Advances in Social Sciences 社会科学前沿, 2024, 13(4), 580-586
Published Online April 2024 in Hans. https://www.hanspub.org/journal/ass
https://doi.org/10.12677/ass.2024.134338

Hans汉斯

ChatGPT冲击下的敏感个人信息安全风险与规制路径

管　彤

东北财经大学萨里国际学院，辽宁　大连

收稿日期：2024年3月5日；录用日期：2024年4月20日；发布日期：2024年4月29日

摘　要

以ChatGPT为代表的生成式人工智能作为科技史上的一次重大变革，在显著提升人们工作效率的同时，也隐藏着不可忽视的风险。ChatGPT的运作需要大量数据作为支撑，其中就涉及敏感个人信息的合法处理问题，引发了信息获取合法合规性、算法黑箱、泄漏与非法利用、监管责任竞合等相关风险，给敏感个人信息保护带来新的挑战。故而有必要重构敏感个人信息风险规制路径，在强调知情同意动态性、赋予信息主体算法解释权的同时，也要构建起有效的泄漏防护机制与监督责任机制，以促进生成式人工智能个人信息法律治理体系的完善。

关键词

ChatGPT，生成式人工智能，敏感个人信息，风险规制

Sensitive Personal Information Security Risks and Regulatory Paths under the Impact of ChatGPT

Tong Guan

Surrey International Institute, Dongbei University of Finance and Economics, Dalian Liaoning

Received: Mar. 5th, 2024; accepted: Apr. 20th, 2024; published: Apr. 29th, 2024

Abstract

As a major change in the history of science and technology, generative artificial intelligence represented by ChatGPT not only significantly improves people's work efficiency, but also hides risks

文章引用: 管彤. ChatGPT 冲击下的敏感个人信息安全风险与规制路径[J]. 社会科学前沿, 2024, 13(4): 580-586.
DOI: 10.12677/ass.2024.134338

that cannot be ignored. The operation of ChatGPT requires a large amount of data as support. Among them, the legal processing of sensitive personal information is involved, which raises related risks such as legal compliance of information acquisition, algorithm black box, leakage and illegal use, and competition of regulatory responsibilities, bringing new challenges to the protection of sensitive personal information. Therefore, it is necessary to reconstruct the risk regulation path of sensitive personal information. While emphasizing the dynamics of informed consent and giving the information subject the right to interpret the algorithm, it is also necessary to construct an effective leakage protection mechanism and supervision responsibility mechanism to promote the improvement of the legal governance system of personal information of generative artificial intelligence.

Keywords

ChatGPT, Generative Artificial Intelligence, Sensitive Personal Information, Risk Regulation

1. 引言

2022 年 11 月，美国 OpenAI 科技公司发布人工智能聊天机器人 ChatGPT。ChatGPT 上线第一周用户数量即突破百万，两个月内吸引全球活跃用户数量破亿，引发广泛讨论热潮。与分析式人工智能不同的是，ChatGPT 是一款基于大型语言和强化学习微调训练模型的生成式人工智能(AIGC)机器人。前者主要从大量数据中寻找隐藏模式并形成一定预测，而后者则通过学习海量的数据进而生成新内容[1]。生成式人工智能的运行以数据为基础，大量数据在模型训练、内容形成等阶段均起到关键作用。然而，生成式人工智能对数据的大规模需求可能导致其对数据的过度挖掘，与此同时，不透明的处理规则也极易形成“算法黑箱”，给个人信息保护造成了很大困难。特别地，敏感个人信息因其风险内容的广泛性，如果被泄漏或非法使用，可能会导致信息主体的人格尊严及其人身、财产权利面临严重的侵犯风险[2]。因此，如何识别和规制 ChatGPT 在其运行过程中可能引发的敏感个人信息安全法律风险，成为大数据时代敏感个人信息保护的关键所在。

