# 平衡 AIGC 服务的安全与发展
## ——议《生成式人工智能服务管理暂行办法》对 AIGC 服务的监管与促进

作者：杨迅 | 象玉婷 | 夏雨薇 | 任家谊 | 杨蕾

7 月 10 日，国家互联网信息办公室(“**国家网信办**”)连同其他六部门共同颁布了《AIGC 服务管理暂行办法》(“**《暂行办法》**”)，该法将于 8 月 15 日正式生效。《暂行办法》在算法监管的基础上进一步规范了 AIGC 服务(“**AIGC 服务**”)。《暂行办法》从 4 月 11 日其征求意见稿(“**《征求意见稿》**”)发布到正式颁布，仅 3 个月，如此迅速的立法进程体现了监管部门对 AI 技术发展的态度——重视技术发展，但也顾及发展带来的风险。

本文以 AI 技术发展为锚点，介绍《暂行办法》的立法亮点、《暂行办法》体现的监管原则、《暂行办法》与其他法律的衔接、以及《暂行办法》下有关 AIGC 服务监管的热点问题，进而试探讨安全与科技发展间矛盾却共生的互动关系。

## 一. 《暂行办法》立法亮点概述

《暂行办法》从征求意见稿到正式颁布虽然只有仅仅 3 个月，但是诸多修订体现出政府对 AI 发展的重视，以及试图平衡发展与安全的努力。

### 1. 主管机关

在主管机关方面，与《征求意见稿》由国家网信办独家发布相比，《暂行办法》由国家网信办与国家发展和改革委员会(“**国家发改委**”)、中华人民共和国教育部、中华人民共和国科技部、中华人民共和国工信部、中华人民共和国公安部以及国家广播电视总局局长联合发布。这意味着 AI 发展的法律规范不仅仅是出于国家网信办一家的监管，一定程度上综合了国家发改委、科技部、工信部等部门的意愿；不仅有这网络安全的考虑，也兼顾统筹发展和技术进步的方针。

如您需要了解我们的出版物，
请联系：

Publication@llinkslaw.com

**2. 适用范围**

与《征求意见稿》相比，《暂行办法》限缩了适用范围：(1)《暂行办法》将管辖范围限定在利用人工智能技术提供服务，而将研发人工智能技术排除在外；(2)《暂行办法》将生成式人工智能的生成物限定在“文本、图片、声音、视频”，而不包括代码，也是有意排除为计算机编程目的的研究性使用；和(3)《暂行办法》第 2 条还对其适用范围做出排除性规定，即利用 AIGC 服务从事新闻出版、影视制作、文艺创作等活动另有规定的，依照该等领域的规定，实际上也是排除了为生产制作目的使用 AIGC 技术情形。

可见，《暂行办法》将其管辖范围限定在直接面向一般公众的人工智能内容生成服务，而排除为研究、生产目的的 AIGC 技术的应用。

**3. 地域管辖**

原则上，《暂行办法》管辖在中国境内的 AIGC 服务。比如，根据《暂行办法》第二条第三款，“研发、应用生成式人工智能技术但未向境内公众提供服务”的，被明确排除出《暂行办法》的监管适用范围。

但是，针对境外实施的，对境内公众有影响的 AIGC 服务，《暂行办法》也给出了对策。《暂行办法》第 20 条新增了国家网信部门对于相关机构的通知义务，即“对来源于中华人民共和国境外向境内提供 AIGC 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。”该规定旨在增强对境外向境内提供 AIGC 服务的监管，即对于境外实施的 AIGC 服务，有损中国利益的，可能采取断开连接、停止解析等方式，制止其在中国的使用。

**4. 鼓励创新**

相比于《征求意见稿》，《暂行办法》更鼓励 AIGC 多行业、多领域的创新应用以及企业、科研机构等专业机构在内容创作、资源平台建设、风险防范、规则制定的等方面开展探索和国内、国际协作。

尤其是，《暂行办法》将《中华人民共和国科学技术进步法》(“**《科技进步法》**”)新增为其制定的法律依据，即《暂行办法》将作为《科技进步法》的下位法和特别法，体现《科技进步法》下“……促进科学技术进步，发挥科学技术第一生产力……”的基本宗旨，这也就为《暂行办法》从征求意见稿以监管为主的规则向正式稿监管与促进并重，安全与发展兼顾的转变打下基础。

**5. 加强个人信息保护**

在个人信息保护方面，《暂行办法》明确了 AIGC 服务提供者的个人信息保护责任，细化了个人信息保护在 AIGC 领域的落地。比如，《暂行办法》第 11 条明确要求服务提供者不得收集非必

要个人信息，同时，新增不得非法留存能够识别使用者身份或非法向他人非法提供使用者的使用记录的要求，本文第三部分第 2 条将详细展开。

此外，《暂行办法》亦新增了对 AIGC 服务进行安全评估和监督检查的相关机构和人员对个人信息和隐私的保密义务，要求其对在履行职责中知悉的个人隐私和个人信息依法予以保密，不得泄露或者非法向他人提供。一方面，体现出 AIGC 服务的监管与《个人信息保护法》下个人信息保护要求的互动，另一方面也体现出生成式人工智能在个人信息保护领域可能涉及较高的合规风险。

