武漢大學學報(哲学社会科学版)
JOURNAL OF WUHAN UNIVERSITY(Philosophy and Social Sciences)
第79卷第4期 2026年7月
Vol. 79 No. 4 July 2026 138~150
DOI:10.14086/j.cnki.wujss.2026.04.012

法律场景下大模型幻觉的类型化规制
——以公权力机关的应用风险为核心

冯玉军 沈鸿艺

摘 要 随着中国人工智能技术的不断发展,法律场景下的大模型幻觉现象愈发普遍。通过访谈调研发现,在公权力机关应用大模型的现实场景中,已经出现了包括法条幻觉、事实幻觉与涵摄幻觉在内的多类典型幻觉。在对公安机关、检察院和法院的醉酒危险驾驶类案件决策场景的专项测试中,发现幻觉集中性呈现在一至两个决策点上。考虑到错误大小、决策模式以及问责模式的转变,法律场景下产生的幻觉有必要进行类型化规制。规制思路应当从场景风险分级的基本原则切入,通过权利影响的处分性和自由裁量权两个维度的组合,形成"高风险、中风险、低风险"的划分格局。依据风险等级与来源针对性地设计"前端预防—中端监督—后端问责"的三阶段规制路径,实现分类分级的精细化治理。

关键词 大模型幻觉;公权力机关;人工智能应用;风险分级;幻觉规制;自由裁量权
中图分类号 D926 **文献标识码** A **文章编号** 1672-7320(2026)04-0138-13
基金项目 国家重点研发计划重点专项项目(2022YFC3301903);中国法学会2024年度部级法学研究课题(CLS-2024-D23)

社会治理智能化是我国的一项长期规划。继2017年《国务院关于印发新一代人工智能发展规划的通知》提出开发智能政务、智慧法庭等具体应用场景的要求后,国家公权力机关积极出台促进人工智能和具体业务融合的制度文件,如2025年出台的《政务领域人工智能大模型部署应用指引》(以下简称《指引》),2022年出台的《最高人民法院关于规范和加强人工智能司法应用的意见》(以下简称《司法应用意见》)等。然而,大模型在法律场景下应用的关键挑战是幻觉(Hallucination)的爆发性出现,在美国2023年发生的"Mata v. Avianca"案件中,两位律师因在提交法庭的文件中引用ChatGPT生成的虚假案例而被处以5000美元罚款。幻觉问题一定程度上关系到人工智能是否可以被信任,从而对大模型在法律领域应用的范围产生影响。

因此,上述趋势自然凸显出如下问题:公权力机关在应用大模型时是否会出现幻觉?幻觉又是否会对公权力使用的正当性产生风险?在大模型应用带来显著效益的情况下,是否有可能通过规制幻觉从而降低风险?为此,本文通过访谈调研和专项测试的方法,考察现有法律场景下大模型幻觉的实践图景,继而通过阐明规制幻觉的必要性,澄清划分场景风险等级的基本原则,最终构建类型化的规制方案。

一、法律场景下大模型幻觉的实践图景

为深入理解法律场景下大模型幻觉的现实性,需对目前公权力机关部署的真实模型产生幻觉的基本情况进行调查。本部分在对法律场景下的幻觉展开描述性分类的基础上,通过定性访谈了解幻觉问题的真实性,继而使用定量方法对主流模型进行专项测试,以完整呈现法律场景下大模型幻觉的实践图景,为风险的判断提供认识论基础。

（一）基于法律推理的幻觉分类

幻觉的通用定义是指人工智能系统生成的看似是事实的错误信息。这类信息主要有错误性和迷惑性两个特征，前者指生成的信息本身是错误的，后者指生成的信息呈现出看似是事实的表象，令人难以立即察觉。在计算机科学中，主要是通过将输出的信息与源信息对比，来判断二者的关系是相互矛盾还是无法验证，从而区分出内在幻觉和外在幻觉[1](P3)。全国网络安全标准化技术委员会发布的《政务大模型应用安全规范》（以下简称《规范》）认为，幻觉是指大模型在生成内容时，输出看似合理但实际与事实不符、与用户输入逻辑不一致或虚构的信息。综上可知，幻觉的定义取决于我们选择哪些正确的材料与输出内容相对比，在法律场景中，具有正确性的三段论推理可以成为描述幻觉的基本框架。

在法律场景中应用大模型不可避免地会在三段论推理的框架下进行。按照传统概念法学的观点，法律规范/规则是一个逻辑严谨、用语确切的融贯体系，对于待裁决的事实纠纷，法官的任务就是从大、小前提演绎式地推出结论，在这一推论过程中，只要前提为真，则结论也为真[2](P33)。全面观之，一个正确的法律结论来源于三个部分，即在大前提中须有正确的法条，在小前提中须正确地概括事实，在结论中须正确地展开法条与事实的涵摄。因此，我们可以将幻觉具体划分为法条幻觉、事实幻觉以及涵摄幻觉三种。法条幻觉是指大模型无法根据案件事实寻找到正确的法条；事实幻觉是指大模型无法根据输入的材料准确提炼出事实内容；涵摄幻觉是指大模型无法在法律和事实之间进行正确的涵摄推理。大模型可以在某个环节或者全部环节发挥作用，一旦输出错误，就可以称之为产生了法律场景下的幻觉。这一分类为下文对实践图景的观察提供了描述性的框架。

（二）针对幻觉开展的访谈调查

为全方位考察大模型在被公权力机关应用时所产生的幻觉情况，本次访谈调查的范围涵盖法院、检察院和行政机关。总体上，幻觉现象真实地存在于法律应用的场景中。

1. 面向法院：对民事案件辅助系统开展的调查

笔者首先考察的对象是已应用在部分法院的民事案件辅助系统，访谈的对象是开发此系统的企业工作人员。该系统为法官提供了从案件材料录入到判决文书生成的全流程辅助。其中，“判项梳理”模块涉及司法辅助的核心环节，包含事实审查、事实认定与判决认定三个步骤，因而访谈着重于这一模块。在受访者提供的劳动争议案例中，系统产生了典型的法条幻觉和事实幻觉。

第一，针对养老金损失赔偿问题，该系统虚构了法条。系统输出的内容是：“根据《最高人民法院关于审理劳动争议案件适用法律问题的解释（三）》第一条，用人单位未为劳动者办理社会保险手续且社会保险经办机构不能补办，导致劳动者无法享受社会保险待遇的，人民法院应当受理。”该输出存在以下两处错误：一是现行有效的司法解释的序号应该是（一）而非（三）。系统出现错误的原因是没有正确地区分正在生效的文件和已被废止的文件，即《最高人民法院关于审理劳动争议案件适用法律问题的解释（一）》与《最高人民法院关于审理劳动争议案件适用法律若干问题的解释（三）》。二是根据有效的司法解释，系统输出的法条内容遗漏或错写了“劳动者”“为由”“要求用人单位赔偿损失发生的纠纷”等内容。第二，该系统在对基础事实的概括中产生了幻觉。在事实认定部分，针对“原告提起诉讼的时间”，系统错误地将“2025年7月”输出为“2025年11月6日”，这种幻觉可能会影响法官对诉讼时效的计算。

