国际法视野下 ChatGPT 对数字安全的挑战及对策

公惟韬

（复旦大学，上海 200433）

摘　要：作为《全球安全倡议》的重要延伸，数字安全具有国家、社会和个人三重维度的内涵，目的是维护国家安全，弥合数字鸿沟以促进数字资源平等利用，并加强个人隐私保护。OpenAI 公司发布的 ChatGPT 是生成式人工智能的标志性进展，但在算法、算力和数据层面，ChatGPT 却分别存在威胁国家安全、数字资源利用不平等和泄露个人隐私的数字安全风险。目前，各国生成式人工智能发展水平不一，既有的国际法体系罕有对生成式人工智能的专门性规范，难以为相关技术水平落后的国家提供必要的保护。应完善数字安全国际法律制度，并从国家、社会和个人维度构建数字安全国际法体系，在算法、算力和数据层面完善对虚假信息的识别、保障数字弱势群体权益、落实“知情—同意”原则，平衡技术发展与安全，保护国家安全，弥合数字鸿沟，加强个人隐私保护，落实《全球安全倡议》，应对 ChatGPT 等生成式人工智能对数字安全的挑战。

关键词：ChatGPT；数字安全；人工智能；风险；监管自治

中图分类号：D996　　**文献标识码：**A　　**文章编号：**1006-1894（2025）02-0063-12

一、引言

2022 年 11 月，OpenAI 公司发布人工智能聊天机器人 ChatGPT，引发各界广泛关注。ChatGPT 可以根据用户问题通过自然语言进行回复，是生成式人工智能的标志性进展。但是，ChatGPT 的运行与使用却存在输出内容虚假、违法且价值不中立的算法风险、数字资源分配不平等的算力风险和泄露个人隐私的数据风险，对数字安全构成挑战。在 ChatGPT 推出后不到半年的时间，百度推出“文心一言”大模型。“文心一言”大模型的深度学习与运行模式同 ChatGPT 相似（赵广立，2023），可能存在同样的问题，DeepSeek 也是如此。

2022 年 4 月，习近平在博鳌亚洲论坛上首次提出《全球安全倡议》。自该倡议提出以来，安全概念的内涵不断丰富和发展，涵盖国家安全、社会安全和个人安全等在内的数字安全是该倡议的应有之意。2023 年 10 月，我国网信办发布《全球人工

作者简介：公惟韬，复旦大学法学院博士生，研究方向：国际法、科技法。

基金项目：国家社会科学基金重大项目“构建人类命运共同体国际法治创新研究”（项目编号：18ZDA153）；国家社会科学基金重点项目“‘人类命运共同体’国际法理论与实践研究”（项目编号：18AFX025）。

智能倡议》，呼吁提升人工智能技术的安全性。目前，相关的国际法罕有专门针对生成式人工智能对数字安全损害的直接规制，多为经济层面的间接规制，且均为软法，执行力低；加之各国生成式人工智能发展水平不一，部分技术落后的国家可能没有能力发现并防止相关技术造成危害，面临数字安全保护困境。对此，应完善既有国际法规则，从监管与行业自治两个角度研究、制定规则，确保生成式人工智能的发展和运用在伦理、规则和法律的约束下进行，并以《全球安全倡议》《全球人工智能倡议》为引领，构建数字安全国际法体系，应对生成式人工智能的挑战。

二、数字安全的内涵

《全球安全倡议》和《全球人工智能倡议》提出坚持共同、综合、合作和可持续的安全观，尊重各国主权、领土完整，坚守联合国宪章宗旨和原则，重视各国合理安全关切，通过对话协商以和平方式解决国家间的分歧和争端，统筹维护传统领域和非传统领域安全。其中，深化信息安全领域国际合作、加强人工智能等新兴科技领域国际安全治理、预防和管控潜在安全风险是重点方向。在数字化快速发展的背景下，包括人工智能安全在内的一系列超出传统网络安全范畴的安全新挑战接踵而至，倒逼网络安全升级为数字安全，将影响国家安全、社会安全乃至个人安全的遍布各类数字化场景的安全风险纳入其中。

（一）数字安全内涵的国家维度：维护国家安全

进入工业化时代以来，国家间竞争总体上经历了从贸易竞争到产业竞争再到技术竞争的发展路径。目前，数字全球化加速发展致使全球数字扩张的无序性加剧，当社会围绕技术来组织时，技术力量就是社会中权力的主要形式（芬伯格，2005）。在人工智能领域，人工智能技术犹如一把“双刃剑”，既打开了生活的无限可能空间，也映射出数字时代国家竞争节奏和大国竞争的复杂性。以 ChatGPT 为例，ChatGPT 诞生后就受到了美国政府的支持。

