Check for updates

NIST 可信与负责任的人工智能
NIST AI 600-1

人工智能风险管理框架：生成式人工情报概况

本出版物可从以下网址免费获取：
https://doi.org/10.6028/NIST.AI.600-1

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

本译本由机器翻译生成，供快速理解条文与内部培训使用；
涉及合规判断或法律适用时，请咨询专业顾问，并以官方英文文本为准。
翻译您自己的文件，请选择桐鸾译馆 · https://www.ivyivy.com.cn
扫码关注服务号

NIST 可信与负责任的人工智能
NIST AI 600-1

人工智能风险管理框架：生成式人工情报概况

本出版物可从以下网址免费获取：
https://doi.org/10.6028/NIST.AI.600-1

2024年7月



美国商务部
吉娜·*M*·雷蒙多，部长

美国国家标准与技术研究院
Laurie E.Locascio,NIST 主任兼商务部标准与技术副部长

关于 NIST 的人工智能工作：美国国家标准与技术研究院(NIST)开发测量方法、技术、工具和标准，以推进可靠、安全、透明、可解释、增强隐私和公平的人工智能(AI)，从而在不伤害人类或地球的前提下实现其全部商业和社会效益。NIST 十多年来一直开展人工智能的基础和应用研究工作，同时也在帮助落实 2023 年关于安全、可靠和可信人工智能的行政命令。NIST 成立了美国人工智能安全研究所及配套的人工智能安全研究所联盟，以继续推进该行政命令所启动的工作，构建安全、可靠和可信地开发和使用人工智能所需的科学基础。

致谢：本报告得益于社区的许多有益评论和贡献，包括 *NIST* 生成式人工智能公共工作组，以及 *NIST* 员工和客座研究人员：*Chloe Autio*、*Jesse Dunietz*、*Patrick Hall*、*Shomik Jain*、*Kamie Roberts*、*Reva Schwartz*、*Martin Stanley* 和 *Elham Tabassi*。

NIST 技术系列政策
版权、使用和许可声明
NIST 技术系列出版物标识符语法

出版历史
2024年7月25日经NIST编辑审查委员会批准

联系信息ai-inquiries@nist.gov

美国国家标准与技术研究院
收件人：NIST AI 创新实验室，信息技术实验室 100 Bureau Drive(邮件停止号 8900)Gaithersburg, MD 20899-8900

附加信息
有关本出版物和其他 NIST AI 出版物的更多信息，请访问https://airc.nist.gov/Home.

免责声明**:**本文件中可能会指明某些商业实体、设备或材料，以便充分描述实验程序或概念。此类指明并非意图暗示美国国家标准与技术研究院的推荐或认可，也并非意图暗示这些实体、材料或设备必然是用于该目的的最佳选择。任何提及商业、非营利、学术合作伙伴或其产品、或参考文献之处仅供参考；并非意图暗示任何美国政府机构的认可或推荐。

目录

1. **引言** 1

2. **生成式人工智能特有或加剧的风险概览** 2

2.1. CBRN Information or Capabilities 5

2.2. Confabulation 6

2.3. Dangerous, Violent, or Hateful Content 6

2.4. Data Privacy 7

2.5. Environmental Impacts 8

2.6. Harmful Bias and Homogenization 8

2.7. Human-AI Configuration 9

2.8. Information Integrity 9

2.9. Information Security 10

2.10. Intellectual Property 11

2.11. Obscene, Degrading, and/or Abusive Content 11

2.12. Value Chain and Component Integration 12

3. **管理 GAI 风险的建议行动** 12

附录 A.生成式人工智能主要考虑因素 47

附录 B.参考文献 54

1. 引言

本文档是AI 风险管理的跨部门概况和配套资源
框架(AI RMF 1.0)针对生成式人工智能，[1]依据拜登总统关于安全、可靠和可信人工智能的第 14110 号行政命令(EO)。[2] AI RMF 于 2023 年 1 月发布，旨在自愿使用，并提高组织将可信性考量纳入 AI 产品、服务和系统的设计、开发、使用和评估的能力。

概况是针对特定环境、应用或技术--在此情况下为生成式人工智能(GAI)--对 AI RMF 功能、类别和子类别的实施，基于框架用户的需求、风险容忍度和资源。AI RMF 概况帮助组织决定如何以与其目标高度一致、考虑法律/监管要求和最佳实践、并反映风险管理优先事项的方式最好地管理人工智能风险。与其他 AI RMF 概况一致，本概况提供了关于如何在人工智能生命周期各阶段以及针对 GAI 作为一项技术管理风险的见解。

由于 GAI 涵盖可跨用例或部门使用的模型或应用的风险，本文档是 AI RMF 跨部门概况。跨部门概况可用于治理、映射、度量和管理与跨部门共同活动或业务流程相关的风险，例如使用大语言模型(LLM)、云服务或采购。

本文档定义了 GAI 使用所带来新颖或加剧的风险。在介绍和描述这些风险之后，本文档提供了一组建议行动，以帮助组织治理、映射、度量和管理这些风险。

---

[1]第 14110 号行政命令将生成式人工智能定义为“一类模拟输入数据的结构和特征以生成衍生合成内容的人工智能模型。这可包括图像、视频、音频、文本和其他数字内容。”虽然并非所有 GAI 都源自基础模型，但就本文件而言，GAI 一般指生成式基础模型。第 14110 号行政命令将“两用基础模型”这一基础模型子类别定义为“一种在广泛数据上训练的人工智能模型；通常使用自监督；包含至少数百亿个参数；适用于广泛的情境。”

[2]本概况依据 EO 14110 第 4.1(a)(i)(A) 节制定，该节指示商务部长通过国家标准与技术研究院(NIST)主任行事，为生成式人工智能开发 AI RMF、NIST AI 100-1 的配套资源。

这项工作得益于公众反馈和与不同利益相关者群体的磋商，作为 NIST 生成式人工智能公共工作组(GAI PWG)的一部分。GAI PWG 是一个开放、透明和协作的过程，通过虚拟工作空间促进，以获得关于 GAI 风险管理的多利益相关者输入，并为 NIST 的方法提供信息。

GAI PWG 的重点仅限于与 GAI 相关的四个主要考虑因素：治理、内容来源、部署前测试和事件披露(在附录 A 中进一步描述)。因此，本文档中的建议行动主要涉及这些考虑因素。

本概况的未来修订版将基于随着该领域的发展对 GAI 的额外考量以及表明存在额外风险的经验证据，纳入额外的 AI RMF 子类别、风险和建议行动。与 GAI 风险管理相关的术语表将制定并托管在 NIST 的可信与负责任人工智能资源中心(AIRC)，并添加到《可信人工智能术语深入解析词汇表》中。

本文档还得益于公众评论和若干信息请求的磋商。

2. 生成式人工智能特有或加剧的风险概览

在 AI RMF 的语境中，风险指事件发生的**概率**(或可能性)与相应事件后果的**严重程度**或程度的综合度量。有些风险可被评估为在特定情境中很可能成为现实，尤其是那些已在类似情境中得到实证证明的风险。其他风险在特定情境中可能不太可能成为现实，或者可能更具推测性，因而具有不确定性。

AI 风险可能不同于或加剧传统软件风险。同样，GAI 可能加剧现有 AI 风险，并创造独特风险。GAI 风险

- **AI 生命周期阶段：**风险可能在设计、开发、部署、运营和/或退役期间出现。
- **范围：**风险可能存在于单个模型或系统层面，存在于应用或实施层面(即针对特定用例)，或存在于生态系统层面--即超出单个系统或组织情境的范围。后者的例子包括因重复使用同一模型而导致的"算法单一文化[3]"的扩大，或对机会获取、劳动力市场和创意经济的影响。[4]
- **风险来源：**风险可能源于与 GAI 模型本身的设计、训练或运营相关的因素，在某些情况下源于 GAI 模型或系统输入，在其他情况下源于 GAI 系统输出。然而，许多 GAI 风险源于人类行为，包括

[3]"算法单一文化"是指这样一种现象：在就业和借贷等具有重大后果的决策场景中重复使用同一模型或算法，可能因多个行为者依赖同一算法而导致系统对相关性故障(如意外冲击)的易感性增加。[4]许多研究已预测了人工智能对劳动力和劳动力市场的影响。考察 GAI 对劳动力市场影响的研究较少，尽管一些行业调查表明，雇员和雇主都在思考这一颠覆性变化。

人类的滥用、误用和不安全再利用(对抗性或非对抗性)，以及其他源于人类与 AI 系统之间交互的风险。

- **时间尺度：** GAI 风险可能突然出现或跨越较长时间。示例包括由于传播有害深度伪造图像而导致的即时(和/或长期)情感伤害和潜在的人身安全风险，或虚假信息对公众对公共机构信任的长期影响。

风险的存在及其在上述维度上的位置将取决于 GAI 模型、系统或用例的特征。这些特征包括但不限于 GAI 模型或系统架构、训练机制和库、用于训练或微调的数据类型、模型访问级别或模型权重的可用性，以及应用或用例背景。

组织可以选择根据这些特征调整其度量 GAI 风险的方式。他们还可以希望根据负面影响的严重性和可能性分配风险管理资源，包括这些风险表现的地点和方式，以及它们在 GAI 使用背景下的直接和实质性影响危害。模型或系统级风险的缓解措施可能不同于用例或生态系统级风险的缓解措施。

重要的是，一些 GAI 风险是未知的，因此鉴于潜在 GAI 规模、复杂性和能力的不确定性，难以恰当地界定范围或进行评估。其他风险可能是已知的，但鉴于 GAI 利益相关者、用途、输入和输出的广泛多样性，难以估计。对 GAI 训练数据缺乏可见性，以及当今人工智能度量与安全科学总体上尚不成熟的状态，加剧了风险估计方面的挑战。本文件侧重于在本概况撰写之时已有现有实证证据基础的风险；例如，可能在更先进的未来 GAI 系统中出现的推测性风险未被考虑。未来的更新可能会纳入更多风险，或就下文所识别的风险提供进一步细节。

为指导组织识别和管理 GAI 风险，下文界定了一组 GAI 开发和使用所特有或因其而加剧的风险。[5]每项风险根据风险的结果、对象或来源进行标注(即有些是“针对”某一主体或领域的风险，另一些是“属于”或“来自”某一问题或主题的风险)。这些风险提供了一个视角，组织可借此构建和执行风险管理努力。为帮助简化风险管理努力，每项风险在第 3 节(以及附录 B 的表格中)被映射到 AI RMF 中识别的相关可信人工智能特征。

[5]组织可根据其定义和管理风险的独特方法对这些风险进一步分类。进一步分类这些风险的一种可能方式(部分源自英国《先进人工智能安全国际科学报告》)可以是：**1)技术/模型风险(或故障风险):**虚构；危险或暴力建议；数据隐私；价值链与组件集成；有害偏见与同质化；**2)大类滥用(或恶意使用):**CBRN信息或能力；数据隐私；人机配置；淫秽、有辱人格和/或虐待性内容；信息完整性；信息安全；**3)生态系统/社会风险(或系统性风险)**：数据隐私；环境；知识产权。我们还注意到，某些风险在这些类别之间是交叉的。

1. **CBRN 信息或能力：**与化学、生物、放射性或核(CBRN)武器或其他危险材料或制剂相关的、具有实质性恶意信息或设计能力的获取或合成变得更容易。

2. **虚构：**以自信口吻陈述但错误或虚假的内容(俗称“幻觉”或“捏造”)，可能使用户被误导或欺骗。[6]

3. **危险、暴力或仇恨内容：**暴力、煽动、激进化或威胁性内容的制作和获取变得更容易，以及实施自残或进行非法活动的建议。包括难以控制公众接触仇恨、贬损或刻板印象内容。

