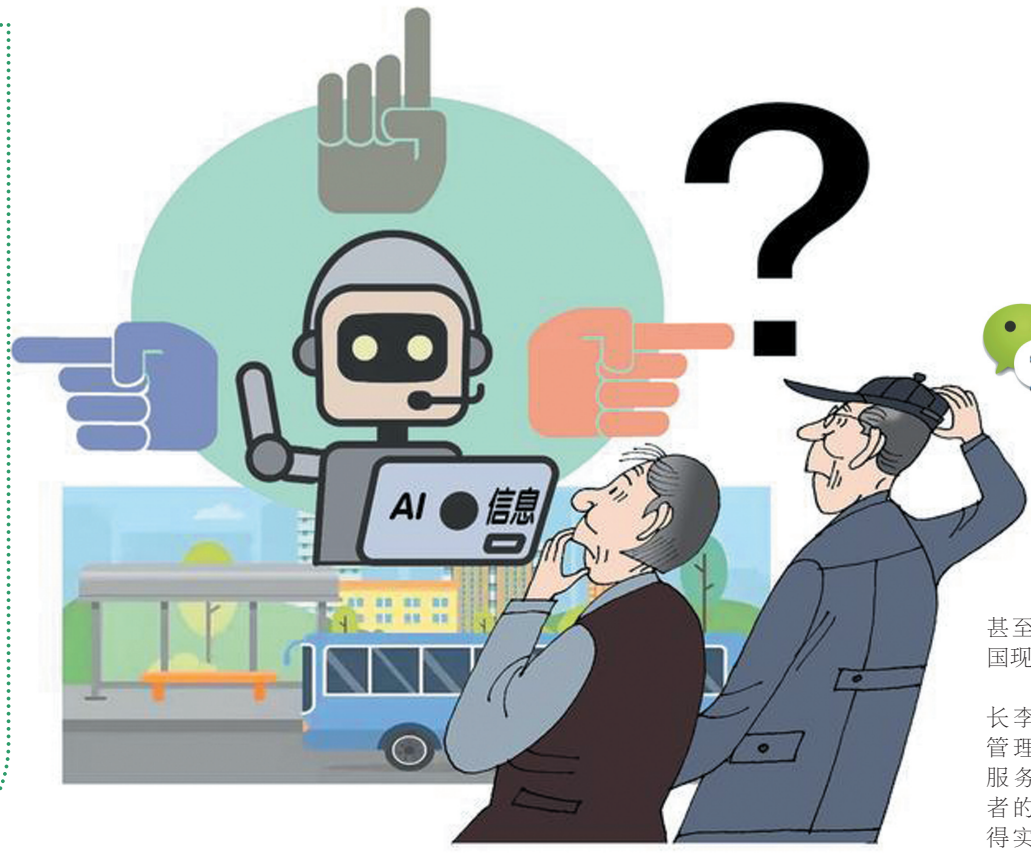
AI为啥会一本正经地胡说八道？

随着生成式人工智能深度融入我们的生产生活，“有问题问AI”已经成为不少人的习惯，但AI带给我们便利的同时，也伴随着一些问题，比如有人就发现，AI提供的内容并不那么准确，有时甚至会出现严重的错误。当AI一本正经地“胡说八道”，就是AI“幻觉”的典型表现。

北京的律师黄贵耕就被AI“幻觉”搞得挺闹心，他无端被AI捏造了多项不实罪名，职业名誉也因此受损。这到底是怎么回事？



现象1 律师被捏造罪名，竟是“AI幻觉”惹祸

黄贵耕是北京的一名律师，去年5月，他经朋友介绍，成为一起刑事案件的代理律师。被告人家属为了了解他的情况，就在百度平台检索相关信息，不料平台AI竟自动生成了多条关于黄贵耕律师的虚假负面内容。

“比如说搜黄贵耕律师办理的案件，百度跳出一大堆，关于黄贵耕涉嫌威胁法官、私刻印章、介绍贿赂等。当事人家属搜到之后，就感到很震惊。这个律师怎么存在这么多污点。”黄贵耕告诉记者，这些不实内容均与他本人无关。起初，他还以为是有人在恶意诋毁他。“我当时都没有想到是百度AI自动生成的，我还怀疑是同行，因为当时我的当事人还请了另外一个律师。”

后来，被告人家属告知，相关内容是搜索时百度AI自动生成的。黄贵耕核查发现，所谓恐吓法官被罚款3万元一事，实际是某媒体报道的发生在某地的新闻中，“律师黄某”因判决结果与预期不符，对承办法官进行侮辱、恐吓，黄某被法院罚款3万元。百度AI则将新闻中的“律师黄某”，直接替换成了“黄贵耕”。

