君合专题研究报告

2023 年 5 月 16 日

与科学怪人共舞

——简析生成式人工智能的立法与监管

2023 年 2 月 18 日，Elon Musk 在推特上发帖，声称要发展“TruthGPT”。与一不留神就胡说八道的 ChatGPT 相对，TruthGPT 的目标是挖掘和生成尽可能无偏见的真实信息[1]。

ChatGPT 和 TruthGPT 都是生成式人工智能（Generative AI，下称“**生成式 AI**”）的典型前沿代表，能够创造内容和想法，包括对话、故事、图像、影片和音乐等等。由生成式 AI 联想到玛丽·雪莱笔下的科学怪人[2]，人类创造了他，但更重要的是如何做好直面和控制他的充分准备。未来已来，本文将带着读者从法律和技术两个角度速览全球主要市场的生成式 AI 的立法和监管。

一、主要市场的生成式 AI 立法概览

1. 中国

国家网信办于 2023 年 4 月 11 日推出《生成式人工智能服务管理办法（征求意见稿）》（下称“**征求意见稿**”），旨在促进生成式 AI 健康发展和规范应用。面向中国境内公众提供生成式 AI 服务的，应当符合法律法规、社会公德、公序良俗的要求。征求意见稿设置了转致性条款，监督检查和法律责任需归拢到我国网络安全和数据隐私保护领域的三部基础性法律，即《网络安全法》、《数据安全法》和《个人信息保护法》。

与生成式 AI 技术和应用相关的，还有此前颁布的两项部门规章，《互联网信息服务算法推荐管理规定》和《互联网信息服务深度合成管理规定》。《互联网信息服务算法推荐管理规定》提及了生成合成类的算法推荐技术，要求对算法生成合成的信息做显著标识。从事互联网新闻信息服务的，不得生成合成虚假新闻信息。《互联网信息服务深度合成管理规定》则更加细致地规定了包括智能对话、智能写作、人脸生成、人脸操控、姿态操控等具有生成或显著改变信息内容功能的深度合成服务应当遵守的要求，包括以显著标识的形式向公众提示，建立健全用户注册、算法机制机理审核、科技伦理审查、信息发布审核、数据安全、个人信息保护、反电信网络诈骗、应急处置等管理制度，加强训练数据管理、技术管理等等。

2. 美国

作为 ChatGPT 的诞生地，紧随中国发布征求意见稿，4 月 13 日，美国商务部下设的国家远程通信和信息管理局（NTIA）发布了一项有关 AI 可归责性政策的征求意见通知（AI Accountability Policy

[1] https://truthgpt.gitbook.io/truthgpt-whitepaper/ai-chatbot/overview

[2] 《科学怪人》（Frankenstein）是英国作家玛丽·雪莱在 1818 年创作的长篇小说，主要描写了生物学家弗兰肯斯坦通过拼凑尸块和电击创造了一个怪物，起初怪物心地善良，后来却因屡屡遭受人们的嫌恶而变得扭曲，对人充满仇恨，滥杀无辜。最后，这个怪物与他的创造者弗兰肯斯坦一起走向了毁灭。

Request for Comment）[3]，以期集思广益，编制包括生成式 AI 在内的 AI 系统问责政策。重点问题包括，当生成式 AI 融合于下游产品时，应当如何向用户解释该下游产品运用了生成式 AI，且该运用是可信赖的。征求意见通知指出，传统的 AI 审计已无法覆盖生成式 AI 的威胁，例如信息扭曲、虚假信息、深度伪造、隐私入侵等等。正如 2 月美国总统拜登签署的行政令中要求的，联邦政府部门设计和使用 AI 技术时，应摒除其中的偏见，保护公众不受算法歧视和威胁。[4]

美国联邦现行立法和行政文件，大多是关于 AI 而非专门针对生成式 AI 的规定。《国家人工智能倡议法》（National Artificial Intelligence Initiative Act of 2020）提出应警惕具有自主意识或者不受控制的 AI，建立可信赖的 AI 系统。根据该法，美国国家标准与技术研究院（NIST）负责建立全面的 AI 发展规范，其中就包括其在 2022 年 8 月发布的第二版《风险管理框架》（Risk Management Framework）。白宫科学和技术政策办公室（Office of Science and Technology Policy）于 2022 年 10 月发布了《人工智能权利法案蓝图》（Blueprint for an AI Bill of Rights），其中列举的算法歧视保护措施包括主动的公平性评估，使用有代表性的数据，谨慎采用人口特征数据，设计无障碍使用功能，开展差异测试，开展算法影响性评估并向公众公布等等。