**6. 完善“避风港”制度**

《征求意见稿》中针对违法的生产内容，对 AIGC 服务提供者施以“举报-过滤”的类避风港监管原则，也在《暂行办法》得以完善。《暂行办法》要求 AIGC 服务提供者“承担网络信息内容生产者责任。”但是，在 AIGC 服务中，输出的内容实际上是预设的算法和用户输入的指示的综合产物。在这种情况下，要求 AIGC 服务提供者对内容承担绝对的责任是不切实际的。因此与《征求意见稿》类似，《暂行办法》参照网络平台运营者的责任制度，设定“避风港”机制，即一旦 AIGC 服务提供者发现违法内容，就应及时“*采取停止生成、停止传输、消除等处置措施*”，并新增向有关主管部门报告的义务。

此外，针对使用者利用 AIGC 服务从事违法活动的情形，在《暂行办法》第 14 条的规定下，AIGC 服务提供者需要向违法使用者采取警示、限制功能、暂停或者终止向其提供服务等处置措施。

由此，AIGC 服务提供者负有较为积极的对违法内容的监管和审查义务。与传统避风港制度下，平台采取措施的义务由用户通知“平台”删除违法内容作为触发，转变为 AIGC 服务提供者主动向 AIGC 服务使用者发出警示，这似乎意味着 AIGC 服务提供者需要采取技术手段和管理手段，监控其 AIGC 服务的实际使用。

## 二. 《暂行办法》下对 AIGC 管控的基本原则

相对于美国针对人工智能发展较为宽松的态度，《暂行办法》更接近于欧盟的分级监管的策略，重视人工智能发展带来的潜在安全问题，同时兼顾发展的需要。

**1. 平衡发展与安全**

平衡发展与安全一直以来都是网络安全立法的关键议题。随着 AIGC 服务的快速发展，在为人们的生活、工作、学习带来新机遇的同时，也产生了传播虚假信息、侵害个人信息权益、侵犯知识产权、数据安全、算法歧视和技术垄断等问题。相比于《征求意见稿》，《暂行办法》第 3 条提出“国家坚持发展和安全并重、促进创新和依法治理相结合的原则，采取有效措施鼓励生成式人工智能创新发展。”《暂行办法》第二章单设“技术发展与治理”章节，为 AIGC 的发展和安全提供

具体指引，推动 AIGC 基础设施和公共训练数据资源平台建设，鼓励人工智能基础技术的自主创新，并明确了训练数据处理活动和数据标注的要求。

OpenAI 于 3 月 15 日发布的 GPT-4 技术报告[1]，重燃了人们关于人工智能及其伴随风险的讨论。位于美国波士顿的生命未来研究所在一封公开信中呼吁暂停训练比 GPT-4 更强大的人工智能系统至少 6 个月，技术领导者(包括 Elon Musk 和 Steve Wozniak)、著名学者(包括 Yoshua Bengio 和 Stuart Russell)均签署了该公开信，敦促人工智能试验研究按下暂停键。[2]公开信中提议"人工智能实验室和独立专家应该利用这个暂停，共同制定和实施一套共享的用于超级人工智能设计和开发的安全协议。"虽然该提议的初衷是为了保障人工智能的安全性，但我们应该预见到维持人工智能的安全性不应留给人工智能实验室和独立专家，也不是一件可以一次性完成的事情。它需要监管机构的参与，以协调各个方面的一致性和持续性。

在人工智能立法方面，各个国家的监管机构以及国际组织都在做出努力并取得一定进展。

- 欧盟: 2019 年 4 月发布《可信赖 AI 的伦理准则》; 2022 年 11 月通过《数字市场服务法》; 2023 年 6 月 14 日《人工智能法案》(草案)在欧洲议会通过，预计 2023 年底生效;
- 美国: 2019 年 2 月发布行政令《维护美国在人工智能领域的领导地位》; 2022 年 10 月白宫科技政策办公室发布《人工智能权利法案蓝图》; 在州一级的层面上，一些州已经引入了解决算法危害的立法，包括加利福尼亚州[3]和康涅狄格州[4];
- 英国: 2021 年 9 月发布《国家人工智能战略》; 2023 年 3 月发布《人工智能白皮书》;
- 中国: 2017 年国务院发布《新一代人工智能发展规划》; 2022 年 3 月发布《互联网信息服务算法推荐管理规定》; 2022 年 11 月发布《互联网信息服务深度合成管理规定》;
- 联合国教科文组织: 2021 年发布《人工智能伦理问题建议书》;
- 经济合作与发展组织: 2019 年发布《关于人工智能设计国际标准的建议》。

其实，即使暂停了超级人工智能系统的开发，一些大语言模型系统的主要风险，如虚假信息、歧视等风险也仍然存在。AIGC 最大的风险实际上并不在于其本身，而是在于由于薄弱的或者不切合的 AIGC 管控导致的 AIGC 服务提供者以不负责任的方式开发和使用 AI 而造成的经济和社会损害。人工智能需要的不是发展暂停，而是治理，不发展是最大的不安全。

