生成式人工智能新规六大特点及合规要点

作者：傅鹏、俞沁

近年来，基于大型语言模型（简称“**大模型**”）衍生的生成式人工智能技术与服务“军备竞赛”，在全球范围内开展地如火如荼。过去数月中，我国诸多大型科技企业和中早期科创企业也纷纷加入该赛道，形成“百模大战”的赛跑格局，对技术路径选择和真实市场需求进行了大量有益探索。

在此过程中，一方面，监管机构与诸多企业频繁互动，对国内生成式人工智能行业现状开展了调研，在一定程度上引导业界达成了“中国一定要有自主可控的大模型”的共识。另一方面，自主可控的优质大模型需要更大量的用户交互、更丰富的场景试用，因此，对监管能否创设有利于模型衍生、训练的制度赛道，能否采取更灵活机动的行业管理机制提出了更高的要求。

基于前述背景，主管机关对 2023 年 4 月发布的《生成式人工智能服务管理办法（征求意见稿）》（简称“**征求意见稿**”）进行了修订，由国家互联网信息办公室（简称“**网信办**”）携手其余六部门于 2023 年 7 月 10 日发布了《生成式人工智能服务管理暂行办法》（简称“**《暂行办法》**”），《暂行办法》将于 2023 年 8 月 15 日施行。相较于征求意见稿，《暂行办法》更显著地突出了**拥抱人工智能发展、融合现有制度框架、务实设计监管思路**的趋势。生成式人工智能企业也需相应理解合规思路，有针对性地完善合规体系建设。

第一部分：新规特点与理解要点

1. 监管准则：“发展”基调和“包容审慎”原则

《暂行办法》充分体现了拥抱生成式人工智能技术的总体基调。例如，《暂行办法》由七部门联合发文。其中，除征求意见稿的主笔起草部门网信办外，还加入了国家发展和改革委员会（简称“**发改委**”）、中华人民共和国教育部、中华人民共和国科学技术部（简称“**科技部**”）、工业和信息化部（简称“**工信部**”）、中华人民共和国公安部（“**公安部**”）、国家广播电视总局（简称“**广电总局**”）等部门。网信办、工信部、公安部作为类似技术赛道传统意义上的主要监管单位，携手发布生成式人工智能监管新规，是对生成式人工智能已有监管框架的延续（例如他们也是《互联网信息服务深度合成管理规定》的发文单位）。

此外，科技部作为年初国务院机构改革重组的重点单位，承担推动健全新型举国体制、优化科技创新全链条管理、促进科技成果转化、促进科技和经济社会发展相结合等职能。特别地，科技部主导起草的《科技伦理审查办法（试行）（征求意见稿）》与生成式人工智能高度相关。举例而言，该文件明确规定从事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。基于大模型的生成式人工智能产品原理复杂，实现算法透明度的难度相对较大，训练数据及参数调优对生成结果影响巨大，科技伦理审查制度和合规要求将在大模型企业的合规实践中占据重要位置。

发改委作为联合发文单位，体现了生成式人工智能产业与数据要素市场的高度相关性以及可预期的高度融合趋势。发改委是《关于构建数据基础制度更好发挥数据要素作用的意见》的主导单位，其下辖的国家数据局负责协调推进数据基础制度建设，统筹推进数字中国、数字经济、数字社会规划和建设。生成式人工智能需要大量数据驱动训练与优化，应用场景也需与数据要素利用高度融合与协同，数据要素市场的建立、活跃以及良性监管，将从数据源头促进生成式人工智能技术迭代和产品更新。

教育部、广电总局，作为相关行业或领域监管部门，其职权行为往往体现为对融合生成式人工智能技术或服务的部分传统行业的监管。例如，广电总局制定广播电视、网络视听节目服务管理政策，进行行业管理，该行业不免集中了多类生成式人工智能服务的应用场景。再如，教育领域也是当前生成式人工智能服务最主要的应用领域之一，相当一部分生成式人工智能服务的研发与高校紧密绑定，是高校科技成果转化的重要场景。

