Open Journal of Legal Science 法学, 2024, 12(5), 3390-3396
Published Online May 2024 in Hans. https://www.hanspub.org/journal/ojls
https://doi.org/10.12677/ojls.2024.125481

Hans汉斯

生成式人工智能视域下个人信息保护的风险路径与算法治理

保洁玉

湖南工业大学法学院，湖南 株洲

收稿日期：2024年4月2日；录用日期：2024年4月17日；发布日期：2024年5月31日

摘　要

生成式人工智能的涌现使得个人信息保护面临着日益严峻的风险。本文探讨了生成式人工智能技术的算法逻辑以及在个人信息保护中产生的风险，分析了个人意思自治空间被压缩、个人信息频繁泄露、数据合规外延扩大、主体权利价值被削弱等风险路径。算法作为规则的正当性仍需深度论证，但其在个人信息保护制度中已初见雏形，完善个人信息保护制度中的算法规则对保护主体信息权益有重要作用。在算法治理进路上，我们以算法可解释性为契机重构算法二元治理进路，打开生成式人工智能场景下算法治理的切口，以此实现个人信息保护和流通的平衡。

关键词

生成式人工智能，个人信息保护，算法解释，算法治理

Risk Paths and Algorithmic Governance of Personal Information Protection in the Perspective of Generative Artificial Intelligence

Jieyu Bao

Law School of Hunan University of Technology, Zhuzhou Hunan

Received: Apr. 2nd, 2024; accepted: Apr. 17th, 2024; published: May 31st, 2024

Abstract

The emergence of Generative Artificial Intelligence (AI) has put the protection of personal infor-

文章引用: 保洁玉. 生成式人工智能视域下个人信息保护的风险路径与算法治理[J]. 法学, 2024, 12(5): 3390-3396.
DOI: 10.12677/ojls.2024.125481

mation at increasing risk. This paper discusses the algorithmic logic of generative AI technology and the risks arising from personal information protection, analyzes the risk paths of compression of the space of personal autonomy, frequent leakage of personal information, expansion of the extension of data compliance, and weakening of the value of the subject's rights. The legitimacy of algorithms as rules still needs to be demonstrated in depth, but it has taken shape in the personal information protection system, and improving the algorithmic rules in the personal information protection system plays an important role in protecting the subject's information rights and interests. On the way of algorithmic governance, we take interpretability of algorithms as an opportunity to reconstruct the algorithmic binary governance approach, and to open the cut of algorithmic governance in generative artificial intelligence scenarios, so as to realize the balance between the protection and circulation of personal information.

Keywords

Generative Artificial Intelligence, Personal Information Protection, Algorithmic Interpretation, Algorithmic Governance

1. 引言

2022 年底，ChatGPT 的问世使得生成式人工智能被推上风口浪尖，这项技术彻底改变了人们对数字时代生活的固有认知。海量数据库支撑下的生成式人工智能给用户带来惊喜的同时，个人信息保护问题逐渐浮现。本文先尝试分析生成式人工智能算法逻辑与其引发的个人信息保护风险，从个人信息的本质出发，进一步探明作为规则的算法在个人信息保护制度中的体现，并结合算法的可解释性特征，尝试在算法治理层面重构个人信息保护的可行性方案。

2. 生成式人工智能算法逻辑与其引发的个人信息风险

(一) 生成式人工智能的算法逻辑

生成式人工智能是指基于算法、模型、规则生成文本、图片、声音、视频、代码等内容的技术[1]。ChatGPT 作为生成式人工智能具有代表性的算法模型，它使用的是基于 Transformer 架构的预训练与微调模型，能够提高模型的性能，并实现更准确的语义理解和语句结构。这种模型使用了大规模的语料库进行预训练，然后在微调阶段通过优化超参数来优化模型性能和训练效果，并通过验证或人工评估来检查模型质量，算法模型的应用是生成式人工智能开发的里程碑。

