一版责任编辑:蓝韵飞 刘凌云 黄新 二版责任编辑:高驰弘 联系电话:0571-85310479 邮箱:zjrb@8531.cn

AI闯祸,由谁买单,如何治理

——“AI幻觉”调查

调查与思考

■ 本报记者 张蓉 杨渐
通讯员 魏澜 赵媛媛

问题

- AI赋能千行百业之际,“AI幻觉”也在入侵人们的日常生活,带来真实的法律与伦理挑战,成为社会治理的隐形风险源。

调查

- AI一本正经地胡说八道,这是“AI幻觉”的典型表现。“AI幻觉”不仅会导致个人信息被“张冠李戴”,还会在专业领域对用户造成误导,引发信任危机,污染信息环境,带来更多潜在的社会风险。
- “AI幻觉”本质上是统计概率驱动的“合理猜测”,主要成因包括推理逻辑的泛化困境、数据偏差以及人机互动的刻板误伤。
- 在当前的大语言模型技术架构下,“AI幻觉”无法完全消失。因此,平台和用户都要承担注意义务和标识义务,既为它系上安全带,也贴上警示标签。

思考

- 数据优化是缓解“AI幻觉”的源头举措,应探索建立具有权威性的公共数据平台,提升训练数据质量,建立良好的数据生态。
- “以技治技”是解决“AI幻觉”问题的首选路径,应采取多种防治措施,加强验证与核查,建立更加多维、可靠的安全墙。
- 针对“AI幻觉”,更有意义的治理创新在于建立技术价值观。AI平台在开发中就应回应多方利益诉求,将负责任创新、可控创造性等伦理价值融入工程师的头脑、植入大模型的代码。

你是否有这样的经历:遇到问题向AI求助,它迅速给出看似专业而合理的解答,可一经推敲,其中的事实、数据、结论,要么信息错位,要么无中生有。

AI一本正经地“胡说八道”,这是“AI幻觉”的典型表现。

在生成式人工智能技术快速发展和普及之际,“AI幻觉”频频产生,包括捏造案例、虚构数字、在图像生成中出现不合理元素等,不仅使个人信息被“张冠李戴”,还在医疗、金融、教育等专业领域严重误导大众认知,带来真实的法律与伦理挑战,成为社会治理的隐形风险源。

面对悄然入侵日常生活的“AI幻觉”,如何应对?AI闯祸,由谁买单,如何治理?

人与AI对簿公堂

检索信息、回答问题、辅助办公、生成音视频,甚至陪伴聊天,观察当代人类生活日常,AI越来越成为不可或缺的部分。

据《第57次中国互联网络发展状况统计报告》,截至2025年12月底,我国已有748款生成式人工智能服务完成备案,用户规模达到6.02亿人,占全国人口的42.8%。这意味着,每三个人中就有一人使用过AI。

AI赋能千行百业之际,也埋下了一个“暗雷”——AI幻觉。此前,清华大学新闻与传播学院新媒体研究中心和人工智能学院就市场上多个热门大模型进行事实性幻觉评测,发现幻觉率超过19%。

前不久,杭州就宣判了两起涉“AI幻觉”案。

第一起,是全国首例“AI幻觉”引发的侵权案。作为一名高考生的哥哥,小梁在网上查询一所高校的报考信息时,AI不仅生成了错误内容,还信誓旦旦地表示内容有误将赔偿10万元。小梁便将该AI的研发公司告上了法庭。

杭州互联网法院在审理中认为,AI的“承诺”不构成平台的意思表达,且开发者已采用可行的技术手段力求降低错误,并提醒告知了风险,判决驳回诉讼请求。

另一起,则是因传播“AI幻觉”内容引发的不正当竞争。自媒体从业者小李发布了一篇由AI生成的文章,将无关公司包装成知名企业的重要子公司。该企业认为文章损害其商业信誉与竞争利益,将小李诉至法院。