3. 欧盟

欧盟现行 AI 立法主要集中在传统 AI 模型而非生成式 AI 模型上，但已逐步触及生成式 AI 的问题。

2020 年 2 月 19 日，欧盟委员会发布的《人工智能白皮书》[5]强调，欧洲人工智能法律框架的核心，是建立“可信赖的生态体系”，保护基本权利（即《欧洲联盟基本权利宪章》），消费者权益等。

2021 年 4 月公布的《人工智能法案》草案[6]对禁止类 AI、高风险类 AI，以及与人类互动的 AI，生成或操纵图片、声音、视频的 AI 有所规定。有消息称，为应对 ChatGPT 所带来的问题，该提案将再次修订，变更部分定义和监管类型：增加对“大型生成式 AI 模型”部署者和用户的直接监管，包括：（1）透明度问题的监管；（2）风险管理；（3）非歧视条款适用于“大型生成式 AI 模型”开发商；（4）内容审核规则。[7]

2022 年 9 月 28 日，欧盟委员会发布的《人工智能责任指令》提案规定了 AI 引发的侵权损害责任，以使受害者在使用 AI 产品时可依法获得侵权损害救济。

二、监管到底在应对什么？——科学怪人的真面目

有关生成式 AI 的威胁已有很多讨论，包括但不限于虚假、歧视和偏见的信息，数据隐私问题、公共安全、知识产权归属等等问题。诸此威胁回溯本质，主要可归结于数据来源和算法模型。

1. 数据来源合法性

AI 模型训练中很重要的一环机器学习，分为监督学习和无监督学习。用于监督学习的训练数据经过筛选和标注，有明确的方向和目的；而无监督学习的训练数据是未经筛选或标注的原始数据，且未

[3] https://www.federalregister.gov/documents/2023/04/13/2023-07776/ai-accountability-policy-request-for-comment#footnote-80-p22439

[4] 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/

[5] https://commission.europa.eu/system/files/2020-02/commission-white-paper-artificial-intelligence-feb2020_en.pdf WHITE PAPER On Artificial Intelligence - A European approach to excellence and trust

[6] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206 Proposal for a Regulation Of The European Parliament And Of The Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts

[7] 《ChatGPT 引发欧美关于人工智能监管的辩论》https://mp.weixin.qq.com/s/SqvHM8nQbwqJ-ecQH8F7Dg

给定明确的方向和目的，而是让算法自行厘清原始数据的内在关联，从而逐步形成一定的模型方向和目的。无监督学习也并非野蛮生长，从无标注数据中发现隐含关联的功能可以运用于欺诈检测、金融分析。[8]

以汽车材料研发为例，运用生成式设计实现汽车零部件的减少及减重，有助于研发更省油的汽车。根据输入的参数，比如生产时间、材料、制造成本等，系统可生成多种方案，设计师经研究比较后做出设计优化处理。[9]系统在生成多种方案的过程中，可能需要生成式 AI 模型自己去互联网上检索学习学术文章、研究评论，根据检索结果多次推演试验，并自行检测学习和推演中的错误，查询资料自我修正，以期达到输入参数的要求。机器自行在公开互联网检索搜寻和抓取信息的方式，就有可能绕开网页的隐私政策、用户协议、知识产权声明和密码保护，导致数据来源违法或侵权。为应对这一问题，不少企业和非营利组织已自行开发解决工具，例如微软为确保数据来源合法性而开发的工具 Datasheets for Datasets 可以记录机器学习模型的训练和评估数据，监控数据的动机、组成、收集、收集、标签、使用目的和数据维护过程。

此外，从数据主体权益保护上看，算法的复杂性和黑箱特性导致开发者或服务提供者无法向数据主体充分解释数据处理的过程和目的。并且，不同于静态孤立地从某数据库、服务器中修改或删除数据主体的个人信息，从 AI 模型中修改或删除某个数据主体的个人信息，则可能影响整个模型的运行，所以数据主体的个人信息修改权、删除权也难以保障。[10]