人工智能技术承载的价值包含政治价值，技术研发者为人工智能技术选择的数据、编写的代码都是技术研发者政治观念与价值取向的无形反映，这种无形性也导致相关的价值取向更趋隐蔽化、全域化、复杂化和动态化。此外，人工智能生成的信息存在被伪造的可能，这种技术在合成虚假视频方面的应用极易煽动网民的非理性情绪，甚至动摇公众对政府公信力的信任（余杰，2022）。因此，人工智能技术不仅凸显相关国家的数字科技地位，而且可以达到为其数字霸权优势地位服务的目的。无论是由此导致的价值取向冲突，还是违法犯罪活动，都将损害他国的国家安全。因此，捍卫国家安全是数字安全国家维度的核心内涵。

（二）数字安全内涵的社会维度：弥合数字鸿沟

数字鸿沟是指工业化社会和发展中社会之间互联网接入的差异，它关系到每个

国家信息贫富差距，意味着那些能够参与和不能参与数字生活、能够使用和不能使用全套数字资源的人之间的差异（Spectar，2000）。在加速推进信息数字化的背景下，每个人每天都会生成海量的数字信息，包括生产、生活的运行轨迹和交往范式，以及身份、行为、关系和言语数据，它们塑造了人的数字属性和社会的数字生态，由此构成的数字资源成为数字经济时代的社会核心资源。数字资源作为数字化时代国家及其居民赖以生存的重要资源，属于自然资源和社会经济资源组成的物质实体外延，对此类数字资源的合理分配以及由于不能合理分配而形成的数字鸿沟问题是传统平等理论在数字空间的延伸，关乎社会安全与稳定。

然而，由发达国家一手建立的既有国际秩序客观上限制了发展中国家的技术发展，因为先进技术往往被发达国家垄断，导致发展中国家生产技术落后。国际电信联盟（ITU）在《2024 年事实和数字》报告中指出，低收入国家目前只有 27% 的人能够使用互联网，而高收入国家有 93% 的人能够使用互联网。此外，在其他方面，如利用人工智能技术的能力，发达国家和发展中国家的差距更大（贾泽驰，2019）。这种“鸿沟”如果不能弥合，将可能导致数字殖民主义，也就是发达国家对发展中国家的数字掠夺和控制，并通过其大型数字技术公司以游说、基础设施投资以及向发展中国家捐赠硬件和软件等方式，迫使发展中国家形成有利于发达国家的国家政策导向，并借此对发展中国家人民的日常生活进行捆绑和垄断（王淑敏，2021）。因此，弥合数字鸿沟，应对发达国家的技术垄断，保护发展中国家平等的数字技术发展权利是数字安全社会维度的核心内涵。

（三）数字安全内涵的个人维度：加强隐私保护

《全球数据安全倡议》要求发展人工智能应坚持以人为本理念，以增进人类共同福祉为目标，并以保障社会安全、尊重人类权益为前提。这是因为以人为本是数字安全的出发点和落脚点，也是数字安全相关规则在发展中应解决的核心问题。所谓以人为本，并非一味强调安全而忽视隐私保护与数据开发利用的平衡，而是既遵循数字经济共享共用共建的基本思想，又充分考虑和尊重个体在大数据时代的隐私保护，因为只有个体数据安全有了保障，数据产业的发展才有根基。

就人工智能而言，个人信息数据存在被人工智能获取而产生个人隐私泄露的风险。一方面，人工智能处理大数据的强能力可能使已被匿名化的数据与其他信息联系从而达到识别信息主体和去匿名化的效果，加大侵犯隐私的风险；另一方面，在隐私保护体系中，数据主体的“知情—同意”是一项至关重要的权利，但是人工智能无处不在的监测和收集能力可能导致相关数据的收集未经过数据主体的同意，破坏隐私保护，以至于个人数据安全问题已成为社会的普遍焦虑（华劼，2023）。因此，通过保护个人数据来保障个人人身、财产安全具有必要性，这样才能维护公共秩序与社会福祉，提升人们对自身个人数据的掌控感和对数据治理的参与度。可见，隐私保护是数字安全个人维度的核心内涵。

三、ChatGPT 的本质及对数字安全的挑战

（一）ChatGPT 的本质

一般认为，ChatGPT 属于生成式人工智能。2023 年 8 月 15 日，我国开始施行的《生成式人工智能服务管理办法》将生成式人工智能定义为基于算法、模型、规则生成文本、图片、声音、视频、代码等内容的技术。此外，构成人工智能的三大要素是算法、算力和数据，ChatGPT 作为生成式人工智能也由这三大要素构成。