因此，上述多部门的联合发文，特别是发改委及几家行业主管部门的参与，体现出监管在对生成式人工智能应用多维度理解的前提下，以“包容审慎”态度塑造监管格局、制度赛道，并力求实现行业良性“发展”的整体基调。生成式人工智能领域技术和服务监管具有特殊性，其应用落地场景变迭极快，应用场景跨行业跨专业特点较为明显，行业参与者拟适用的技术路径和服务模式也在探索中迅速更新。因此，一方面，监管需要以相对“审慎”的基调设定合规要求和监管思路；另一方面，政策层面已经奠定“发展”的主基调，因此需要强调“包容”，让技术的应用随着业务和场景在审慎的监管环境下先行落地和推进。“包容审慎”这一关键词，也总体贯彻于《暂行办法》相较于征求意见稿的若干重要修订中，详见后文分析。

2. 管理逻辑：“分类分级”监管方式

《暂行办法》明确提出生成式人工智能数据源以及相关技术与服务应用的“分类分级监管”思路。具体地，《暂行办法》对生成式人工智能“基础设施”之一的训练数据源有简略规定，提到应推动公共数据分类分级有序开放，扩展高质量的公共训练数据资源。这既是《中华人民共和国数据安全法》对数据处理活动分类分级原则性要求在公共数据应用领域的体现，也反映为数据驱动的生成式人工智能服务分类分级监管的入口维度。此外，《暂行办法》也明确规定，国家有关主管部门将针对生成式人工智能技术特点及其在有关行业和领域的服务应用，完善与创新发展相适应的科学监管方式，制定相应的分类分级监管规则或者指引。

从产业发展的视角看，如何对生成式人工智能进行更精准的“分类分级监管”，是值得期待的重要话题。欧洲《人工智能法案》（Artificial Intelligence Act）的监管思路深度渗透、贯穿“风险分级”的概念，是我国建设、完善该领域下分类分级监管制度的有益参考。《人工智能法案》采取“基于风险水平”的风险分级导向监管模式；例如，其以用途、功能、场景作为单元，按照风险从低到高，将对人工智能的监管分为“低风险或无风险人工智能”“需要履行一定透明度义务的人工智能”“高风险人工智能”“无法接受的人工智能”四类。对于“低风险或无风险人工智能”，技术或服务提供者不需要承担特别的义务，或者可以自愿承担若干义务、进行一些标识活动。对于“需要履行一定透明度义务的人工智能”，被要求在人工智能产生的信息以及人工智能算法透明度方面履行一些法定义务。对于“高风险人工智能”，典型如对人事聘用进行决策，对医疗用药提供建议和决策等场景，提供者需

要履行严格的法定义务，也需要事先进行一系列评估。对于“无法接受的人工智能”，典型如对于自然人的社会信用状态进行一般性的打分排名，则是禁止的。上述监管方式，有利于在保障人工智能技术发展和场景积累的同时，准确识别和监管真正具有风险的应用及场景，值得借鉴和参考。

3. 把握实质：生成式人工智能真正的监管目标

征求意见稿规定，利用生成式人工智能生成的内容应当真实准确，应采取措施防止生成虚假信息，即俗称的“结果保真”。但实际上，该领域不论从技术原理角度（典型如基于 Transformer 模型的“猜字猜词式”生成路径，并不是传统意义上的“数据库+检索词”原理），还是从应用场景角度（典型如图片生成、视频换脸、小说生成，本身就不意形成真实的输出结果），都确实难以做到、也并不必须实现“结果保真”。这一点也是征求意见稿征求意见阶段，诸多专家、学者和业界从业人员讨论最为密集的“痛点”之一。

基于此，《暂行办法》不再提及“结果保真”的相关要求，而专注于对训练数据源合法合规性的要求、着力于强调训练及生成结果质量的提高以及对人工标注工作合规性的引导，还原和精准定位了生成式人工智能监管的真正方向。此外，《暂行办法》还专门提及，国家对利用生成式人工智能服务从事新闻出版、影视制作、文艺创作等活动另有规定的，从其规定。这也为精准监管生成式人工智能辅助该等领域部分场景下的“虚构创作”行为，铺设了可适用的制度通道。

