权、产权相关法律适应，AI对齐及信用风险企业规制，以及如何配合监管都是大模型供给者和使用者需要考虑的问题。

SECTION 002

## 不同领域内应对方案

### 1.不同领域、地域内企业规制发展的思考

在生成式人工智能发展进入快车道的当下，我国政府为应对生成式人工智能可能产生的法律问题，会在不同领域制定更为详细的规制。例如目前在医疗行业，国家药品监督管理局医疗器械技术审评中心和国家药监局分别发布了《人工智能医疗器械注册审查指导原则》《人工智能医用软件产品分类界定指导原则》及《深度学习辅助决策医疗器械软件审评要点（试行）》，这对医用行业人工智能的发展有了很好的深入指导作用。同时，在不同地域之间出台相应地方性法规或工作文件也是未来趋势之一，例如《上海市促进人工智能产业发展条例》、《深圳经济特区人工智能产业促进条例》、《上海市数据条例》、《北京市促进通用人工智能创新发展的若干措施》及《北京市加快建设具有全球影响力的人工智能创新策源地实施方案》（2023-2025）等。

随着人工智能技术高速迭代，为满足不同领域内“大模型供给和大模型使用企业”（以下简称**“企业”**）规制制定的专业及高效，政府与各领域企业之间的合作会日渐密切。在此情形下，衍生出多种对企业规制发展模式的思考，比如“政府主导规制制定+企业适应”，这是当下较为主流的规制制定模式；或是“政府牵头+专业平台方拟定+企业配合”三方联动的官方或私有化部署合作；还有“行业内龙头企业牵头+政府配合制定规制”模式，例如互联网大厂、国内大型银行在牵头制定产品规则时就有过相关成功经验。政企合作模式为各个业务场景下的规则制定提供了清晰的指引，为确立不同领域中的主体责任和义务提供了有力的支撑。

**2.企业内部规制发展趋势**

在生成式人工智能技术快速创新的当下，政府的规制制定倾向于更加精细化，行业化和地域化。为应对这种情形，企业内部应当对国内不同场景下（不同业务板块或不同地域）的规制做出适应，例如可以通过培养专业法规人员建立内部规制库，用以更好的在各类规制框架下发展业务；通过与第三方专业法规平台合作等方式向用户提供服务；或是生成式人工智能企业之间联合制定自律公约、搭建技术交流平台，这个层面我们可以参照国际相关企业处理经验。

例如，2023年7月美国OpenAI、微软、谷歌、Anthropic、Meta、亚马逊和Inflection等七家AI领域巨头宣布成立行业组织“前沿模型论坛”，该论坛成立具体发展目标如下：

（1）促进人工智能安全研究，推动前沿模型发展，最大限度降低风险，实现对AI能力与安全的独立、标准化评估，确保相关技术安全、可靠、并处于人类控制之下；

（2）确定负责前沿模型开发和部署的最佳做法，向公众普及AI技术的性质、能力、限制与影响；

（3）与政府政策制定者、学者、民间社会和公司合作，就论坛的设计与合作方式展开磋商，分享有关AI信任和安全风险的知识，促进企业与政府间信息共享；

（4）用主动合规避免被动监管，借此换取更大的主动开发空间，在一定程度上参与规则的制定；

（5）支持和推动现有政府、七国集团（G7）、经合组织（OCED）、欧美贸易、技术理事会（TTC)和多边倡议在人工智能方面有关风险、标准与社会影响的工作；

（6）建立顾问委员会，以代表不同背景和观点，帮助指导论坛的战略和优先事项；

(7) 在不同领域内，针对当前社会最大挑战，如气候变化减缓和适应、早期癌症检测和预防，以及网络威胁应对，支持开发相关应用。

总体来看，美国七家人工智能“领头羊”做出的承诺旨在提升安全、保障和可信任三个方面， 而“安全、可靠和可信”大概率将会成为未来全球人工智能领域的行业标准。国内技术企业在内部规制制定上，可以参照国际相关经验，促进国内生成式人工智能生态良性发展。

SECTION 003

## 全面配合监管机构要求

企业在经营过程中，不仅应考虑盈利及技术服务本身，还应当考虑背后的风险与法益，为了更好的适应监管机构的相关要求，企业对于生成式人工智能技术的监管可以融合技术与法律进行如下思考：