在属性上，ChatGPT 应属于聊天机器人，但其运行逻辑又不同于传统聊天机器人，后者的回答多已预先写入数据库，用户提出问题后，传统聊天机器人从预先写入数据库的回答中调取相应答案，且传统聊天机器人并不具备上下文学习能力，一旦用户问题超出预设范围或者提出与前述问题存在关联的问题，就可能出现答非所问的情形。ChatGPT 则是基于 3,000 亿个单词的语料，通过 Transformer 算法形成高达 1,750 亿参数的 GPT-3 大语言模型与用户对话；ChatGPT 的语料源自约 45TB 的互联网公开数据和数字书籍，这些海量知识成为 ChatGPT 的自然语言模型和智能化数据基底，使其拥有回答各类问题的基础能力（朱光辉和王喜文，2023）。

此外，ChatGPT 还包含“亿万人工标注”的数据，对其注入人类偏好知识，使其了解提问者的意图，并能评判什么是好的回答、什么是不好的回答，如比较详细的回答是好的，带有歧视性内容的回答是不好的（陆小华，2023）。深度神经网络的学习能力和模型的参数规模呈正相关，ChatGPT 的高参数规模语言模型使其能够精准描述涉及的广泛话题，又能以拟人化的姿态应对用户的连续提问，并通过上下文学习及时补充回答。同时，ChatGPT 还会在与用户对话的过程中进行强化学习。大量用户实践经验表明，ChatGPT 在对用户意图的理解上和结果的准确性、完成度和易用性上都达到了前所未有的高度（陆小华，2023）。

（二）ChatGPT 对数字安全的挑战

如前所述，数字安全是数字时代与数字化相关的一切安全要素、行为和状态的集合，具有国家维度的国家安全保护、社会维度的数字鸿沟弥合和个人维度的隐私保护等内涵，ChatGPT 则分别在算法、算力和数据 3 个方面对数字安全的 3 个维度内涵形成挑战。

1. 算法层面：ChatGPT 对国家安全的挑战

2023 年 3 月 14 日推出的搭载 GPT-4 大语言模型的 ChatGPT，在 GPT-3 大语言模型的 1,750 亿参数规模上再次提升，而人类大脑皮层的突触总数约 100 万亿左右，一旦 GPT-4 实现 100 万亿参数规模，从仿生学的角度来说，意味着它将可能达到与人类大脑神经触点规模的同等水平。也就是说 GPT-4 大语言模型的思想和创作能力几乎堪比人类大脑（朱光辉和王喜文，2023）。但是，目前 ChatGPT 缺乏全面的伦理道德过滤机制，可能输出违法信息，完全依赖其进行决策将产生国家安全风险。

尽管在开发 ChatGPT 时，OpenAI 公司已经极力避免输出带有算法歧视、偏见或侮辱、攻击、色情、暴力及血腥导向的内容，但实践证明，如果使用者有意通过设置命令和禁止要求，仍可能诱导 ChatGPT 输出违法信息，致使 ChatGPT“越狱”，最终突破研发者对其设置的道德伦理及法律底线（樊雪寒，2023）。甚至在网络犯罪方面，ChatGPT 可以直接根据用户的要求编写代码，让“网络小白”变成“资深黑客”，助长犯罪，危害国家安全。

此外，ChatGPT 人工标注数据的目的不是监督其内容的真实性，甚至 ChatGPT 在回答时会提供编造的虚假信息（邓宏光和王雪璠，2024）。若直接以 ChatGPT 的回答作为判断依据而不进行核查，很有可能会导出错误、不全面的结论。更有甚者，部分西方国家存在着利用高效隐蔽和灵活多变的信息传播手段对我国进行意识形态渗透和价值分化，规训受众价值，使其对主流价值产生厌恶和偏见，进而在潜移默化中削弱社会主义核心价值观的凝聚力和引领力。这在人工智能领域也是一样，如果 ChatGPT 被其背后政府裹挟，给出虚假的、煽动性的回答，也可能出现公众舆论被操纵的后果。

在数字时代，数字空间对用户自由选择权产生了极高的控制力和重塑力，即使在 2015 年，平台通过用户画像对用户的平均可预测程度已高达 93%（巴拉巴西，2017）。通过预测，平台便会基于此对用户进行相关商品的推荐，达到盈利的目的，而用户自认为进行的自由购物行为其实受到平台大数据计算的预判和引导，甚至还会由于这类推荐导致价值认同方面的信息同质化和极端化。ChatGPT 提供的回答同样存在上述问题，因为技术研发者为人工智能技术选择的数据、编写的代码都是技术研发者政治观念与价值取向的反映，用户与 ChatGPT 的每一次互动都会使人工智能习得人类的观念和行为。与此同时，人工智能也在影响人类的观念与行为，周而复始，一旦存在伦理风险，该风险就会在不断的交互作用中呈现演进发展的动态化特征，技术研发者观念的渗透也在此过程中更加难以为人们觉察（余杰，2022）。OpenAI 公司虽然声称 ChatGPT 价值中立，但是目前 ChatGPT 尚未对我国开放，而且其训练的数据绝大部分是英文，中文数据占比较少（王建磊和曹卉萌，2023）。ChatGPT 可能的确无法表达自己的立场和价值观，但训练有素的大语言模型一定可以表达其背后能够控制数据来源及知识立场的那些人的立场与价值观。由于人工智能技术具有自我强化的倾向，一旦存在偏见的数据被用于机器学习，看似客观中立的算法运行结果也将带有偏见，人工智能的回答也将继承人类社会的原始偏见并不断循环扩散，使偏见被进一步复制与强化，由此形成的价值冲突与恶意引导将对国家安全产生消极影响。