2. 面向检察院：对刑事案件辅助系统开展的调查

笔者继而考察了已部署于部分检察院的刑事案件辅助系统，访谈的对象是开发此系统的企业工作人员。目前系统同样为检察官提供了从案件审查到文书生成的全流程辅助。其中，负责证据审查、审查认定和审查意见任务的系统模块为检察官提供案件事实和法律适用的参考。访谈着重于审查意见模块，观察大模型在量刑任务中的表现。在受访者提供的某盗窃案中，系统在量刑认定部分产生了法条幻觉。系统输出的内容是：“根据《中华人民共和国刑法》（以下简称《刑法》）第264条的规定，盗窃数额较大（2000元以上）的量刑起点为6个月有期徒刑。”此内容存在两处错误：第一，在盗窃数额部分，《刑法》第

264 条没有规定“数额较大”这一概念的具体金额。第二,在量刑起点部分,系统所列举的法条的真实内容也并未包括量刑起点的具体刑期。因此,法条错误使得公务人员难以采信经由模型推理得出的最终刑期。

3. 面向行政系统:对政务问答和行政审批系统开展的调查

在行政系统中,各种业务场景间具有较大差异。笔者根据《指引》将实践中普遍采用的智能问答和行政审批两类场景作为研究对象,分别访谈了某市数据局的相关负责同志和某科技公司的总经理。对于智能问答场景,某市数据局的同志表示,在最初将系统接入 DeepSeek 模型时,他们首要考虑的便是法条幻觉。为保证系统能够获得准确的源头信息,他们对政策库本身建立了严格的审核发布与动态巡检机制,公务人员通过管理甲数据库的信息录入、审核和发布流程,使得系统能够在 5 分钟内将新信息同步到乙数据库中。对于行政审批场景,某科技公司总经理认为法律援助智能审核系统存在产生涵摄幻觉的风险。审批的基本流程是:系统首先根据材料对申请人是否符合法定条件进行预审核并提前给出通过或不通过的结论,继而交由工作人员进行二次确认。系统需提前审核评估申请事项是否属于援助范围、申请人的经济状况是否达到标准等内容。这一过程本质上是将具体案件事实归入抽象法律规范之下的涵摄推理。然而,自动化预审系统主要是按照规则的文义进行审批,难以应对超出文义外的情形。例如,对于涉及申请人继承问题的场景,系统很难判断申请人是否属于“经济困难”,从而影响是否对申请人进行法律援助的判断。

从以上访谈来看,大模型在法律场景下会产生包括法条幻觉、事实幻觉与涵摄幻觉在内的多类典型幻觉。为系统地评估由幻觉带来的风险的普遍性与严重程度,下一部分将以某典型的法律推理任务为例对主流模型进行专项测试。

(三)专项测试

本部分将选取来自行政机关、检察院、法院的真实“醉酒危险驾驶”案例,用以测评中国现有的性能较优的 DeepSeek、千问、豆包三种模型。根据访谈,市级机关很少拥有针对专有业务场景来微调整个大模型的资源,主要通过提示词的形式将通用模型融入具体业务场景。由此,采用测评通用模型的方法能够一定程度上展现幻觉的基本分布。

1. 取样范围

本文将醉酒危险驾驶案件作为测试的典型案例。主要原因在于:第一,此类案件的数量一直高居刑事犯罪前列。公权力机关在分配智能化资源时,首先倾向于选择体量大的案件类型,例如重庆法院优先研发了醉驾案件的智能专审平台,威宁自治县检察院构建了基于大模型的醉驾案件全流程智能化辅助系统。第二,此类案件能够涉及行政机关、检察院、法院三个公权力机关的业务,因而能较为全面地展现不同机关应用大模型时产生幻觉的场景。

本文的测试案例均来源于北大法宝行政执法、检察文书以及司法案例数据库。北大法宝数据库相较于各地方的行政公开网站、12309 中国检察网以及中国裁判文书网,能够集成海量案例、提供标签标记并支持高级检索。因此,选择北大法宝数据库可以保证案例搜集的全面性和搜索的精准性。对案件的筛选标准是:第一,由于《最高人民法院、最高人民检察院、公安部、司法部关于办理醉酒危险驾驶刑事案件的意见》(以下简称《意见》)自 2023 年 12 月 28 日起施行,应当选择审理结束时间为此日期之后的案例。第二,案件需具有足够的案件细节和证据,因而在案件审理单位部分应选择基层机关。其余的检索条件是:(1)在行政执法数据库中,案件名称检索“行政处罚决定书/不予处罚决定书”,全文检索“醉酒”,主题分类选择为“公安”。(2)在检察文书数据库中,文书类型选择为“起诉书”以及“不起诉决定书”,全文检索“醉酒”,案由选择为“危险驾驶罪”。(3)在司法案例数据库中,案由设定为“危险驾驶罪”,全文检索“醉酒”,审理程序限定为“一审”,文书类型限定为“判决书”。在检索到的公安、检察院、法院的文书中分别随机抽取 5 个完整案例。

2. 实验设计

本实验的核心目标是观察大模型在分别接收“案件信息”和“案件信息与决策逻辑”的两种情况下，是否会在法律推理中产生幻觉，并分析这些幻觉的基本分布。设计两种输入条件的原因在于，我国是大陆法系国家，公权力机关的审查逻辑主要由成文法规定。因此，通过预先提供某一类型案件的法律决策逻辑，可以更好地测评模型能力。具体的测试方法为，只保留各类文书中当事人信息和案件事实部分作为案件信息，然后将“案件信息”和“案件信息与决策逻辑”分别上传至各类大模型，要求其做出判断并续写文书，开启深度思考功能，共测试90次。

对于决策逻辑而言，确定法律真实的逻辑往往属于“堆垛式”逻辑，即两个或者两个以上的三段式堆垛起来，每一三段式的结论为下一三段式的前提[3](P88-89)。公务人员的核心需求是判断某一决策是否具有法律上的依据。本实验首先需要根据相关法律归纳出公安机关、检察院、法院在醉酒危险驾驶案件上堆垛的三段论决策逻辑，划分出决策点。举例而言，“是否存在情节显著轻微情形”是一个决策点，这一决策点又根据《意见》第12条第一款规定的五项中再划分出“血液酒精含量不满150mg/100ml的”等五个决策点。决策逻辑不仅被应用在输入信息中，还被用来观察大模型的输出内容在这些决策点上是否存在幻觉。对于公安机关而言，需要决定是否刑事立案。对于检察院而言，需要决定是否起诉。对于法院而言，需要决定是否定罪以及量刑内容。