某种程度上，《暂行办法》试图平衡人工智能的安全与发展的价值取向。不同于互联网发展之初，相对粗放的，缺少监管的野蛮生长状态，在 AIGC 领域，政府试图以监管促发展: 明确监管要求的基础上，对具体的落地制度采取相对灵活的安排，避免法律的滞后制约 AIGC 服务的超前发展。

[1] https://cdn.openai.com/papers/gpt-4.pdf#:~:text=This%20technical%20report%20presents%20GPT-4%2C%20a%20large%20multimodal,as%20dialogue%20systems%2C%20text%20summarization%2C%20and%20machine%20translation.
[2] https://futureoflife.org/open-letter/pause-giant-ai-experiments/
[3] https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240AB331
[4] https://www.courant.com/2023/03/04/ct-governments-ai-use-is-already-extensive-raising-equity-and-privacy-concerns-a-proposed-bill-would-add-oversight/

## 2. 包容审慎和分级监管

作为平衡发展与安全的原则的具体体现，《暂行办法》第 3 条提出“对 AIGC 服务实行包容审慎和分类分级监管”。分类分级监管在行政资源有限的情况下是高效的。它既能够集中资源管控危害性大的高风险人工智能，又为低风险人工智能的发展和应用留出空间。

目前国内并无相应的分类分级监管规则的指引，但是对 AIGC 算法的备案要求体现了分级管理的思路。即仅有具有舆论属性和社会动员能力的 AIGC 服务的算法才需要备案。具体见本文第三部分第 3 节。

此外，欧盟的《人工智能法案》(草案)的分类分级措施可以提供参考。欧盟《人工智能法案》(草案)将人工智能系统的风险程度分为不可接受的风险、高风险、有限风险和最低风险。[5]

[5] https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

| | 定义 | 监管 |
|---|---|---|
| **不可接受的风险** | 不可接受风险的人工智能系统被认为对人类构成威胁，包括:<br>对人或特定弱势群体的认知行为操纵:例如鼓励儿童危险行为的声控玩具;<br>社会评分:根据行为、社会经济地位或个人特征对人进行分类;<br>实时和远程生物识别系统，如面部识别; | 禁止使用。 |
| **高风险** | 对安全或基本权利产生负面影响的人工智能系统将被视为高风险，分为两类<br>1) 在欧盟产品安全立法范围内的产品中使用人工智能系统，包括玩具、航空、汽车、医疗设备和电梯。<br>2) 属于八个特定领域的人工智能系统，必须在欧盟数据库中注册:<br>a) 自然人的生物识别和分类;<br>b) 关键基础设施的管理和运行;<br>c) 教育和职业培训;<br>d) 就业、工人管理和获得自营职业的机会;<br>e) 获取和享受基本私人服务、公共服务和福利;<br>f) 执法;<br>g) 移民、庇护和边境管制管理<br>h) 法律解释和法律适用方面的援助。 | 所有高风险的人工智能系统都将在投放市场之前以及在其整个生命周期内接受评估。 |
| **有限风险** | 使用人工智能时，该类别的产品和服务的使用会引发透明度问题。 | 风险有限的人工智能系统应符合最低透明度要求，使用户能够做出明智的决策。 |
| **最低风险** | 除上述三类风险外的人工智能系统。 | 没有特殊的干预和审查制度。 |

2023 年 6 月 14 日通过的《人工智能法案》(草案)新纳入了关于 ChatGPT 和 GPT-4 等生成式人工智能系统的规则[6]，首次为其所谓的“基础模型”给出了明确的定义。“基础模型”，指在大规模的广泛数据上进行训练，为输出的通用性而设计，并可以适应各种不同任务的 AI 系统模型，如 ChatGPT、Bard 或 Stable Diffusion。欧盟为这些基础模型设立了(1) 所有基础模型的最低标准;(2) 高风险场景中具体应用的具体规则;(3) 人工智能价值链上的合作和信息交流规则，即开发者、部署者和(专业)用户之间的合作和信息交流规则。根据该提案，生成式人工智能系统的开发者(如 OpenAI)若想向欧盟用户提供其模型，就必须遵守某些最低标准，包括透明度义务、真实性义务及公开训练数据集中涉及版权的信息。

[6] 《人工智能法案》Article 28 b

## 三. 《暂行办法》与其他法律的衔接

《暂行办法》作为 AIGC 服务领域的行政规章，其很多制度的落实和执行依赖于相关法律和行政法规。

### 1. 生成式人工智能兴起所引发的知识产权保护与争议

《暂行办法》要求训练数据不侵犯知识产权，而判断是否侵犯知识产权则依赖于具体的相关《著作权法》规定和其它知识产权有关法律。

AI 技术有关的知识产权争议，主要集中于以下三点: (1) AIGC 权属问题; (2) 模型训练的侵权风险; 以及(3) AIGC 的侵权风险。我们在*《风起云涌的 AIGC: 监管、知识产权与算法安全——中篇: AIGC 之知识产权》二. AIGC 作品的权属*中已经对上述(1)的问题试进行过探讨，本文主要考虑 AIGC 与模型训练以及与用户互动中可能存在的侵权风险与保护思路。