相较于传统算法模型，ChatGPT 突破传统人工智能的核心在于其不只是依靠机器学习，同时也掌握了大量的人工标注来修改、核对机器学习所得出的结论，在机器学习的基础上通过人工标注的方式双线性推动人工智能突破式进化，开创了“机器学习 + 人工标注”这一算法模型，这也是生成式人工智能的典型特征[1]。将“人工标注”这一传统数据筛选方式应用至 ChatGPT 中，是 ChatGPT 算法模型的不同之处，可以更好地向用户反馈生成式人工智能的数据接受和学习功能。由此可见，“机器学习 + 人工标注”作为 ChatGPT 算法模型运行过程中的技术内核，这种技术组合模型本质上是为了提升 ChatGPT 的智能性，以达到生成式人工智能流畅地同人类交流的目的，但这同时也会导致其中存在算法黑箱和算法偏见，从而增加了个人信息保护的法律风险。

(二) 生成式人工智能激化个人信息保护风险

生成式人工智能能够“记住”与用户对话中的个人信息，并用于模型训练和内容生成，它本身具备收集、储存和使用个人信息的功能。解构生成式人工智能的技术逻辑可以发现，生成式人工智能的工作主要分为数据收集、数据训练、模型迭代、数据应用四个阶段，均离不开个人信息的收集与分析，为个人信息保护带来了诸多不可预知的风险。

1) 算法黑箱压缩个人意思自治空间

个人意思自治在个人信息保护制度中的集中表现即是知情同意原则的适用。知情同意原则一直作为个人信息处理的合法性基础，我国《个人信息保护法》第 17 条对知情同意原则做出了明确规定。然而在生成式人工智能的开发和应用过程中，知情同意原则受到挑战，致使个人意思自治空间被压缩，其背后的关键在于算法黑箱的复杂性和未知性。

算法“黑箱”本质上归属于技术“黑箱”，“其特点是部分人知道，另一部分人不一定知道”[2]。这一特性意味着其内部机制只为部分精通科技的群体所知悉，而对于外行人，则多数维持在一种未知或不可解之境。特别是在算法驱动的程序阶段中，相关技术极为复杂，常超出普通用户的理解范畴。尤其在生成式人工智能领域，数据的搜集与利用环节暗含算法的这种“黑箱”现象[3]。以用户注册账户为例，当系统提示个人信息收集并引导用户点击“下一步”时，此行为默认构成了授权同意。但是，尽管隐私政策的告知在形式上存在，实质上却往往仅停留在对采集信息种类的揭示上，并未详尽阐释处理细节及算法决策逻辑。这种做法恰恰映射了技术细节对非专业群体的黑箱效应。由此可见，在以数据收集和繁复算法为技术逻辑的 ChatGPT 空间里，能够接触到生成式人工智能运作算法机理的信息收集者和信息处理者成为了“可以了解或得到解释”的部分人，“算法黑箱”的存在实质上剥夺了用户对于个人信息收集和处理过程的知情权，压缩了用户意思自治的空间[4]。

2) 数据聚集增加个人信息泄露风险

生成式人工智能可能给个人信息保护带来重大挑战，其中个人信息泄露风险尤为突出。据 ChatGPT 开发者 OpenAI 的首席技术官米拉·穆拉蒂表示，生成式人工智能的模型开发需要海量的个人信息，这些信息在对模型进行训练的同时，也存在着被泄露的风险。微软已宣布禁止公司员工向 ChatGPT 分享公司敏感数据。亚马逊公司律师也警告员工“不要与 ChatGPT 分享任何亚马逊的机密信息，因为被分享的信息极大可能被用于算法模型的训练而被泄露”[5]。除此之外，在模型训练过程中大量的个人信息被收集、存储且可能未被及时删除，生成式人工智能拼凭借巨大的数据库和较强的算法功能，对此前输入的个人信息进行分析匹配，在对话过程中将用户个人信息发送给相关用户。例如用户通过 ChatGPT 对自己的亲人进行画像，此过程势必会输入亲人的个人信息，ChatGPT 对这些信息进行关联性分析之后可能发送给其认为与此亲人有密切关系的其他用户，这无疑会将信息主体的合法权益置于风险之地。