### 1.从数据监管方面

在应对我国现有数据监管有关立法的同时，企业也应构建内部监管机制。首先，企业内可建立统一数据库内审体系，保存训练模型的迭代和修改过程，备份有关企业在向公众发布新的大型语言模型产品(或更新版本)前评估风险和安全性所遵循政策和程序的详细描述，形成向内定期核查，向外配合行政司法机关检查的闭环；其次，企业内部可确定是否形成前期数据池管理、中期技术架构及数据挖掘、后期人工智能工具应用及用户管理的全流程监管体系，企业可在数据池管理中构建解释数据的来源、抓取方式以及企业如何识别、评估、审查和选择模型数据来源等信息，布局并完善人工智能工具监管生态搭建；最后，针对数据泄露及违规风险，企业应建立问题数据响应与报告机制，机制中应留存企业是如何识别、评估和测试出大型语言模型在生成内容时所包含的真实个人信息，以及采取了什么措施用以避免对真实个人的虚假误导或诋毁性陈述风险，同时，企业对于发现违法内容或发现使用者利用人工智能服务从事违法活动的，应当在响应与报告机制中设置及时警示、限制功能、停止生成、停止运输、终止服务等处置措施，保存有关记录，及时采用模型优化训练完成整改并向相关主管部门报告。

**2.从算法监管方面**

在生成式人工智能“算法推荐”方面，我国于 2023年3月1日生效了《互联网信息服务算法推荐管理规定》，这意味着企业算法推荐服务者应当建立健全算法机制机理审核，在重要事项上应该对输出内容进行全面审核，并采取有效措施提高生成内容的准确性和可靠性，且不得设置诱导用户进行消费等违法违规或违背伦理道德的算法模型；同时，在互联网信息服务算法备案系统内会有更完善的企业服务者信息体系，企业应在体系内记录更全面的信息，包括但不限于名称、服务形式、算法类型、应用领域、自评估报告等内容，所以企业应提前做好相关调整。其中，具有舆论属性或者社会动员能力的生成式人工智能企业，需开展安全评估和算法备案。