3. 实验结果

通过实际的测试发现，幻觉和争议会集中在1-2个决策点上产生。在此，笔者刻意区分了争议以及幻觉，原因在于，在自由裁量权极大的决策点（如犯罪情节轻微的判断）中，输出的内容并不能按照适用法律错误判断为幻觉，因为这一类输出本质上仍旧存在法律依据且法条、事实和涵摄均正确，因而这一类情况称之为“争议”。以下将对大模型在公安机关、检察机关和法院场景下的实验结果分而述之。

第一，对于公安机关的刑事立案决定，幻觉集中于“情节显著轻微”决策点的判断上，决策逻辑提示降低幻觉发生率。首先，40%的测试案例输出了刑事立案的错误结论，错误主要来源于法条幻觉（未搜索到《意见》第12条第一款）导致的“情节显著轻微”认定错误。正确的判断链条可以概括为：血液酒精含量不满150mg/100ml的，且不具有《意见》第十条从重规定情形的，可以认为情节显著轻微、危害不大，依照《刑法》第13条（不认为是犯罪），或者《中华人民共和国刑事诉讼法》第16条（不追究刑事责任）的规定处理。而测试案件中当事人血液酒精含量区间在44-142.65mg/100ml之间，按照此条款正确的答案应当是“不立案”。但大模型没有搜索到正确法条，因而错误地作出了“立案”的决定。其次，决策逻辑提示显著降低了结论的错误率。具体而言，在输入决策逻辑的测试案例中，93.3%的案例对于刑事立案的问题输出了正确结论。而相比之下，在没有决策逻辑的测试中，仅有26.7%的案例输出的结论正确。

第二，对于检察机关的起诉决定，争议集中于“犯罪情节轻微”决策点的判断上，相对不起诉比率较高。从结果上看，46.7%的测试案例输出了不起诉的结论，且其中超90%的判断是相对不起诉。对比抽取的5例真实案件的结果中只有1例为相对不起诉，可见大模型输出的不起诉比率较高。原因在于大模型选择了自由裁量权非常大的决策点——犯罪情节轻微（《意见》第13条）。在大模型的输出结果中，比较典型的是针对“同检刑诉[2025]68号案件”的不起诉原因：“本院综合考量：1.驾驶动机：无恶意驾驶意图；2.醉酒程度：血液酒精含量157.71mg/100ml，虽超出150mg/100ml，但无从重情节；3.机动车类型：普通小型轿车，社会危害性较低；4.道路及行驶情况：凌晨5时许行驶，车流量较小，未造成实际危害后果；5.认罪悔罪表现：到案后如实供述，自愿认罪认罚，悔罪态度明确。”然而在现实中，如果行为人已经满足了犯罪全部要件，那么检察机关按照此条得出相对不起诉结论的可能性较小，这体现了大模型的判断与公务人员的实际判断之间的显著差异。

第三，对于法院的审判决定，幻觉集中于“从重”和“缓刑”两个决策点的判断上。首先，40%的案件在“从重”这一决策点上的判断出现了幻觉，其次，20%的案件在“是否缓刑”的判断上出现了幻觉或者争

议。在从重判断中法条幻觉出现较多,主要原因是大模型将已废止的《关于办理醉酒驾驶机动车刑事案件适用法律若干问题的意见》文件中的“血液酒精含量达到200mg/100ml以上”作为从重的情节之一,而新法条《意见》中已经不再将酒精含量作为从重情节考察。对于缓刑这一决策点的判断有4例存在争议,较为典型的是大模型针对“(2025)内0802刑初237号案件”作出的缓刑判断,豆包根据《刑法》第72条缓刑条件中的“再犯风险”输出:“被告人虽有从宽情节,但具有多次违法犯罪前科,再犯罪风险较高,不符合没有再犯罪的危险的缓刑适用条件。”DeepSeek和千问还适用了《意见》第14条规定中“其他情节恶劣的情形”。典型的输出为:“被告人梁某某具有逃避公安机关依法检查的行为,虽未达到‘暴力’抗拒程度,但结合其多次违法犯罪的前科劣迹,综合评判属于‘其他情节恶劣的情形’,故不宜适用缓刑。”

(四)幻觉形成的技术成因

由以上实践可知,大模型在法律场景下生成幻觉的现象十分普遍。事实上,大语言模型的幻觉很难被彻底消除[4](P191)。从技术成因上看,生成模型的结构、训练数据的质量以及验证机制的缺失均会对幻觉的生成产生影响。

首先,概率驱动的生成架构存在缺陷。语言模型的定义为:描述任意字符串可能性的概率分布。通过语言模型,可以预测文本中接下来可能出现的单词,计算出对文本进行哪些更改概率更高,从而提供拼写或文法更正建议[5](P699)。其原因是,大语言模型是生成模型——即预测每个词后面的下一个词是什么——其任务并非检索和调用以往的准确信息[4](P186)。注意力分布的微小偏移可能在下游层引发连锁反应,因为每一层都建立在前一层的推理之上。这种累积效应最终可能产生结构化的错误信息。而对精确性和引用准确性要求极高的法律场景,正是这类错误输出的高风险领域[6](P13)。也就是说,目前生成模型的目标并不是“真实地呈现”,而是“流畅地呈现”,从而导致了幻觉的发生。

其次,训练数据存在认知局限。大模型需要在大量的文本数据中进行训练。然而,目前法律文书的质量参差不齐,并且常常缺少形成判断的完整决策逻辑。用于训练大模型的数据量巨大,其往往来自不同且未经验证的来源,这增加了包含相互矛盾或低质量文本的可能性[6](P13)。但对于法律而言,一个逻辑上或者术语上的使用错误很有可能会对推理的结果产生决定性的影响。在上述的专项测试中,大模型在“从重”这一决策点上频繁出错的原因可能在于大量的训练数据是根据旧法条(将200mg/100ml作为从重条件之一)得出的文书,而应用新法条(200mg/100ml已经不再作为从重条件)得出的文书在训练数据中占比较小。

最后,验证机制的缺失问题。大模型训练中的测试集旨在评估模型预测的准确性,它必须与用来训练模型的数据集分开。评估的核心是判断模型在测试集上所做的预测是否准确,以及这些预测能否支撑更优质的决策[7](P575)。对于保证大模型生成语言的流畅度而言,所有的训练数据都是正确的。但是法律场景并非仅要求流畅度,因而更关键的问题是谁来定义“正确的”法律推理。事实上,现有法律文书的主要结构是列举相关的法条然后直接给出结论,大部分文书并没有呈现出具体的决策逻辑。现实的法律文书若直接被选择为测试集,大模型显然易产生各类幻觉。