杭州市滨江区法院认定,被告以获取商业利益为目的发布文章,明知内容源于AI、可能失实,却没尽到审核与显著标识的义务,判决被告发布声明以消除影响,并赔偿3万元。

两个案例相继落槌,既折射出“AI幻觉”的蔓延趋势,也让这一新生烦恼及其引发的现实危害、带来的法律挑战等,进一步走入公众视野。

随着“AI幻觉”的危害从虚拟空间逐渐延伸到现实生活,人与AI对簿公堂的现象,也在国内外其他地方接连上演。

在北京,律师黄贵耕通过某AI检索时发现,自己被“移花接木”、塑造成“威胁法官、伪造印章、伪造担保函”的违法形象。愤怒中,黄贵耕以侵犯名誉权为由,将AI开发公司诉上法庭。

德国的“AI幻觉”第一案也在进行中。这是一起针对AI聊天机器人Grok的诉讼,原告Campacte.V.协会指控Grok在回答用户提问时,输出了损伤自己公信力的不实答案。

有人被AI“移花接木”个人信息,成了违法分子;有人借助AI问诊,结果却延误了疾病治疗;有人依据AI建议投资理财,造成了亏损;还有人将AI生成的虚假案例援引进起诉书……类似情况屡见不鲜。

对此,中国计量大学法学院教授、副院长冀瑜认为,“AI幻觉”不仅会成为名誉“杀手”,还在医疗、法律、金融等专业领域对用户造成误导,引发信任危机,带来更多潜在的社会风险。

从长远来看,若不对“AI幻觉”进行有效治理,它还会成为虚假和错误信息的温床。“大量充斥着‘AI幻觉’的内容涌入互联网,如虚构的历史等,可能加剧虚假信息传播的雪球效应,使后代产生认知偏差、刻板印象,甚至被篡改记忆,并污染下一代模型的训练数据。”首例“AI幻觉”案承办法官、杭州互联网法院跨境贸易法庭庭长肖芃说。

争议纠纷不断,指向一个潜藏的共性问题:AI“闯祸”后,该由谁负责?

肖芃认为,AI生成的不准确信息是否构成侵权不能一概而论,重点要审视的是AI的开发者、使用者是否存在过错。而想厘清过错,必须先弄清楚——“AI幻觉”从何而来。

AI为何会“胡说八道”

“AI幻觉”,有时像上演“历史穿越剧”,有时会创造出“学术鬼打墙”,还可能摆出“法律迷魂阵”……它们往往细节精致,但底层逻辑却破碎得如万花筒中的斑驳光影。

看似“聪明绝顶”的AI,为何会“胡说八道”?

调查中,多位受访专家指出,“AI幻觉”本质上是统计概率驱动的“合理猜测”。出现幻觉的根源,在于AI的思维方式与人类存在本质不同。

“目前,生成式人工智能服务往往依托大语言模型。”传播大脑副总经理兼首席技术官张健长期从事大模型开发,他将大语言模型比喻成一只“能说话但不理解语言”的鹦鹉,它预设的思考方式像一场“高级文字接龙”,通过对文本、图片等输入信息进行编码和计算来预测下一个词元,从而具备文本生成和对话能力。

正因过度依赖参数化记忆,模型难以处理训练集外的复杂场景。面对用户五花八门的需求,AI就像新手走迷宫,出现推理逻辑失灵等问题。

黄贵耕的遭遇就很典型。详细核查后,黄贵耕发现,“威胁法官”和“伪造印章”的情节出自新闻报道,但AI将匿名处理的“律师黄某”“法律顾问黄某”篡改成他;“伪造担保函”的内容出自最高检发布的案例,AI又将当事人“黄某平”替换成他的名字。