2. 算力层面：ChatGPT 导致的数字鸿沟挑战

根据中国互联网络信息中心发布的最新报告，我国的网民规模和互联网普及率呈现城乡二元格局。其中，老年人群、贫困人群和文化层次较低人群分享数字红利

的能力明显低于其他群体（郑智航，2023）。保护这些数字弱势群体的权利是数字安全的应有之义，也是传统平等价值在数字空间的新变体。ChatGPT 等高新数字技术导致的数字鸿沟现象使得发展中国家尤其是“低联网国家”的人民受制于信息本身的稀缺性和个人在获取、掌握及运用信息技术方面的不同，导致他们在使用信息传播技术和网络技术时存在客观差异，以至于无法平等分享数字红利（杨洸和杜丽洁，2022）。

就教育领域而言，ChatGPT 虽然可以拓宽学生的知识面和加深对各类问题的理解，但是其双刃剑效应也逐渐显现。由于 ChatGPT 能自主整合并输出不同于原始资料的内容，使得它完全有能力代替学生完成论文写作（邓建鹏和朱怿成，2023）。这将极大地影响教育生态，因为测验、论文评估的将不是学生的能力和知识，而是 ChatGPT 用户的提问水平和 ChatGPT 的应答能力；同时这也不利于平等竞争环境的构建，能够使用 ChatGPT 的学生可能相对于不能使用的学生具有了特殊优势，造成数字资源利用不平等的现象。在教育以外的其他领域，如在企业界，如果企业使用 2023 年 3 月 22 日微软发布的内嵌 ChatGPT 大语言模型的 Copilot 办公系统，就能够自动总结、改写、翻译和润色 Word 文字作品，自动生成 PPT 作品，并能够利用 Copilot 办公系统引导 Excel 文件的公式运用；Copilot 系统还可以对 Outlook 的邮件内容和专业性进行调整，让企业在对外交流尤其是在与外籍客户交流时，避免不必要的语言错误，改善企业形象。然而，对于受条件所限而不能使用相关办公系统的企业以及其他数字弱势群体来说，他们将无法分享 Copilot 办公系统的数字红利，无法通过生成式人工智能提高效率，面临数字资源分配不平等的困境。

3. 数据层面：ChatGPT 对隐私保护的挑战

ChatGPT 的用户反馈强化学习模式原理可以通过索尔克大学教授 Sejnowski 的实践予以解释。在 Sejnowski 询问 ChatGPT 步行穿过英吉利海峡的世界纪录时，ChatGPT 的回答是 18 小时 33 分钟，但是，事实上一个人是无法步行跨过英吉利海峡的，ChatGPT 如此回答的原因是询问者使用了动词“步行”而不是“游泳”；但如果用户针对 ChatGPT 的回答进行纠正，并告诉 ChatGPT 步行穿过英吉利海峡是不可能的，那么 ChatGPT 就会学习并在以后再遇到类似问题时避免相应的错误（Sejnowski，2023）。这暴露一个潜在的问题，那就是用户的反馈、纠正将被 ChatGPT 学习并载入数据库。无论用户反馈与纠正的信息内容是否涉及个人隐私，ChatGPT 都会“记住”并用于后期的模型训练和内容生成。

在世界范围内，“知情—同意”原则一直是传统个人信息数据处理的主要合法性基础。在数字社会中，大数据平台会通过算法了解用户的偏好兴趣和喜怒哀乐，并从海量用户行为数据中挖掘、分析、处理和生成体现用户全貌及作为生活方式数字化表达的用户画像，并赋予抽取的用户画像以数据商业价值。针对商家抽取的用户画像涉及的信息安全风险，我国《个人信息保护法》赋予个体知情权、决定权等个人信息