2. 敏感个人信息界定及法律保护现状

2.1. 敏感个人信息的界定

通过《中华人民共和国个人信息保护法》(以下简称《个人信息保护法》)第 28 条第 1 款，我国采取“概括 + 列举”的模式，首次对敏感个人信息的概念作出界定。然而在具体实践中，《个人信息保护法》对其的一般概述仍然稍显笼统，很难直接应用于司法解释。针对这一问题，有学者指出，敏感个人信息的界定需经过“法律基准－具体维度－具体判定标准”三个循序渐进的深化阶段，结合场景融入与场景抽离的双重途径[3]。亦有学者认为，应当采取法定标准，以人格尊严为核心，兼顾人身、财产安全及未成年人标准，并根据特定情景进行判断[4]。总体而言，在敏感个人信息的界定问题上，美国学者尼森鲍姆的“场景完整性理论”被国内学者普遍接受并发展。笔者认为，场景理论的运用强调在某一情景中考量个人信息的“敏感性”大小，使风险评估更为准确，从而实现敏感个人信息的动态界定，确有其必要性。

具体实践中，场景因素将促成敏感个人信息的转化，此类敏感个人信息具有即时确定的性质。在此

前提下，ChatGPT 的应用很有可能涉及敏感个人信息的大量收集，这主要与信息存在状态与认知处理技术有关。关于信息存在状态，通常而言，单一状态的个人信息，由于可识别性较弱、可利用性较差，其权益受到侵害的风险也较小；ChatGPT 获取海量数据的行为客观上增大了信息汇聚密度，信息之间得以相互联系并补充，从而使个人信息处理风险增加。关于认知处理技术，ChatGPT 通过强大算法实现对信息的认知与处理，从而对海量数据进行快速分析，并据此生成新内容输出。在上述因素的共同作用下，即使个人信息本身并不具备敏感性，亦很有可能在 ChatGPT 收集或处理过程中转化为敏感个人信息。

2.2. 敏感个人信息保护现有法律规定

2.2.1. 强调合理利用而非绝对禁止

不同于《民法典》对私密信息的保护关注其隐私性、禁止非法披露和对外公开，《个人信息保护法》强调对敏感个人信息处理行为的规制[5]。主要原因在于，私密信息涉及信息主体的隐私权，立法关注信息主体对其非公开的控制力，除涉及公共利益外倾向于严格保护；相比之下，敏感的个人信息不必然是私密的，对于那些非私密但敏感的个人信息，如能得到合理利用则具有极高的公共价值。因此，我国并没有完全禁止敏感个人信息的处理，而是制定了特殊处理规则。这一点与绝大多数域外法有所不同。例如，欧盟 GDPR 第 9 条规定敏感数据处理应遵循“一般禁止 + 例外”原则，对敏感数据的处理采取严格限制，并不提倡对其合理利用。

2.2.2. 限定敏感个人信息处理前提

我国目前对敏感个人信息的保护，核心在于对其处理前提的限定。首先，特定目的要求敏感个人信息的处理必须遵守法律规定或合同约定，其目的明确可进一步细分为收集阶段的目的明确和使用阶段的使用限制[6]。其次，充分必要性的衡量主要参考比例原则，对个人信息保护影响进行评估。从一般公益、重要公益到极端重要公益，可处理的个人信息敏感程度依次递进，以确保个人信息的敏感性与公共利益的重要性协调一致[7]。域外法中，亦提倡使用比例原则进行必要性检验。例如，瑞典数据保护局“瑞典 GDPR 处罚第一案”称，采用面部识别对学校出勤情况进行评价并不符合比例原则，应当选择侵害权益更小的方式。最后，严格保护措施采取“事前”视角，实质上是针对敏感个人信息泄漏或非法使用风险而实施的防御手段，采取时应将个人信息处理目的等多种因素纳入考虑范围。

2.2.3. 遵循“特别告知 + 单独同意”规则

在知情同意方面，相比于一般个人信息处理时的“告知 + 同意”，我国对于敏感个人信息处理的规范更为严格，采取“特别告知 + 单独同意”规则。根据《个人信息保护法》第 29 条，信息主体对于其敏感个人信息的处理必须给予单独同意；根据第 30 条，处理敏感个人信息时，必须另行告知处理的必要性以及对个人权益的影响。这一规则的实施不仅有利于提高个人信息处理者的义务要求，规范敏感个人信息的收集行为；亦能通过单独授权的方式，使信息主体意识到授权所引发的风险，强化其权利保护意识[8]。