二、法律场景下规制幻觉的必要性

在公权力机关使用人工智能的背景下,幻觉作为一种新型的错误对法治产生了系统性的风险。讨论规制的必要性应当回答三种可能的质疑:第一,幻觉是一种低级的错误,就如同书记员错误一样,无需上升到理论层面对其规制。第二,人工智能处于辅助地位,因为有人在监督,所以错误也可以很快得到解决。第三,即使错误没有及时得到解决,也可以通过后续的问责机制来处理。以下将逐一进行回答。

(一)错误大小的转变:从个别到系统

公权力机关使用的人工智能如果产生幻觉,最大的危险在于错误的规模化,即从传统意义上的个别性风险转化为一种系统性风险。在完全依赖人工办案的阶段,错误源于个别公务人员在特定案件中产

生的偏差。由于公务人员处理案件的主要模式是逐案处理，影响范围比较有限。但是幻觉源于模型本身的内在缺陷，且模型的使用具有规模效应，影响范围显著扩大。

目前公权力机关使用人工智能的布局，呈现出高级别部门或者地区统一部署，下级部门或者地区接入或试验的模式。在法院系统，最高人民法院于2022年出台《司法应用意见》，于2024年在新闻发布会上发布“法信法律基座大模型”并指出希望推进其融入全国法院“一张网”。在检察系统，最高人民检察院在2025年启动了检察机关智能化设备建设以及试点工作。在政务系统，《指引》指出有条件的中央或省级部门可统一部署智能算力资源、人工智能大模型，面向下级单位提供人工智能大模型服务。因此，人工智能的使用已经呈现出从个别试点向全国推广的趋势。如果在未来全国业务系统均使用一种模型，那么幻觉产生的风险是巨大的。从现实的案例来看，国外存在由于系统的错误应用导致大规模诉讼的先例。比如，荷兰税务机关依赖自动化决策识别“福利欺诈现象”，但该算法将国籍作为一种错误的预测特征，很多受害者被要求追溯偿还大笔款项，导致一些受害者因经济压力而健康状况恶化、精神失常甚至家庭破碎[8](P162)。此类事件表明，大规模使用的算法一旦产生错误，可能导致严重的系统性风险。而幻觉是指人工智能提供的错误信息，属于一种系统上的缺陷，在未来同样可能产生规模性的影响。

（二）决策模式的转变：从人类主导到橡皮图章

公务人员在决策时倾向于接受算法输出的结果，这使得幻觉产生的风险进一步扩大。有学者提出了“准自动化”的概念，即人类作为一种基本的橡皮图章式机制被纳入一个完全自动化的决策系统[9](P104)。例如，在欧盟的边境安全检查中，警察使用算法分析乘客姓名记录和社交媒体数据以识别潜在的犯罪分子。虽然警察在形式上拥有决策权，但实际上由于时间紧迫（欧盟边境警察平均只有12秒来作出决定），警察往往依赖算法的输出而非独立地进行判断[9](P111)。

针对“人工智能仅为辅助，人类监督足以纠正幻觉”的质疑，目前的实验显示人可能无法对算法进行有效的监督。在人机协同的框架下，主要有自动化偏见、选择性遵从以及判断力萎缩三种心理机制。自动化偏见是指人机互动中形成的对自动化或算法的过度依赖[10](P164)。选择性遵从是指当人工智能的建议与决策主体预先存在的刻板印象相匹配时，决策者更倾向于选择性地采纳这些建议[8](P154)。判断力萎缩是指随着算法决策技术的应用，越来越多的价值判断被交给机器去计算和预测，却减少了人类对法律价值的理解和感知、对案情实质和其他相关因素的考量[11](P38)。基于此，采用人工审核的方案无法完全保证幻觉被纠正。

（三）问责模式的转变：从清晰问责到责任分散

责任的分配是保障公务人员尽力履职的重要制度约束。在完全依赖人工办案的阶段，公权力机关采用“权责统一”的清晰问责机制。然而，当人工智能不仅能够给予公务人员决策建议，同时还拥有部分监督权时，决策责任开始分散，导致问责制逐渐失去效力。

人工智能正在作为监督工具被推行，“向人工智能生成的结果靠拢”可能是公务人员更安全的工作方式。根据《关于充分运用智能化手段推进政法系统顽瘴痼疾常治长效的指导意见》，政法系统正积极运用人工智能完善执法司法内部监督机制。《司法应用意见》也提出了需加强人工智能辅助司法管理、支持案件裁判偏离度预警等意见。然而，让人工智能自动生成判决、根据大数据矫正法律决定的偏差等做法势必导致审判主体的复数化，造成“算法支配审判”的事态，使得法官无从负责，对法官办案的结果也很难进行切实有效的问责[12](P125-127)。另外，人工智能拥有部分监督权导致人在回路原则难以落实，因而单纯强调人在回路无法实质性解决责任的分散问题。人在回路的本质是需要人类在决策过程中不过度依赖算法，强调有意义的人类监督。据此，人类应当实质性地对决策结果负责，因为从理论上而言，只有负责决策的人同样会受到其决策的影响，才能提高决策者对决策结果的谨慎程度[11](P38)。但是目前面临的状况却是人工智能也可以反过来监督人类的活动，从而使得责任问题变得复杂。一项对于医

疗事故责任的研究显示,一旦系统达到足够高的性能,人工智能的结果将成为标准,偏离该标准的人类医生需要对由此造成的损害承担责任,这可能导致人类只是表象性地参与到回路中,而实质性地受到算法建议的限制,结果就是将人类逐步排除出回路[13](P458)。由此可见,原有的问责机制陷入了困境。

三、划分场景风险等级的基本原则

强调规制幻觉的必要性并不意味着所有产生幻觉的场景都需要进行规制,规制幻觉首先需要对场景的风险等级进行划分。根据本文的实践调查,很多场景并不会产生非常高的风险。对于行政机关、检察院、法院的法律推理而言,具有明确标准的、自由裁量权极小的判断部分,大模型的准确率非常高。即使存在幻觉,人也可以依据法条快速判断并直接予以纠正,因而风险并没有预想中的严重。在行政系统中,智能问答虽可能出现幻觉,但由于其显著的AI生成提示和准确的数据库使得风险较低。在理论上,已有学者在算法规制中提出了场景化原理这一重要的划分原则[14](P150)。由此,我们需要一种方案对不同幻觉场景产生的风险程度进行精细化划分,从而更有针对性地配置规制资源。