而“伪造公司印章”，实际是媒体报道过的另一起案件：“法律顾问黄某”伪造公司印章试图追讨1000余万元债务。百度AI把“法律顾问黄某”直接替换成了“黄贵耕”。

黄贵耕表示，AI会生成这么恶劣的信息，这突破了普通人的想象，难以相信这是假的。如果这种行为得不到遏制的话，很多人都会成为它的潜在受害对象。

于是，黄贵耕律师决定起诉北京百度网讯科技有限公司，案件正在审理中。目前，百度搜索已不再显示AI生成的关于“黄贵耕律师犯罪”等虚假信息。

业内2 AI一本正经地“胡说八道”，有多种原因

类似的事件还有不少，近日，一位网友吐槽，自己的手机不断接到陌生电话，对方开口就问是否卖猪。起初他以为是恶作剧，后来才发现，这些电话竟然来自豆包AI的搜索结果，他的号码被AI错误标注为某养殖场的联系方式。AI为何会产生这样的“幻觉”？这种“幻觉”能否避免呢？

某人工智能公司首席AI专家王占一表示，AI的幻觉现在主要指的是大模型的幻觉，大模型在它的生成、回答的过程中，看似生成了很多合理的、很流畅、很自然、语义语法、也没问题的回答，但是它在事实性、真实性上面是有问题的，可能比较自信，一本正经在说什么，但是其实是不对的。

当AI一本正经地“胡说八道”时，便是“AI幻觉”的典型表现。由于AI给出的回答读着通顺、语气笃定，用户往往很难判断其真伪。那么，AI为什么会产生这种幻觉呢？

王占一认为：第一个方面是在数据层面，训练的数据有可能质量差、数据缺失或时效性不够，以及这个问题可能模型没有学过。第二个也是最关键的，模型本身的实现机制。由于其本身是一种概率模型，它依赖于前面的上文去推理下一个词，这个过程中可能会产生一些“幻觉”，原因是模型回答时更关注流畅性，而未关注事实性。第三个，在推理和处理上下文时，若上下文过长会对模型的能力提出挑战，用户的输入和指令可能前后矛盾，模型可能无法判断选择哪个，从而导致识别错误。

分析3 AI出现“幻觉”导致侵权事件发生，谁担责？

AI“幻觉”导致虚假信息频发，有时候甚至会侵犯名誉权等，面对这类问题，我国现行法律法规有哪些明确规定？

中国政法大学网络法学研究所所长李怀胜表示，《生成式人工智能服务管理暂行办法》规定，生成式人工智能服务提供者要承担网络信息内容生产者的一个法律地位，而且生成的内容不得实施或者用来侵害他人人格权、名誉权、个人信息权益等相关权益。如果相关的内容造成他人损害，就按照民法典的一般侵权的过错责任原则来进行相关的审查。

专家表示，为规范人工智能发展，我国已初步形成以《中华人民共和国民法典》等法律为基础、《生成式人工智能服务管理暂行办法》为核心规范，并辅以《人工智能生成合成内容标识办法》等多层次制度框架的AI生成内容法律规制体系，在一定程度上能够为规制AI“幻觉”提供基本依据。

当前技术条件下，AI生成内容难以做到百分之百准确，是由于技术发展的不完善造成的，但这并不意味着AI服务提供者可以免责，作为平台，需承担以下法律义务。

中国社会科学院法学研究所研究员支振锋表示，平台对于法律禁止的“有毒”、有害、违法信息负有严格审查义务。AI服务提供者应当依法承担网络信息内容生产者责任，履行网络信息安全义务。对于涉及国家安全、社会公共利益以及他人合法权益保护的违法信息，平台不得以“系AI生成”为由免除审查责任。

专家表示，平台要以显著方式向用户提示AI生成内容可能不准确，对可能生成的有害信息提供事前预防和事后救济途径。

支振锋指出，在事前预防层面，平台应尽可能做到功能可靠性的基本的注意义务，采取同行业通行的技术措施提高生成内容的准确性。事后救济层面，如果平台发现或接到投诉知悉存在一些AI生成的不实信息内容后，应当及时检验生成内容，并在确认侵权后第一时间采取删除内容、禁止传播等必要措施，防止损害扩大，平台是可以有作为的，也是有可能会承担相应的法律责任的。

据半月谈