“就因同为律师,同样姓黄,AI就挪用了信息,进行‘移花接木’。”黄贵耕叹息,按照这种逻辑推理,每个人都可能成为AI的受害者。

引发“AI幻觉”的另一重因素,则是数据偏差。在“养成”AI的过程中,“投喂”海量数据是关键环节。AI“吸收”的原始数据不足、存在错误或本身带有偏见,都会让其生成错误的观点和事实,陷入信息误区。

“把AI比作一个学生,数据污染就像给学生看了错误的教科书,自然会导致‘胡说八道’。”暨南大学网络空间安全学院教授翁健认为。他透露,目前,AI的训练数据大多来自开源网络,而互联网上充斥着大量未经验证的垃圾信源,大模型难免会“学习”一些虚假数据。

同时,还有不法分子恶意进行“数据投毒”。

前不久,央视“3·15”晚会就曝光了“AI投毒”的黑灰领域——生成式引擎优化(GEO)行业,通过在网上系统地、定向地投放大量虚假信息“喂”给大模型,进而操控大模型输出答案,让某个品牌、产品或信息源获得流量和影响力。

此外,“AI幻觉”成因中还有一个极易被忽视的因素——人机互动的刻板误伤。用户在提问过程中,也为AI输入了信息,无形中干预着它的回答。

张健解释,从底层规则来看,大模型被强制要求必须有回应,而训练过程的缺陷导致AI存在讨好用户的倾向。“如果预设的问题是错的,AI不会认为是用户出错,而是生成迎合用户需求的内容,并编造虚假的例证或看似科学的术语来支撑自己的假说。”

厘清“AI幻觉”的三重成因后,再回看杭州宣判的两个案例,判决结果虽看似不同,实则为AI的服务提供、内容创作与传播,划定了统一的责任边界——平台和用户都要为“AI幻觉”承担标识义务和注意义务,既要贴上警示标签,也要系上安全带。

安全带该怎么系,也不能一刀切。肖芃认为,对于法律禁止的有毒、有害、违法信息等,平台要严防死守;对于技术能防的错,尽力去防。同时,用户也要审慎查验,尤其是对涉及他人、其他主体的事实性信息,要在能力范围内核实清楚。

人机共处,怎样控制风险

基于当前的技术条件,专家普遍认为,“AI幻觉”或许难以消灭。它像是对智能社会的反讽与警醒,迫使我们回答一个问题——人类如何应对信息生成速度远超验证能力的时代。

当算法可以流畅地编织谎言,破解之道或许不在于追求完美的技术,而是构建起一套与之匹配的数据生态、防治体系与伦理底线,形成风险可控的人机协同机制。

其一,让AI吸收健康有效的数据,建立良好的数据生态。

数据优化是缓解“AI幻觉”的源头。目前,AI背后的语料库以企业自建自用为主,数据来源各有偏好,内容良莠不齐。

多位受访专家建议,在保护数据隐私和信息安全的前提下,应建立跨机构的数据协同机制,实现机构信息、政府公开数据、学术科研数据等数据共享,探索建立权威性的公共数据平台,减少单一信源的失真问题。同时,AI开发主体须提升训练数据质量,对训练数据进行清洗和验证,确保训练数据的准确性和可靠性。

在浙江,杭州国家语料库正加快建设步伐,其目的就是为大模型训练提供易获取、高质量、规模化、低成本的语料资源。放眼全球,国内外机构也已发布多款具备行业广泛影响力的AI语料库,覆盖通用、医疗、政务、方言、代码等多个核心赛道。

其二,加强验证与核查,建立更加多维、可靠的安全墙。

“以技治技”是解决“AI幻觉”问题的首选路径。调查中,记者发现,多家主流人工智能厂商已采取多种防治措施,试图降低幻觉。其中,使用检索增强(RAG)技术成为不少厂商的选择之一,“它相当于外置的知识库,能从大量外部数据库中检索相关信息,并结合这些信息进行生成,提高内容的准确性。”张健介绍。