生成式人工智能，一种基于大量数据聚集和“机器学习 + 人工标注”算法的技术，虽然具有不断学习和改进模型的能力，以提高文本准确性，但其自身仍存在设计不当、错误存储系统和算法漏洞等问题。这些问题不仅使得生成式人工智能面临安全隐患，还可能导致个人信息泄露。

3) 模型交互训练扩大数据合规外延

生成式人工智能在技术逻辑上不同于传统人工智能之处在于它存在模型交互训练阶段。在传统人工智能的工作环境下，用户只要同意其隐私政策即视为人工智能需对用户的所有信息进行保密，人工智能的数据合规范围是相对固定的。但是在生成式人工智能的大模型交互训练阶段，用户与人工智能交互产生的数据存在被运用到模型迭代训练的可能性，用户初次允许获取的数据范围与模型迭代训练的交互数据范围并不完全相同，因此原则上不能以用户初次的同意权来限制交互数据的适用范围，此举会扩大个人信息主体权利扩张，然而 OpenAI 的隐私政策中并未明确阐释对此类数据的合规处理方式，对模型交

互训练产生的数据合规风险仍不容小觑。

我国的数据合规法律构架，以网络安全法、数据安全法及个人信息保护法为三大柱石，形成涵盖网络、数据与个人信息全域的合规体系。然而，在生成式人工智能领域，关于数据合规性的考量尚未充分拓展至模型互动训练阶段，目前主要集中在数据来源的合规性风险评估。尽管 OpenAI 声明宣示将采取匿名化、加密等措施加强数据安全性，但公开报道揭示数据泄露的隐患依旧存在。模型交互训练已然在实质上扩大了数据合规的外延，需对生成式人工智能的数据合规体系进行更深入的研究与制度设计，从而有效遏制和预防潜在的数据安全漏洞。

4) 风险控制路径弱化主体权利价值

我国个人信息保护的实践中，构建了以权利保障与风险治理双轨并行的路径。在“权利保护路径”中，立足于信息主体的自决权，遵守确认权利、行使权利至救济权利的逻辑链条，彰显以人为本的法治精神，属于以个人为中心的法律范式[6]。而“风险控制路径”则着眼于多元方法对个人信息潜在风险进行洞察和把控，致力于平衡公民权利保护与合法利益保障的关系，其本质上削弱了信息权的绝对化，更显现出其作为策略工具的一面。这两种路径相互补充，共同筑起了个人信息保护的堡垒。

在国内学术界探究个人信息保护的理论框架时，以风险管理为视角构建的思考模式显著差异于欧盟范式，关键区别在于多数学者偏向选择风险控制途径作为权利主导路径的替代选项，而非互补机制。虽然个人信息保护肩负起保障社会公正等集体价值的重任，然而确保个人自主性与尊严不容置疑地屹立于其核之。诚然，在实践中，权利框架可能在全面维护个人信息权益上显得力不从心，涉及价值平衡方面的短板也难以忽视，但个人对信息的实质控制权仍然具有无可替代的价值。从风险角度出发的功效论分析在某种程度上与作为伦理边界的权利保护存在显著张力。这是因为，个体信息权的侵害并非可以简单量化或以程度衡量的事件，其中包含了复杂且深刻的道德考量。我国学者主张将风险控制路径替代权利控制路径来构建个人信息保护制度，在主体权利保护和个人价值方面仍存在风险。

3. 生成式人工智能视域下个人信息保护与算法之间的逻辑重塑

(一) 算法作为规则的内部证成

本质上，算法是一系列以代码为表达方式、以程序为载体、以解决特定问题为目标，并根据数据、环境等因素变化而变化的独立于人脑运行的步骤方法[7]。算法本身并不具有规则效力，但是在这个“一切被数据化的时代”，算法已经充斥人们生活的方方面面，为规制数据的合理合法使用，算法规则化受到了强烈呼吁。有的学者将算法的规则化演进过程称之为“算法的法律归化”，该理论源自于“技术归化”理论，意指通过法律手段对算法进行转化和驯服，使其融入现代社会生活，成为具有使用价值的可控之物。

