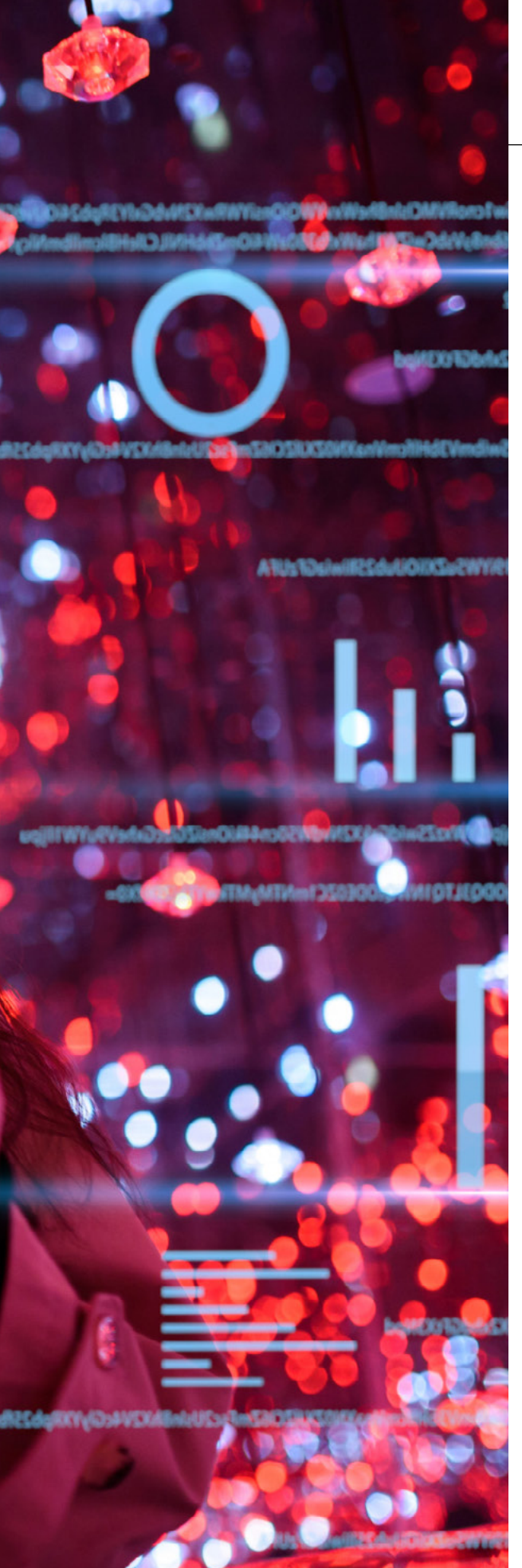
提要：打造负责任的生成式AI，防范合规、数据、用户以及价值风险，需要在生成式AI基础模型的设计之初就未雨绸缪，并在其全生命周期中持续领航匡正的系统性工程。



防范生成式AI四大风险

文 陈泽奇

当今，ChatGPT的诞生激发了前所未有的创造浪潮，也让公众直观地感受到生成式AI的力量。但缺少规范的技术如同一个尚未辨知善恶的新生儿，在带给人类极大帮助的同时，也潜藏着新的风险。

新加坡《海峡时报》就曾报道，某科技企业工程师在借助ChatGPT处理工作时，无意中将产品核心数据输入其中，导致商业机密外泄。越来越多的ChatGPT用户也发现，在与AI对话时会收获一些细节丰满的故事，甚至言之凿凿的文献参考，但实际上这些并非真实信息，而是生成式AI的“模型幻想”（model hallucination）。

生成式AI风险何在

对企业而言，生成式AI将改变工作方式、重塑商业模式。越来越多的企业已开始积极探索相关应用，以期提升创新效率、实现高质增长，但生成式AI的风险同样需要引起重视。

首当其冲的就是合规风险，它贯穿于模型设计、搭建、使用各个阶段，并会产生长远的效应。比如生成式AI基于学习需要而对用户数据的留存、分析是否侵犯了个人和商业隐私以及相关数据保护法案；它在借鉴创意作品（如画作）的过程中，是否侵犯了作者版权和著作权；由此产出的作品又是否可用于商业用途；甚至使用生成式AI本身，是否违反了部分国家和地区的法律法规。