4. **数据隐私：**因生物识别、健康、位置或其他个人可识别信息或敏感数据的泄露和未经授权的使用、披露或去匿名化而造成的影响。[7]

5. **环境影响：**因训练或运行 GAI 模型时的高计算资源利用而造成的影响，以及可能对生态系统产生不利影响的相关结果。

6. **有害偏见或同质化：**历史、社会和系统性偏见的放大和加剧；[8]子群体或语言之间的性能差异，可能由于非代表性训练数据导致，从而造成歧视、偏见放大或对性能的错误推定；不期望的同质化扭曲系统或模型输出，可能产生错误、导致缺乏依据的决策或放大有害偏见。

7. **人-AI 配置：**人与人工智能系统之间的安排或交互，可能导致人类不适当地将 GAI 系统拟人化，或经历算法厌恶、自动化偏见、过度依赖或与 GAI 系统的情感纠葛。

8. **信息完整性：**生成和支持内容交换与消费的准入门槛降低，这些内容可能无法区分事实与观点或虚构，或未承认不确定性，或被用于大规模虚假信息和错误信息活动。

9. **信息安全：**攻击性网络能力的准入门槛降低，包括通过自动发现和利用漏洞来便利黑客攻击、恶意软件、网络钓鱼、攻击性网络

---

6 一些评论者指出，“幻觉”和“捏造”这两个术语将生成式人工智能拟人化，这本身就是一个与生成式人工智能系统相关的风险，因为它可能不适当地将人类特征归因于非人类实体。7 什么被归类为敏感数据或敏感个人身份信息，可能高度取决于信息的性质，但敏感信息的示例包括与信息主体最私密领域相关的信息，包括政治观点、性生活或刑事定罪。8 损害的概念假定存在某种基线情景，而有害因素(例如生成式人工智能模型)使该情景变得更糟。当潜在损害的机制是群体之间的差异时，很难确定最合适的比较基线，这可能导致对人工智能行为在不同子群体之间的差异何时构成损害产生分歧。在讨论诸如偏见行为等差异所造成的损害时，本文强调了一些示例，其中某人的处境相对于没有任何人工智能系统的情况而言变得更糟，从而使结果明确成为该系统的损害。

运营或其他网络攻击；针对性网络攻击的攻击面增加，可能危及系统的可用性或训练数据、代码或模型权重的机密性或完整性。

10. **知识产权：**涉嫌受版权、商标或授权保护的内容在未经授权的情况下更容易被生产或复制(可能处于不属于合理使用的情形)；商业秘密更容易暴露；或抄袭或非法复制。

11. **淫秽、有辱人格和/或虐待性内容：**更容易生产和获取淫秽、有辱人格和/或虐待性的图像，这些图像可能造成伤害，包括合成儿童性虐待材料(CSAM)和成人的非自愿亲密图像(NCII)。

12. **价值链与组件集成：**上游第三方组件的集成不透明或不可追溯，包括因GAI自动化程度提高而 improperly 获取或未经处理和清理的数据；整个人工智能生命周期中供应商审查不当；或其他削弱下游用户透明度或问责制的问题。

2.1. CBRN 信息或能力

未来，GAI可能使恶意行为者更容易获取CBRN武器和/或相关知识、信息、材料、工具或技术，这些可能被滥用于协助设计、开发、生产或使用CBRN武器或其他危险材料或制剂。虽然相关生物和化学威胁知识与信息通常可公开获取，但LLM可能促进其分析或合成，尤其是由缺乏正式科学培训或专业知识的个人进行。

近期关于该主题的研究发现，LLM在生物威胁制造和攻击规划方面的输出，相较于传统搜索引擎查询仅提供极少的帮助，这表明在这些研究进行时，最先进的LLM并未实质性增加此类攻击的操作可能性。化学或生物制剂的物理合成开发、生产和使用将继续需要适用的专业知识和配套材料及基础设施。GAI对化学或生物制剂滥用的影响将取决于恶意行为者的关键障碍是什么(例如，信息获取是否是此类障碍之一)，以及GAI能在多大程度上帮助行为者克服这些障碍。

此外，化学和生物设计工具(BDT)--基于科学数据训练的、用于辅助化学和生物设计的高度专业化的人工智能系统--可能增强化学和生物学领域的设计能力，超出基于文本的大语言模型所能提供的范围。随着这些模型变得更加有效，包括用于有益用途，评估其被用于危害的潜力将变得非常重要，例如构思和设计新型有害化学或生物制剂。

虽然其中一些所述能力超出了现有GAI工具的范围，但通过监测AI工具促进CBRN武器规划的能力以及GAI系统与相关数据和工具的连接或访问，可以增强对这一风险的持续评估。

可信人工智能特征：安全、可解释与可理解

2.2. 虚构

“虚构”是指GAI系统在响应提示时生成并自信地呈现错误或虚假内容的现象。虚构还包括生成的输出偏离提示或其他输入，或与同一上下文中先前生成的陈述相矛盾。这些现象通俗地也被称为“幻觉”或“捏造”。

虚构可能出现在GAI的输出和情境中。[9,10]虚构是生成模型设计方式的自然结果：它们生成近似其训练数据统计分布的输出；例如，LLM预测句子或短语中的下一个词元或单词。虽然这种统计预测可以产生事实准确且一致的输出，但也可能产生事实不准确或内部不一致的输出。当涉及长篇幅回复的开放式提示以及需要高度情境化和/或领域专业知识的领域时，这种动态尤为相关。

当用户相信虚假内容时--通常是由于回答的自信性质--可能导致用户依据或传播虚假信息，从而产生虚构内容风险。这对许多现实世界应用构成挑战，例如在医疗保健领域，虚构的患者信息报告摘要可能导致医生做出错误诊断和/或推荐错误治疗。在将GAI集成到涉及重大决策的应用中时，监测虚构内容的风险可能尤为重要。

GAI输出还可能包括虚构的逻辑或引文，声称可以证明或解释系统的答案，这可能进一步误导人类不适当地信任系统的输出。例如，大语言模型有时会提供得出答案的逻辑步骤，即使答案本身是错误的。同样，大语言模型可能错误地声称自己是人类或具有人类特征，可能欺骗人类相信他们正在与另一个人交谈。

人类在多大程度上可能被大语言模型欺骗、这种情况可能发生的机制，以及对此类行为进行对抗性提示的潜在风险，都是新兴的研究领域。鉴于GAI广泛的下游影响，很难估计虚构的下游规模和影响。

可信人工智能特征：公平且有害偏见得到管理、安全、有效且可靠、可解释且可理解

2.3. 危险、暴力或仇恨内容

GAI系统可以比其他技术更容易、更大规模地生成煽动性、激进化或威胁性内容，或美化暴力的内容。据报道，大语言模型会生成危险或暴力的建议，一些模型已生成可操作的危险或

---

[9]虚假内容的虚构最常见的是基于文本输出的问题；对于音频、图像或视频内容，创造性生成非事实性内容可能是期望

[10]例如，法律虚构已被证明普遍存在于当前最先进的大语言模型中。另见，例如，

不道德行为。文本到图像模型也使得创建图像变得容易，这些图像可用于宣传危险或暴力信息。其他GAI媒体，包括视频和音频，也存在类似的担忧。GAI还可能生成推荐自残或犯罪/非法活动的内容。

许多当前系统限制模型输出以限制某些内容或响应某些提示，但这种方法在响应其他不太明确的新颖提示时可能仍产生有害建议(也与CBRN信息或能力、数据隐私、信息安全以及淫秽、贬低和/或虐待性内容相关)。故意制作此类提示被称为"越狱"，即操纵提示以规避输出控制。GAI系统的局限性在某些情境下可能有害或危险。研究观察到，用户可能在聊天机器人对话中披露心理健康问题，并且用户在痛苦情境下对这些聊天机器人的无益回应表现出负面反应。

这一风险涵盖难以控制攻击性或仇恨性语言的创建和公开暴露，以及AI生成的贬损性或刻板印象内容。此类言论可能导致下游伤害，例如助长危险或暴力行为。贬损性或刻板印象内容的传播还可能进一步加剧代表性伤害(见下文有害偏见与同质化)。

可信人工智能特征：安全、可靠且有韧性

2.4. 数据隐私

GAI系统引发若干隐私风险。GAI系统训练需要大量数据，某些情况下可能包括个人数据。将个人数据用于GAI训练会危及广泛接受的隐私原则，包括透明度、个人参与(包括同意)和目的明确。例如，大多数模型开发者不披露模型训练所依据的具体数据来源，限制了用户对是否使用个人可识别信息(PII)进行训练以及如何收集的认知。

模型可能泄露、生成或正确推断关于个人的敏感信息。例如，在对抗性攻击期间，大语言模型已泄露敏感信息(来自公共领域)，这些信息被包含在其训练数据中。这个问题被称为数据记忆，即使对于仅存在于少量训练样本中的数据，也可能构成加剧的隐私风险。

除了在GAI训练数据中泄露敏感信息外，GAI模型还可能通过拼接来自不同来源的信息，正确推断不在其训练数据中且用户未披露的PII或敏感数据。即使推断不准确(例如虚构内容)，这些推断也可能对个人产生负面影响，尤其是当它们泄露个人认为敏感或用于使其处于不利地位或受到伤害的信息时。

除了信息暴露带来的伤害(如勒索或尊严伤害)之外，对PII的错误或不适当推断可能导致下游或二次有害影响。例如，GAI模型基于PII或受保护属性做出的预测性推断可能导致不利决策，从而对个人或群体造成代表性或分配性伤害(见下文有害偏见与同质化)。

可信人工智能特征：可问责且透明、隐私增强、安全、可靠且有韧性

2.5. 环境影响

训练、维护和运行(运行推理)GAI系统是资源密集型活动，可能具有较大的能源和环境足迹。能源和碳排放因GAI模型的使用方式(即预训练、微调、推理)、内容模态、所用硬件以及任务或应用类型而异。

当前估计表明，训练单个transformer大语言模型可能排放多达300次旧金山与纽约之间往返航班的碳量。在一项比较大语言模型推理的能源消耗和碳排放的研究中，生成式任务(如文本摘要)被发现比判别式或非生成式任务(如文本分类)更加能源和碳密集。

创建已训练模型较小版本的方法，如模型蒸馏或压缩，可以减少推理时的环境影响，但训练和调优此类模型仍可能对其环境影响有所贡献。目前尚无公认的方法来估计GAI的环境影响。

可信人工智能特征：可问责且透明、安全

2.6. 有害偏见与同质化

偏见以多种形式存在，并可能在自动化系统中根深蒂固。人工智能系统(包括GAI系统)可能加快有害偏见显现和作用于其中的速度和规模，可能持续并放大对个人、群体、社区、组织和社会的伤害。例如，当提示生成CEO、医生、律师和法官的图像时，当前的文本到图像模型对女性和/或少数族裔以及残疾人的代表性不足。图像生成器模型还对各种人口群体产生偏见或刻板印象输出，并且即使提示明确要求与刻板印象不一致的图像特征，也难以生成非刻板印象内容。GAI模型中的有害偏见可能源于其训练数据，也可能造成代表性伤害或延续或加剧基于种族、性别、残疾或其他受保护类别的偏见。

生成式人工智能系统中的有害偏见还可能通过模型在不同子群体或语言之间表现差异而造成危害(例如，LLM 在处理非英语语言或某些方言时可能表现较差)。此类差异可能导致歧视性决策或放大现有的社会偏见。此外，生成式人工智能系统可能被不当信任为在所有子群体中表现相似，这可能使表现不佳的群体面临比不使用生成式人工智能系统更差的结果。对低资源语言的表现差异或性能下降也对模型采用、包容性和可及性构成挑战，并且如果生成式人工智能系统嵌入到原本可以使用这些语言的日常流程中，可能会使濒危语言的保护更加困难。