二十多年前，数据时代特征初现，学术界就存在“代码就是法律”的断言，现今如果把代码换成算法，在逻辑上也能成立[8]。规则作为权力的逻辑基础，它可以表达、预测权力的理性化方式，从规则角度观察算法意味着关注算法的表现形态。

规则包括正式规则和非正式规则，算法在形式上以计算机代码的“成文”形式呈现，根据“波兰尼悖论”的分析框架，算法亦可被划分为正式与非正式两类。传统算法属于正式规则，具有明确的可解释逻辑，人类能理解其规则内容及决策形成的影响方式；相比之下，在第三次人工智能浪潮中，机器学习算法作为非正式规则缺乏可解释性。无论是算法的创建者、使用者还是受影响者(个人信息主体)，均难以明确解释生成式人工智能作出算法决策的原因和过程[9]。生成式人工智能场景下的算法是一个动态的技术，通过模型迭代而不断变化，现有的法律和政策正推动生成式人工智能的算法实现规则化，并推动算法与平台规则、用户协议、隐私政策等成文规则进行整体性协调。

(二) 个人信息保护中算法规则的构成分析

生成式人工智能环境中，个人信息需具备“可识别”特征，并应受到算法技术的限制[10]。识别性是个人信息的自然特征，算法性则是其技术特征。有学者提出，在数字时代，个人信息权的客体范围、法律属性以及权力内容都根植于算法识别[11]。算法是数字时代个人信息无法回避的构成因素。分析算法规则在个人信息保护制度中的构成有助于了解现阶段个人信息保护的偏倚，并在算法治理路径中予以完善。

在我国的法律规范中，算法的表述主要面向自动化决策、数据处理、算法推荐技术、深度合成技术。我国近年来加强了在个人信息保护制度层面的算法治理研究与立法工作。《个人信息保护法》第 24 条第 3 款即是对算法自动化决策设定的规范条文。2022 年《算法推荐管理规定》生效，该规定细化了互联网信息服务领域包括服务规范、用户权益保护、监督管理等方面的算法推荐。算法的进步增加了个人信息的重要性，然而算法在个人信息保护制度中的角色尚不明确，基于算法对个人信息的保护亦未形成体系化的法律规范[12]。

在当前生成式人工智能时代背景下，通过立法手段健全个人信息保护体系已成为一种趋势。然而，也有学者提出，或许可以探索一种法治的补充形式——自治进路，从而寻找解决个人信息保护规则体系不足的可能途径。这包括认可算法中特有的个人信息保护规则，即平台规则，具有一定的法律效力[13]。在实施《个人信息保护法》及其他个人信息保护法规设定的具体义务时，平台的信息保护政策在规范个人信息处理者行为方面发挥着不可小觑的功效，有学者认为，算法平台规则具备准法律规范的属性。例如在 ChatGPT 应用场景下，OpenAI 的隐私政策是 ChatGPT 用户在注册时关注的关于个人权益保护的首要参考。平台在全面考量的基础上，针对特定的算法场景对个人信息保护规则进行细化，使得算法与个人信息保护更加契合。由此可见，平台规则已经在执行类似准法律规则的职能，平衡个人信息的保护与利用。

4. 生成式人工智能视域下个人信息保护风险的算法治理进路

算法治理绝不应将算法视为一种独立于社会生活的技术，借由韦伯对形式理性与实质理性的区分原理来阐明算法治理模型之建构，需要同时兼顾形式理性和权力理性，一方面需要推动算法它本身在形式理性上体现的计算、效率、客观的特点，另一方面需要算法发挥权力理性，协助社会权力确保生成式人工智能的决策过程和结果的公平性、有责性[14]。