4. 明确管辖：监管边界逐渐清晰

对于面向何种用户的生成式人工智能是监管的范围，面向何种用户的服务不是监管的范围，《暂行办法》在基本保持征求意见稿的框架同时，作出了更加清晰的规定。首先，利用生成式人工智能技术**向中华人民共和国境内公众**提供生成文本、图片、音频、视频等内容的服务，适用《暂行办法》。其次，《暂行办法》强调了如下例外：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，不适用本办法的规定。

上述管辖边界的规定与目前实践中的做法相契合。实践中，由于算法备案与安全评估（详见下文介绍）办理时间长，其中特定的环节的文书资料准备也具备一定难度，一部分服务提供者为尽早上线产品、开展场景测试与迭代，往往选择面向境外用户提供服务（其筛选用户的典型方式是阻止来自中华人民共和国境内的 IP 地址注册或使用，或仅允许境外手机号进行注册）。这一“面向海外”的服务形态，确实不落入当前《暂行办法》的监管管辖范围内。因此，《暂行办法》厘定监管边界，一定程度上为希望尽快上线、仅面向海外用户的服务提供者提供了制度上的便利通道。

与“面向海外”的服务形态不在明确监管范围之内相对，海外“面向境内”的生成式人工智能服务已经受到了监管的关注。具体地说，实践中存在部分生成式人工智能服务提供者通过 API 直接调用境外大模型，用以提供特定服务、技术或适用已成型的数据资源的情形。该等操作易产生大模型自身合规风险、数据出境风险等一系列合规问题。针对这一情况，《暂行办法》明确规定，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。综合上述监管背景，从境内已有实践讯息（典型如金融机构）来看，选择与境内大模型提供者（特别是已经完成算法备案的大模型提供者）进行技术合作、接口调用，将会人工智能服务提供者成为更为合规、稳妥的选择。

此外，《暂行办法》规定，外商投资生成式人工智能服务行业，应当符合外商投资相关法律、行政法规的规定；体现了人工智能服务行业与现有的外商投资监管框架的接轨要求。尽管从当前监管架构来看，我国外商投资负面清单没有一般性禁止外商投资生成式人工智能服务，但实践中，特别是典型的“to C”（面向用户的）生成式人工智能服务，相对容易在融合性应用场景下，涉及对外资有一定限制或禁止行业领域并受限于相关细分行业的外资准入要求。存在外资限制或禁止要求的资质典型如 B25 类增值电信业务经营许可证，《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》，《广播电视节目制作经营许可证》等。综上所述，外国投资者如拟在中国经营或投资生成式人工智能行业，需明确，一方面，外资并非被一般性禁止进入生成式人工智能行业，另一方面，受限于生成式人工智能具体服务最终交付的模式形态，外国投资者宜实质性关注相关服务或产品是否触发其他融合应用场景下的准入资质要求、该等准入资质要求是否禁止外商投资或者对外资股权比例有所限制。

5. 注重务实：训练数据义务和模型改进义务的可行性回归

征求意见稿对生成式人工智能服务提供者提出了非常高的与训练数据源有关的义务，并规定，用于生成式人工智能产品的预训练、优化训练数据，应**能够保证**数据的真实性、准确性、客观性、多样性。实践中，一方面，数据是否“真实”“准确”“客观”“多样”本身没有统一的标准，难以进行绝对把握或判断；另一方面，即使就特定性质的数据而言，可能实现对其准确性的相对控制，但在大模型的行业应用语境下，企业对训练数据的量级需求极大（且训练数据量级及相应形成的参数量级甚至成为衡量大模型效果的最重要指标之一），如要求企业对训练数据的“真实性、准确性、客观性、多样性”进行穷尽的审查并作出“保证”，事实上对企业作出了过重的合规要求。

基于上述考虑，《暂行办法》将监管重点放在训练数据来源合法、不侵权以及质量提升等方面，强化了合规要求可行性。具体地，其要求服务提供者使用具有合法来源的数据和基础模型；涉及知识产权的，对数据源的使用不得侵害他人依法享有的知识产权；涉及个人信息的，对数据源的使用应当取得个人同意或者符合法律、行政法规规定的其他情形；采取有效措施提高训练数据质量，增强训练数据的真实性、准确性、客观性、多样性。实践中，生成式人工智能行业企业可以通过从数据交易所以及产生数据或文字资料的行业性供应商采买，以获取一些公共数据或公开数据，从而在一定程度上确保数据来源合法性。企业也可以在训练过程中，对测试的生成结果进行适当监控、测算或核验，逐步改善和提高训练数据的真实性、多样性。