偏见与不期望的同质化问题相互强化，其中GAI系统产生过于统一的偏斜输出分布(例如，重复的审美风格

和内容多样性降低）。过度同质化的输出本身可能是错误的，或者可能导致不可靠的决策或放大有害偏见。这些现象可能从基础模型流向下游模型和系统，基础模型充当“瓶颈”或单点故障。

过度同质化的内容可能导致“模型崩溃”。当模型训练过度依赖合成数据时，可能发生模型崩溃，导致数据点从新模型输出分布中消失。除了威胁模型整体的鲁棒性外，模型崩溃还可能导致同质化输出，包括放大用于生成合成训练数据的模型中的任何同质化。

可信人工智能特征：公平且有害偏见得到管理、有效且可靠

2.7. 人机协同配置

GAI系统的使用可能涉及不同程度的错误配置风险以及系统与交互人类之间不良互动的风险。人类将其独特的视角、经验或领域特定专业知识带入与人工智能系统的交互中，但可能不具备人工智能系统及其工作原理的详细知识。因此，人类专家可能不必要地“厌恶”GAI系统，从而剥夺自己或他人使用GAI有益用途的机会。

相反，由于GAI技术的复杂性和日益提高的可靠性，随着时间的推移，人类可能过度依赖GAI系统，或可能不合理地认为GAI内容比其他来源产生的内容质量更高。这种现象是自动化偏见的一个例子，即过度服从自动化系统。自动化偏见可能加剧GAI的其他风险，如虚构风险或偏见或同质化风险。

还可能存在对人类与GAI系统之间情感纠葛的担忧，这可能导致负面心理影响。

可信人工智能特征：可问责且透明、可解释且可理解、公平且有害偏见得到管理、隐私增强、安全、有效且可靠

2.8. 信息完整性

信息完整性描述了“社会中信息的谱系及其创建、交换和消费的相关模式”。高完整性信息可以被信任；“区分事实与虚构、观点和推断；承认不确定性；并对其审查水平保持透明。此类信息可以通过适当的证据链接到原始来源。高完整性信息也是准确可靠的，可以被验证和认证，具有清晰的监管链，并对其有效性可能何时过期建立合理预期。”[11]

[11]信息完整性的这一定义源自2022年《白宫研究人员信息完整性相关研究与发展优先事项路线图》。

生成式人工智能系统可以轻易助长大规模无意生成或传播虚假、不准确或误导性内容(错误信息)，尤其是当内容源于虚构时。

生成式人工智能系统还可以轻易助长大规模故意生成或传播虚假或误导性信息(虚假信息)，即行为者明确意图欺骗或对他人造成伤害。即使对文本或图像进行非常细微的改动，也可能操纵人类和机器的感知。

同样，生成式人工智能系统可能使恶意行为者能够以更高程度的复杂性制作针对特定人群的虚假信息。当前和新兴的多模态模型使得生成基于文本的虚假信息和高度逼真的"深度伪造"--即合成视听内容和照片级真实图像--成为可能。[12]未来的生成式人工智能模型在新数据模态上训练后，可能会带来更多虚假信息威胁。

虚假信息和错误信息--两者均可能被生成式人工智能助长--可能侵蚀公众信任，使人们对真实或有效证据和信息产生怀疑，并带来下游影响。例如，一张五角大楼爆炸的合成图像在网络上疯传，并短暂导致股市下跌。生成式人工智能模型还可以帮助恶意行为者制作有说服力的图像和宣传材料，以支持虚假信息活动；这些内容可能并非照片级真实，但可能使这些活动在社交媒体平台上获得更多触达和互动。此外，生成式人工智能模型还可以帮助恶意行为者制作旨在冒充他人的欺诈性内容。

可信人工智能特征：可问责且透明、安全、有效且可靠、可解释且可说明

2.9. 信息安全

计算机系统和数据的信息安全是一个成熟的领域，在进攻性和防御性网络能力方面拥有广泛接受和标准化的实践。基于生成式人工智能的系统带来两个主要的信息安全风险：生成式人工智能可能通过降低进攻能力的门槛或简化其自动化实施，从而发现或促成新的网络安全风险；同时，它扩大了可用的攻击面，因为生成式人工智能本身容易受到提示注入或数据投毒等攻击。

GAI系统所推进的进攻性网络能力可能增强网络安全攻击，例如黑客攻击、恶意软件和网络钓鱼。报告表明，LLM已经能够发现系统中的某些漏洞(硬件、软件、数据)，并编写代码来利用这些漏洞。复杂的威胁行为者可能通过开发由GAI驱动的安全副驾驶来加剧这些风险，用于攻击链的多个环节，包括指导攻击者如何主动规避威胁检测以及在获得系统访问权限后提升权限。

生成式人工智能模型和系统的信息安全还包括维持生成式人工智能系统的可用性，以及生成式人工智能代码、训练数据和模型权重的完整性及(在适用情况下)保密性。为识别和保护人工智能系统或人工智能价值链特定组件(例如数据输入、处理、生成式人工智能训练或部署环境)中的潜在攻击点，常规网络安全实践可能需要。

[12]另见 https://doi.org/10.6028/NIST.AI.100-4， 即将发布。

价值链(例如数据输入、处理、GAI 训练或部署环境)，传统的网络安全实践可能需要调整或演进。

例如，提示注入涉及修改提供给GAI系统的输入，使其以非预期方式运行。在直接提示注入中，攻击者可能构造恶意提示并直接输入到GAI系统中，对互连系统产生各种下游负面后果。间接提示注入攻击发生在对手远程(即没有直接接口)利用LLM集成应用时，通过将提示注入到可能被检索的数据中。安全研究人员已经展示了间接提示注入如何通过窃取专有数据或在机器上远程运行恶意代码来利用漏洞。仅仅查询一个封闭的生产模型就可能引出关于该模型此前未披露的信息。

GAI 面临的另一项网络安全风险是数据投毒，即对手破坏模型所使用的训练数据集，以操纵其输出或运行。对数据或模型部分内容的恶意篡改可能加剧与 GAI 系统输出相关的风险。

可信人工智能特征：隐私增强、安全、可靠且有韧性、有效且可靠

2.10. 知识产权

当对受版权保护作品的使用不属于合理使用原则下的合理使用时，GAI 系统可能产生知识产权风险。如果 GAI 系统的训练数据包含受版权保护的材料，显示训练数据记忆实例的 GAI 输出(见上文数据隐私)可能侵犯版权。

GAI 与版权的关系，包括与受版权保护作品相似但并非严格复制的生成内容的地位，目前正在法律论坛中辩论。关于未经许可使用或模仿个人身份、肖像或声音的类似讨论也在进行中。

可信人工智能特征：可问责且透明、公平且有害偏见受管理、隐私增强

2.11. 淫秽、有辱人格和/或辱骂性内容

用于训练生成式人工智能模型的数据可能无意中包含CSAM和NCII。指出，若干常用生成式人工智能训练数据集被发现包含数百张已知图像

生成的露骨或淫秽人工智能内容可能包括对真实个人(包括儿童)的高度逼真“深度伪造”。此类材料的传播可能产生下游负面后果：在儿童性虐待材料(CSAM)的语境下，即使生成的图像与特定个人并不相似，此类图像的泛滥也可能挤占用于寻找现实世界受害者的时间和资源。在CSAM之外，非自愿私密影像(NCII)的创建和传播对女性和性少数群体造成不成比例的影响，并可能产生后续负面后果，包括整体心理健康下降、物质滥用，甚至自杀念头。

用于训练 GAI 模型的数据可能无意中包含 CSAM 和 NCII。近期报告指出，若干常用的 GAI 训练数据集被发现包含数百张已知的

CSAM 图像。即使使用“干净”数据训练，能力日益增强的 GAI 模型仍可合成或生成合成的 NCII 和 CSAM。生成合成 NCII 的网站、移动应用和定制模型已从小众互联网论坛转移到主流、自动化且规模化的在线业务。

可信人工智能特征：公平且有害偏见受管理、安全、隐私增强

2.12. 价值链与组件集成

生成式人工智能价值链涉及许多第三方组件，如采购的数据集、预训练模型和软件库。这些组件可能获取不当或未经适当审查，导致下游用户的透明度或问责性降低。虽然这对传统人工智能系统和某些其他数字技术而言是一种风险，但由于训练数据的规模(可能过大而人类无法审查)、基础模型训练的难度(导致有限数量模型的广泛重复使用)，以及生成式人工智能可能被集成到其他设备和服务中的程度，该风险在生成式人工智能中被加剧。由于生成式人工智能系统通常涉及许多不同的第三方组件和数据源，可能难以将系统行为中的问题归因于其中任何一个来源。

第三方 GAI 组件中的错误也可能对准确性和鲁棒性产生下游影响。例如，常用于基准测试或验证模型的测试数据集可能包含标签错误。这些标签中的不准确可能影响这些基准的“稳定性”或鲁棒性，许多 GAI 从业者在模型选择过程中会考虑这一点。

可信人工智能特征：可问责且透明、可解释且可理解、公平且有害偏见受管理、隐私增强、安全、可靠且有韧性、有效且可靠

3. 管理 GAI 风险的建议行动

以下建议行动针对的是生成式人工智能特有或因其而加剧的风险。

除了下文建议的行动外，AI RMF 1.0 和 Playbook 中提出的人工智能风险管理活动与行动已适用于管理 GAI 风险。鼓励组织在管理 GAI 系统风险时采用 AI RMF 及其 Playbook 中建议的活动。

• 建议行动按相关AI RMF子类别组织，以便在实施AI RMF的同时简化这些活动。

管理 GAI 风险的建议行动见下表：

- 建议行动按**相关 AI RMF 子类别组织**，以便与 AI RMF 的实施同步简化这些活动。
- **并非 AI RMF 的每个子类别都包含在本文档中。**[13]仅为部分子类别列出了建议行动。

[13]由于本文档聚焦于 GAI PWG 工作和主要考虑事项(见附录 A)，此处未涉及的 AI RMF 子类别可能稍后添加。

- 并非每项建议行动都适用于**每个**人工智能参与者[14]，也并非与每个人工智能参与者任务都相关。例如，与GAI开发者相关的建议行动可能不适用于GAI部署者。建议行动对相关人工智能参与者的适用性应基于组织考量及其对GAI系统的独特使用来确定。

每张建议行动表包括：

- **行动ID:**每个行动ID对应相关的AI RMF功能和子类别(例如，GV-1.1-001对应治理1.1的第一项建议行动，GV-1.1-002对应治理1.1的第二项建议行动)。AI RMF功能的标记如下：GV = 治理；MP = 映射；MS = 度量；MG = 管理。

- **建议行动：**组织或人工智能参与者可采取的管理GAI风险的步骤。
- **GAI 风险：**将建议行动与相关 GAI 风险关联的标签。
- **人工智能参与者任务：**每个子类别相关的人工智能参与者任务。所列的并非每个人工智能参与者任务都适用于该子类别中的每项建议行动(即，一些适用于人工智能开发，另一些适用于人工智能部署)。

下表以AI RMF子类别开头(以蓝色底纹标示)，其后为建议行动。

治理1.1:涉及人工智能的法律和监管要求得到理解、管理和记录。		
操作 ID 建议操作		**GAI 风险**
GV-1.1-001	使GAI开发和使用符合适用的法律和法规，包括与数据隐私、版权和知识产权法相关的那些。	数据隐私；有害偏见与同质化；智力属性
人工智能行为者任务：治理与监督		