3. ChatGPT 引发的敏感个人信息安全风险

《个人信息保护法》的出台使我国敏感个人信息保护的法律框架更为系统完善，然而，ChatGPT 带来了新的挑战。作为大数据时代下生成式人工智能的最新成果，ChatGPT 虽然还未在我国得到广泛应用，但目前投入研发类 ChatGPT 技术的科技公司并不在少数，生成式人工智能已成为该领域的发展趋势。由于生成式人工智能在海量数据收集、处理和保存阶段均极有可能涉及个人信息，一旦这些个人信息的敏感性达到了将其界定为敏感个人信息的程度，则有必要关注 ChatGPT 在运行的各个阶段是否遵循敏感个人信息保护的处理前提与相关法律规定。此类问题集中体现在敏感个人信息获取合法合规性、处理规则

透明度、信息保存利用情况、监管主体与机制方面。

3.1. 信息获取合法合规性风险

知情同意作为个人信息收集的基本原则，在敏感个人信息问题上有着更为严格的规定，应得到特别关注。然而，OpenAI 公司公布的隐私政策中没有涉及其在数据收集阶段如何对敏感个人信息进行特殊保护，也没有提及知情同意原则的严格贯彻。值得注意的是，在场景理论的支持下，个人信息存在内在勾连性，而 ChatGPT 在运行的各个阶段均有个人信息收集的行为，伴随个人数据混同收集风险，这在一定程度上增加了个人信息的敏感性。

ChatGPT 在多阶段的个人收集过程中存在较大的敏感个人信息获取合法合规性风险。第一，在模型预训练阶段，ChatGPT 主要依靠网络爬虫技术对互联网上的海量数据进行收集，此类数据中不乏敏感个人信息的存在。这些个人信息被反复迭代学习，几乎完全脱离了知情同意原则，构成对敏感个人信息的侵权。尽管 OpenAI 表示，其收集信息还是会遵循 Robots (爬虫排除标准)协议，但是目前此协议只属于行业标准而无法律效力，无法形成对 ChatGPT 信息获取行为的有力规范。第二，在用户注册阶段，OpenAI 有权在用户创建账户时随时对其姓名、联系方式等账户信息和个人社交媒体信息进行收集，否则用户将无法获得完整服务。这种被迫同意的行为，实质上是对用户信息进行不当收集的表现[9]。第三，在正式使用阶段，用户与应用的对话信息将被自动收集，而此行为并未特别告知用户和取得用户的单独同意。

3.2. “算法黑箱”风险

算法黑箱在生成式人工智能的语境下，主要指其算法模型的复杂性使人们无法理解算法的目标和意图，也不清楚数据的具体处理过程。也就是说，用户很难得知 ChatGPT 使用已有数据和个人信息做出判断并生成相关内容的背后原理。并且，无监督学习状态下的 ChatGPT 将采用“自注意力机制”进行数据收集和处理，即使是开发设计者也难以理解其获得结果的过程。

ChatGPT 运行与目的限制原则和公开透明原则相违背，这是其在处理阶段引发敏感个人信息安全风险的最主要原因。根据我国相关法律规定，个人信息处理必须遵循以上两大原则。一方面，《个人信息保护法》针对敏感个人信息处理目的做出特殊规定，即“特定的目的”。例如，出于游戏“防沉迷”的需要，厂商被允许合理处理十四岁以下未成年人的信息；在股票交易中，金融机构有权对用户金融账户信息进行处理。然而，ChatGPT 处理规则的不透明性，使用户难以判断个人信息的处理是否超越了既定目的。另一方面，公开透明原则对信息处理透明度做出了更加直接的规定，即要求信息处理者公开披露个人信息处理的具体细节。ChatGPT 却凭借深度学习技术形成了“算法黑箱”，在系统输入的数据和输出的结果之间构筑起人们无法洞悉的“隐层”，具体处理细节无从得知[10]。

3.3. 信息泄漏与非法利用风险

敏感个人信息涉及信息主体的人格尊严和人身、财产权利，一旦遭到泄漏或非法利用，则很可能使信息主体权益暴露在重大侵害风险之下。2023 年 6 月 28 日，针对 OpenAI 公司的第一起集体诉讼就与个人信息的非法处理有关，原告指控其在生成式人工智能开发和运营中非法收集和使用了包括儿童信息在内的数以亿计的互联网网户个人信息，侵犯了信息主体的财产权、隐私权等法律权利。由此可见，ChatGPT 在具体实践中的确存在泄漏或非法利用敏感个人信息的现象，易引发相关风险。