应当按照何种明确的原则来划分场景的风险等级?本部分将以与人工智能规制相关的风险理论为依据,提出通过综合权利影响的处分性、自由裁量权的大小两个维度判断场景所属的风险等级,以此作为下一章分级规制的基础。根据欧盟《人工智能法》第三条第2款,风险是指损害发生的可能性和损害严重程度的结合。这一域外的理论划分对处理我国人工智能的风险问题有其借鉴意义。衡量风险有两个维度,一是损害的严重性,二是损害实际发生的可能性。以下将分别论述在公权力机关应用大模型的场景下,如何具体化这两个维度以明晰划分风险等级的标准。

(一)以权利影响的处分性衡量损害严重性

在现有的人工智能规制体系中,权利影响被认为是评估损害严重性的关键因素,这有其合理之处,但此理论仍需依据中国的情况进一步改造。根据欧盟《人工智能法》第6条,是否对自然人基本权利造成显著损害风险是高风险场景的判断原则之一。该法第27条要求所有高风险系统都需进行基本权利影响评估。按照欧盟《人工智能法》附件三的内容,公权力机关应用的人工智能主要归属于高风险类型。典型的使用场景包括生物识别、关键基础设施、基本公共服务和福利、执法、移民、司法等。已有学者指出,公权力运用决策型算法,会直接对个人的基本权利造成影响。例如,以"秒批"为代表的自动行政算法极大地提升了行政效率,但也在一定程度上侵犯了公民的人身自由和知情权[15](P92)。

然而,权利影响这一概念过于模糊,需用处分性这一标准加以具体化。有欧盟学者指出,权利风险评估的关键就在于对严重性的评估,但当前评估模型往往依赖于强度、严重程度、幅度等模糊概念,这些表述本质上是对严重性概念的同义重复,缺乏实质性的量化标准[16](P1)。在政务领域,有学者提出依据行政行为分类标准对场景的风险进行分类,他根据行为是否产生法律效果将其分为行政事实行为和行政法律行为,这二者对相对人权利义务的影响程度区别较大,前者属于低风险场景,后者属于中风险场景或者高风险场景[17](P102-103)。这一标准可以扩展概括为权利影响的处分性标准。对于处分性的判断,取决于公权力行为能否直接对公民的权利义务产生法律上的效果。行政处分,并不是指行政机关所实施的全部行为,而是指带有权力性并且能够使相对人的权利义务发生具体变化的法律行为[18](P114)。其作为核心标准的优越性在于:第一,处分性与权利影响高度相关,对于行政机关,直接行政行为自然对权利的影响程度极高,对于法院、检察院,判决结果、起诉决定等均是对公民权利的直接处分,因而这些决策有着极高的风险。第二,分类标准具有清晰性,是可操作性强的界定标准。

基于此,根据权利影响的处分性,可以将场景划分为高权利影响和低权利影响两个类别。高权利影响主要是指直接影响公民权利义务的法律场景,典型的场景有法院系统的审判行为、检察系统的起诉行为以及行政系统的给付、执法等行为。低权利影响主要是指不直接影响公民权利义务的法律场景,典型的场景有智能问答、电子卷宗自动分类等场景。

（二）以自由裁量权的大小衡量损害发生可能性

损害发生的可能性维度应当通过自由裁量权标准来衡量。缘由在于幻觉对公民权利产生损害的可能性取决于其被纠正的可能性，幻觉的重要表征是迷惑性，如果在大模型输出幻觉的情况下，人类无法穿透模型给出的表面答案并及时纠正，那么损害发生的可能性就非常高。在法学研究中，自由裁量权的理论与被纠正的可能性高度相关，原因在于自由裁量权的大小代表着在某些场景下是否有正确答案。自由裁量权是公权力的决策主体在面对两个及以上合法的方案中进行选择的权力[19](P16)，这代表着在这些场景中没有唯一正确的答案，客观上只存在公权力主体采纳的最终答案。自由裁量权广泛存在于公权力机关的决策过程中，不仅司法机关享有，行政机关甚至立法机关也享有[20](P17)。在自由裁量权很小的场景中，如果人类可以直接识别幻觉并予以纠正，那么幻觉就无法对法律场景造成风险。然而，如果在自由裁量权较大的场景中出现幻觉，结合前文提及的自动化偏见、选择性遵从以及判断力萎缩三种心理机制，幻觉产生风险的可能性就极大。

衡量自由裁量权大小，并非简单地将一个存在自由裁量权的场景全部划为高风险，而应当按照具体案件类型的决策点进行精细化划分。欧盟《人工智能法》是将司法机关使用人工智能的整体领域“一刀切”地划分为高风险类别的典型，其附件三第8(a)项规定：拟由司法机关或其代表使用，以协助司法机关研究和解释事实和法律，并将法律适用于一组具体事实，或以类似方式用于替代性争议解决的人工智能系统属于高风险人工智能系统。然而，一旦深入到具体案件的处理之中，会发现这种观点错误地限缩了人工智能的使用范围。在公权力机关适用法律的很多场景中，只有一个正确答案。以醉酒危险驾驶案件的决策逻辑为例，审查犯罪嫌疑人的血液酒精含量是否大于或等于80mg/100ml，五年内是否曾因饮酒后驾驶机动车被查获或者受过行政处罚、是否在高速公路驾驶等判断均属于自由裁量权极小的场景。而典型的自由裁量权较大的决策点就是对犯罪情节是否轻微的判断，需要综合考虑犯罪嫌疑人驾驶的动机和目的、醉酒程度、机动车类型、道路情况、行驶时间、速度、距离以及认罪悔罪表现等因素，决定是否相对不起诉或免予刑事处罚。欧盟将整体场景判断为高风险的错误在于，其忽视了得出法律结论需要多个推理链条，而在某些推理链条中存在“唯一正确”的决策点，因而欧盟不适当地限缩了人工智能的使用场景。

基于此，本文根据自由裁量权大小将场景划分为两类。第一类是高自由裁量权场景，指公权力机关在决策中有多种合法的选择，无法通过简单的规则套用完成。典型场景包括刑事案件中的量刑、检察机关的相对不起诉决定等。由于幻觉能够因其表面的合理性影响决策过程且难以被校验，因而损害发生的可能性较高。第二类是低自由裁量权场景，指公权力机关在决策中只有一种正确答案，可以通过简单的规则套用直接输出，几乎没有个人判断的空间。典型场景包括法院立案庭的批量化诉讼案件等。由于大模型输出的内容可以被校验和纠正，损害发生的可能性较低。

（三）场景风险的三级划分格局

对公权力机关使用人工智能的场景进行划分，可以通过评估权利影响的处分性衡量损害严重性，通过评估自由裁量权的大小衡量损害实际发生的可能性，最终通过两个维度的组合形成“高风险、中风险、低风险”的场景划分格局（见表1）。