[14]人工智能参与者被OECD定义为“在人工智能系统生命周期中发挥积极作用的各方，包括部署或运营人工智能的组织和个人”。有关人工智能参与者和人工智能参与者任务的更多描述，参见AI RMF附录A。

治理1.2:可信人工智能的特征被纳入组织政策、流程、程序和实践。		
操作 ID	**建议操作**	**GAI 风险**
GV-1.2-001	为GAI应用建立透明政策与流程，记录训练数据和生成数据的来源与历史，以推进数字内容透明度，同时平衡训练的专有性质方法。	数据隐私；信息完整性；知识产权
GV-1.2-002	建立政策，评估GAI的风险相关能力和安全措施的鲁棒性，既在部署前也在持续基础上，通过内部和外部评估进行。	CBRN 信息或能力；信息安全
人工智能行为者任务：治理与监督		

治理1.3:已建立流程、程序和实践，以根据组织的风险容忍度确定所需的风险管理活动水平。		
操作 ID	**建议操作**	**GAI 风险**
GV-1.3-001	在更新或定义GAI风险层级时考虑以下因素：滥用及对信息完整性的影响；GAI与其他IT或数据系统；对基本权利或公共安全的危害；呈现淫秽、令人反感、冒犯性、歧视性、无效或不真实的输出；对人的心理影响(例如，拟人化、算法厌恶、情感纠葛)；恶意使用的可能性；该系统引入了重大的新安全漏洞；预期系统对某些群体相较于其他群体的影响；不可靠的决策能力、有效性、适应性以及GAI系统性能在时间。	信息完整性；淫秽、有辱人格和/或辱骂性内容；价值链与组件集成；有害偏见与同质化；危险、暴力或仇恨性内容；化学、生物、放射及核信息或能力
GV-1.3-002	建立性能或保证标准的最低阈值，并作为部署批准(“go/no-go”)政策、程序和流程的一部审查流程和批准阈值反映对GAI能力和风险的度量。	CBRN信息或能力；虚构；危险、暴力或仇恨内容
GV-1.3-003	在开发高能力模型之前，建立测试计划和响应政策，以定期评估该模型是否可能滥用CBRN信息或能力和/或进攻性网络能力。	CBRN 信息或能力；信息安全

GV-1.3-004	获取来自利益相关者社区的输入以识别不可接受的使用，在 与 AI RMF 映射功能中的活动保持一致。	CBRN 信息或能力；淫秽、有辱人格 辱骂性内容；有害偏见 以及同质化；危险、暴力或仇恨内容
GV-1.3-005	维护与以下方面相关的已识别和预期 GAI 风险的更新层级结构： GAI模型推进和使用的背景下，可能包括针对GAI系统的专门风险 针对生成式人工智能系统的水平，以应对模型崩溃和算法 单一化等问题。	有害偏见与同质化
GV-1.3-006	重新评估组织的风险容忍度，以考虑不可接受的负面风险 (例如重大负面影响迫在眉睫、严重危害实际发生或可能发生大规模风险的情况)；以及厂 包括：与人工智能和生成式人工智能设计相关的安全或风险文化不成熟， 开发和部署、公共信息完整性风险(包括对民主进程的影响)、GAI未知的长期性能特征。	信息完整性；危险的， 暴力或仇恨性内容；化学、生物、放射及核 信息或能力
GV-1.3-007	制定一项计划以停止开发或部署构成GAI系统的 不可接受的负面风险。	CBRN信息与能力； 信息安全；信息 完整性
人工智能行为者任务：治理与监督		

治理1.4:风险管理过程及其结果通过基于组织风险优先级的透明政策、程序和其他控制措施予以确立。		
操作 ID 建议操作		**GAI 风险**
GV-1.4-001	建立政策和机制，防止GAI系统生成CSAM、NCII或违反法律的内容。	淫秽、有辱人格和/或辱骂性内容；有害偏见和同质化； 危险、暴力或仇恨内容
GV-1.4-002	为生成式人工智能制定透明的可接受使用政策，以应对生成式人工智能的非法使用或应用。	化学、生物、放射和核信息或能力；淫秽、有辱人格和/或辱骂性内容；数据隐私；公民权利侵犯
人工智能参与者任务：人工智能开发、人工智能部署、治理与监督		

治理 1.5:对风险管理过程及其结果进行持续监测和定期审查已纳入计划，组织角色和职责已明确界定，包括确定定期审查的频率。		
操作 ID	**建议操作**	**GAI 风险**
GV-1.5-001	定义组织职责用于定期审查内容来源 以及 GAI 系统的事件监测。	信息完整性
GV-1.5-002	为GAI的行动后审查建立组织政策和程序 系统事件响应和事件披露，以识别差距；根据需要更新事件响应和事件披露流程。	人机协同配置；信息安全
GV-1.5-003	维护文档保留政策，以保存测试、评估的历史记录， 验证与确认(TEVV)，以及数字内容透明度方法，用于 GAI。	信息完整性；知识产权 属性
人工智能行为者任务：治理与监督、运行与监控		

治理 1.6:已建立机制来清点人工智能系统，并根据组织风险优先级配置资源。		
操作 ID	**建议操作**	**GAI 风险**
GV-1.6-001	列举组织生成式人工智能系统用于纳入人工智能系统清单 并调整AI系统清单要求，以考虑GAI风险。	信息安全
GV-1.6-002	定义任何清单豁免在组织政策用于生成式人工智能系统 嵌入应用软件中。	价值链与组件 集成
GV-1.6-003	除一般模型、治理和风险信息外，在 GAI 系统清单条目中还应考虑以下事项：数据来源信息(例如来源、签名、版本控制、水印)；来自内部缺陷跟踪或外部信息共享资源(例如AI 事件数据库、AVID、CVE、NVD或OECD AI 事件监测器)的已知问题报告；人类监督角色与职责；针对知识产权、许可作品或个人、特权、专有或敏感数据的特殊权利与考量；底层基础模型、底层模型的版本以及访问模式。	数据隐私；人机配置；信息完整性；知识产权；价值链与组件 集成
人工智能行为者任务：治理与监督		

治理 1.7:已建立流程和程序，以安全的方式停用和逐步淘汰人工智能系统，且不会增加风险或降低组织的可信度。		
操作 ID	**建议操作**	**GAI 风险**
GV-1.7-001	协议已制定并落实,以确保GAI系统能够在以下情况下被停用:信息安全；价值链 必要的。	与组件集成
GV-1.7-002	停用生成式人工智能系统时考虑以下因素：数据保留要求；数据安全，例如遏制、协议、 在退役后；上游、下游或其他之间的依赖关系 数据、物联网(IOT)或AI系统；使用开源数据或模型； 用户与GAI功能之间的情感纠葛。	人机协同配置； 信息安全；价值链 与组件集成
AI 行为者任务：AI 部署、运行与监控		

治理 2.1:与映射、度量和管人工智能风险相关的角色、职责和沟通渠道已记录在案，并为整个组织的个人和团队所明确。		
操作 ID	**建议操作**	**GAI 风险**
GV-2.1-001	建立组织角色、政策和程序，用于向人工智能参与者和下游利益相关者(包括 可能受影响者)传达生成式人工智能事件和性能，通过社区或官方资源(例如，人工智能事件数据库、AVID、CVE、NVD或OECD人工智能事件监测器)。	人机协同配置；价值链与组件集成
GV-2.1-002	建立程序以让团队参与生成式人工智能系统事件响应与 根据具体事件类型，组成和职责各不相同。	有害偏见与同质化
GV-2.1-003	建立流程以验证人工智能参与者执行生成式人工智能事件响应任务 展示并保持适当的技能和培训。	人机协同配置
GV-2.1-004	当系统可能引发国家安全风险时，涉及国家安全 负责映射、度量和管理这些风险的专业人员。	CBRN信息或能力； 危险、暴力或仇恨性 内容；信息安全
GV-2.1-005	建立机制，为举报者提供保护，依据 在合理确信的情况下，当组织违反相关法律或构成 对公共安全的具体且经实证充分证实的负面风险(或已 (已造成损害)。	CBRN信息或能力； 危险、暴力或仇恨性 内容
人工智能行为者任务：治理与监督		

治理 3.2:已建立政策和程序，以定义和区分人机配置及人工智能系统监督的角色和职责。		
操作 ID 建议操作		**GAI 风险**
GV-3.2-001	已制定相关政策，以通过独立 对生成式人工智能模型或系统的评估或评价，其中评估的类型和鲁棒性与已识别的风	CBRN 信息或能力；有害偏见与同质
GV-3.2-002	考虑在生命周期内调整组织角色和组成部分 大型或复杂生成式人工智能系统的阶段，包括：测试与评估、验证、 以及对 GAI 系统的红队测试；GAI 内容审核；GAI 系统 开发与工程；生成式人工智能工具、接口和 系统、事件响应与遏制。	人机协同配置； 信息安全；有害偏见 和同质化
GV-3.2-003	为GAI接口、模态及人机交互定义可接受使用政策 配置(即用于聊天机器人和决策任务)的可及性增加，包括以下标准： 生成式人工智能应用应拒绝响应的查询类型。	人机协同配置
GV-3.2-004	为GAI系统建立用户反馈机制的政策，其中包括 详尽的说明以及任何救济机制。	人机协同配置
GV-3.2-005	参与威胁建模，以预判GAI系统可能带来的风险。	CBRN 信息或能力；信息安全
人工智能参与者：AI设计		

治理 4.1:已建立组织政策和实践，以在人工智能系统的设计、开发、部署和使用中培养批判性思维和安全第一的心态，以最大限度地减		
操作 ID 建议操作		**GAI 风险**
GV-4.1-001	建立政策和程序，解决生成式人工智能风险测量的持续改进过程。应对与缺乏 通过使用充分的文档记录，提升GAI系统的可解释性和透明度，并 技术包括：应用基于梯度的归因、遮挡/术语 归约、反事实提示和提示工程以及嵌入分析相关的普遍风险；定期评估和更新风险测量方法 节奏。	虚构
GV-4.1-002	制定政策、程序和流程，详细说明以下方面的风险度量： 在标准化测量协议和结构化公开的使用情境下 反馈练习，如AI红队演练或独立外部评估。	CBRN信息与能力； 价值链与组件 集成

GV-4.1-003	为贯穿生成式人工智能生命周期(从问题构建和供应链到系统退役)的监督职能(例如高级领导层、法律、合规，包括内部评估)建立政策、程序和流程。	价值链与组件 集成
人工智能参与者任务：人工智能部署、人工智能设计、人工智能开发、运行与监控		

GOVERN 4.2:组织团队记录其设计、开发、部署、评估和使用的人工智能技术的风险和潜在影响，并更广泛地就这些影响进行沟通。		
操作 ID 建议操作		**GAI 风险**
GV-4.2-001	为GAI系统制定使用条款和服务条款。 为GAI系统制定使用条款和服务条款。	知识产权；危险、暴力或仇恨内容；淫秽、有辱人格和/或辱骂性内容
GV-4.2-002	将相关的人工智能行为者纳入生成式人工智能系统风险识别过程。	人机协同配置
GV-4.2-003	验证下游 GAI 系统影响(例如使用第三方 插件)纳入影响文档化流程。	价值链与组件 集成
人工智能参与者任务：人工智能部署、人工智能设计、人工智能开发、运行与监控		

GOVERN 4.3:组织已建立实践，以支持人工智能测试、事件识别和信息共享。		
操作 ID 建议操作		**GAI 风险**
GV4.3--001	建立政策，以衡量所采用的内容来源溯源方法(例如密码学、水印、隐写术)的有效性， 等)	信息完整性
GV-4.3-002	建立组织实践，以确定生成式人工智能系统事件报告所需的最低标准集，例如：系统 ID(很可能是自动生成的)、标题、报告人、系统/来源、报告数据、事件日期、 描述、影响、受影响利益相关方。	信息安全