权利。在欧美，以欧盟《一般数据保护条例》和美国《加州消费者隐私法案》为代表的一系列涉及数字空间个人信息保护的立法也都作出了类似规定（徐娟，2023）。在数字经济背景下，知情权的范围还包括平台有义务告知用户与其数据将被处理的一切相关资讯，并只有在获得用户同意后才能处理相关数据。虽然 ChatGPT 和大数据平台的运行模式不尽相同，但是如前所述，采用用户反馈学习模式的生成式人工智能客观上需要采集用户提供的信息，因此应落实“知情—同意”原则作为规范生成式人工智能采集用户信息的基础，防止被作为学习及未来回答内容源的用户对话数据及其中的个人信息泄露。

然而，OpenAI 公司一方面声称 ChatGPT 不会记住也不会主动提供用户的任何信息，另一方面又表示 ChatGPT 与用户对话的数据需要存储在 OpenAI 公司及其使用的云服务提供商的数据中心（邓辉，2023）。这是因为，ChatGPT 运用的生成式人工智能技术本身就要求其收集、储存和使用对话数据，这样才能实现用户反馈强化学习，从而通过和用户对话的方式促进 ChatGPT 迭代。虽然 OpenAI 公司具有对 ChatGPT 的内容源进行人工标注的技术能力，但是 ChatGPT 包含的亿万人工标注数据的标注目的只是为了对内容是否符合人类的习惯和偏好进行评判，并非对相关信息进行实质性筛选和人工监督过滤，导致 ChatGPT 在自主学习训练与输出的过程中几乎不受监督，存在将含有个人信息的问答内容作为模型训练的基础语料并作为回答内容输出的可能，极大地增加了隐私泄露的个人安全风险。

四、ChatGPT 等生成式人工智能挑战下数字安全国际保护的困境及出路

（一）困境：缺乏对 ChatGPT 等生成式人工智能数字安全的国际规制

目前，对于 ChatGPT 及类似生成式人工智能，虽然有相关的区域性实践和双边条约，如欧盟 2024 年 8 月 1 日正式生效的人工智能法案，但在最广泛的各国政府合作层面相关的国际法规制较少，2021 年 11 月 24 日在联合国教科文组织第 41 届大会上 193 个会员一致通过了《人工智能伦理问题建议书》，是第一个涉及人工智能伦理问题的全球性协议。该协议书之所以提出，是考虑到人工智能技术虽然对人类大有助益并惠及所有国家，但也会引发根本性的伦理关切。例如，人工智能技术可能内嵌价值观并加剧偏见，导致歧视、不平等，造成数字鸿沟。由此，人工智能技术会加深国家内和国家间现有的鸿沟和不平等。为维护正义、信任和公平，以便在公平获取人工智能技术、享受这些技术带来的惠益和避免受其负面影响方面不让任何国家和任何人掉队。直至 2024 年 3 月 21 日，第一个关于人工智能的全球决议《抓住安全、可靠和值得信赖的人工智能系统带来的机遇，促进可持续发展》才在联合国通过，呼吁构建安全、可靠和可信赖的人工智能系统。

当前国际法的人工智能规制主要依托国际软法，从执行力角度来说取决于国家自身是否有意愿实施。但是，由于各国生成式人工智能技术水平不同，可能面临部分国家想要规制人工智能而没有能力的情形。此外，目前对 ChatGPT 及类似生成式人工智能技术的国际规制更多是间接性的，如欧盟相关法律中的人工智能规则更多是促进欧盟作为市场平台对投资者的吸引力，缺乏对涉及的数字空间中出现的各种风险的规范，难以应对数字安全面临的挑战。

（二）出路：构建人工智能数字安全国际法体系

1. 人工智能数字安全国际法体系的构建路径

针对数字安全来自 ChatGPT 及类似生成式人工智能的挑战，由于各国人工智能相关技术的发展水平不同，部分技术水平欠发达国家相关领域的规范可能是空白，而各国国内法层面关于人工智能数字安全的规范也不尽相同，由此可能会导致部分行为在一国合法而在其他国家会被规制的困境。这不利于技术的长期稳定发展，也无法形成便捷的物联网。此外，部分发达国家为了维护自身在技术领域的利益与优势地位，还有可能会滥用长臂管辖原则，或以意识形态、国家安全等为缘由，隔离和挤压新兴技术体和新兴技术国家（李盛竹，2011）。因此，构建统一的人工智能数字安全国际法体系具有必要性。

目前，针对生成式人工智能这一新兴事物，尚无专门的普遍性国际条约，虽然如前所述存在双边和区域性条约，但数量不多，涉及的国家数量也有限，各国对人工智能相关技术的态度也不同，如欧盟主要关注个人信息保护与数据安全，美国则力主尽可能让数据自由流动，以充分发挥美国的既有优势（Drahos，2005）。因此，直接在国际层面形成统一的国际规范存在一定的难度。对此，可首先效仿防范网络安全犯罪的《塔林手册》，由学者就涉及人工智能所有可能存在的法律问题进行梳理，在数字主权、数字资源合理分配和个人信息保护方面形成具有指导性而非约束性的示范性规则，保护平等与和谐的数字安全观，规范“知情—同意”原则，保障数字弱势群体并达成防范虚假和不良信息的国际共识，作为建构人工智能数字安全国际法体系的基础。然后，各国可以此示范性规则为指引，规范双边和区域条约关于人工智能数字安全的相关约定与实践，进而通过这些国际实践逐渐将示范性规则的指导意见转化为习惯国际法，并在此基础上逐步形成统一的、具有执行力的人工智能数字安全国际法规范。