《暂行办法》另一务实修订是取消了征求意见稿对模型修改的时间限制。征求意见稿规定，对于运行中发现的、用户举报的不符合征求意见稿要求的生成内容，除采取内容过滤等措施外，应在 **3 个月内**通过模型优化训练等方式**防止其再次生成**。从技术维度来看，受限于大模型技术机理设置，对其开展的优化训练并不一定能够完全或精准阻拦或隔断特定具体内

容的再次生成。从行业角度来看，不同技术机理的大模型，通过定向调优或定向训练来精准、完全杜绝生成特定内容的难度，可能存在巨大差异。征求意见稿颁布后，一些从业者也对此作出了讨论，认为在 3 个月内通过优化模型防止再次生成某些特定结果的技术可行性有限。基于此，《暂行办法》采取更务实的路径，规定提供者发现违法内容的，应当及时采取停止生成、停止传输、消除等处置措施，采取模型优化训练等措施进行整改，并向有关主管部门报告。此类更具可行性的务实规定也使得企业有动力持续投入合理资源、渐进式落实合规要求。

6. 衔接融合：注意与现有监管工具顺利接轨

征求意见稿阶段，其已经意在利用现有制度，作为生成式人工智能在“投放市场”之前的监管抓手；但在规则文本层面，存在一些具体衔接和关联的细节差距。例如，征求意见稿规定，利用生成式人工智能产品向公众**提供服务前**，应当按照《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》向国家网信部门申报安全评估，并按照《互联网信息服务算法推荐管理规定》（简称“**《算法推荐管理规定》**”）履行算法备案和变更、注销备案手续。《暂行办法》修改为，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，**应当按照国家有关规定**开展安全评估，并按照《算法推荐管理规定》履行算法备案和变更、注销备案手续。

这一修改在制度层面上形成了更准确的衔接。《算法推荐管理规定》的制度是，具有舆论属性或者社会动员能力的算法推荐服务提供者应当在**提供服务之日起十个工作日内**通过互联网信息服务算法备案系统进行备案。《具有舆论属性或社会动员能力的互联网信息服务安全评估规定》规定需要开展安全评估的五种情形中，前**两种情形是在上线或者功能增设前**提交安全评估报告，后三种情形是**自相关情形发生之日起 30 个工作日内**提交安全评估报告。因此，两个文件并非所有场景都要求在向公众提供服务前完成备案和评估。《暂行办法》的规定进行了更准确的描述，避免了制度间衔接的立法技术性难题。

需要特别注意的是，在过去一段时间里，如果典型的生成式人工智能服务（在实践层面，更多地体现为可感知的基于大模型的生成式服务）移动应用程序需要在应用市场上架，部分应用市场经营者可能会要求上架的移动应用的运营者事先完成算法备案和安全评估，客观上与《算法推荐管理规定》的要求并不完全吻合。因此，在《暂行办法》于 8 月份正式施行后，实践中对于移动应用程序完成算法备案和安全评估的时点把握尺度以及 APP 上架审核要求是否会有所调整，也是值得观察的一个方面。

此外，根据我们对算法备案和安全评估的实际操作经验，两项手续的准备工作相对繁杂；特别是算法备案所需准备的材料，对申请人的内部制度的丰富性、体系性、完备性提出了较高的要求。依相关规定需要进行算法备案和安全评估的生成式人工智能服务提供者，应在研发阶段尽早准备，特别是在制度建立方面“未雨绸缪”，方能在更大程度上增强及时完成算法备案和安全评估的可预见性，更好地匹配产品在境内公开上架的时间节点，避免陷入冗长的“内测”或被迫仅仅面向海外用户运营。

下图是算法备案与安全评估的实务关注点：

第二部分：合规控制点概览

生成式人工智能服务的提供者，目前视具体形态，在相当部分场景下也需要遵守《算法推荐管理规定》《互联网信息服务深度合成管理规定》（简称**“《深度合成管理规定》”**）的要求。下表综合《暂行办法》《深度合成管理规定》和《算法推荐管理规定》，简要列示合规控制点：