#### (1) 模型训练行为

《暂行办法》第 7 条明确规定，“*提供者应当依法开展预训练、优化训练等训练数据处理活动，涉及知识产权的，不得侵犯他人依法享有的知识产权。*”算法模型训练需要庞大的数据库做支持，训练素材一般来自于公开互联网，不可回避地包括到著作权保护的作品(“**作品**”)。从模型训练行为本身的目的考虑，这种使用需涵盖作品的实质内容或几乎所有内容，极易存在侵犯作品复制与修改权的风险。

模型训练行为所引发的著作权纠纷在域外已有实例。今年 1 月，有作者向利用扩散模型等算法进行图片生成的公司 Stability AI, LTD,等提起集体诉讼(“**Stability 案**”)，指控其 “*未经作者同意，使用作品以进行机器学习、人工智能等….的训练*”[7]。类似地，上月 7 号 Sarah Silverman 等人也对 OpenAI, INC.等提起著作权等侵权之诉(“**OpenAI 案**”)，认为 OpenAI 公司 “*在语言模型训练中未经原告同意复制了原告的作品，且由于 OpenAI 的语言模型的运作不能缺少其模型中留存的、从他人作品中提取出的信息表达部分，OpenAI 模型构成侵权的衍生作品*”。[8]

在一些场合，模型训练中对素材的使用可能被视为合理使用。在我国现行的《著作权法》中，被诉侵权行为，如符合《著作权法》第 24 条的合理使用情形，同时不影响作品的正常

[7] Andersen et al v Stability AI Ltd. et Compliant ¶155“Defendants had access to but were not licensed by Plaintiffs or the Class to train any machine learning, AI, or other computer program, algorithm, or other functional prediction engine using the Works”

[8] Silverman Et Al V. Openai, Inc. Et Al ¶56-57 “OpenAI made copies of Plaintiffs’ books during the training process of the OpenAI Language Models without Plaintiffs’ permission, …Because the OpenAI Language Models cannot function without the expressive information extracted from Plaintiffs’ works (and others) and retained inside them, the OpenAI Language Models are themselves infringing derivative works…”
”

使用，且没有不合理地损害著作权人的合法权益的，属于法律允许的合理使用行为。就模型训练行为而言，似乎可以从“合理使用”角度考虑。

模型训练似乎类似于《著作权法》第 24 条下“学习使用”，即“为个人学习、研究或者欣赏，使用他人已经发表的作品”。但是，此处的个人指的是“自然人”，而非“人工智能”，机器的学习虽然在概念上相应于个人学习，但在法律含以上与之不同。

那么，模型训练行为是否属于《著作权法》第 24 条下的“适当引用”?一般而言，“适当引用”仅允许引用部分作品内容以评价作品，或说明其他问题。在模型训练行为中，AIGC 服务提供者可以主张，模型训练所使用作品，仅只为了让模型更了解人类语言、绘画或其他作品创作中的模式，比起对作品进行“引用”和“替代”的，算法模型更多是对相应的作品进行编码与训练。例如，利用扩散模型进行图片生成时，其原理主要包含正向扩散与逆向扩散，在对原始输入图片解码后，使用马尔卡夫链模型，计算原始输入图片到纯高斯噪声状态变化的概率分布，并让机器通过计算逆推正向扩散的反过程，重现图片。

AIGC 服务提供者利用“适当引用”的主要不足之处在于：模型训练行为使用与转化的图片是海量的，引用占比极大，且难以剔除其商业目的与属性。但另一方面，模型训练行为确与一般的侵权行为不同，其服务于机器算法，不是《著作权法》意义上典型的作品复制与传播行为；同时，模型训练行为未必影响到了原作者的利益，甚至不会触及原作品的流通与传播渠道，特别是对于图像等内容来说，训练前提就是将其转化为编码语言，这种使用似乎并不会削弱原作品在原有流通渠道中所强调的艺术价值。

另一种意见认为，训练中的使用属于“转化性使用”。“转化性使用”并非是我国著作权的概念，是在美国 1994 年最高院的判例 *Campbell v. Acuff-Rose Music*.中的确立的一种合理使用情形。其通过合理使用的四要件分析[9]，认为在判断某一作品使用行为是否属于合理使用的，应考虑其使用性质，如果其没有重复原作品，而赋予了原作品新的价值属性的，不属于侵权。“转化性使用”能被豁免的原因在于，作品的新旧价值互不干扰，均能鼓励推动创新。而我国最高院在 2011 年发布的《关于充分发挥知识产权审判职能作用推动社会主义文化大发展大繁荣和促进经济自主协调发展若干问题的意见》(“**《若干意见》**”)中也提到，“在促进**技术创新和商业发展**确有必要的特殊情形下，*考虑作品使用行为的性质和目的*、*被使用作品的性质*、*被使用部分的数量和质量*、*使用对作品潜在市场或价值的影响等因素*，如果该使用行为既不与作品的正常使用相冲突，也不至于不合理地损害作者的正当利益，可以认定为合理使用。”《若干意见》中提出可考虑的四个合理使用因素，恰好是美国法律下合理使用四要件的判断要素。

近年来，我们也发现，已经有部分企业在著作权纠纷的案件中提出了“转化性使用”作为支持合理使用的辩护意见[10]。而在(2015)沪知民终字第 730 号中，一审法院使用了四要件的标