2. 宏观层面的构建内容：从国家、社会和个人维度规范生成式人工智能

（1）国家维度：构建数字主权保护国家安全

针对人工智能可能导致对国家安全的挑战，需要加强国家应对相关挑战的自主性与规范性，构建数字主权。数字主权就是将数据作为主权的一部分，其概念最早来自欧盟。在 2020 年 7 月的《欧洲数字主权》研究报告中，数字主权被定义为在数字世

界中独立行动的能力，并可从保护机制和积极发展数字创新工具的自主性和规范性来理解，一是欧盟通过数字基础设施建设提升数字技术竞争力，自主掌控战略性技术和价值链的能力；二是欧盟对数字空间监管以及对数字标准进行规范的权力（闫广和忻华，2023）。由此可见，通过在数字世界中建立管辖权，确定数字边界重新控制数据和数字基础设施，行使数字主权，才能以更加积极、更加自主的姿态参与数字地缘博弈，应对在非对称技术竞争中的价值冲突，保护国家安全（刘玲霞，2024）。

（2）社会维度：促进数字技术发展的权利

在数字主权的前提下遵循多元共治、合作普惠原则，处于数字技术和市场优势地位的国家应当承担引领责任，通过合理的制度安排，为在互联网的传输与使用中处于劣势、与别国存在信息差距和使用数字资源不平衡的国家提供必要的数字技术援助。应发挥区域贸易协定的作用，在逆全球化趋势下，区域贸易协定将有助于发展中国家确保更大的市场和竞争力。此外，还可参照《数字经济伙伴关系协定》（DEPA）的数字包容性条款，引导缔约各方合作共赢、共同发展，消除技术和基础设施障碍，平等获取发展数字技术的权利，应对发达国家与发展中国家之间的技术鸿沟。

（3）个人维度：设置个人信息保护总则

数字贸易的活力源于信息技术的发展，尤其是跨境流动的信息，这些信息大部分是个人信息，应平衡隐私保护与贸易，避免关涉个人信息安全的安全例外成为变相限制国际贸易的条款（Ciuriak and Ptashkina，2020）。此外，考虑到目前各国相关数字技术发展不一，个人信息保护的技术不健全，应在数字安全国际法体系中设置总则性的个人信息保护规则，并对不同种类的个人信息进行分类，增加提升个人对数据掌控的自主性规则并完善和强化通知与救济制度，开展多维联动的全球造法活动，加强隐私保护（刘笋和佘佳亮，2023）。

3. 微观层面的构建内容：从算法、算力和数据层面规范生成式人工智能

（1）算法层面：完善虚假和违法信息的识别，防范国家安全风险

针对 ChatGPT 这类生成式人工智能算法存在提供虚假数据的可能，应引导企业将防止虚假、错误的信息加入语言训练模型，并利用人工标注技术，引导 ChatGPT 回答问题的偏好标准，避免在收到关于客观事实的提问时给出虚构的答案。此外，针对 ChatGPT 这类生成式人工智能可能会受到用户诱导而生成违反道德、法律的回答，应加强各国的政府监管，必要时可对人工智能生成结果进行审核，完善对违法、不良信息的识别和阻断机制，抵御潜在的道德风险与危害，维护良好的网络生态环境。

就 ChatGPT 这类生成式人工智能对人的自由选择和价值判断的影响，一方面，实质上生成式人工智能技术是一种对人类已有知识、发现、认知等的“再现”，但这种“再现”更多是在依据算法模型水平进行信息提取基础上的“再现”；目前这种“再现”水平还不够高，不具有“专才”水平，担心 ChatGPT 这类人工智能会取代人类是不必要的。另一方面，用户的思维能力仍存在受到 ChatGPT 背后所含价值