将一个场景认定为高风险，需要使其同时满足高权利影响和高自由裁量权两个条件，即场景具有对公民权利的处分性强且裁量空间大的基本特点。在上文的实际调查中，典型场景有公权力机关“审查犯罪情节是否轻微”“决定拘留时间和罚款额度”等。更进一步需要追问的是：这一类场景是由于何种深层次的原因导致幻觉现象不可以被容忍？原因在于这类场景非常依靠人类决策的亲历性知识，人工智能产生的决策偏差会对实质正义造成潜在的威胁。人类可以捕捉超出文本形式的多种信息，在亲历性方面更具优势。就理论方面而言，亲历性原则是自由心证的基础。在法庭的辩论程序中，法官可以通过提问和观察的方法，感知诉讼参与人肢体、表情、动作所包含的隐形语言，通过多姿多彩的现场材料获取更

表 1　公权力机关应用场景的风险等级划分

权利影响 / 自由裁量权	高权利影响	低权利影响
高自由裁量权	第一象限:高风险场景 特征:处分性强且裁量空间大 典型示例:审查犯罪情节是否轻微/拘留时间和罚款额度	第二象限:中风险场景 特征:裁量空间大但处分性弱 典型示例:经济运行大模型/城市信用状况洞察
低自由裁量权	第三象限:中风险场景 特征:处分性强但裁量空间小 典型示例:血液酒精含量的标准/责任年龄的判断	第四象限:低风险场景 特征:处分性弱且裁量空间小 典型示例:智能问答/电子卷宗自动分类

多的案件信息,从而构建心证[21](P81)。就实践方面而言,亲历性审查能够起到纠正系统偏差的重要作用。在一桩羁押案件中,当浙江金华检察院开发的社会危险性审查模型系统基于同案犯未归案、未退赃等客观数据,将黄某的社会危险性评估为较高等级时,检察官通过亲历性审查,发现了“其妻子系高位妊娠待产,继续羁押可能导致黄某妻子无人照料”“犯罪事实供述稳定,且有主动联系家属筹措赔款的意愿”等更为全面的信息,最终变更强制措施,黄某随后全额退赃[22]。这一案件展现了检察官的直接经验如何纠正人工智能产生的错误。因此,对此类场景进行规制时,必须坚决捍卫人的裁量主体地位。

中风险场景有两个类别,一是满足高权利影响、低自由裁量权条件的“高影响—低裁量场景”,比如行政秒批、审核犯罪嫌疑人是否满足刑事责任年龄等人工智能应用场景。二是满足低权利影响、高自由裁量权条件的“低影响—高裁量场景”,主要是用于辅助整体决策(非直接针对特定个体)的各类模型,例如厦门市数据管理局在2025年度人工智能应用场景机会清单中提及的经济运行大模型、AI+城市信用状况洞察等。这些均是对非个人的情况进行整体评估的场景。这两类场景的共同特点是决策结果均在人类的可控范围内,因而同属中风险类别。“高影响—低裁量场景”的风险主要来源于幻觉导致的准确性问题。比如大模型错误地将犯罪嫌疑人的血液酒精含量超过200mg/100ml作为醉酒危险驾驶罪的从重条件之一,对于普通的民众而言,可能难以识别这一幻觉,但是对于拥有丰富办案经验的公务人员而言,可以较为容易地识别幻觉并对其进行纠正。“低影响—高裁量场景”产生的风险主要来源于决策所依据的信息存在幻觉这一问题,这种错误是否可以被纠正往往取决于决策程序本身的设计。只要遵循原有对决策的讨论程序,人工智能幻觉就不太可能对决策结果产生重要影响。因此,两类幻觉均处于人类的可控范围之内。

将一个场景认定为低风险,需要使其同时满足低权利影响和低自由裁量权的条件,即场景具有对公民权利的处分性弱且裁量空间小的基本特点。在目前的实践中,主要包括公权力机关对外和对内的两种业务场景。对外的场景主要是指公权力机关开放给公众的公共服务类系统,比如行政机关、检察院、法院的线上智能问答系统等。对内的场景主要是指在公权力机关系统内用于提升管理效能的内部管理类活动,比如电子卷宗自动分类、法律文书智能校对等。这两类场景的共同特点是对公民的权利义务几乎没有影响,因而我们无需过分担心。但是在这些场景中的幻觉仍有可能造成公共舆情以及内部程序低效的风险,如果大模型在对外业务场景中产生幻觉,公权力机关就要面临人民群众的质疑。如果大模型在对内业务场景中产生幻觉,机关的内部效率就会受到影响,从而背离提高效率的最初目标。

四、基于风险分级的幻觉规制路径

为确保幻觉不会产生系统性的影响,在借鉴现存的算法规制方案的基础上,本文将依据风险等级与来源针对性地设计“前端预防—中端监督—后端问责”的三阶段规制路径(见表2),旨在将风险防控的原

则具体细化为技术性要求和制度化配置，实现分类分级的精细化治理。

表2 基于风险分级的幻觉规制路径

风险 / 规制等级 / 阶段	高风险	中风险		低风险	
		裁量权大	裁量权小	公共服务类	内部管理类
前端预防	思辨性设计	参与程序	准确性测试+人工审核	准确性测试+转接要求	内部反映程序
中端监督	监督机关、行业内最高机关	行业内省级机关		同级网信部门	本单位党委（党组）
后端问责	监督机关、行业内最高机关、部署者、提供者	行业内省级机关、部署者、提供者		同级网信部门、部署者、提供者	本单位党委（党组）、部署者、提供者

（一）高风险场景中对幻觉的规制路径

在高风险场景中，规制幻觉的核心逻辑是坚决捍卫人在决策中的裁量主体地位，严格将大模型限定在辅助性角色中，以此构建贯穿前端、中端、后端的全链条强干预机制。

在前端预防层面，重点是构建具有“思辨性”的算法设计缺陷防范机制，将人工智能的功能严格限定于“去偏差”的程序性辅助，禁止人工智能输出任何形式的结论性判断。实证研究表明，强迫法官在作出判决前写下他们的推理是一种有效的“去偏差程序”[23]（P83），因此可以将人工智能设计成帮助人类纠正偏见的系统，从而促进公务人员思考。具体而言，大模型在高风险场景中的正当性在于其作为“批判性思维工具”的定位，实现这一点的核心是给公务人员提供正反两面的信息，帮助他们深入思考并重新审视自己的结论。需要明确的是，思辨性设计是在做出法律判断过程中的程序性标准，目前最常见的算法设计缺陷防范机制是在算法设计阶段明确要求其包含必要的安全措施[24]（P170）。主管机关应在《互联网信息服务算法推荐管理规定》要求的基础上，对法律行业的高风险场景增加思辨性设计这一强制审查要求。