个人信息保护作为一项法律问题，应从算法治理层面上探究可行进路，从法律“应然”回归算法“实然”，算法治理是在对算法本质性特征进行系统分析和学术抽离的基础上搭建起的一套完整的权利义务体系，算法可解释性作为算法归责的重要依据，由此衍生的算法解释权可以作为打开生成式人工智能场景下算法治理的切口，推动算法与个人信息保护从剥离走向融合。

(一) 算法解释权在算法治理中的地位

算法解释权的设立是为了应对与个人信息相关的算法黑箱问题。算法解释权包括广义和狭义两种理解。广义算法解释权指的是法律授予个人与算法决策相关的各种权利，例如拒绝纯自动化决策的权利。而狭义的算法解释权则指个人有权要求对任何算法决策进行解释的权利[15]。我国《个人信息保护法》第 7 条、第 24 条和第 44 条的规定都体现了算法解释权的一般性主张，为了更好的搭建以算法解释权为基点的算法治理体系，本文采用广义上的算法解释权界定。

算法解释权是一种新型权利，是指数据主体有权获得解释和挑战算法决策的权利，其实质是立法者设计的促进算法透明度的个性化措施。算法解释权能够满足规制算法决策、保障信用主体权益的双重需求，对构建有效的算法治理制度的特殊作用在于破解算法模型的不可解释性，若在立法层面赋予个人信息主体算法解释权，要求个人信息处理者履行算法解释义务，能够较好地依据算法应用的多元化场景有

力保护公众个人信息免受算法黑箱的不当侵害。

(二) 以算法解释权为基点的算法治理制度建构

1) 构建二元共治的算法治理模型

算法作为平台内部治理和外部纠纷解决的规则，算法治理涉及众多主体之间的交互监管，落实算法解释权在个人信息保护方面的功效，需要盘活算法解释权有效实施的配套制度，兼具内部和外部视角，构建法治和自治协同推进的二元共治算法治理模型。

其一，构筑内部算法治理机制。内部算法治理机制主要面向生成式人工智能技术的开发者，其集中表现形式为企业研发，企业是算法的研发者和应用者，在促进算法创新及健康发展中发挥着至关重要的作用，企业作为生成式人工智能的开发者、个人信息流转的平台，自身应该及时介入对算法的监管并形成行业自律，形成行业自律公约，有效完成企业合规计划。算法解释作为实施算法问责的重要机制，可以直接反应企业内部监管效力，企业应积极落实个人信息处理者的算法解释义务，强化内部技术团队配置，聘请专业人员担任算法系统准确性的守门人，确保个人信息主体有权获得对算法决策的解释。

其二，布局外部算法监管机制。外部监管意味着应当通过立法重新配置算法治理的权力体系，明确除企业外，政府、行业组织、公众多元治理的主体地位，尤其应当确立人工智能治理专业委员会等机构的实体性属性，赋能该委员会在实质层面的统一领导地位[15]。同时，还应以立法方式为黑箱算法设计者和控制者建立责任追究机制，为个人信息主体遭受侵害提供维权渠道。通过内部治理，外部监管的算法治理模型布局“政府监管、社会监督、公民维权、企业自治”的算法治理格局[16]。

2) 明确信息处理者的算法解释义务

《个人信息保护法》第五章专门规定了个人信息处理者的义务，传统的个人信息处理者的义务一般包含个人信息安全保障的一般义务、合规审计义务、信息安全评估与补救义务、大型互联网平台的“守门人义务”以及个人信息保护负责人制度。但是算法黑箱的存在，增加了个人信息处理过程的场景性和不确定性，因此在生成式人工智能场景下，需要对信息处理者额外设定算法解释义务。

算法解释义务的设立根据可从现有关于算法的规定中厘清。2021 年《互联网信息服务算法推荐管理规定》规定了个人信息处理者的算法解释义务。如果个人信息主体认为个人信息处理者存在利用算法屏蔽信息、过度推荐的情况，对保护其个人信息权益造成重大影响，有权要求算法推荐服务提供者进行解释，并采取相应改进或补救措施[17]。根据《通用数据保护条例》第 22 条规定，个人信息主体有权利不受仅依靠算法分析的自动化决策的限制，个人信息处理者只能在特定情况下使用自动化决策，并应当实施适当的措施保护个人信息主体的权利、自由和正当化利益[18]。