敏感个人信息泄漏风险。有效的存储措施是防范个人信息泄漏的重要保证。然而，OpenAI 公司目前并未做出其存储用户个人信息情况的相关说明，用户无法得知其个人信息库的存储方式，对信息存储期限亦一无所知。并且，个人信息在泄漏之后对信息主体造成的损害常常伴有延后性，这不仅增加了信息主体事后举证的难度，也不利于信息的敏感性界定和风险评估[11]。

敏感个人信息非法利用风险。根据 OpenAI 公司公布的隐私政策，用户在应用 ChatGPT 时提供的所有个人信息，公司均有权在法律规定内使用并被允许将其提供给第三方。然而，“法律规定内”的承诺只对 OpenAI 公司本身具有规范性，一旦信息转移到第三方，OpenAI 公司便失去了对个人信息的掌控，进而引发个人信息扩散范围和用途失控等风险。

3.4. 多监管主体造成的责任竞合风险

我国近年来形成了由《民法典》《个人信息保护法》《网络安全法》等构成的人工智能治理框架，在类 ChatGPT 生成式人工智能出现广泛应用趋势之后，又于 2023 年出台《生成式人工智能服务管理暂行办法》，进一步明确有关生成式人工智能的监管主体责任。ChatGPT 通过深度学习生成内容，这包括文本、图像、声音、视频以及虚拟场景等多种信息形式的技术应用，故而我国在目前的治理框架下以各部门的专业性作为依托，在不同领域由不同主体实施监管并制定相应政策。

敏感个人信息在 ChatGPT 的运行各阶段随时面临着不同风险，其对于监管机制的效率要求也相应地高于其他一般个人信息。多监管主体有着全面规范的优势，但也产生了一定程度上的责任竞合风险，影响了其对于 ChatGPT 可能引发的各类风险的规制效率。具体来看，责任竞合风险可能造成监管的竞争与推诿[12]。监管竞争，即各个监管主体对于某些事项的潜在利益争相制定政策并积极推动执法，易造成不同规则之间的冲突；监管推诿，即各个监管主体因某侵权事件的处理复杂性、涉及面广和潜在利益不足等，可能选择避而不谈或相互推诿责任。当前的监管机制由于责任竞合风险的存在，不仅无法为敏感个人信息的保护提供及时有效的支持，并且可能导致治理资源的分配不均，从而影响公共利益。

4. 敏感个人信息风险规制路径

4.1. 构建动态的知情同意框架

面对 ChatGPT 的冲击，知情同意的作用正在逐渐虚化。然而，作为个人信息收集与处理限制的基本原则，知情同意依然不可或缺，当务之急是加强该制度的约束机制[13]。在 ChatGPT 收集与处理敏感个人信息这一过程中，信息在不同阶段处于不同场景之下，收集阶段告知并获得用户同意的既定目的，不一定在处理阶段继续适用。因此，应当针对敏感个人信息面对的不同场景构建起“动态”的知情同意框架。

一方面，“知情”需在具体场景中进行，做到持续性披露，同时注意披露的有效性。持续性披露义务区别于传统意义上的一次性告知义务，要求类 ChatGPT 生成式人工智能应用在信息收集阶段做到充分告知，并且在后续的信息处理阶段，如有目的或风险状况的变化，依然需要及时告知用户，充分保障信息主体知情权。有效的披露，具体指尽量避免信息过载带来的负面影响，需要确保不遗漏关键信息的同时，也避免过于频繁地告知敏感度较低的信息。为了引起信息主体的关注，还应当采取弹出窗口等醒目的方式强化告知效果。

另一方面，“同意”也需在具体场景中进行，做到明确同意，同时强化同意要件。与一般个人信息处理适用推定同意和默示同意的规定有所不同，敏感个人信息必须得到信息主体的明确同意[14]。类 ChatGPT 生成式人工智能应用可以在满足敏感个人信息处理前提的情况下对其进行合理利用，但必须在每一阶段征得信息主体明确的单独同意。同时，还应根据个人信息的敏感程度来划分个人信息种类，在处理高度敏感的个人信息时，必须设定更加严格的同意要件。例如，如果需要收集和处理儿童的个人信息，应当取得其监护人的明确同意[15]。