在中端监督层面，算法接口应当开放给外部监督单位和行业内的最高机关。在传统的法律监督体系中，除了本系统内部的监督外，检察院可以根据《中华人民共和国人民检察院组织法》第20条监督法院的诉讼活动。公民、法人或者其他组织可以根据《中华人民共和国行政诉讼法》第二条通过向法院提起行政诉讼的方式监督行政机关职权的行使。对于高风险场景应当按照原本的监督格局对模型本身进行监督。涉及重要法益或公共安全的算法应用应当为监管者提供监管上的入口或其他便利条件。这种方法在网络安全监测、区块链信息服务等方面均有相关的立法[24]（P174）。监督机制必须能够审查算法在实际运行中是否恪守了“辅助思考”而非“输出答案”的角色定位，一旦发现算法可能干扰公务人员的独立性判断，就可以按照传统的监督方式提起法律监督。

在后端问责层面，如果在高风险场景中大模型幻觉产生了严重后果，就需要启动问责程序。问责的前提是需要审查算法是否超出了“辅助”这一定位，如果算法被设计为直接对高风险场景进行结论性输出且这一结论直接或间接作为公务人员的考核标准，那么就需要启动后续问责程序。归责需要追溯到算法设计和监督链条上的相关主体，部署者、提供者应承担主要责任，若负有监督职责的主体未能及时识别和纠正算法已超越辅助定位这一问题，也应承担相应的监管责任。另外，对人工智能辅助下决策错误案件与一般决策错误案件的问责机制有所不同，原因在于对错误结果产生实质性影响的主体不同。在一般决策错误的案件中，决策结果由公权力主体独立作出，归责主体清晰。而当对人工智能辅助决策下的错误案件问责时，最终决策结果是否受到人工智能错误的影响取决于系统本身的设计。如果模型本身也承担纠正公务人员错误的任务，那么系统本身的部署者和提供者也需要根据具体情况承担责任。

（二）中风险场景中对幻觉的规制路径

在中风险场景中，规制幻觉的核心逻辑是通过对决策进行程序性审核来确保大模型生成信息的准

确性,这需要构建贯穿前端、中端、后端的全链条,以及中等强度的干预机制。

在前端预防层面,第一,幻觉导致的准确性问题是“高影响—低裁量场景”的主要风险来源,因而在设计时应当遵循两种途径,一是采用真实的测试集对算法进行准确性测试,二是必须设置人工审核的程序。借鉴《规范》的测试思路,应当核查是否制定清晰、具体的测试标准和指南,对测试结果的分布和一致性进行分析,选择具有相关领域知识和经验的测试人员,确保测试结果的准确性。由于低裁量场景中的待判断问题本身就具有清晰的答案,因而准确性测试是可行的规制方案。同时,由于此场景对人的权利有十分重要的影响,在公权力机关使用系统作出具体的决策行为前,应当将“人作为审核主体”作为强制性要求。第二,幻觉导致的决策信息错误这一问题是“低影响—高裁量场景”的主要风险来源,本质上可以通过将原始数据公开给系统内有权参与决策的主体来解决。在真实的决策场景中,大模型有可能偏离原始数据,错误地报告问题,那么一旦将原始数据予以公开,并设计各类主体的参与程序,就相当于增强了其他主体的质疑能力,从而保证了决策的程序正义。

在中端监督层面,算法接口应当向行业内省级机关开放。相较于高风险场景,中风险场景的算法接口无需开放给外部监督机构和行业内最高机关,只需开放给行业内省级机关。原因在于,第一,省级机关具备足够的专业能力处理中风险场景。由于中风险场景要么具有准确规则,要么是提供非直接处分性的信息支持,省级机关具备制定本省实施细则的专业能力,可以对算法是否符合业务要求作出准确判断。第二,省级机关直接管辖本省范围内的下属单位,可以防止属地管理的权责失衡。“属地管理”权责失衡的困局直接表现为基层治理中权力和责任的不对等,上级职能部门将大量工作,特别是风险性大的、棘手的工作借“治理重心下移”之名转嫁给基层[25](P44)。将此监督责任赋予省级机关,其可以统筹本省范围内的相关资源,符合在算法监督场景下权责匹配的要求。

在后端问责层面,省级机关和部署者、提供者均需承担相应的责任。在此须进一步对部署者和提供者之间的责任分配予以明确。在实践中,部署者和提供者的关系是民法中的合同关系,需要按照双方约定的义务承担责任。按照民法缔约过程所遵循的动态体系论方法,部署者需在采购环节说明业务场景和验收要求,而提供者作为技术的设计者,需要承担技术性说明的义务以使部署者合理预见或控制风险[26](P38-39)。省级机关应当实质性地审核部署者和提供者是否履行了双方的义务,如果未进行实质性审核,那么省级机关同样需要承担责任。

(三)低风险场景中对幻觉的规制路径

低风险场景包含公共服务类和内部管理类场景。这两类场景的共同特点是对公民权利义务几乎没有影响。然而,在这些场景中的幻觉仍旧可能造成公共舆情以及内部程序低效的风险,需要设计与风险相匹配的幻觉控制机制。

在前端预防层面,第一,公共服务类场景的主要风险来源于幻觉对群众的误导问题,在模型使用量大的情况下易引发公共舆情。在设计算法时需要采用两种方案,一是对各类问题进行准确性测试。公共法律问答场景往往会出现法条幻觉,可以通过使用引入外部知识库的检索增强生成(RAG)技术来保证人工智能生成内容的准确性[27](P71)。二是保证便携的人工接管。借鉴《规范》的规定,当输出准确率下降到设定的阈值时(比如不能低于95%)支持快速切换至人工服务。第二,内部管理类场景的主要风险来源于幻觉导致的内部程序低效问题,错误的输出可能导致内部信息传输延迟、资源浪费等问题。在前端设计时应当配置通畅的内部反映程序,缘由在于一线业务人员对系统效果的感知最为敏锐,通过将意见反馈机制集成到技术流程中可以保障信息高效传递从而实现对问题的快速整改。

在中端监督层面,公共服务类场景的监督应当归口于同级的网信部门,内部管理类场景的监督归口于本单位的党委(党组)。公共服务类场景的主要风险来源是舆情,而网信部门的主要职责之一是组织、协调网上宣传工作。因此,网信部门在职权上和技术能力上适合监督这类场景。将内部管理类场景的监督归口于本单位的党委(党组)可以更有效地为基层减负。《整治形式主义为基层减负若干规定》要求

规范政务移动互联网应用程序管理等形式主义，党委（党组）切实履行主体责任。其中的原因在于，公权力机关的内部事务具有经济学意义上的外部性，为建成系统需要公权力机关投入较多资源。内部的效率低下不仅影响本单位的效率，同样影响整个业务链条的运行。党委（党组）有能力也有职权对内部管理类场景进行监督。