算法解释义务的范围也并非不分畛域，个人信息处理者履行义务不需要提供复杂的数学解释和繁冗的技术细节为个人信息主体增加理解难度，而应以清晰简明的语言解释特定自动化决策系统的基本逻辑、使用的数据类型及权重，只需要证实与具体算法决策结果有关的有用信息。尽管现阶段学界对算法解释义务的对象和衡量标准还无法达成一致，但该义务的设立已经实质性地促进了个人信息处理者更公平、负责、审慎地做出算法决策，与个人信息主体的算法解释权紧密合作保护个人的尊严。

5. 结语

生成式人工智能为人类生活和社会带来了诸多便利，但用户个人信息泄露事件频繁发生，暴露了个人信息保护及风险防范机制诸多不足，例如隐私政策不透明、数据安全措施不足、平台数据攻防存在风险、数据泄露风险、缺乏监管和执行机制等。算法在个人信息保护中扮演着至关重要的角色，是确保数据安全和隐私保护的支柱，“算法的法律归化”逐渐走入人们视野，其进程存在正当性。我们深切认识到个人信息安全面临着日益复杂的挑战，需要跨学科合作、法律监管与技术创新相结合的综合治理机制。

未来的研究和实践应当致力于建立更加健全的法律法规框架、加强数据隐私保护技术研发，以及促进人工智能技术与个人信息保护的良性互动，从而实现个人信息保护与科技发展的动态平衡，确保社会的信息安全与隐私权益得到最佳的保障与平衡。

参考文献

[1] 周学峰. 生成式人工智能侵权责任探析[J]. 比较法研究, 2023(4): 117-131.

[2] 刘艳红. 生成式人工智能的三大安全风险及法律规制——以 ChatGPT 为例[J]. 东方法学, 2023(4): 29-43.

[3] 谭九生, 范晓韵. 算法“黑箱”的成因、风险及其治理[J]. 湖南科技大学学报(社会科学版), 2020, 23(6): 92-99.

[4] 仇筠茜, 陈昌凤. 基于人工智能与算法新闻透明度的“黑箱”打开方式选择[J]. 郑州大学学报(哲学社会科学版), 2018, 51(5): 84-88+159.

[5] 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.

[6] 斜晓东. 风险与控制: 论生成式人工智能应用的个人信息保护[J]. 政法论丛, 2023(4): 59-68.

[7] 丁晓东. 论个人信息法律保护的思想渊源与基本原理——基于“公平信息实践”的分析[J]. 现代法学, 2019, 41(3): 96-110.

[8] 张新平. 算法与法律的冲突及其化解[J]. 法律科学(西北政法大学学报), 2024, 42(2): 135-146.

[9] 胡凌. 作为规则的推荐算法: 演进与法律治理[J]. 图书情报知识, 2023, 40(1): 45-55.

[10] 贾开. 人工智能与算法治理研究[J]. 中国行政管理, 2019(1): 17-22.

[11] 蒋舸. 作为算法的法律[J]. 清华法学, 2019, 13(1): 64-75.

[12] 彭诚信. 重解个人信息的本质特征: 算法识别性[J]. 上海师范大学学报(哲学社会科学版), 2023, 52(3): 68-81.

[13] 张凌寒. 自动化决策与人的主体性[J]. 人大法律评论, 2020(2): 20-48.

[14] 范明志, 吕一川. 论算法中的个人信息保护规则体系[J]. 数字法治, 2024(1): 93-111.

[15] 丁晓东. 基于信任的自动化决策: 算法解释权的原理反思与制度重构[J]. 中国法学, 2022(1): 99-118.

[16] 张欣. 算法解释权与算法治理路径研究[J]. 中外法学, 2019, 31(6): 1425-1445.

[17] 何新新, 徐澜波. 个人信息处理者的自动化决策解释义务研究[J]. 学习与实践, 2022(8): 79-87.

[18] 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.