4.2. 打开信息处理的“算法黑箱”

提高生成式人工智能处理规则的透明度，不仅有利于个人信息主体了解其信息处理过程，也有利于

监管部门规范信息处理行为和处理意图。在具体实践中，相较于解释 ChatGPT 所生成内容，或是依据“浴缸型透明”原则公开数据源代码和训练集，公开其算法模型的原理更具实用性。原因在于，一方面目前以 ChatGPT 为代表的生成式人工智能在技术发展方面仍具有不完善性，其生成内容的可解释性一般较低，算法黑箱难以得到完全消解；另一方面，人们受限于认知能力，数据源代码和训练集的公开对其了解 ChatGPT 算法几乎没有实际意义。算法模型原理的解释避免了以上问题，成为目前打开 ChatGPT 信息处理“算法黑箱”的最优解。

从信息主体权益角度来看，算法模型原理的解释对于个人信息权益的丰富亦有所启示。在类 ChatGPT 生成式人工智能的冲击下，个人信息保护不仅可以从强化知情决定权、更正补充权、删除权等传统权利入手，也可以借助新型权利的授予。算法解释权能够维护受算法自动化决策影响的权利人之合法权益，在商业自动化决策领域有所应用。此权利要求算法设计者解释算法结果的决策过程、运行原理等事项，符合当前对于 ChatGPT 算法模型原理解释的需要。鉴此，应当通过扩展个人信息权益的方式，赋予信息主体算法解释权，从而要求 ChatGPT 人工智能服务提供者披露算法的运行机制，实现算法的透明化。

4.3. 敏感个人信息脱敏与泄漏防护

技术治理是互联网时代敏感个人信息保护的有效措施。立法可以制定一系列原则性规定，确立敏感个人信息处理应遵循的技术标准，例如规定 ChatGPT 在处理敏感个人信息时做到脱敏处理。个人信息脱敏的方法有多种，一种常见的做法是去标识化处理，它通过删减或变更某些数据来阻断信息和具体个人之间的关联，有效降低了信息的敏感性，从根本上减少了敏感个人信息被泄露和非法使用的可能性。但是，有关去标识化程度的相关规定需要在立法中得到特别关注。原因在于，过于彻底的去标识化可能损害数据的质量，从而限制信息的效用；而不彻底的去标识化处理又会增加敏感个人信息通过相互关联重新获得识别性的可能[16]。

在防范敏感个人信息泄漏方面，去标识化等脱敏处理只能降低信息泄漏对于信息主体的危害风险，而无法做到直接防护。由于敏感个人信息在收集、存储、使用等阶段均有泄漏可能，全流程的泄漏防护机制有其设立必要性。另外，针对 OpenAI 公布的隐私协议中其有权将个人信息提供给第三方这一条款，可以在其遵循以下要求的前提下保留。首先，OpenAI 只有在法律事由的基础上，才能合法或依约将敏感个人信息传输给特定的第三方。其次，敏感个人信息的传输必须符合强相关性和最小必要原则，即传输的敏感个人信息要尽可能少且与第三方需求目的紧密相关。最后，应在信息被传输前为其添加追踪溯源标记，使用加密传输方法，并且对信息泄漏风险进行实时监测。

4.4. 完善监管主体责任机制

在监管机制层面，应将企业自治、司法审查与行政监管相结合。尽管目前生成式人工智能的监管主体众多，但基本都集中在行政层面，例如国家市场监督管理总局、国家互联网信息办公室、工业和信息化部等。考虑到 ChatGPT 等生成式人工智能所具有的专业性和复杂性特征，应在监管机制内纳入更多相关主体。企业作为人工智能技术的投资者、开发者，本身具有其社会责任；加之其拥有的资金、技术、人才市场等多方面的资源将大大增强治理的效能，因此将企业自治纳入监管机制内具有合理性。司法审查相较于行政监管，更具稳定性和统一性，能够更好地推动社会树立正确的价值观念、巩固正确的行为规范，重塑监管治理的价值目标，因此也应被纳入监管机制内。