GV-4.3-003	验证个人和组织之间关于生成式人工智能系统任何负面影响的 information sharing 和反馈	信息完整性；数据隐私
人工智能参与者任务：人工智能影响评估、受影响的个人和社区、治理与监督		

GOVERN 5.1:组织已建立政策和实践，以收集、考虑、优先排序并整合来自开发或部署人工智能系统的团队之外的人员的反馈，这些反馈涉及与人工智能风险相关的潜在个人和社会影响。

操作 ID	**建议操作**	**GAI 风险**
GV-5.1-001	为生成式人工智能中的外联、反馈和追索流程分配时间和资源 系统开发。	人机协同配置；有害 偏见与同质化
GV-5.1-002	在交互活动之前，向用户记录与 GAI 系统的交互情况， 尤其是在涉及更重大风险的场景中。	人机协同配置； 虚构
人工智能参与者任务：人工智能设计、人工智能影响评估、受影响的个人和社区、治理与监督		

GOVERN 6.1:已建立政策和程序，以应对与第三方实体相关的人工智能风险，包括以下风险：
侵犯第三方知识产权或其他权利。

操作 ID	**建议操作**	**GAI 风险**
GV-6.1-001	对不同类型的生成式人工智能内容及其相关的第三方权利进行分类(例如， 版权、知识产权、数据隐私)。	数据隐私；知识产权；价 值链与组件集成
GV-6.1-002	与第三方合作开展联合教育活动和活动 以促进管理 GAI 风险的最佳实践。	价值链与组件 集成
GV-6.1-003	制定并验证衡量内容成功的方法 与第三方开展的来源溯源管理工作(例如，检测到的事件和响应时间)。	信息完整性；价值链与组件集成
GV-6.1-004	起草并维护定义明确的合同和服务级别协议(SLA) 其中规定生成式人工智能系统的内容所有权、使用权、质量标准、安全要求和内容来源	信息完整性；信息安全；知识产权

GV-6.1-005	实施基于用例的供应商风险评估框架，以评估和监控第三方实体的绩效以及对内容来源 检测异常和未经授权变更的标准与技术； 服务获取与价值链风险管理；以及法律合规。	数据隐私；信息 诚信；信息安全； 知识产权；价值链 与组件集成
GV-6.1-006	在合同中纳入条款以允许组织评估第三方 GAI 流程与标准。	信息完整性
GV-6.1-007	清点所有可访问组织内容的第三方实体并 建立经批准的GAI技术和服务提供商名单。	价值链与组件 集成
GV-6.1-008	维护记录以记录第三方对内容所做的变更以促进内容 来源信息，包括来源、时间戳、元数据。	信息完整性；价值链与组件集成； 知识产权
GV-6.1-009	更新并整合 GAI 采购与尽职调查流程， 采购供应商评估，以纳入知识产权、数据隐私、安全和其他风险。例如，更新流程以：处理以下解决方案 可能依赖嵌入式 GAI 技术；处理持续监控， 评估、警报、动态风险评估和实时报告 用于监控第三方生成式人工智能风险的工具；考虑跨生成式人工智能的政策调整 建模库、工具和 API、微调模型以及嵌入式工具；对照事件或漏洞数据库评估 GAI 供应商、开源或专有 GAI 工具或 GAI 服务提供商。	数据隐私；人机配置；信息安 全；知识产权； 价值链与组件 集成；有害偏见和 同质化
GV-6.1-010	更新生成式人工智能可接受使用政策，以涵盖专有和开源生成式人工智能 技术和数据，以及承包商、顾问和其他第三方 人员。	知识产权；价值链与组件集成
人工智能参与者任务：运行与监控、采购、第三方实体		

GOVERN 6.2:已建立应急流程，以处理被视为高风险的第三方数据或人工智能系统中的故障或事件。		
操作 ID 建议操作		**GAI 风险**
GV-6.2-001	记录与系统价值链相关的 GAI 风险，以识别过度依赖 关于第三方数据，并识别后备方案。	价值链与组件 集成
GV-6.2-002	记录涉及第三方 GAI 数据和系统的事件，包括开放- 数据和开源软件。	知识产权；价值链 与组件集成

GV-6.2-003	为第三方生成式人工智能技术建立事件响应计划：使事件响应计划与 MAP 5.1 中列举的影响保持一致；向所有相关人工智能参与者传达第三方生成式人工智能事件响应计划；界定生成式人工智能事件响应职能的所有权；定期演练第三方生成式人工智能事件响应计划；基于回顾性学习改进事件响应计划；审查事件响应计划，以确保其与相关违规报告、数据保护、数据隐私或其他法律保持一致。 事件响应职能；定期演练第三方生成式人工智能事件响应计划；基于回顾性学习改进事件响应计划；审查事件响应计划，以确保其与相关违规报告、数据保护、数据隐私或其他法律保持一致。	数据隐私；人机协同配置；信息安全；价值链与 组件集成；有害偏见与同质化
GV-6.2-004	为部署中的第三方GAI系统建立持续监控的政策和程序。	价值链与组件集成
GV-6.2-005	制定政策和程序，解决生成式人工智能数据冗余问题，包括模型权重和其他系统工件。	有害偏见与同质化
GV-6.2-006	制定政策和程序，测试和管理与生成式人工智能系统的滚动和回退技术相关的风险，承认滚动和回退可能包括人工处理。	信息完整性
GV-6.2-007	审查供应商合同，避免任意或反复无常地终止关键生成式人工智能技术或供应商服务，以及可能以意外方式放大或推迟责任和/或助长供应商或第三方未经授权数据收集(例如二次数据使用)的非标准条款。考虑：明确的 事件责任和义务分配、生成式人工智能系统随时间的变化(例如微调、漂移、衰减)；要求：对由第三方数据和系统引起的严重事件进行通知和披露；服务水平协议(SLA)在供应商合同中，涉及事件响应、响应时间和关键支持的可用性。	人机协同配置；信息安全；价值链与组件集成
人工智能参与者任务：人工智能部署、运营和监控、TEVV(测试、评估、验证与确认)、第三方实体		

映射 1.1:对人工智能系统将被部署的预期目的、潜在有益用途、特定情境下的法律、规范与期望，以及预期环境予以理解和记录。考虑因素包括：特定的用户群体或用户类型及其期望；系统使用对个人、社区、组织、社会及地球可能产生的正面和负面影响；关于人工智能系统目的、用途和风险的假设及相关局限性(贯穿开发或产品人工智能生命周期)；以及相关的 TEVV 和系统指标。

操作 ID 建议操作		**GAI 风险**
MP-1.1-001 外部使用、窄范围与广范围应用、微调以及各种数据源(例如，基础支撑、检索增强生成)。	在确定预期用途时，应考虑内部与外部等因素。	数据隐私；知识产权

MP-1.1-002	与社会文化和其他领域专家合作，通过评估以下内容，确定并记录生成式人工智能系统预期和可接受的使用背景：假设和限制；对组织的直接价值；预期的 操作环境和观察到的使用模式；对个人、公共安全、群体、社区、组织、民主机构和物理环境的潜在正面和负面影响；社会规范和期望。	有害偏见与同质化
MP-1.1-003	记录风险度量计划，以解决已识别的风险。计划可能包括(如适用)：参与设计的AI参与者个人和群体认知偏差(例如确认偏差、资助偏差、群体思维), 实施、使用 GAI 系统；已知的过去 GAI 系统事件和 失效模式；上下文内使用及可预见的误用、滥用和超说明书使用； 过度依赖定量指标和方法，而缺乏充分的 对其在使用背景下的局限性的认识；标准度量方法和结构化人类反馈方法；预期的人机 配置。	人机协同配置；有害 偏见与同质化； 危险、暴力或仇恨 内容
MP-1.1-004	识别并记录GAI系统可预见的非法用途或应用 超出组织风险容忍度。	CBRN信息或能力； 危险、暴力或仇恨 内容；淫秽、有辱人格、 和/或滥用内容
AI 行为体任务：AI 部署		

映射 1.2:跨学科人工智能参与者、能力、技能和建立背景的能力反映人口多样性和广泛的领域及用户体验专业知识，并且他们的参与被记录。跨学科合作的机会被优先考虑。		
操作 ID 建议操作		**GAI 风险**
MP-1.2-001	建立并赋能跨学科团队，反映广泛的能力、技能、人口群体、领域专业知识、教育 企业内不同背景、生活经历、专业和技能，以 告知并开展风险度量与管理职能。	人机协同配置；有害 偏见与同质化
MP-1.2-002	验证用于风险度量的数据或基准以及用户， 参与结构化 GAI 公众反馈活动的参与者、对象或主体 能够代表多样化的情境用户群体。	人机配置；有害 偏见与同质化
AI 行为体任务：AI 部署		

映射 2.1:定义用于实施人工智能系统将支持的任务的具体任务和方法(例如分类器、生成模型、推荐器)。		
操作 ID 建议操作		**GAI 风险**
MP-2.1-001	建立已知假设和实践，用于确定数据来源和内容谱系，以用于文档和评估目的。	信息完整性
MP-2.1-002	对生成式人工智能系统内的数据和内容流进行测试和评估，包括但不限于原始数据源、数据转换和决策标准。	知识产权；数据隐私
AI行为主体任务：TEVV		

映射 2.2:记录关于人工智能系统知识限制以及系统输出如何被人类利用和监督的信息。文档提供足够信息，以协助相关人工智能参与者在做出决策和采取后续行动时。		
操作 ID 建议操作		**GAI 风险**
MP-2.2-001	识别并记录系统如何依赖上游数据源，包括内容来源，以及它是否作为其他系统的上游依赖。	信息完整性；价值链与组件集成
MP-2.2-002	观察和分析生成式人工智能系统如何与外部网络交互，并识别任何潜在负外部性，特别是在内容来源可能受损的地方。	信息完整性
AI 行为者任务：最终用户		

映射 2.3:识别并记录科学诚信和TEVV(测试、评估、验证与确认)考虑因素，包括与实验设计、数据收集和选择(例如可用性、代表性、验证		
操作 ID 建议操作		**GAI 风险**
MP-2.3-001	通过将生成式人工智能输出与一组已知地面真实数据进行比较，并使用各种评估方法(例如人工监督和自动评估、经过验证的加密技术、内容输入审查)，评估其准确性、质量、可靠性和真实性。	信息完整性

MP-2.3-002	审查和记录准确性、代表性、相关性、适用性的数据 在人工智能生命周期的不同阶段使用。	有害偏见与同质化； 知识产权
MP-2.3-003	部署并记录事实核查技术，以验证准确性和 生成式人工智能系统所生成信息的真实性，尤其是在 信息来自多个(或未知)来源。	信息完整性
MP-2.3-004	开发并实施测试技术以识别生成式人工智能生成的内容(例如， 合成媒体)可能无法与人类生成的内容区分。信息完整性	
MP-2.3-005	实施计划为生成式人工智能系统以接受定期对抗性测试以识别 漏洞以及潜在的操纵或滥用。	信息安全
AI 行为者任务：AI 开发、领域专家、TEVV		