2. AI 算法模型的透明性、可解释性和可追责性

影响算法模型的透明性、可解释性和可追责性的因素很多。

主观人为上，企业为追求自身利益，将利益企图和行为指向植入算法中，并以算法安全与保密为由，人为构设出不透明的算法黑箱，以避免因程序漏洞、方法不当、违规违法等问题遭受指控。[11]美国人工智能和数据政策中心（Center for Artificial Intelligence and Digital Policy）于今年 3 月向美国联邦贸易委员会(FTC)提交了一项针对 ChatGPT 的投诉，指出随着 AI 模型营利能力增强，开发者可能不再愿意公开 AI 模型。[12]

客观技术上，数据驱动的算法主要采用深度神经网络技术，算法核心部分是通过对数据进行自动学习而自动生成，并非人工设计，人难以知晓具体的学习过程，对算法结果“知其然而不知其所以然”。[13]

以汽车自动驾驶为例，生成式 AI 能够生成任何人类想象到的驾驶场景，甚至涵盖许多极端环境的“边缘案例”。生成式 AI 的大模型的突现能力（即量变引起质变），将有助于模型的“思维链”能力产生，从而展示类似人类的复杂推理能力。生成式 AI 的人类反馈强化学习训练方式，使自动驾驶不断自我纠错、进步。推理能力和自我进步，是否会最终演化出脱离人类意志、违背安全驾驶的决策机制？这将给生成式 AI 以及相关应用场景规则的制定和监管带来巨大的挑战。[14]无独有偶，欧盟委员会于 2021

[8] https://viso.ai/deep-learning/ml-ai-models/ The Ultimate Guide to Understanding and Using AI Models (2023)

[9] https://mp.weixin.qq.com/s/FoF18PXpNtBOGLNXuyXGaA Mixlab 技术前沿：《AI 驱动的生成式设计，如何应用于汽车智能建造？》

[10] https://iapp.org/news/a/generative-ai-privacy-and-tech-perspectives/ Generative AI: Privacy and tech perspectives Katharina Koerner, CIPP/US

[11] https://www.samr.gov.cn/wljys/ptjjyj/202112/t20211210_337980.html 《算法黑箱基本概念及成因》国家市场监督管理总局网络交易监督管理司

[12] https://www.caidp.org/cases/openai/

[13] 同 11

[14] https://www.cnii.com.cn/rmydb/202304/t20230423_465226.html 《以 ChatGPT 为代表的生成式 AI 在自动驾驶领域的应

年 4 月的另一项针对驾驶系统的提案[15]，就对驾驶系统相关的机器产品的制造者、代理、进口商、经销商的责任予以规定。

三、与科学怪人共舞——主要市场生成式 AI 的精细化管理点评

1. 中国

按照现行立法和征求意见稿，生成式 AI 的主管部门以网信部门为首，会同电信、公安、市场监管等部门。未来当生成式 AI 与终端产品相结合，例如用于文化娱乐、医疗、智能汽车等，则相关行业的主管部门也可能根据相关法律法规参与监管。

监管重点上，征求意见稿仅适用于面向中国境内公众的产品，暂不对仍处于实验室训练阶段和非公开试运行的生成式 AI 予以过多限制，鼓励创新。对于面向中国公众的生成式 AI，结合上文讨论的数据来源和算法模型衍生出的各种威胁，就不难理解征求意见稿中各项措施背后的用意，例如保证数据来源合法、真实、客观；维护个人信息主体的合法权益；披露数据描述、人工标注规则、基础算法、技术体系；生成的内容不得虚假、歧视或侵权；采取内容过滤手段、优化训练方式；进行互联网信息服务安全评估和算法备案等等。

征求意见稿的规制对象主要是利用生成式 AI 产品提供聊天和文本、图像、声音生成等服务的组织和个人，包括通过提供可编程接口等方式支持他人自行生成文本、图像、声音等。不过，即使处于实验室训练阶段和非公开试运行阶段受到生成式 AI 产品的法律限制较小，开发者仍然需要遵守网络安全、数据隐私保护相关的法律法规，而且应当为产品推向市场后的合规做充分预案。