观影响的可能，以至于扰乱数字社会运作的秩序伦理。由于科技的超越式进步，法律规则的时滞性弊端将更为凸显。考虑到数字社会中服务民生的义务主体已不再局限于以政府为代表的公权力机关，因此还可通过国际行业自治的力量，在政府监管之外通过更灵活的行业自治条例等软规范，应对新技术应用带来的风险。如在 20 世纪 90 年代互联网开始大规模商用、国家法律监管介入之前，互联网早期的工程师、计算机专家通过运用网络技术规则，建立了一套依据伦理道德原则和技术规则的技术编码网络控制方式，形成了有效的网络空间秩序（李建新，2022）。可见，非政府社会组织、商业机构等网络空间利益相关方其实与政府对网络空间秩序的形成几乎具有同等作用。因此，可以在国家法律介入监管之前，通过行业自治，尽可能多地引入不同的价值观念，并建立内含自治伦理的普遍技术编码体系，从算法层面预防恶意引导致使的价值冲突对国家安全的危害。

（2）算力层面：保障数字弱势群体，防范数字鸿沟风险

ChatGPT 及其他类似的生成式人工智能能够通过其强大的数据运算能力，提高人类工作和生活效率。但是，由于技术等各类原因，老年人、贫困地区和文化层次较低的人群可能难以在分享数字红利时获得信息公平权益的保障。数字弱势群体依法平等享有使用公共信息资源及其服务的权利，并有资格享有数字空间的公共信息及其服务。ChatGPT 及其他类似的生成式人工智能技术不应被少数“绝对强者”垄断，从而产生阶级固化和技术壁垒。

针对 ChatGPT 及其他类似的生成式人工智能可能造成的平等使用权“数字鸿沟”，有必要在技术之外辅以制度性的法律回应。具体来说，可借鉴法国、英国、芬兰等国家已正式纳入规范性法律文本的上网权，作为平等使用生成式人工智能技术的具体权利制度的理论基础。此外，以教育领域为例，如担心学生利用 ChatGPT 及类似人工智能作弊完成课程论文，对于学校来说，回归传统线下的监考制度将是一项积极的应对措施，同时应进一步加强学术审核和学术不端行为的问责机制及规章的惩罚执行力度。此外，还应通过第三方的技术手段建立有效的、具有针对性的检测体系，采用定期审核、评估、验证算法机制，分辨文本是由人类撰写还是由 ChatGPT 及其他类似人工智能生成，通过此类方式防止学生利用生成式人工智能造成教育领域的不平等现象。

（3）数据层面：落实“知情—同意”原则，防范隐私泄露风险

无论是 ChatGPT，还是其他生成式人工智能，在其研发和运营过程中，应严格落实企业的数据合规主体责任，实现个人信息利用和保护的平衡。首先，将个人信息保护作为企业数据合规体系的重要内容。其次，对于像 ChatGPT 这样采取用户学习反馈模式的生成式人工智能，在基础算法上无法避免对用户问答内容的采集，应首先征得用户的同意，遵守“知情—同意”规则；对于经用户同意而采集后的用户问答内容，要求企业强化个人信息甄别能力，采取隐私计算等技术过滤受法律保护的个人信息，

可采取“机器 + 人工”双重安全审核机制，强化输出内容管理（位涛涛，2024）。再次，对互联网已经公开的个人信息，也应当充分运用技术措施和管理制度予以核实。最后，在收到关于他人个人信息的提问时，应谨慎回答，从而确保个人信息处理的合法性。

五、结语

在大数据背景下，数字安全是国家、社会和个人健康发展的基础，数字技术也正在成为信息时代的核心战略资源，其背后存在的数字安全风险亟需解决。对于ChatGPT 及其他类似的生成式人工智能可能存在的危害数字安全的风险，如提供的信息不真实、不合法且存在歧视偏见、数字资源分配不平等和非法处理个人信息等，考虑到各国技术发展水平不同，应针对目前既有数字安全国际规则体系的不足，引导各国政府和企业保护个人信息，公平分配数字资源，保护数字弱势群体使用生成式人工智能的权利，平衡好技术发展与安全的法律和伦理天平，进而在此基础上构建数字安全国际法体系，保护国家安全、弥合数字鸿沟并加强隐私保护。

参考文献：

[1] [美] 艾伯特—拉斯洛·巴拉巴西 . 爆发 : 大数据时代预见未来的新思维 [M].2017 年版 . 马慧译 , 北京 : 北京联合出版公司 ,2017.

[2] [美] 安德鲁·芬伯格 . 技术批判理论 [M]. 韩连庆 , 曹观法译 , 北京 : 北京大学出版社 ,2005.

[3] 邓辉 . 妥善应对 ChatGPT 带来的个人信息保护挑战 [N]. 民主与法制时报 ,2023–02–22,(3).

[4] 邓建鹏 , 朱怿成 .ChatGPT 模型的法律风险及应对之策 [J]. 新疆师范大学学报 (哲学社会科学版),2023,44(5).

[5] 邓宏光 , 王雪璠 . 生成式人工智能虚假信息治理的风险与应对 [J], 理论月刊 ,2024,(9).

[6] 樊雪寒 .ChatGPT 的数据安全问题引发关注 [N]. 第一财经日报 ,2023–02–27,(A04).