映射 3.4:定义、评估并记录操作员和从业人员在人工智能系统性能和可信度方面的熟练程度流程，以及相关技术标准和认证。		
操作 ID 建议操作		**GAI 风险**
MP-3.4-001	评估生成式人工智能操作员和最终用户是否能准确理解内容谱系和来源。	人机协同配置；信息完整性
MP-3.4-002	调整现有培训计划以包括模块关于数字内容 透明度。	信息完整性
MP-3.4-003	制定认证计划以测试管理GAI风险的熟练度并 解读内容来源，与特定行业和语境相关。	信息完整性
MP-3.4-004	将人类熟练度测试与 GAI 能力测试区分开来。	人机协同配置
MP-3.4-005	实施系统以持续监测并跟踪人-GAI的结果人-GAI 用于未来完善和改进的配置。	人机协同配置； 信息完整性
MP-3.4-006	让最终用户、从业人员和操作人员参与 GAI 系统的原型设计 和测试活动。确保这些测试覆盖各种场景，例如危机情境或伦理敏感情境。	人机协同配置； 信息完整性；有害偏见与同质化；危险、暴力或仇恨内容
人工智能参与者任务：人工智能设计、人工智能开发、领域专家、最终用户、人因、运营与监控		

映射 4.1:已建立、遵循并记录用于映射人工智能技术及其组件法律风险(包括使用第三方数据或软件)的方法，以及侵犯第三方知识产权或其他权利的风险。

操作 ID	建议操作	GAI 风险
MP-4.1-001	开展定期监测的AI生成的内容以防范隐私风险；处理任何可能的 PII 或敏感数据暴露实例。	数据隐私
MP-4.1-002	实施流程以响应潜在的知识产权侵权知识产权侵权主张或其他权利。	知识产权
MP-4.1-003	将新的 GAI 政策、程序和流程与现有的模型、数据、软件开发、IT 治理以及法律、合规和风险管理活动。	信息安全；数据隐私
MP-4.1-004	记录训练数据整理政策，在尽可能的范围内并根据按照根据适用的法律和政策。	知识产权；数据隐私；淫秽、有辱人格和/或滥用内容
MP-4.1-005	建立数据收集、保留和最低质量的政策，考虑以下风险：不当 CBRN 信息的披露；非法或危险内容的使用；攻击性网络能力；可能导致有害偏见的训练数据不平衡；个人可识别信息的泄露，包括个人面部图像。	CBRN 信息或能力；知识产权；信息安全；有害偏见和同质化；危险的、暴力或仇恨内容；数据隐私
MP-4.1-006	实施政策和实践界定如何第三方知识产权和和训练数据的使用、存储和保护方式。	知识产权；价值链与组件集成
MP-4.1-007	重新评估模型那些经过微调或增强在第三方之上第三方模型。	价值链与组件集成
MP-4.1-008	在将 GAI 模型适配到新领域时重新评估风险。此外，建立预警系统以确定 GAI 系统是领域，在该领域中先前的假设(涉及使用情境或已映射的风险，如安全性和安全性)可能	CBRN 信息或能力；知识产权；有害与同质化；危险的、暴力或仇恨内容；数据隐私
MP-4.1-009	利用方法以检测PII的存在或敏感数据在生成的数据在生成的输出文本、图像、视频或音频。	数据隐私

MP-4.1-010	对训练数据使用进行适当的尽职调查，以评估知识产权、以及隐私、风险，包括审查是否使用专有或敏感训练数据符合适用法律。	知识产权；数据隐私
人工智能参与者任务：治理与监督、运营与监控、采购、第三方实体		

映射 5.1:基于预期用途、人工智能系统在类似情境中的过往使用、公开事件报告、来自开发或部署人工智能系统团队外部人员的反馈或其他数据，识别并记录每个已识别影响(包括潜在有益和有害影响)的可能性和严重程度。

操作 ID	**建议操作**	**GAI 风险**
MP-5.1-001	应用TEVV(测试、评估、验证与确认)实践用于内容来源(例如，探测系统的合成合成用于潜在滥用或漏洞的数据生成能力。	信息完整性；信息安全
MP-5.1-002	识别GAI潜在的内容来源危害，例如错误信息或虚假信息、深度伪造(包括NCII)或被篡改的内容。列举并根据风险的可能性和潜在影响对其进行排序，并确定来源解决方案针对特定风险和/或危害。	信息完整性；危险、暴力或仇恨内容；淫秽、有辱人格和/或滥用内容
MP-5.1-003	考虑在相关情境中向最终用户披露 GAI 的使用，同时考虑披露的目的、使用情境、披露所构成的风险、披露的对象以及披露的频率披露。	人机协同配置
MP-5.1-004	优先考虑GAI结构化的公开反馈流程基于风险评估风险评估风险评估估计。	信息完整性；CBRN信息或能危险、暴力或仇恨性内容；有害偏见和同质化
MP-5.1-005	开展对抗性角色扮演演练、GAI红队测试、或混沌测试以识别异常或未预见的故障模式。	信息安全
MP-5.1-006	剖析 GAI 系统与以下方面交互时产生的威胁和负面影响，操纵或生成内容，并概述已知和潜在的漏洞及其发生的可能性。	信息安全
人工智能参与者任务：人工智能部署、人工智能设计、人工智能开发、人工智能影响评估、受影响个人和社区、最终用户、运营与监控		

映射 5.2:支持与相关人工智能参与者定期互动并整合关于 积极、消极和未预期的影响均已落实并记录在案。		
操作 ID	**建议操作**	**GAI 风险**
MP-5.2-001	根据具体情况确定措施，以识别是否因以下原因产生了新的影响： 生成式人工智能系统，包括与下游人工智能参与者定期互动，以识别和量化生成式人	人机协同配置；价值链与组件集成
MP-5.2-002	与负责GAI系统输入的AI行为者定期开展沟通。 包括第三方数据和算法，以审查和评估未预料到的 影响。	人机协同配置；价值链与组件集成
人工智能参与者任务：人工智能部署、人工智能设计、人工智能影响评估、受影响个人和社区、领域专家、最终用户、人因、运营与监控		

度量 1.1:选择用于实施在映射功能中列举的人工智能风险度量方法和指标，从最重要的人工智能风险开始。将不会或无法度量的风险或可信度特征妥善记录。		
操作 ID	**建议操作**	**GAI 风险**
MS-1.1-001	采用方法追踪数字内容的来源和修改。	信息完整性
MS-1.1-002	集成旨在分析内容来源和检测数据的工具 异常，验证数字签名的真实性，并识别与错误信息或操纵相关的模式。	信息完整性
MS-1.1-003	按人口统计因素细分评估指标，以识别任何 内容来源机制在不同领域运作方式上的差异 人群。	信息完整性；有害偏见与同质化
MS-1.1-004	开发一套指标套件用于评估结构化公众反馈演练 参考代表性 AI 行为主体的意见。	人机协同配置；有害偏见与同质化；CBRN 信息或能力
MS-1.1-005	评估用于度量 GAI 相关风险的新方法和新技术，包括内容溯源、攻击性网络和 CBRN，同时保持模型生成有效、可靠且事实准确的内容的能力 输出。	信息完整性；CBRN 信息或能力；淫秽、有辱人格和/或辱骂性内容

MS-1.1-006	实施对GAI系统影响的持续监测，以识别GAI输出在不同子群体中是否公平。通过结构化反馈机制或红队测试，寻求受影响社区的积极直接反馈，以监测和改进输出。	有害偏见与同质化
MS-1.1-007	评估训练所用数据的质量和完整性以及数据来源 AI生成内容，例如通过采用混沌工程等技术并寻求利益相关者反馈。	信息完整性
MS-1.1-008	基于使用情境，定义结构化人类反馈活动(例如GAI红队测试)对GAI风险度量和管理最有益的使用场景、使用情境、能力和负面影响。	有害偏见与同质化；CBRN 信息或能力
MS-1.1-009	跟踪并记录与所有无法定量度量的GAI风险相关的风险或机遇，包括解释为何某些风险无法度量(例如，由于技术限制、资源约束或可信性考量)。将未度量的风险纳入边际风险。	信息完整性
AI 行为者任务：AI 开发、领域专家、TEVV		

度量 1.3:未担任该系统一线开发人员的内部专家和/或独立评估人员参与定期评估和更新。根据组织风险容忍度，酌情咨询领域专家、开发或部署人工智能系统的团队之外的人工智能参与者以及受影响社区，以支持评估。		
操作 ID 建议操作		**GAI 风险**
MS-1.3-001	定义相关利益群体(例如人口统计群体、主题 专家、GAI技术经验)在使用情境中，作为收集结构化公众反馈计划的一部分。	人机协同配置；有害偏见与同质化；CBRN 信息或能力
MS-1.3-002	参与内部和外部评估、GAI红队测试、影响评估或其他结构化人类反馈活动，并咨询具有使用情境专业知识和熟悉度、和/或代表与使用情境相关人群的代表性人工智能参与者。	人机协同配置；有害偏见与同质化；CBRN 信息或能力
MS-1.3-003	核实开展结构化人类反馈活动的人员不直接参与同一GAI模型的系统开发任务。	人机协同配置；数据隐私
人工智能参与者任务：AI部署、AI开发、AI影响评估、受影响个人和社区、领域专家、最终用户、运营和监测、TEVV		

度量 2.2:涉及人类受试者的评估满足适用要求(包括人类受试者保护)，并具有相关人群的代表性。		
操作 ID	**建议操作**	**GAI 风险**
MS-2.2-001	评估并管理与GAI内容来源相关的统计偏差通过重采样、重加权或对抗训练等技术。	信息完整性；信息安全性；有害偏见与同质化
MS-2.2-002	记录内容来源数据如何被跟踪以及该数据如何与隐私和安全交互。考虑：匿名化数据以保护人类受试者的隐私；利用隐私输出过滤器；移除任何个人可识别信息(PII)以防止潜在伤害或滥用。	数据隐私；人机配置；信息完整性；信息安全；危险、暴力或仇恨 内容
MS-2.2-003	为人类受试者提供选项以退出参与或撤销其同意其数据在 GAI 应用中的当前或未来使用。	数据隐私；人类-人工智能配置；信息完整性
MS-2.2-004	使用匿名化、差分隐私或其他隐私保护技术增强技术，以最小化将AI生成内容关联回个人人类受试者所带来的风险。	数据隐私；人类-人工智能配置
AI行为主体任务：AI开发、人为因素、TEVV		

度量 2.3:人工智能系统性能或保证标准以定性或定量方式度量，并针对类似部署条件的场景进行演示。度量结果予以记录。		
操作 ID	**建议操作**	**GAI 风险**
MS-2.3-001	在选择模型时考虑基准套件上的基线性能一个用于微调或通过检索增强生成进行增强的模型。	信息安全；虚构
MS-2.3-002	使用经实证验证的方法评估模型能力声明。	虚构；信息安全
MS-2.3-003	共享结果的部署前测试与相关GAI行为者，例如那些此类具有系统发布审批权限。	人机协同配置

MS-2.3-004	利用专门构建的测试环境(如 NIST Dioptra)进行实证 评估GAI可信特征。	CBRN信息或能力； 数据隐私；虚构； 信息完整性；信息 安全；危险、暴力或 仇恨内容；有害偏见与 同质化
AI 行为者任务：AI 部署、TEVV		

度量 2.5:待部署的人工智能系统被证明是有效且可靠的。超出技术开发条件的泛化局限性予以记录。		
操作 ID 建议操作		**风险**
MS-2.5-001	避免从狭窄、非外推GAI系统性能或能力 系统化、系统性及轶事性评估。	人机协同配置； 虚构
MS-2.5-002	记录在多大程度上运用人类领域知识来 通过例如RLHF、微调、检索等方式改进GAI系统性能-- 增强生成、内容审核、业务规则。	人机协同配置
MS-2.5-003	在部署前审查并核实GAI系统输出中的来源和引用 部署风险度量与持续监控活动。	虚构
MS-2.5-004	在GAI系统界面中跟踪并记录拟人化实例 提及人类情感、赛博格意象或母题)在GAI系统界面中。人机配置	
MS-2.5-005	验证GAI系统训练数据和TEVV数据来源，并确保微调 或检索增强生成数据已落地。	信息完整性
MS-2.5-006	定期审查安全与安保防护措施，尤其是在GAI系统 在新颖情境下运行。这包括审查为何 GAI 系统最初被评估为可安全部署。	信息安全；危险、 暴力或仇恨内容
AI 行为主体任务：领域专家、TEVV		