2. 美国

美国有关 AI 系统（包括生成式 AI）发展和监管的政府机构包括白宫办公室的科学和技术办公室(OSTP)，国家人工智能倡议办公室（NAIIO），商务部下设的美国国家标准技术研究所（NIST）、国家电信和信息管理局（NTIA）等等。行业性的 AI 系统规制散见于近十年的不同政府部门文件中，例如美联储发布的算法模型风险的研讨，平等就业机会委员会（EEOC）有关 AI 歧视的应对等等。

监管重点上，同各国遇到的状况类似，政策制定者们试图在鼓励生成式 AI 发展与确保 AI 的可信赖性之间权衡摸索，难点包括如何实施 AI 可信赖机制，厘清 AI 生命周期和价值链，以及如何将以上目标标准化。今年 4 月底，美国联邦贸易委员会（FTC）、消费者金融保护局（CFPB）、司法部民权司（DOJ）和平等就业机会委员会（EEOC）四个部门联合发布声明，将重点从人工智能可能产生偏见的数据集、大模型的透明度、系统设计的前提假设这三个方向，监管其潜在的歧视风险。[16]

监管手段上，4 月 13 日国家电信和信息管理局（NTIA）发布的征求意见通知中提及了合规性评估、偏见审计、算法影响性评价、黑箱对抗审计等等。初步理解，这类技术性手段也适用于实验室训练阶段的生成式 AI。

3. 欧盟

欧盟委员会于 2018 年根据《欧洲人工智能战略》（European Strategy for Artificial Intelligence）建立了高水平人工智能专家组("AI HLG")。该专家组

用》作者：赛迪研究院刘胜语、赫荣亮

[15] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0202 Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on machinery products

[16] https://www.ftc.gov/system/files/ftc_gov/pdf/EEOC-CRT-FTC-CFPB-AI-Joint-Statement%28final%29.pdf Joint Statement on Enforcement Efforts Against Discrimination and Bias in Automated Systems

的分为两个工作组，分别负责 AI 伦理指南，和政策与投资战略[17]，致力于落实欧盟的各项 AI 战略。此外，该专家组与欧盟 AI 联盟、欧盟各成员国一起共同探讨落实 AI 监管。

虽然欧盟尚未专门对生成式 AI 形成立法，但是其立法思路值得参考。专家组于 2020 年 7 月发布的一项文件[18]提出应当分行业（公共部门、卫生、制造业和物联网）对 AI 予以不同程度的监管，这和正在修改讨论中的欧盟《人工智能法案》对生成式 AI 区分不同级别风险予以监管有异曲同工。

除了分行业监管，还有分阶段监管。欧盟委员会未来的立法设想中，责任分配的对象应当是最适合风险控制的主体，比如 AI 系统开发者最有义务在开发阶段控制风险，但 AI 系统投入使用后最有义务控制风险的主体则是部署者。终端用户和受到 AI 系统侵害的其他相关方的请求主体也需要立法时细细斟酌。类似观点在欧盟学界也有提出，负责模型预训练的开发者，负责精调的人员，AI 产品的用户，收到 AI 生成内容的用户，所应受到的监管和保护程度应当加以区分。[19]这一全流程的监管需要谨慎平衡科技创新和保护用户安全之间的界限，笔者非常期待立法公布后的外界反映。

翁亚军　合伙人　电话：86-21 2208 6264　邮箱地址：wengyj@junhe.com
蒋柠蔚　律　师　电话：86-21 2208 6186　邮箱地址：jiangnw@junhe.com

[17] https://ec.europa.eu/futurium/en/system/files/ged/concept_note_on_the_ai_hlg_0.pdf CONCEPT NOTE The High-Level Expert Group on Artificial Intelligence

[18] Sectoral Considerations on the Policy and Investment Recommendations https://futurium.ec.europa.eu/sites/default/files/2020-07/Sectoral%20Considerations%20On%20The%20Policy%20And%20Investment%20Recommendations%20For%20Trustworthy%20Artificial%20Intelligence_0.pdf

[19] https://blogs.law.ox.ac.uk/blog-post/2023/03/regulating-chatgpt-and-other-large-generative-ai-models#:~:text=The%20EU%20is%20at%20the%20forefront%20of%20efforts,cover%20AI%20%28Digital%20Services%20Act%2C%20Digital%20Markets%20Act%29. Regulating ChatGPT and Other Large Generative AI Models, Philipp Hacker, Andreas Engel, Marco Mauer