责任编辑 / 徐荔 E-mail:fzbfzsy@126.com

近年来，随着人工智能技术的不断进步与发展，ChatGPT、豆包等生成式AI大语言模型越来越多地运用于日常生活中。不可否认的是，一系列生成式AI产品的出现，使得我们的生活更加便利，但与此同时，AI“一本正经地胡说八道”、生成虚假有害信息的现象也屡见不鲜。由此产生的问题是，当AI“说瞎话”造成实际损害结果后，如何确认相应的责任承担主体。这不仅关乎公众权益的保护，同时也是生成式AI产业发展中必须解决的问题。

AI一本正经胡说八道，谁该负责？

聚焦虚假信息生成后的责任归属

□ 薛昊 华东政法大学

AI“编造事实”的情况并非个案

2023年，美国有位律师在处理航空诉讼案件时，让ChatGPT帮忙检索类似判例。AI收到指令后一口气列出了六个“判例”，其中案号、法院、裁判等要素都齐全，看似言之凿凿。该律师并未核实就将其写入法律文书并提交法庭。

然而，事后查明，这六个“判例”都是AI凭空编造的虚假案例。最终，该律师被法院罚款5000美元。

与之相似的是在2026年5月，江苏一名消费者用AI软件预约某家餐厅的座位，AI给他生成了一个“预约成功”的凭证，但是事实上，那家餐厅并不支持线上预约，因此导致消费者白跑一趟。

此类现象在人工智能技术领域被称为“AI幻觉”，具体是指：模型在缺乏事实依据的情况下，编出了语言流畅、逻辑通顺，但同现实对不上的内容。这主要是因为，AI本质上是依托大数据进行概率预测的统计模型，并不是个明辨是非、独立思考的个体。易言之，它无法“知道”自己所说的是真是假。

现如今，生成式AI越来越深地扎进我们的日常生活与商业活动里，因为“AI幻觉”而引发的现实纠纷也越来越多。

因此，厘清这类虚假信息造成危害后果后，法律责任究竟由谁承担，具有紧迫的现实意义。

事实上，面对生成式AI可能带来的诸多风险，我国在法律与制度层面，也已作出积极回应。

2023年，国家互联网信息办公室出台了《生成式人工智能服务管理办法》（征求意见稿），明确要求服务提供者采取有效措施提升训练数据质量，防止生成虚假、有害信息，并规定“提供者的行为构成违反治安管理行为的，依法给予治安管理处罚；构成犯罪的，依法追究刑事责任”。

这也意味着，AI生成虚假信息绝非无人问津的法律真空地带，而是有明确的规范指引和责任追究路径。只是在具体个案中，责任究竟该由何者承担、具体规则如何认定等问题，仍需要进一步厘清。

责任不能简单归于“AI本身”

面对AI生成虚假信息引发的损害，一种常见却存在误区的看法是：“这是AI自己犯的错，让AI负责”。但这一看法有待商榷。因为想要AI“背锅”，承担责任的前提是需判断其是否符合承担责任的基本要件。

从刑法理论上看，要承担刑事责任，主体需要具备辨认与控制能力，即行为主体能够明白自己行为的性质与后果，并能自主的控制行为。但是生成式AI的运行特点决定了它不具备这一能力：其作出的所有决定，都是算法在海量数据中运算出的结果，并不包含类似人类的主观想法与自由意志。

此外，AI没有独立的财产，无法对其采用监禁、罚金等传统刑罚手段，因此即便将其认定为责任主体，也无法实现刑罚的目的。

因此，将责任全推给“AI本身”，实质上会导致责任承担的实际落空。真正需要讨论的，是AI背后的开发者、服务提供者与用户等自然人主体。

责任划分的关键：谁在哪个环节掌握实际控制权

将责任归于开发者、服务提供者或用户中的某一方并“一刀切”处理，同样有失公允。生成式AI从模型开发设计到最终生成结论，往往需经数据收集、算法运算、结论输出等多个环节，不同主体在不同环节对系统的实际支配程度存在明显差异。责任认定应立足于各主体在具体环节对风险的控制能力予以区分。