合规控制点类型	合规控制点内容
用户注册与管理	• **规则公开**：制定和公开平台管理规则、公约、服务协议。 • **实名认证**：落实真实身份信息认证制度。 • **投诉处理**：设置便捷、有效的用户投诉、举报入口；公布处理流程和反

	馈时限。 • **服务协议**：与使用者签订服务协议，明确双方权利义务。 • **用户引导与防沉迷**：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，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。 • **用户违法行为管理**：发现使用者利用生成式人工智能服务从事违法活动的，应当依法依约采取警示、限制功能、暂停或者终止向其提供服务等处置措施，保存有关记录，并向有关主管部门报告。
算法机制机理提升	• **反不良模型**：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。 • **反歧视**：在算法设计、训练数据选择、模型生成和优化等过程中，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。 • **反垄断与反不正当竞争**：尊重知识产权、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。 • **禁止不合理差别待遇**：不得根据消费者的偏好、交易习惯等特征，利用算法在交易价格等交易条件上实施不合理的差别待遇等违法行为。 • **反不良关键词**：不得将违法和不良信息关键词记入用户兴趣点或者作为用户标签并据以推送信息。 • **透明度改进**：基于服务类型特点，提升服务透明度，提高生成内容的准确性和可靠性。
生成内容治理	• **不得生成违法内容**：不得生成煽动颠覆国家政权、推翻社会主义制度，危害国家安全和利益、损害国家形象，煽动分裂国家、破坏国家统一和社会稳定，宣扬恐怖主义、极端主义，宣扬民族仇恨、民族歧视，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 • **审核义务**：采取技术或者人工方式对输入数据和合成结果进行审核，建立健全用于识别违法和不良信息的特征库，记录并留存相关网络日志。 • **尊重他人权益**：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。 • **停止违规内容与模型优化**：提供者发现违法内容的，应当及时采取停止生成、停止传输、消除等处置措施，采取模型优化训练等措施进行整改，并向有关主管部门报告。 • **推荐管理**：加强算法推荐服务版面页面生态管理，建立完善人工干预和用户自主选择机制，在首页首屏、热搜、精选、榜单类、弹窗等重点环节积极呈现符合主流价值导向的信息。
建立健全辟谣机制	发现利用深度合成服务制作、复制、发布、传播虚假信息的，应当及时采取辟谣措施，保存有关记录，并向网信部门和有关主管部门报告。
标识要求	• **标识义务**：对使用其服务生成或者编辑的信息内容，应当采取技术措施添加不影响用户使用的标识，并依法依规保存日志信息。 • **显著提示**：提供智能对话、合成人声、人脸生成、沉浸式拟真场景等具有生成或者显著改变信息内容功能服务的，应当在生成或者编辑的信息内容的合理位置、区域进行显著标识；提供非前述深度合成服务的，应当提供显著标识功能，并提示使用者可以进行显著标识。
训练数据合法性	• **合法来源**：应使用具有合法来源的数据和基础模型。 • **知识产权保护**：涉及知识产权的，不得侵害他人依法享有的知识产权。 • **个人信息保护**：涉及个人信息的，应当取得个人同意或者符合法律、行政法规规定的其他情形。 • **数据质量提升**：应采取有效措施提高训练数据质量，增强训练数据的真实性、准确性、客观性、多样性。

人工标注	制定符合要求的清晰、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，提升尊法守法意识，监督指导标注人员规范开展标注工作。
协助监管	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。

结语：

基于大模型的生成式人工智能技术及服务很可能构成各行各业的“基础设施”，几乎所有传统服务模式及形态都可能在其全部或部分环节利用大模型进行迭代与改变，即所谓基于大模型的“重做浪潮”。在这一次变革迭代周期，妥善设计监管框架，创造优势制度赛道，有利于激发竞争活力，提升垂直领域中企业的合规可触达性。而赛道参与者也需要准确理解合规要求，将有限的合规资源投入在关键的合规控制点，助力业务快速发展。