华东政法大学刑事法学院副教授房慧颖长期研究数据法治。她建议,平台应构建精细化的治理体系,对生活资讯、娱乐消息等普通信息,适当简化审核流程;但对涉及违法犯罪等可能影响公民声誉的敏感信息,必须使用权威数据库,或进行多重交叉信源验证。

在监管层面,肖芃建议,需在AI平台上线前,进行有针对性的对抗性训练等检测,并设定相关行业标准,细化对其训练数据、技术措施等要求。

其三,提高技术开发应用的价值基准。

针对“AI幻觉”,更有意义、更可持续的治理创新在于建立正确的技术价值观。

在省社科院法学所所长王鉴岫看来,算法权已成为一种新型社会权,因此AI平台在开发过程中就须回应多方利益诉求,将负责任创新、可控性等伦理价值融入工程师的头脑、植入大模型的代码。

例如,倡导存在争议结论不生成、无法溯源的信息不生成、超出模型认知边界的内容不生成等原则,推动大模型从追求生成流畅度向确保内容可靠性转型;或建立分级置信提示制度,按照高可信、须核实、推测性结论等,进行分类标注,加强输出内容的透明度。

这注定是一个AI蓬勃生长的时代。“AI幻觉”的治理之路,需要技术的迭代和探索,也是人与AI如何共处、怎样协同的一场修行。

正如多位受访专家所说,驾驭AI的关键,不是让AI“无所不能”,而是让人类更有判断力。唯有持续强化技术研发、健全规范体系、完善校验机制,方能遏制幻觉风险,推动科技向善。

树立和践行正确政绩观学习教育省委督导组培训会议召开

本报杭州4月13日讯(记者 张熙锦)13日下午,我省召开树立和践行正确政绩观学习教育省委督导组培训会议。会议深入学习贯彻习近平总书记重要讲话和重要指示批示精神,认真落实党中央和省委部署要求,对学习教育指导督导工作进行动员培训。

会议指出,省委决定,对12个地方和单位派出省委督导组,目的就是针对性地督促指导相关地方和单位扎扎实实开展学习教育,推动解决好部分地方和单位存在的政绩观深层次问题,确保学习教育各项任务到底到边。相关地方和单位要全力支持配合督导,以此为契机,切实解决矛盾问题。

会议强调,省委督导组要深入理解把握中央和省委精神要求,统一思想认识、明确任务要求,确保督导工作取得实实在在的成效。要吃透精神、把准方向,高站位扛起使命担当。深刻领会树立和践行正确政绩观的时代要求,锚定“走前列、作示范”标准,在深学、真查、实改上下功夫见成效;要对标对表、真督实导,高质量推动任务落实。督促持续深化学习研讨、查准查实突出问题、真刀真枪抓好问题整改、破立并举推进建章立制、办实事促发展解难题;要找准定位、优化方法,高水平加强指导督导,做到督帮一体、同题共答,切实提升督导工作质效,做到以点促面、抓县促市,推动解决共性问题,做到分类指导、靶向施策,注重结合实际情况和特点找准切入点发力点。

全国无党派人士考察团来浙调研

本报杭州4月13日讯(记者 沈吟)4月13日,全国无党派人士考察团围绕“完善有利于全国统一大市场建设的体制机制”,赴浙江开展年度重点考察调研并举行座谈交流。

考察团成员表示,浙江是中国革命红船起航地、改革开放先行地、习近平新时代中国特色社会主义思想重要萌发地,见证了党的创新理论的探索发展和国家改革开放的重大成就,在推动经济社会高质量发展方面历来走在全国前列,在融入全国统一大市场方面也积累了丰富经验。考察团将在全面总结提炼浙江服务全国统一大市场建设好经验好做法基础上,力争提出高质量、有价值的政策建议,为建设强大国内市场、加快构建新发展格局贡献智慧力量。