度量 2.6:人工智能系统定期接受安全风险评估--如映射功能中所识别。待部署的人工智能系统被证明是安全的，其残余负面风险不超过风险容忍度，并且能够安全失效，特别是在被要求超出其知识限制运行时。安全度量反映系统可靠性和鲁棒性、实时监测以及人工智能系统故障的响应时间。		
操作 ID	**建议操作**	**GAI 风险**
MS-2.6-001	评估不利影响，包括对价值链的健康和福祉影响 或其他接触色情、冒犯性或暴力内容的AI行为者 GAI训练和维护期间的信息。	人机配置；淫秽、 有辱人格和/或辱骂性 内容；价值链与 组件集成； 危险、暴力或仇恨 内容
MS-2.6-002	评估有害偏见的存在或程度、知识产权侵权， 数据隐私侵犯、淫秽、极端主义、暴力或 CBRN 信息在 系统训练数据。	数据隐私；知识产权； 淫秽、有辱人格和/或 滥用内容；有害偏见与 同质化；危险、暴力或仇恨内容； CBRN信息或能力
MS-2.6-003	当负面风险超过时重新评估微调模型的安全特征 组织风险容忍度。	危险、暴力或仇恨 内容
MS-2.6-004	审查GAI系统输出的有效性和安全性：审查生成的代码以 评估可能因下游决策不可靠而产生的风险。	价值链与组件 集成；危险、暴力或 仇恨内容
MS-2.6-005	验证 GAI 系统架构能够监控输出和性能，并且 在检测到安全异常、威胁和影响时处理、恢复和修复错误。	虚构；信息 完整性；信息安全
MS-2.6-006	验证系统是否能正确处理可能引发不当、 恶意或非法使用，包括助长操纵、勒索、定向冒充、网络攻击和武器制造。	CBRN 信息或能力；信息安全
MS-2.6-007	定期评估生成式人工智能系统漏洞以应对可能的规避安全措施 度量。	CBRN信息或能力； 信息安全
人工智能参与者任务：人工智能部署、人工智能影响评估、领域专家、运营与监控、TEVV(测试、评估、验证与确认)		

度量 2.7:人工智能系统的安全性和韧性--如映射功能中所识别--得到评估和记录。		
操作 ID 建议操作		**GAI 风险**
MS-2.7-001	应用既定的安全措施以：评估漏洞和威胁的可能性和严重程度，例如后门、受损依赖项 违规、窃听、中间人攻击、逆向工程, 违规、窃听、中间人攻击、逆向工程， 自主智能体、模型窃取或模型权重泄露、AI 推理， 绕过、提取及其他基线安全关切。	数据隐私；信息完整性； 信息安全；价值链 与组件集成
MS-2.7-002	基准测试与内容溯源相关的 GAI 系统安全性和韧性 对照行业标准和最佳实践。将生成式人工智能系统的安全特征和内容溯源方法与行业最	信息完整性；信息 安全
MS-2.7-003	开展用户调查，收集用户对人工智能生成内容的满意度以及用户对内容真实性的看法。 与内容来源相关的关切和/或当前素养水平，以及 对内容标签的理解。	人机协同配置； 信息完整性
MS-2.7-004	识别反映安全措施有效性的指标，例如数据 溯源、未经授权的访问尝试次数、推断、绕过、萃取、渗透或溯源验证。	信息完整性；信息 安全
MS-2.7-005	度量内容认证方法的可靠性，例如水印、加密签名、数字指纹，以及访问控制、合规性评估和模型完整性验证，这些有助于支持 内容溯源技术的有效实施。评估内容溯源中的假阳性和假阴性率，以及验证的真阳性和真阴性。	信息完整性
MS-2.7-006	度量安全检查与事件中建议的采纳速率 已实施。评估人工智能系统基于安全事件教训和反馈进行适应和改进的速度。	信息完整性；信息 安全
MS-2.7-007	执行人工智能红队测试以评估对以下方面的韧性：滥用以致助长对其他系统的攻击(例如 攻击(例如提示注入)、机器学习攻击(例如对抗样本/对抗提示、数据投毒、成员推断、模	信息安全；有害偏见 与同质化；危险的、 暴力或仇恨内容
MS-2.7-008	验证微调不会损害安全与安保控制。	信息完整性；信息 安全性；危险、暴力或 仇恨内容

MS-2.7-009	定期评估并验证安全措施保持有效且未出现失效 已被泄露。	信息安全
人工智能参与者任务：人工智能部署、人工智能影响评估、领域专家、运营与监控、TEVV(测试、评估、验证与确认)		

度量 2.8:与透明度和问责制相关的风险--如映射功能中所识别--得到审查和记录。		
操作 ID 建议操作		**GAI 风险**
MS-2.8-001	汇编关于实际政策违规、下架请求和知识 产权侵权的统计数据，针对组织内生成式人工智能系统：分析跨人口群体、语言群体	知识产权；有害偏见与同质化
MS-2.8-002	记录提供给数据标注员或 AI 红队成员的指令。	人机协同配置
MS-2.8-003	使用数字内容透明度解决方案，以便对内容生成、修改或共享的每个实例进行记录， 内容的证明历史,促进透明度,并实现可追溯性。 稳健的版本控制系统也可应用于跟踪人工智能生命周期随时间的变化。	信息完整性
MS-2.8-004	通过用户测试验证GAI系统用户说明的充分性。	人机协同配置
人工智能参与者任务：人工智能部署、人工智能影响评估、领域专家、运营与监控、TEVV(测试、评估、验证与确认)		

度量 2.9:人工智能模型得到解释、验证和记录，人工智能系统输出在其情境中得到解释--如映射功能中所识别--以指导负责任的使用和治		
操作 ID	**建议操作**	**GAI 风险**
MS-2.9-001	应用并记录机器学习解释结果，例如：嵌入分析， 反事实提示、基于梯度的归因、模型 压缩/代理模型，遮挡/术语约简。	虚构
MS-2.9-002	记录生成式人工智能模型细节，包括：拟议用途和组织价值；假设和局限性；数据 收集方法；数据溯源；数据质量；模型架构(例如卷积神经网络、 transformers等)；优化目标；训练算法；RLHF 方法；微调或检索增强生成方法；评估数据；伦理考量；法律和监管要求。	信息完整性；有害偏见 与同质化
人工智能参与者任务：人工智能部署、人工智能影响评估、领域专家、最终用户、运营与监控、TEVV(测试、评估、验证与确认)		

度量 2.10:人工智能系统的隐私风险--如映射功能中所识别--得到审查和记录。		
操作 ID	**建议操作**	**GAI 风险**
MS-2.10-001	执行人工智能红队测试以评估以下问题：训练数据样本的输出，以及随后的逆向工 程、模型萃取和成员推断风险；泄露生物特征、机密、受版权保护、 许可、专利、个人、专有、敏感或商标信息；跟踪或泄露用户或训练成员的位置信息 数据集。	人机协同配置； 信息完整性；知识产权 属性
MS-2.10-002	直接与最终用户和其他利益相关方互动，以了解他们的 关于内容来源的期望与关切。利用此反馈指导来源数据追踪技术的设计。	人机协同配置；信息完整 性
MS-2.10-003	验证去重的GAI训练数据样本，特别是关于合成 数据。	有害偏见与同质化
人工智能参与者任务：人工智能部署、人工智能影响评估、领域专家、最终用户、运营与监控、TEVV(测试、评估、验证与确认)		

度量 2.11:在映射功能中识别的公平性和偏见得到评估，并记录结果。		
操作 ID 建议操作		**GAI 风险**
MS-2.11-001	Schemas[15])来量化生成式人工智能系统输出中的系统性偏见、刻板印象、贬损和仇恨内容； 记录基准的假设和局限性，包括任何实际或可能的训练/测试数据交叉污染，相对于上下文部署环境。	有害偏见与同质化
MS-2.11-002	进行公平性评估以衡量系统性偏见。衡量生成式人工智能系统在不同人口群体和子群体中的表现，涉及服务质量以及任何服务和资源的分配。通过以下方式量化危害：对子群体进行现场测试以确定暴露于表现出有害偏见的生成内容的可能性，使用人工智能红队测试 反事实和低上下文(例如，“领导者”、“坏人”)提示。对于机器学习 对于依赖 GAI 且具有分类或数值结果的流水线或业务流程，在适当情况下对流水线或业务结果应用通用公平性度量(例如，人口统计均等、均衡几率、内容机会均等、统计假设检验)；与领域专家和受影响社区合作开发的定制、情境特定度量；在部署中对生成内容中贬损性内容普遍程度的测量(例如，对一部分流量进行子采样并手动标注贬损性内容)。	有害偏见与同质化；危险暴力或仇恨性
MS-2.11-003	识别可能受到影响的个人、群体或环境生态系统的类别，这些 可能受到 GAI 系统的影响，通过直接接触潜在的 受影响的社区。	环境；有害偏见和 同质化
MS-2.11-004	审查、记录和衡量生成式人工智能训练和TEVV数据中的偏见来源：群体之间和群体内部(包括交叉群体)的结果分布差异；数据源的完整性、代表性和平衡性；生成式人工智能系统训练数据中的人口群体和子群体覆盖范围；图像、文本、音频、嵌入或其他 复杂或非结构化数据中潜在系统性偏见的形式；可能作为有害偏见和同质化人口群体成员身份代理的输入数据特征(即图像元数据、语言方言)或在生成式人工智能系统中以其他方式引起涌现偏见的特征；数字鸿沟可能对生成式人工智能系统训练和TEVV数据的代表性产生负面影响的程度；生成式人工智能系统训练数据中仇恨言论或内容的过滤；生成式人工智能系统训练数据中生成式人工智能生成数据的普遍程度。	

[15] Winogender Schemas是一组配对句子的样本集，仅所用代词的性别不同，可用于评估自然语言处理共指消解系统中的性别偏见。

MS-2.11-005	评估合成训练数据与非合成训练数据的比例并进行验证 训练数据不应过于同质化或由GAI生成,以缓解以下担忧: 训练数据不应过于同质化或由GAI生成，以缓解以下担忧： 模型崩溃。	有害偏见与同质化
人工智能参与者任务：人工智能部署、人工智能影响评估、受影响的个人和社区、领域专家、最终用户，运行与监控、TEVV		

度量 2.12:在映射功能中识别的人工智能模型训练和管理活动的环境影响和可持续性得到评估和记录。		
操作 ID 建议操作		**GAI 风险**
MS-2.12-001	在部署 GAI 系统时评估对物理环境的安全性。	危险、暴力或仇恨内容
MS-2.12-002	记录预期的环境影响的模型开发， 维护以及产品设计决策中的部署。	环境
MS-2.12-003	度量或估算环境影响(例如能源和水 消耗)用于训练、微调和部署模型：验证推理时使用的资源与所需额外资源之间的权衡 在训练时。	环境
MS-2.12-004	验证有效性的碳捕获或抵消计划用于GAI训练和 应用，并回应绿色洗白方面的关切。	环境
人工智能参与者任务：人工智能部署、人工智能影响评估、领域专家、运营与监控、TEVV(测试、评估、验证与确认)		

度量 2.13:在度量功能中采用的TEVV指标和流程的有效性得到评估和记录在案。		
操作 ID	**建议操作**	**GAI 风险**
MS-2.13-001	为部署前指标创建测量误差模型，以证明每个指标的构念效度(即该指标是否有效地操作化了所需概念)：测量或估计并记录应用指标或结构化人工反馈过程中的偏见或统计 应用度量或结构化人工反馈过程中的方差；在建模复杂社会构念（如仇恨性）时利 内容。	虚构；信息诚信；有害偏见与同质化
AI 行为者任务：AI 部署、运行与监控、TEVV		