[9] 使用目的与特点、被使用作品的性质、被使用部分的比例与实质性、对市场或原作品的价值的影响。
[10] (2021)京 0102 民初 37702 号、(2019)粤 03 民初 2836 号等

准得出合理使用的结论，也被终审法院支持。综上，虽我国暂未确定“转化性使用”这一合理使用概念，但考虑到《若干意见》以及部分司法实务的现状，了解美国法下 “转化性使用”在人工智能领域的应用对 AIGC 服务提供者有很高的实践帮助价值。

早在 OpenAI 案前，USPTO 在 2019 年就人工智能创新的知识产权保护问题向公众征求意见[11]。其中问题 3 是“*AI 算法或处理通过吸收大量受保护的作品以完成训练与进化，是否有任何现存的法律理论或案例可以支持这种使用的合法性*?”OpenAI, LP (同样是 OpenAI 案的被告之一)起草了回复，认为模型训练行为应属于“转化性使用”[12]。其理由在于:

(1) 使用性质: 模型训练行为的使用性质具有高度的转化性。原始作品提供的是独立的娱乐价值，而模型处理行为是为了训练。没有人可以通过研究 AI 系统和 AI 输出内容的形式获取原始作品[13]。在高转化性的使用行为下，公司是否盈利、是否有商业目的也不会决定性影响前述判断。

(2) 使用部分: 模型训练行为几乎使用了原作品的所有部分，但这种使用并不被公众可及，不会替代原作品的使用。且完整使用对模型训练行为本身是必要的，这决定了模型训练后的完整度和实用性。

(3) 使用影响: 模型训练行为不会影响原作品的市场，语料库受众对象是机器，而非人类，语料库使用原作品并不会导致原作者的受众减少。

值得关注的是，今年 1 月的 Stability 案已经有了一些进展。在 4 月 18 日被告提出的撤诉申请(motion to dismiss)中，被告在前置性说明内强调，Stable Diffusion 复制与记忆图片并不是为了分发，而是为了对图画中的数百万个参数进行开发完善。[14]模型训练行为是否属于合理使用或是侵权行为，希望 Stability 案以及 OpenAI 案能给我们一个答案。

### (2) AIGC 的侵权风险

除了模型训练行为以外，AIGC 服务的合法性与使用者更为相关。输出内容是否可能因为与受保护的作品“实质性相似”+“存在接触可能性”而被认为构成著作权侵权。如果构成的，谁才是侵权人?AIGC 服务使用者还是提供方?

《暂行办法》并没有明确说明 AIGC 知识产权侵权的问题，但强调了服务提供者应当承担网络信息内容生产者的责任(第 9 条)。在《网络信息内容生态治理规定》规定下，网络信息内

11 《Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation》 https://www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf

12 《Comment Regarding Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation Docket No. PTO–C–2019–0038 Comment of Open AI, LP Addressing Question 3》

13 “nobody looking to read a specific webpage contained in the corpus used to train an AI system can do so by studying the AI system or its outputs.”

14 Andersen et al v Stability AI Ltd. et, Amended Motion to Dismiss, ¶2 “*To be clear, training a model does not mean copying or memorizing images for later distribution. Indeed, Stable Diffusion does not “store” any images. Rather, training involves development and refinement of millions of parameters that collectively define—in a learned sense—what things look like. Lines, colors, shades, and other attributes associated with innumerable subjects and concepts.*”

容生产者不得损害他人合法权益。网络信息内容的生产者，不得制作、复制、发布违法信息，还需要采取措施抵制违法行为与违法内容。可以认为，知识产权应作为考虑因素。

从著作权侵权角度而论，AIGC 服务提供者需有效避免 AIGC“实质性相似”以及“存在接触可能性”的可能。“接触可能性”对 AIGC 服务提供者而言一般难以回避，毕竟模型训练涉及数量范围很广。但有趣地是，正因为 AIGC“集百家所长”，其与单个作品的表达之间的相似程度被极大削弱。AIGC 服务提供者也许可以宣称，AIGC 更倾向于“风格相似”而非“表达相似”的新的作品。

Stability 案中，原告也在起诉状中承认“插入多种图像后…，…被告软件中基于特定指令输出的图片与任何素材数据都无法高度匹配。”[15]这段话也在被告的撤诉申请中屡遭引用作为抗辩。在 AI 技术不断发展的今天，AI 技术提供者也可以对模型与算法本身进行限制，从而进一步降低其输出数据与训练数据之间的相似程度，从而规避侵权风险。

## 2. 生成式人工智能的个人信息保护问题

如何最大发挥个人信息的经济效益并平衡个人权益保护，对任何 AIGC 服务提供者来说都是难题。早在今年 4 月份，路透社爆出意大利个人数据保护局宣布禁止使用 ChatGPT，因其涉嫌违法收集小于 13 岁的儿童的个人信息。目前，ChatGPT 在意大利地区已经恢复使用，但在进入聊天界面前设置了弹窗，用户必须确认其已满 18 岁，或已满 13 岁且获得父母/监护人的同意。就在不久前的 7 月 13 日，美国联邦贸易委员会召开听证会，对 OpenAI 公司存在的隐私安全、消费者权益相关的潜在风险进行调查。