在后端问责层面，在公共服务类场景中，若因算法输出错误引发公众投诉或舆情关注，同级网信部门作为监督归口单位，应负责启动调查，其问责焦点在于部署单位是否履行了“准确性测试”与“人工转接”等前端义务。责任形式主要为行政性督促，如责令限期整改、公开说明情况等，核心目的在于恢复公众信任。在内部管理类场景中，本单位的党委（党组）作为监督的核心主体，应依据党内法规与内部管理规定，对其因算法应用失当影响运行效率的情形进行处置。对于负有责任的党的领导干部，可以参照《中国共产党问责条例》等相关规定，对其采取阶梯化的处理方式。可根据危害程度及具体情况，决定是否对其问责，或采用通报、诫勉等程度较轻的责任方案。

在法律场景下公权力机关应用大模型时会出现幻觉现象，但并不意味着所有产生幻觉的场景都需要同等强度的规制，因而规制的边界划定尤为重要。相较于欧盟“一刀切”地将司法机关在具体案件中使用人工智能的场景列入高风险范围，本文认为在幻觉问题的处理上，通过衡量权利影响的处分性和自由裁量权的大小来判断场景的风险等级更具合理性。这一方案的优势在于最大化地利用了人工智能对问题进行高效判断的能力，又将自由裁量权较大、高权利影响的判断完全交由人类审慎地裁断，可避免产生系统歧视、橡皮图章、责任分散等问题。未来，随着人工智能与法律场景之间的深度融合，法学理论界和实务界应当深入研究的核心问题是如何将人类决策和机器决策的优势各取所长。唯有如此才能有效提升智能化治理的能力，创造更高水平的数字正义。

（论文调研写作期间，华南理工大学法学院曾庆醒老师和何骁同学给予了关键启发，接受访谈的政府部门和公司在调研中也给予了大力支持）

参考文献

[1] Ziwei Ji, Nayeon Lee, Rita Frieske et al. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 2023, 55(12).

[2] 侯学勇.从法律规范的可反驳性到法律知识的不确定性——法律论证中融贯论的必要性.内蒙古社会科学(汉文版)，2008，(1).

[3] 曹士兵.裁判的形成：作为知识、方法、价值的法律科学与应用.北京：法律出版社，2024.

[4] 刘庄，卢圣华.机器能取代法官吗——人工智能、数据科学与法律.北京：北京大学出版社，2025.

[5] 斯图尔特·罗素，彼得·诺维格.人工智能：现代方法.张博雅、陈坤、田超等译.北京：人民邮电出版社，2023.

[6] Youssef Abdel Latif. Hallucinations in Large Language Models and Their Influence on Legal Reasoning: Examining the Risks of AI-Generated Factual Inaccuracies in Judicial Processes. *Journal of Computational Intelligence, Machine Reasoning, and Decision-Making*, 2025, 10(2).

[7] Brandon L. Garrett, Cynthia Rudin. The Right to a Glass Box: Rethinking the Use of Artificial Intelligence in Criminal Justice. *Cornell Law Review*, 2024, 109(3).

[8] Saar Alon-Barkat, Madalina Busuioc. Human-AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice. *Journal of Public Administration Research and Theory*, 2023, 33(1).

[9] Ben Wagner. Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-making Systems. *Policy and Internet,* 2019, 11(1).

[10] 汪庆华.算法透明的多重维度和算法问责.比较法研究，2020，(6).

[11] 陈悦.论自动化行政中算法决策风险的“人在回路”治理模式.行政法学研究，2024，(4).

[12] 季卫东.AI时代的法制变迁.上海：上海三联书店，2020.

[13] Rebecca Crootof, Margot E. Kaminski, W. Nicholson Price II. Humans in the Loop. *Vanderbilt Law Review*, 2023, 76(2).
[14] 丁晓东.论算法的法律规制.中国社会科学,2020,(12).
[15] 林洹民.论人工智能立法的基本路径.中国法学,2024,(5).
[16] Gianclaudio Malgieri, Cristiana Santos. Assessing the (Severity of) Impacts on Fundamental Rights. *Computer Law & Security Review*, 2025, (56).
[17] 张硕.政务人工智能的幻觉应对:基于技术与场景耦合风险分级的梯度化准入规制.电子政务,2026,(2).
[18] 石龙潭.日本行政诉讼救济范围之拓展——"行政处分性"之理论解析.行政法学研究,2017,(3).
[19] 梁迎修.法官自由裁量权.北京:中国法制出版社,2005.
[20] 江必新.论司法自由裁量权.法律适用,2006,(11).
[21] 陈卫东.直接言词原则:以审判为中心的逻辑展开与实现路径.法学论坛,2022,(6).
[22] 史隽,婺检.AI助力,开启人机协同"新引擎"——浙江金华婺城:智能辅助系统让社会危险性审查评估效率提升60%.检察日报,2025-09-10.
[23] Zhuang Liu. Does Reason Writing Reduce Decision Bias? Experimental Evidence from Judges in China. *The Journal of Legal Studies*, 2018, 47(1).
[24] 苏宇.算法规制的谱系.中国法学,2020,(3).
[25] 刘帮成."属地管理"权责失衡的根源与破解之道.人民论坛,2021,(26).
[26] 冯玉军,沈鸿艺.XAI背景下司法人工智能的可解释性义务研究——基于司法真诚的理论视角.北京航空航天大学学报(社会科学版),2025,(4).
[27] 赵静,汤文玉,霍钰等.大模型检索增强生成(RAG)技术浅析.中国信息化,2024,(10).

Categorized Regulation of Large-scale Model Hallucinations In Legal Scenarios

Focusing on the Risks of Large-scale Model Application in Public Authorities

Feng Yujun, Shen Hongyi (Renmin University of China)

Abstract With the continuous development of artificial intelligence in China, large-scale AI model hallucinations have become increasingly prevalent in legal scenarios. Field interviews and investigations reveal that the practical deployment of large-scale models by public authorities has given rise to various typical hallucination categories, namely, statutory provision hallucination, factual hallucination, and subsumption hallucination. Targeted tests on case adjudication involving drunk driving across public security authorities, procuratorates and people's courts demonstrate that such hallucinations tend to converge at merely one or two pivotal decision-making nodes. Given the changes in severity of errors, decision-making patterns, and accountability models, categorical regulation of legal-domain large-scale model hallucinations is urgently warranted, The regulatory approach, starting from the basic principle of scenario-based risk classification, should establish a "high, medium, and low-risk" classification pattern by combining the two dimensions of dispositive impact on legal rights and discretion scope. Targeted at respective risk levels and risk sources, a three-stage regulatory framework of "front-end prevention — in-process supervision —post-hoc accountability" can be formulated to deliver refined, tiered governance by category.

Key words hallucinations in large-scale AI models; public authorities; artificial intelligence applications; risk grading; regulating hallucinations; discretion

■ 作者简介 冯玉军,中国人民大学法学院教授,北京 100872;
沈鸿艺,中国人民大学法学院博士研究生。

■ 责任编辑 李 媛