对于可能存在的责任竞合问题，可以在政府、法院、企业三方协作机制的搭建过程中，通过重新梳理类 ChatGPT 生成式人工智能的具体监管主体得到解决。即在现有的分领域确定监管主体的机制下，兼采“信息数据－算法技术－生成内容”分阶段划分标准，并强调三方监管的协调性，确定不同情景下监

管主体责任的优先级，以预防不同主体之间的监管竞争与监管推诿，降低责任竞合风险。最终构建起责任明确、多层次、分阶段的生成式人工智能监管体系。

5. 结语

ChatGPT 等生成式人工智能的快速发展是一把双刃剑，其作为一项新技术极大便利了人们的生活、工作与学习，但与此同时，此类技术对敏感个人信息安全造成的重大威胁和风险也不容忽视。这种风险体现在 ChatGPT 运行的全过程，具体表现为信息获取合法合规风险、算法黑箱风险、泄漏与非法利用风险、监管责任竞合风险等。虽然我国目前针对生成式人工智能可能引发的法律风险已经初步建立起治理框架，但在知情同意原则、算法解释规定、泄漏防护机制以及监督机制方面存在的漏洞亟待进一步完善。对此，应将敏感个人信息安全风险的动态性纳入考虑，采取动态知情同意、全流程泄漏防护等措施，并要求算法模型的公开解释，以此尽可能消除 ChatGPT 等生成式人工智能引发的各类敏感个人信息安全风险，使其更好地服务于全社会。

参考文献

[1] 於兴中, 郑戈, 丁晓东. “生成性人工智能”与法律: 以 ChatGPT 为例[J]. 中国法律评论, 2023(2): 1-20.

[2] 韩旭至. 敏感个人信息的界定及其处理前提——以《个人信息保护法》第 28 条为中心[J]. 求是学刊, 2022, 49(5): 132-145.

[3] 宁园. 敏感个人信息的法律基准与范畴界定——以《个人信息保护法》第 28 条第 1 款为中心[J]. 比较法研究, 2021(5): 33-49.

[4] 王利明. 敏感个人信息保护的基本问题——以《民法典》和《个人信息保护法》的解释为背景[J]. 当代法学, 2022, 36(1): 3-14.

[5] 朱晓峰, 黎泓玥. 私密信息与敏感个人信息区分保护论[J]. 经贸法律评论, 2023(1): 21-38.

[6] 刘新宇. 中华人民共和国个人信息保护法重点解读与案例解析[M]. 北京: 中国法制出版社, 2021.

[7] 蔡星月. 个人隐私信息公开豁免的双重界限[J]. 行政法学研究, 2019(3): 134-144.

[8] 程啸. 个人信息保护的理解与适用[M]. 北京: 中国法制出版社, 2021.

[9] 郭雪慧. 人工智能时代的个人信息安全挑战与应对[J]. 浙江大学学报(人文社会科学版), 2021, 51(5): 157-169.

[10] 徐凤. 人工智能算法黑箱的法律规制——以智能投顾为例展开[J]. 东方法学, 2019(6): 78-86.

[11] 朱荣荣. 类 ChatGPT 生成式人工智能对个人信息保护的挑战及应对[J/OL]. 重庆大学学报(社会科学版): 1-14. http: //kns. cnki. net/kcms/detail/50. 1023. C. 20230921. 1151. 002. html, 2023-09-21.

[12] 毕文轩. 生成式人工智能的风险规制困境及其化解: 以 ChatGPT 的规制为视角[J]. 比较法研究, 2023(3): 155-172.

[13] 孙清白. 敏感个人信息保护的特殊制度逻辑及其规制策略[J]. 行政法学研究, 2022(1): 119-130.

[14] 丁晓强. 个人数据保护中同意规则的“扬”与“抑”——卡-梅框架视域下的规则配置研究[J]. 法学评论, 2020, 38(4): 130-143.

[15] 郑志峰. 人工智能时代的隐私保护[J]. 法律科学(西北政法大学学报), 2019, 37(2): 51-60.

[16] 万方. 隐私政策中的告知同意原则及其异化[J]. 法律科学(西北政法大学学报), 2019, 37(2): 61-68.