# EPILOGUE 结语

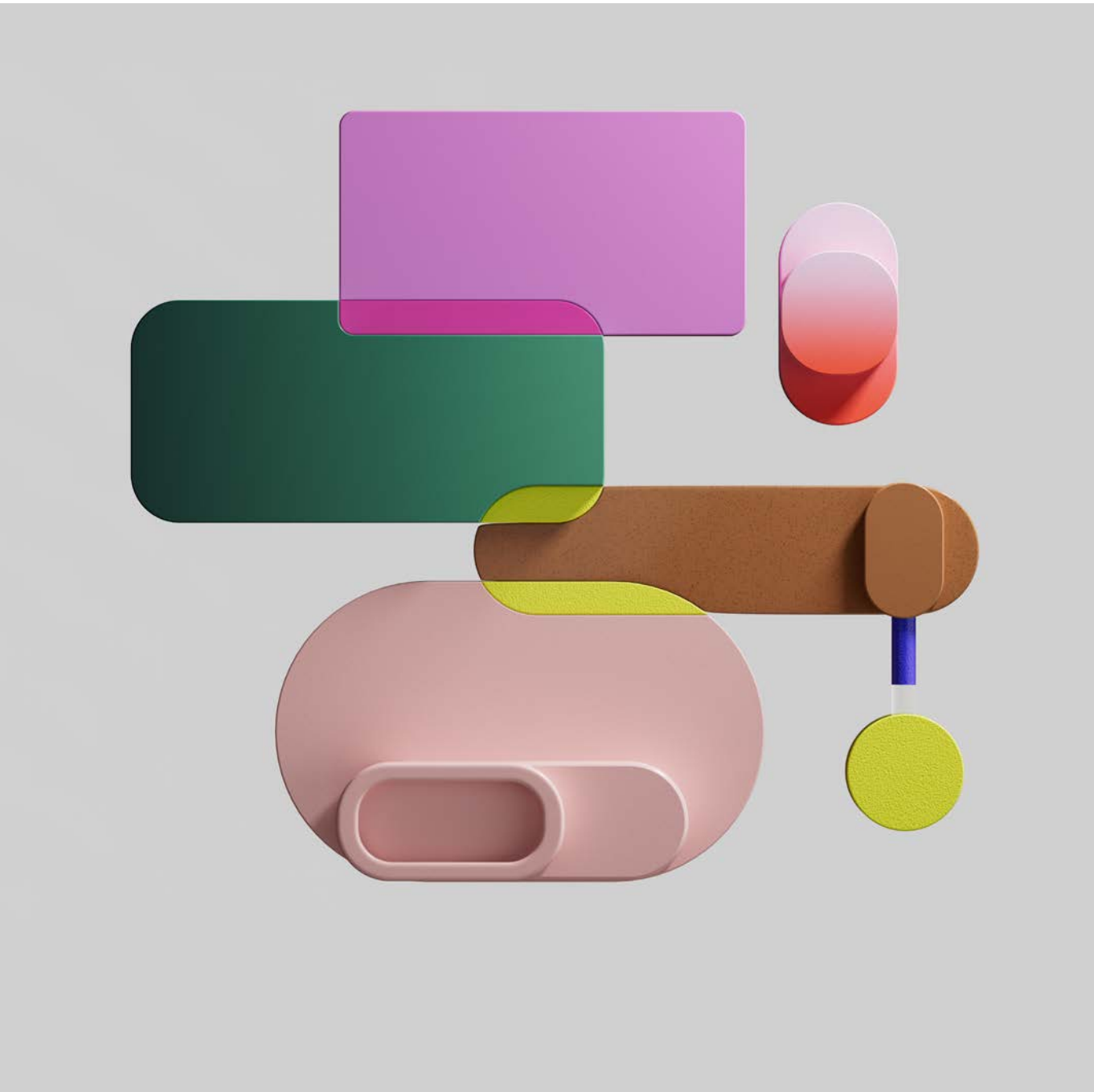
历史经验表明，新技术具有重塑社会的潜力，生成式人工智能产品作为一种新型基础设施，已经逐步进化为全球经济社会发展的重要支撑。新模式下的AI+产业组合正成为当下人工智能主流发展方向：我们可以看到，在创新应用场景中，AI数字人让数字分身、穿越时空对话变为可能；寿险、财险相关领域生成式人工智能技术的运用给保险产品设计带来重大机遇，利好消费者并带来更贴合需求的服务；人工智能在广告和办公方面的进展改变了公众学习、工作与生活的方式。

生成式人工智能技术前期建设需要国家统筹布局、适度超前，但带来的相应安全风险问题却不容忽视。不同国家和行业因过快发展造成的法律问题触目惊心，所以在借鉴相关经验的基础上，我国应进一步重视规避数据安全风险、个人信息安全风险、大模型提供企业、使用企业或相关当事方知识产权风险、伦理道德风险、AI对齐风险以及信任风险等因素，保护好各方当事人的合法权益，防范恶意诱导、侵权及欺诈。我们应在监管领域拥有前瞻性：一方面对生成式人工智能相关技术的高度发展给予大力支持，对于生成式人工智能如训练数据、模型开发等相关责任方的监管不应仅为了服务信息内容安全而施加过多义务，以致影响其作为基础设施的功能研发；另一方面要重视人工智能服务的公平性，确定前中后整个条线各方当事人的责任认定并纳入相关监管，以更好的保护各方权益。

人工智能技术的发展不能以牺牲相关当事方著作权、知识产权、隐私权为代价，而是应当在各方之间找到平衡点以达到可持续发展。国家、企业、相关各方应如何灵活应对，这是我们接下来面临的一个重要挑战。

# THANKS

鸣谢



白皮书课题组在初稿完成后，通过举行闭门研讨会，得到了以下学者和专家的点评和指教（按姓氏拼音顺序）：蔡荣伟、曹红星、董继明、方园、胡宇、王资凯、邢杰、许多奇、杨燕青。课题组在此一并致谢！

课题组对白皮书所有内容负责，任何可能的错误与参与闭门研讨会的学者和专家无关。

本研究还得到了上海市科委“人工智能新型社会实验与治理方法研究与应用示范”项目 (项目编号21511101200 )的支持。支持单位：上海人工智能研究院有限公司。