浙江为建设全国统一大市场积极开展创新探索,2025年全省公平竞争指数提升至86.45,在服务全国统一大市场建设上取得显著成效,树立标杆示范。省领导表示,浙江将以此次考察为契机,深入贯彻落实好全国统一大市场建设“五统一、一开放”要求,更好肩负起经济大省挑大梁的责任担当,为推动高质量发展建设共同富裕示范区取得决定性进展、率先呈现基本实现社会主义现代化生动图景增添强劲动能。

此次考察团由中央统战部组织,全国人大常委会委员、中国科学院院士郭雷等十余位知名专家学者参加。

“国家安全长城行”主题活动举行

本报临海4月13日电(记者 钱祎 金晨)在第11个全民国家安全教育日即将到来之际,13日,由省委国家安全委员会办公室、省国家安全厅联合主办,临海市委、市政府承办的“登江南长城 护国家安全——‘国家安全长城行’”主题活动,在临海“台州府城墙”(江南古长城)揽胜门广场、兴善门广场及沿线展开。

活动以“统筹发展和安全、护航‘十五五’新征程”为主线,通过历史与现实呼应、线上与线下联动、体验与教育融合的形式,深入宣传贯彻总体国家安全观,集中展示新时代国家安全事业发展的新气象,呈现浙江在维护和塑造国家安全的生动实践和工作成效。

本次主题活动以“国家安全长城行”主题展览为主线,同步开展“国家安全长城守护队”授旗仪式、“长城上的国家安全教育课”实地研学、“戚继光抗倭”沉浸式互动体验、“我心中的国家安全”作品征集展示,以及“云上长城万里行”线上推广等系列活动。通过巧妙融合实体展陈、情景演绎与数字技术,将国家安全的宏大叙事融入中华民族的历史现场,让总体国家安全观可感、可知、可及,有力提升全民国家安全意识和素养,为在新征程上筑牢国家安全屏障凝聚浙江力量。

“国家安全长城行”主题展在宁波、金华等地同步举办。

创新赢市场

4月13日,永康市经济开发区浙江千禧龙纤特种纤维股份有限公司车间内,自动化生产线高速运转,工人正加紧赶制来自全国各地的订单。作为高性能纤维领域的深耕者,该公司产品已广泛应用于军工、安全防护、航空航天等高端领域,并逐步拓展至海洋渔业、体育用品、医疗材料等民生应用场景。

本报记者 林云龙 通讯员 谭孝军 摄

浙江今年将建100个“电梯益家”联系点

电梯有了专属管家

本报讯(记者 李洁薇 通讯员 市闻)老旧小区加装电梯后如何管护;高龄电梯出了故障,如何更快得到维修;人被困在电梯里,如何更快救援……这些问题正逐步得到化解。日前,全省“无忧乘梯安民行动”工作交流暨“电梯益家”现场观摩活动在杭州市拱墅区举行。这项已连续第三年列入省市场监管局十大民生实事的行动,取得一系列让群众乘梯更安心的新进展。

活动现场,省市场监管局部署了2026年度“无忧乘梯安民行动”各项任务。到2026年底,全省将建成“电梯益家”服务联系点100处,实现县域全覆盖,确保老百姓“就近就便、快速响应”。

去年9月,拱墅区建成了全国首个“电梯益家”服务联系点,将电梯服务直接延伸到社区。政策咨询、应急救援、技术指导等,居民在家门口就能找到“管电梯的人”。“以前电梯出了问题,要先找物业,物业再找维保,来回折腾。现在联系点就在小区旁边,一个电话,人就到了。”拱墅区一名居民告诉记者。

同时部署的其他民生实事还包括:一年内完成新版96333电梯使用标志更换;新增电梯物联感知3万台以上;推动电梯扫码检验检测率、扫码维保率保持在95%以上;开展电梯科普宣教活动1000场以上。

省市场监管局有关负责人表示,下一步将会同省建设厅等有关部门,大力推广“电梯益家”服务经验,协同推进电梯更新、加装、改造等重点工作,推动企业规范维保、基层安心管梯、群众放心乘梯。