同样可能在生成式AI的生命周期中出现的还有数据风险。生成式AI的运作核心是机器学习，其价值与数据的质量和真实性密切相关。如果一台基础模型长期浸染在存有偏差的数据当中，它就会被这些数据“诱导”，从而输出错误的信息或执行歧视性操作。某些群体特质也会使生成式AI为其打上固化标签，“一刀切”地去执行某些程序，从而失去了应有的公平性。

此外，用户风险也是需要重点关注的一环。事在人为，生成式AI的价值高低，很大程度上取决于使用它的人。我们需要通过一系列的法律法规、流程规范来防止人类有意、无意地使用生成式AI造成负面影响。

价值风险也是企业和用户应该考虑的要点，使用生成式AI是否有违社会、企业和个人的价值文化？例如，《麻省理工学院技术评论》就曾指出，训练一台普通AI模型所消耗的能源，相当于5辆汽车全生命周期排放的碳总量。为此，在部署AI战略的时候，我们必须思考相关碳排放是否会减缓乃至抵消企业的零碳进程，继而有所取舍。

负责任的AI，促进业务增长

令人庆幸的是，大多数的政府部门和企业已经意识到让生成式AI“更有责任感”的重要性，并开始采取积极行动。埃森哲最新全球调研显示，97%的受访高管认为自身企业将受到AI相关监管法案的影响，77%将对于AI的监管列为优先事项。此外，有80%的受访者表示，他们将投入10%或更多的AI总预算，以满足未来的监管要求；69%的受访高管表示，其所在企业已经开始尝试负责任的AI实践，但并未将其作为运营基础。

值得注意的是，不少受访者认为，合规的AI将为其提供额外的竞争优势——43%的高管认为这将提高他们将AI产业化和规模化的能力，36%认为它将为竞争优势和差异化创造机会，41%认为它可以帮助吸引和留住人才。

与之相应，在2022年埃森哲发布的《AI成熟之道：从实践到实效》报告中，我们还发现，有13%的受访企业成功使用AI技术实现了超过50%的收入增长，同时在客户体验（CX）和环境、社会与企业治理（ESG）方面表现出色。相较于普通企业，这些“AI领军者”更加注重从设计企业各个环节的时候就把负责任的AI列为优先事项。

新加坡是金融科技的沃土，不少机构都已开始将AI技术用于信用卡审批、保险理赔等金融服务领域，在大幅提升工作效率的同时，显著降低运营成本。作为监管方，新加坡金融管理局看到了其中隐含的风险——如果放任AI自我学习，可能造成潜在的群体歧视。

为此，新加坡组建了一个由25家机构组成的行业联盟，并在2018年发布了以公平、伦理、问责、透明为核心的“FEAT原则”，为金融机构采用负责任AI提供指导。作为开发小组成员，埃森哲为其打造了用于评估模型公平性的工具包，以保证FEAT原则的执行。

目前，埃森哲与新加坡金融管理局仍保持着密切协作，在整个行业深化推进负责任AI的落地与优化，为金融机构提供建议，并致力于培养一批拥有相关知识和经验的专业人员，鼓励更多科技企业创建符合FEAT原则的AI解决方案。

设计负责任的生成式AI

虽然多数企业已经认识到培育负责任AI的价值，并正致力于此，但只有6%的企业建立了负责任的AI基础原则，并付诸实践。过于传统的组织架构、风险管理框架、生态伙伴、AI人才和考核标准，都是限制企业成功的主要因素。

埃森哲建议，企业可以从原则与治理，风险、政策与管控，技术与支持，文化与培训四个层面入手，通过设计让生成式AI变得更负责任。

原则与治理：在最高管理层的支持下，定义并阐明负责任AI的使命和原则，同时在整个组织中建立清晰的治理结构，以建立对AI技术的信心和信任。

风险、政策与管控：加强对既定原则和现行法律法规的遵守，同时监测未来的法律法规，制定降低风险的政策，并通过定期报告和监控的风险管理框架实施这些政策。

技术与支持：开发工具和技术来支持公平性、可解释性、稳健性、问责制和隐私等原则，并将其构建到AI系统和平台中。

文化与培训：推动领导层将负责任AI提升为一项关键的业务，并为所有员工提供培训，让他们清楚地了解负责任AI原则以及如何将这些原则转化为行动。

在生成式AI的发展道路上，价值与风险并存，关键在于设计、搭建、使用它的企业和个人如何作为。打造负责任的生成式AI，则可以让创新科技的成果合规、安全、平等地惠及每一家企业、每一个人。

陈泽奇
埃森哲大中华区董事总经理、首席数据科学家

业务垂询：accenture.direct.apc@accenture.com