本次出台的《暂行办法》强调了 AIGC 服务提供者的个人信息保护义务，并与《个人信息保护法》实现联动。重点个人信息保护义务如下:

### (1) 模型训练阶段，收集个人信息应当具备合法基础并符合其他法律要求(第 7 条)

本条对应《个人信息保护法》第 13 条等的义务，即要求 AIGC 服务提供者确保模型训练时使用的个人信息均具有合法基础。

如上所述，AIGC 服务语料库的训练数据一般来自于公开互联网，且数据量庞大。追溯所有个人信息主体并取得他们的个人同意几乎没有可能。对于 AIGC 服务提供者而言，能否可以援引《个人信息保护法》第 13 条(四)“*在合理的范围内处理个人自行公开或者其他已经合法公开的个人信息;*”作为处理个人信息的合法基础便成为了关键。该合法基础需要关注信息公开的方式，以及拟进行的处理的性质。

[15] Andersen et al v Stability AI Ltd. et Compliant ¶93 “none of the Stable Diffusion output images provided in response to a particular Text Prompt is likely to be a close match for any specific image in the training data.”

一方面，对公有领域个人信息进行使用，必须需要限于“合理的范围内”还不得“对个人权益产生重大影响”(第 27 条)，否则需要重新获得个人的同意。模型训练行为是否属于“对个人权益产生重大影响”、超出合理范围的处理行为?AIGC 服务提供者可以主张，模型训练行为本身并不针对任何个体，训练过程打破了个人信息的个体颗粒度，反倒不会体现某一个体的特征，对个人权益产生的影响也较小。换言之，这也促使 AIGC 服务提供者尽可能减少模型训练行为对个人权益的影响，例如尽可能减少对原始数据的留存与重现，对训练数据进行必要的去标识化或匿名化处理等，否则 AIGC 服务提供者只能退而求其次，诉诸高成本的“知情同意”合法基础。

另一方面，处理个人信息需要履行告知义务。在难以触达所有用户的公开互联网中，公告似乎是 AIGC 服务提供者能够选择的最可行方式，但公告方式，严格意义上而言不一定能覆盖所有相关个人，其效果也存在不确定性。且值得注意的是，《个人信息保护法》中并没有基于告知难度过高，成本过大显不合理而豁免个人信息处理者告知义务的例外。

**(2) 服务阶段，收集使用个人信息应遵守个人信息保护义务(第 9 条)**

在 AIGC 服务阶段，个人信息保护义务主要有以下几个方面:

a) 知情同意(第 9、11 条)

AIGC 服务过程中收集、对外提供用户个人信息的，应当提供个人信息处理说明并获得用户同意(或满足其他的合法基础)。

b) 最小必要(第 11 条)

《暂行办法》要求 AIGC 服务提供者不得收集非必要个人信息。这可能涉及到服务注册登录阶段以及实际使用阶段，AIGC 服务提供者收集的个人信息需与其所声明的且实际使用的业务场景相适应。

此外，《暂行办法》对于输入信息、使用记录的留存时间进行了明确规定。要求 AIGC 服务提供者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。笔者推测，《暂行办法》可能考虑到了 AIGC 服务提供者留存使用者的使用记录并进一步进行用户画像或进行其他模型训练的风险。当然，这一条并非禁止 AIGC 服务提供者留存任何个人信息，而是强调留存的合法前提在于：用户没有主动要求撤回同意或删除相关信息，且个人信息留存后具有与之适应的处理目的与业务场景。

c) 权利响应(第 11 条)

《暂行办法》对个人信息权利进行了重述，即提供者应当提供及时、便捷有效的响应渠道以响应个人对个人信息的查询、复制、更正、补充、删除等个人信息权利的请求。

前述个人信息保护义务并非《暂行办法》首创，而是《个人信息保护法》下的规定在AIGC服务领域的重述。但是，如何在AIGC服务场景中落实个人信息保护的相关规定，比如就用户向AIGC提供的个人信息，如何回应用户的查询、复制等请求，则还是一个有待实践的问题。

### 3. AIGC服务中的算法服务相关要求

几年来已通过了数部与算法相关的管理规定。从范围较广的"算法服务"，到"深度合成类的算法服务"，再到本文聚焦的"AIGC服务"，有关算法立法在不断精细化。AIGC服务不但需要满足《暂行办法》下的特殊规定，还必须遵守与算法服务相关的一般规定。

**首先，对于算法的披露义务。**参见本文第一、2有关算法透明度、准确性和可靠性要求的介绍。

**其次，对于算法的安全评估和备案义务。**《暂行办法》延续了《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》中对于安全评估与备案义务的规定。也就是说，AIGC服务与其上位概念深度合成类服务，算法推荐服务一致，提供具有舆论属性或者社会动员能力的信息服务的进行备案。有趣地是，作为备案前提，"舆论属性""社会动员能力"的定义却并不明确。根据《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》下的定义，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的，均可以视为具有"舆论属性或社会动员能力"。一般认为，渠道特征是重要的考量因素——即使受众很少，但一旦渠道是较为公开的且面对不特定用户的，都不能排除具有"舆论属性或社会动员能力"的可能。