[7] 华劼 . 人工智能时代的隐私保护——兼论欧盟《通用数据保护条例》条款及相关新规 [J]. 兰州学刊 , 2023,(6).

[8] 贾泽驰 . 联合国发布首份数字经济报告 美中主导全球数字经济 [N]. 文汇报 ,2019–09–06,(6).

[9] 李建新 . 数字社会人权保护的伦理进路 [J]. 河北法学 ,2022,40(12).

[10] 李盛竹 . 跨国公司国际竞争背景中的技术霸权现象——理论回顾与展望 [J]. 社会科学家 ,2011,(9).

[11] 刘玲霞 . 数字主权安全的理论内涵、现实挑战与应对路径 [J]. 陕西师范大学学报 (哲学社会科学版), 2024,53(1).

[12] 刘笋 , 佘佳亮 . 美欧个人信息保护的国际造法竞争 : 现状、冲突与启示 [J]. 河北法学 ,2023,41(1).

[13] 陆小华 .ChatGPT 等智能内容生成与新闻出版业面临的智能变革 [J]. 中国出版 ,2023,(5).

[14] 陆小华 . 智能内容生成在催生什么传播新变局 [J]. 青年记者 ,2023,(3).

[15] 王建磊 , 曹卉萌 .ChatGPT 的传播特质、逻辑、范式 [J]. 深圳大学学报 (人文社会科学版),2023,40(2).

[16] 王淑敏 . 全球数字鸿沟弥合 : 国际法何去何从 [J]. 政法论丛 , 2021,(6).

[17] 位涛涛 . 生成式人工智能的数据风险与规制进路 [J]. 科技管理研究 ,2024,44(22).

[18] 徐娟 . 用户画像中的人权保护 [J]. 河北法学 ,2023,41(2).

[19] 闫广 , 忻华 . 中美欧竞争背景下的欧盟“数字主权”战略研究 [J]. 国际关系研究 ,2023,(3).

[20] 杨洸 , 杜丽洁 . 数字技术与数字鸿沟 : 弥合数字不平等的困境与行动 [J]. 青年记者 ,2022,(22).

[21] 余杰 . 人工智能时代的意识形态风险及其化解 [J]. 思想理论教育 ,2022,(12).

[22] 赵广立 . 百度首席技术官王海峰揭秘 : 文心一言是如何炼成的 ?[N]. 中国科学报 ,2023–03–23,(3).

[23] 郑智航 . 数字人权的理论证成与自主性内涵 [J]. 华东政法大学学报 ,2023,26(1).

[24] 朱光辉 , 王喜文 .ChatGPT 的运行模式、关键技术及未来图景 [J]. 新疆师范大学学报 (哲学社会科学版), 2023,44(4).

[25] Ciuriak, D.,Ptashkina, M.Towards a Robust Architecture for the Regulation of Data and Digital Trade[M]. Waterloo:Centre for International Governance Innovation,2019.

[26] Drahos, P.Developing Countries and International Intellectual Property Standard Setting[J].The Journal of World Intellectual Property,2005,5(5).

[27] Sejnowski, T. J.Large Language Models and the Reverse Turing Test[J].Neural Computation,2023,35(3).

[28] Spectar, J. M.Bridging the Global Digital Divide: Frameworks for Access and the World Wireless Web[J]. North Carolina Journal of International Law and Commercial Regulation,2000,26(1).

The Countermeasures and the Challenges of ChatGPT to Digital Security from the Perspective of International Law

GONG Weitao

Abstract: As an important extension of the Global Security Initiative, digital security has three dimensions including national, social, and individual. The purpose of digital security is to maintain national security, bridge the digital divide for equal use of digital resources, and strengthen the protection of personal privacy. ChatGPT released by OpenAI is a landmark for generative artificial intelligence. However, from the aspects of algorithm, computing power, and data, ChatGPT faces digital security risks that threaten national security, equal utilization of digital resources and personal privacy. At present, the development of generative artificial intelligence technology in various countries is different, and there are few special norms for generative artificial intelligence in the current international law system, which is difficult to provide necessary protection for countries which lack concerning technology. It is necessary to improve the relevant international law system of digital security and build an international law system of digital security from the national, social and personal dimensions. At the algorithm, computing power and data levels, measures should be taken including improving the identification of false information, safeguarding the rights of digital disadvantaged groups, and implementing the "informed consent" principle. This is to balance technological development with security, protect national security, bridge the digital divide, and enhance personal privacy protection, so as to implement the Global Security Initiative and addressing the security challenges posed by ChatGPT and other generative artificial intelligence technologies.

Keywords: ChatGPT; digital security; artificial intelligence; risk; supervision autonomy

（责任编辑：金孝柏）