首先，在开发设计阶段，开发者负有数据质量审查义务。其需要通过海量数据对模型进行训练。若开发者在数据收集、处理环节存在明显疏漏，导致大量虚假、过时信息混入训练数据库，进而致使AI生成失实内容，此环节的责任理应指向开发者。

其次，在算法运算阶段，开发者负有防范系统性风险的义务。生成式AI的算法运算过程存在所谓“算法黑箱”特征，即便开发者自己，往往也难以逆向还原模型得出某一具体结论的完整运算路径。但这并不意味着开发者可以此为由免责。

开发者虽然无法预见AI在某次具体交互中会产生哪一具体错误，但基于行业专业认知，对“算法运算存在生成虚假信息的系统性风险”这一事实是完全可以预见的。若开发者未采取与行业平均水准相当的测试、纠错及防护措施，就可认定其未尽到相应的审慎义务，需对由此产生的虚假信息承担相应责任。

面对“算法黑箱”引发的AI输出不实信息，有一个典型案例：2023年，谷歌人工智能产品Bard发布会中，产品研发者在向大众公开演示天文科普问答时，该AI产品输出了大量关于韦伯太空望远镜的虚假科研成果。

事后调查发现，开发者在训练数据时的审核环节中存在疏漏，导致其他望远镜的科研成果被错误地“张冠李戴”混入其中。

这一案例也印证了前述观点：算法运算过程或许难以完全还原，但只要开发者在测试、审核等环节存在明显的疏忽懈怠，未达到行业应有的谨慎标准，就理应对由此生成的虚假信息所引发的危害结果承担责任。

最后，在结论生成阶段，这一阶段的责任因主体行为性质不同而有所区分。

若用户是被动接受AI生成的失实内容，如前述餐厅预约案例，用户并未诱导系统作出虚假回应，此时责任应由开发者与服务提供者承担。

反之，若用户明知系统可能生成不实内容，仍主动诱导系统生成谣言、诽谤等违法信息并加以利用，则应根据其行为的主观故意与客观危害，依法承担相应的法律责任。

开发者、服务提供者若已尽到合理的安全审核义务，则不应因用户的恶意规避行为而被株连。

普通用户在日常生活中使用AI时应注意哪些内容

在厘清各主体在生成式AI的不同运行阶段中的责任归属的同时，作为AI产品的使用者，也应当树立如下的正确风险防范意识。

其一，当向AI提问有关法律规范、医疗建议、财务数据等专业性、严谨性较强的事项时，不宜直接相信并采纳AI生成的内容，而是需要经过必要的专业人员的核实，确认无误后再进行使用。目前的技术条件下，“AI幻觉”仍然是行业公认难以彻底消除的固有局限，即便是技术最为先进的AI研发团队，也无法保证其开发的AI产品所生成的内容百分之百准确无误。

第二，切勿认为借助AI生成、传播虚假有害信息就能规避法律责任。部分用户存在侥幸心理，觉得“反正这是AI说的话，和我没关系”，于是采取隐晦提问、规避安全审核等方式，故意诱导AI生成谣言、诽谤等违法信息并加以进行传播。这种认识存在明显误区：法律责任的认定，关键在于评价行为人的主观方面以及客观行为所引发的具体损害后果，而非表面上“危害信息是由谁说出来的”。只要用户主观上存在诱导AI生成违法信息的故意意图，客观上造成了危害后果，就应当依法承担相应的法律责任，不会因为“经AI之口”而免责。

结语

生成式AI的不断发展，正在深刻改变社会生产与生活方式，在不久的将来，或许会有更为先进、全能的大语言模型诞生。但技术本身的局限性决定了生成虚假信息这一风险短期内难以彻底消除。厘清开发者、服务提供者与用户在不同环节应承担的法律责任，既是防范此类风险的现实需要，也是推动AI产业规范、健康发展的必然要求。唯有责任边界清晰、各方义务明确，才能真正实现技术创新与风险防控的平衡。