算法备案监管已经提上日程，AIGC 服务提供者可以趁《暂行办法》生效前夕进一步向监管征询了解备案问题。截止8月份，网信办已经发表了4份算法备案清单(http://www.cac.gov.cn/2022-08/12/c_1661927474338504.htm)，以及一份深度合称服务备案清单(http://www.cac.gov.cn/2023-06/20/c_1688910683316256.htm)。已经备案的服务提供者、技术支持者基本仍限于主流的互联网信息服务公司。

**最后，对于生成内容的标识义务。**《暂行办法》要求提供者按照《互联网信息服务深度合成管理规定》对图片、视频等生成内容进行标识。目前对标识的要去主要限于两类。一类是《互联网信息服务深度合称管理规定》第17条中所列举的可能导致公众混淆或者误认的生成内容。至于不落入前述条款的生成内容，提供者不需要进行显著标识，但应当提供给使用者标识功能以供其选择使用。

## 四. 《暂行办法》下的热点问题

对于AIGC服务中经常被关注的问题，《暂行办法》或多或少给予了回应。

### 1. 数据来源的合法性要求和判断

人工智能依赖三大要素，即：数据、算法和算力，而数据则是人工智能研究和发展的基石。为训练出更加准确、可靠的模型，需要大规模、高质量的数据作为养分，以进行数据的分析、预测和决策。《暂行办法》第 7 条对 AIGC 服务提供者在开展预训练、优化训练过程中，数据及基础模型的合法性进行了规制，其要求数据来源必须合法，不得侵害他人依法享有的知识产权及个人信息权益等。

训练模型的数据主要来源于公开数据、自有数据以及第三方数据。对于不同来源及类型的数据合法性判断，所依据的法律及合规措施不尽相同，例如，在生成式人工智能的训练数据中，企业可能使用了通过爬虫技术爬取的公开数据，对于这些数据的使用是否合法，则需要具体结合《反垄断法》以及《反不正当竞争法》等法律法规以及目前司法实践的相关要求，避免对原始数据的权利人或平台数据的开发者造成侵权或其他权益上的损害；又如，训练数据中可能包含了他人的个人信息，对于该等数据的收集及使用，则应当遵守上位法《个人信息保护法》中有关获得个人信息主体的知情同意或其合法基础的相关规定；再如，若企业使用了未经许可他人享有著作权的作品进行数据训练，是否构成《著作权法》规定的“合理使用”范围则值得商榷，正如我们上文所提到的，合理使用认定的过程往往很复杂，且因具体情况而定，特别对于 AI 技术这样的新兴行业，短时间内无论是司法、执法机关亦或是 AIGC 服务提供者，都仍在探索把握。由此则可能具有侵犯他人合法知识产权的风险，企业应当注意规避相关风险。

**2. 算法透明度、准确性和可靠性要求**

算法透明是人工智能治理领域公认的原则，其要求算法所有者对算法的机制、决策过程等进行披露和公示，旨在保护公众的知情权。我国在《关于加强互联网信息服务算法综合治理的指导意见》中，也强调了算法应用应当透明可释的原则，在《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》等规范中亦明确了算法备案、安全评估等监管和问责的手段。

相比于《征求意见稿》，《暂行办法》删除了其 17 条对生成式人工智能提供者“提供可以影响用户信任、选择的必要信息”以及“人工标注数据的规模和类型，基础算法和技术体系等”，而是要求提升 AIGC 服务的透明度(第 4 条第 5 项)以及对算法机理等予以说明(第 19 条)。由此变化可以看出《暂行办法》相对降低了对 AIGC 服务提供者在算法透明度上的要求。其一，实现算法完全透明存在技术上的难度，即机器学习法存在“算法黑箱”，算法的推演不完全依照人类逻辑，会导致部分算法无法被完全解释；其二，算法通常属于企业的核心竞争力，可能构成商业秘密或知识产权，过分的公开披露企业算法的运作机制将可能破坏企业的市场竞争力，影响企业的发展动力。尤其可见，《暂行办法》相较于征求意见稿平衡了公众知情权与企业商业秘密、知识产权保护之间的关系。

同样，相对于《征求意见稿》，《暂行办法》适当降低了对训练数据质量的要求。《征求意见稿》

第 7 条中要求“能够保证数据的真实性、准确性、客观性、多样性”，但事实上，要求训练的数据完全符合真实性、可靠性等要求，对于服务提供者而言往往难以保证。首先，由于加入模型训练的数据很多都源于公开领域，而该部分数据可能良莠不齐，不可避免会有事实性的错误或虚假内容等。其次，即使数据准确，也未必客观，数据本身可能存在社会性的偏差和误解。

《暂行办法》将《征求意见稿》修订为“*采取有效措施，提供训练数据质量、增强训练数据的真实性、准确性、客观性、多样显示器设置专注性*”，则是更加符合 AIGC 服务领域的现实，强调在数据训练过程中行为方式和主管因素。《暂行办法》适度放宽了生成式人工智能数据训练方面的具体合规要求，从一定程度上减轻了服务提供者的相应责任，也是更多的从实际的角度进行了考量。

### 3. 生成内容的责任

服务提供者需要对生成式人工智能生成的内容负有一定的责任。根据《暂行办法》第 21 条，“*提供者违反本办法规定的，由有关主管部门依照《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》、《中华人民共和国科学技术进步法》等法律、行政法规的规定予以处罚；法律、行政法规没有规定的，由有关主管部门依据职责予以警告、通报批评，责令限期改正；拒不改正或者情节严重的，责令暂停提供相关服务。*”如下，我们整理了服务提供者对人工智能生成的内容负有的义务及需要承担的法律责任。

| 责任 | 义务 | 法律责任 |
| --- | --- | --- |
| **合法性责任** | 坚持社会主义核心价值观，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 | 《科学技术进步法》第 112 条<br>违反本法规定，进行危害国家安全、损害社会公共利益、危害人体健康、违背科研诚信和科技伦理的科学技术研究开发和应用活动的，由科学技术人员所在单位或者有关主管部门责令改正；获得用于科学技术进步的财政性资金或者有违法所得的，由有关主管部门终止或者撤销相关科学技术活动，追回财政性资金，没收违法所得；情节严重的，由有关主管部门向社会公布其违法行为，依法给予行政处罚和处分，禁止一定期限内承担或者参与财政性资金支持的科学技术活动、申请相关科学技术活动行政许可；对直接负责的主管人员和其他直接责任人员依法给予行政处罚和处分。<br>《网络安全法》第 70 条 |

| 责任 | 义务 | 法律责任 |
| --- | --- | --- |
| | | 发布或者传输本法第十二条第二款和其他法律、行政法规禁止发布或者传输的信息的，依照有关法律、行政法规的规定处罚。<br>《网络安全法》第 71 条<br>有本法规定的违法行为的，依照有关法律、行政法规的规定记入信用档案，并予以公示。 |
| **准确可靠性责任** | 采取有效措施，提升 AIGC 服务的透明度，提高生成内容的准确性和可靠性。 | 法律、行政法规没有规定的，由有关主管部门依据职责予以警告、通报批评，责令限期改正；拒不改正或者情节严重的，责令暂停提供相关服务。 |
| **网络信息内容生产者责任** | 需满足《网络信息内容生态治理规定》第 4-7 条的要求 | 同合法性责任中的法律责任 |
| **标注责任** | 按照《互联网信息服务深度合成管理规定》对图片、视频等生成内容进行标识 | 法律、行政法规没有规定的，由有关主管部门依据职责予以警告、通报批评，责令限期改正；拒不改正或者情节严重的，责令暂停提供相关服务。 |
| **监管和报告责任** | 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。 | 法律、行政法规没有规定的，由有关主管部门依据职责予以警告、通报批评，责令限期改正；拒不改正或者情节严重的，责令暂停提供相关服务。 |

## 五. 结语

《暂行办法》在吸收有关部门、行业和公众对《征求意见稿》的意见和建议的基础上，更加充分考虑了 AIGC 服务的技术特点和难点，体现了国家坚持发展与安全并重，在鼓励创新、支持新兴 AI 服务领域发展的同时，采取包容审慎的原则，明确了 AI 服务领域的安全底线，同时也为 AIGC 服务企业的发展预留了的空间。这是我国对 AIGC 服务领域立法的一次积极探索，未来可以预见，将会有更加全面的监管治理体系形成。

对于提供 AIGC 服务的企业而言，则需要在开展科技创新的同时，提高自身的合规意识，在数据的开发使用、用户权益保护以及生成内容的准确性、透明性等方面按照《暂行办法》的指引建立良好的合规体系。

如您希望就相关问题进一步交流，请联系：

杨 迅
+86 21 3135 8799
xun.yang@llinkslaw.com

如您希望就其他问题进一步交流或有其他业务咨询需求，请随时与我们联系：**master@llinkslaw.com**


| 上海 | 北京 | 深圳 |
| --- | --- | --- |
| 上海市银城中路 68 号<br>时代金融中心 19 楼<br>T: +86 21 3135 8666<br>F: +86 21 3135 8600 | 北京市朝阳区光华东里 8 号<br>中海广场中楼 30 层<br>T: +86 10 5081 3888<br>F: +86 10 5081 3866 | 深圳市南山区科苑南路 2666 号<br>中国华润大厦 18 楼<br>T: +86 755 3391 7666<br>F: +86 755 3391 7668 |
| **香港** | **伦敦** | |
| 香港中环遮打道 18 号<br>历山大厦 32 楼 3201 室<br>T: +852 2592 1978<br>F: +852 2868 0883 | 1/F, 3 More London Riverside<br>London SE1 2RE<br>T: +44 (0)20 3283 4337<br>D: +44 (0)20 3283 4323 | |


**www.llinkslaw.com**

**Wechat: LlinksLaw**

**本土化资源 国际化视野**

免责声明:
本出版物仅供一般性参考，并无意提供任何法律或其他建议。我们明示不对任何依赖本出版物的任何内容而采取或不采取行动所导致的后果承担责任。我们保留所有对本出版物的权利。

© 本篇文章独家授权威科先行法律信息库发布，未经许可，不得转载。