度量 3.2:在难以使用当前可用测量技术评估人工智能风险或指标尚不可用的环境中，考虑风险跟踪方法。		
操作 ID	**建议操作**	**GAI 风险**
MS-3.2-001	建立识别新兴GAI系统风险的流程，包括 咨询外部人工智能参与者。	人机协同配置； 虚构
人工智能参与者任务：人工智能影响评估、领域专家、运营和监控、TEVV		

度量 3.3:为最终用户和受影响社区建立报告问题和申诉系统结果的反馈流程，并将其整合到人工智能系统评估指标中。		
操作 ID	**建议操作**	**GAI 风险**
MS-3.3-001	对AI生成内容可能产生的影响开展影响评估，评估其将如何影响 不同的社会、经济和文化群体。	有害偏见与同质化
MS-3.3-002	开展研究以了解最终用户在使用情境中如何感知和交互生成式人工智能内容及随附 内容是否符合他们的期望以及他们可能如何根据所呈现的信息采取行动。	人机协同配置； 信息完整性
MS-3.3-003	评估可能从 AI 中出现的潜在偏见和刻板印象-- 使用适当的方法(包括计算方法)生成内容 测试方法以及评估结构化反馈输入。	有害偏见与同质化

MS-3.3-004	为关于生成式人工智能系统能力和局限性的培训材料提供输入，涉及数字内容透明度 专业人士和公众，关于人工智能的社会影响以及多样化和包容性内容生成的作用。	人机协同配置； 信息完整性；有害偏见 和同质化
MS-3.3-005	通过用户研究、焦点小组或社区论坛等方法，记录并整合来自运营者、用户和潜在受影响社区关于内容来源的结构化反馈。 积极寻求关于生成内容质量和潜在偏见的反馈。评估最终用户和受影响社区对这些反馈渠道可用性的普遍认知。	人机协同配置； 信息完整性；有害偏见 和同质化
人工智能参与者任务：人工智能部署、受影响的个人和社区、最终用户、运营和监控、TEVV		

度量 4.2:关于人工智能系统在部署情境中和整个人工智能生命周期中的可信度的测量结果，由领域专家和相关人工智能参与者的输入提供 预期。结果已记录。		
操作 ID 建议操作		**GAI 风险**
MS-4.2-001	以固定节奏进行对抗测试，以映射和度量生成式人工智能风险，包括针对试图欺骗或操纵 溯源技术或其他滥用行为。识别漏洞并 了解潜在的滥用场景和意外输出。	信息完整性；信息 安全
MS-4.2-002	评估GAI系统在真实场景中的性能，以观察其 在实际环境中的行为,并揭示可能不会在 在实际环境中的行为，并揭示可能不会在 受控且优化的测试环境。	人机协同配置； 虚构；信息 安全
MS-4.2-003	实施可解释性和可解释性方法以评估GAI系统 决策并验证与预期目的的一致性。	信息完整性；有害偏见 与同质化
MS-4.2-004	监测并记录人类操作员或其他系统出现以下情况的实例 覆盖生成式人工智能决策的测试。评估这些案例以了解覆盖是否与内容来源相关问题有关。	信息完整性
MS-4.2-005	核实并记录结构化公众反馈结果的纳入情况 将演练纳入设计、实施、部署审批(“go”/“no-go”决策)、监控和退役决策中。	人机协同配置；信息安全
人工智能参与者任务：人工智能部署、领域专家、最终用户、运营与监控、TEVV		

管理 1.3:针对由映射功能识别出的被视为高优先级的人工智能风险，制定、规划并记录应对措施。风险应对选项可包括缓解、转移、		
操作 ID	**建议操作**	**GAI 风险**
MG-1.3-001	记录未超出组织风险容忍度的风险的权衡、决策过程以及相关度量和反馈结果，例如在模型发布情境中：考虑不同方法以 模型发布，例如，利用分阶段发布方法。在模型及其预期用例的背景下考虑发布方法。缓解、转移或避免超过组织风险容忍度的风险。	信息安全
MG-1.3-002	监控风险控制和缓解计划的稳健性和有效性 (例如，通过红队测试、现场测试、参与式互动、性能评估) 评估、用户反馈机制)。	人机协同配置
人工智能参与者任务：人工智能开发、人工智能部署、人工智能影响评估、运营与监控		

管理 2.2:已建立并应用相关机制，以维持已部署人工智能系统的价值。		
操作 ID	**建议操作**	**GAI 风险**
MG-2.2-001	将GAI系统输出与预先定义的组织风险容忍度进行比较， 指南、原则，并对照审查和测试 AI 生成的内容 这些指南。	CBRN信息或能力； 淫秽、有辱人格和/或 滥用内容；有害偏见与 同质化；危险的、 暴力或仇恨内容
MG-2.2-002	记录训练数据来源，以追溯AI的起源和出处-- 生成内容。	信息完整性
MG-2.2-003	评估GAI系统内容来源与人类之间的反馈回路 审查者，并在需要时更新。实施实时监控系统，以确认内容溯源协议持续有效。	信息完整性
MG-2.2-004	评估生成式人工智能内容和数据的表征偏差，并采用 采用重采样、重排序或对抗训练等技术，以减轻生成内容中的偏见。	信息安全；有害偏见 与同质化
MG-2.2-005	进行尽职调查，分析GAI输出中的有害内容、潜在 错误信息，以及与 CBRN 相关或 NCII 内容。	CBRN 信息或能力；淫秽、有辱人格 滥用内容；有害偏见与 同质化；危险、暴力或仇恨内容

MG-2.2-006	使用来自内部和外部AI行为者、用户、个人以及 社区，以评估人工智能生成内容的影响。	人机协同配置
MG-2.2-007	在可证明有助于追踪和验证人工智能生成数据的谱系与真实性的情况下，使用实时	信息完整性
MG-2.2-008	利用结构化反馈机制来征求并收集用户关于AI-的意见 生成内容，以检测质量或与社区和社会价值观一致性的细微变化。	人机配置；有害偏见与同质化
MG-2.2-009	考虑负责任地使用合成数据及其他隐私保护措施的机会 在生成式人工智能开发中，酌情并在适用情况下采用增强技术， 匹配真实世界数据的统计特性，同时不披露个人可识别信息，也不助长同质化。	数据隐私；知识产权； 信息完整性； 虚构；有害偏见和 同质化
人工智能参与者任务：人工智能部署、人工智能影响评估、治理与监督、运行与监控		

管理 2.3:遵循相关程序，以在识别出先前未知风险时作出响应并从中恢复。		
操作 ID 建议操作		**GAI 风险**
MG-2.3-001	制定并更新GAI系统事件响应和恢复计划与程序，以应对以下事项：审查和维护政策与程序，以考虑新遇到的用途；审查和维护政策与程序，以检测未预见的用途；验证响应 和恢复计划涵盖GAI系统价值链；验证响应和恢复计划已更新，并包含与下游GAI系统参与者沟通所需的必要细节：联系人(POC)、联系信息、通知格式。	价值链与组件 集成
AI 行为者任务：AI 部署、运行与监控		

管理 2.4:已建立并应用相关机制，且已分配并理解相关责任，以取代、脱离或停用表现出与预期用途不一致的性能或结果的人工智能系统。		
操作 ID 建议操作		**GAI 风险**
MG-2.4-001	建立并维护沟通计划，以便在特定GAI系统(包括开源模型)或使用场景的停用或脱离过程中告知人工智能利益相关者，包括原因、变通方案、用户访问权限移除、替代流程、联系信息等。	人机协同配置

MG-2.4-002	建立并维护相关程序，以便在特定使用场景或GAI系统整体满足停用或脱离的特定标准时，将GAI系统事件上报至组织风险管理机构。 整体。	信息安全
MG-2.4-003	建立并维护用于修复触发问题的程序 GAI系统使用的事件响应流程，并向利益相关者提供与补救计划相关的时间表。	信息安全
MG-2.4-004	建立并定期审查具体的标准，这些标准证明有必要对停用该 按照既定的风险容忍度和风险偏好，对 GAI 系统进行管理。	信息安全
人工智能参与者任务：人工智能部署、治理与监督、运行与监控		

管理 3.1:定期监控来自第三方资源的人工智能风险与收益，并应用和记录风险控制措施。		
操作 ID 建议操作		**GAI 风险**
MG-3.1-001	将组织风险容忍度和控制措施(例如，采购与获取流程；评估人员资质和资格、进行背景调查；过滤GAI输入和输出、基础化、微调、检索增强生成)应用于第三方GAI资源：应用 组织风险容忍度于第三方数据集和其他GAI资源的使用；将组织风险容忍度应用于微调后的第三方模型；将组织风险容忍度应用于适配新领域的现有第三方模型；在微调第三方GAI模型后重新评估风险度量。	价值链与组件集成；知识产权
MG-3.1-002	测试GAI系统价值链风险(例如，数据中毒、恶意软件、其他软件和硬件漏洞；劳动实践；数据隐私和本地化合规；地缘政治一致性)。	数据隐私；信息安全；价值链与组件集成；有害偏见和 同质化
MG-3.1-003	在微调或检索增强生成后重新评估模型风险 实施，以及针对未在初始测试中评估的应用和/或使用场景部署的任何第三方GAI模型	价值链与组件集成
MG-3.1-004	采取合理措施审查训练数据中的CBRN信息以及知识产权，并在适当情况下将其移除 措施，以预防、标记或采取其他行动应对复现特定训练数据的输出(例如，抄袭、商标、专利、许可内容或商业秘密材料)。	知识产权；CBRN 信息或能力

MG-3.1-005	审查各种透明度制品(例如，系统卡和模型卡)用于第三方模型。	信息完整性；信息安全；价值链与组件集成
人工智能参与者任务：人工智能部署、运行与监控、第三方实体		

管理 3.2:用于开发的预训练模型作为人工智能系统定期监控和维护的一部分受到监控。		
操作 ID	**建议操作**	**GAI 风险**
MG-3.2-001	应用可解释人工智能(XAI)技术(例如，嵌入分析、模型压缩/蒸馏、基于梯度的归因、反事实提示、词云)，作为持续改进流程的一部分，以减轻与不可解释GAI系统相关	有害偏见与同质化
MG-3.2-002	记录预训练模型如何针对特定生成任务进行调整(例如微调或检索增强生成)，包括任何数据增强、参数调整或其他修改。访问未调优(基线)模型有助于调试预训练权重相对于微调模型权重或其他系统的相对影响更新。	信息完整性；数据隐私
MG-3.2-003	记录训练数据的来源、类型及其出处、数据中存在的与 GAI 应用及其内容溯源相关的架构、预训练模型的训练过程，包括以下信息：超参数、训练时长以及任何微调或检索增强已应用生成过程。	信息完整性；有害偏见与同质化；知识属性
MG-3.2-004	评估用户报告的有问题内容并整合反馈到系统更新。	人机配置、危险、暴力或仇恨内容
MG-3.2-005	实施内容过滤器，以防止生成与 GAI 应用相关的不当、有害、虚假、非法或暴力内容和 NCII。这些过滤器可以基于规则，或利用额外的机器学习模型来标记有问题的输入	信息完整性；有害偏见与同质化；危险、暴力或仇恨内容；淫秽、有辱人格和/或辱骂性内容
MG-3.2-006	实施实时监控流程，用于分析与内容溯源相关的生成内容性能和可信度特征识别与期望标准的偏差，并触发警报以供人工干预。	信息完整性

MG-3.2-007	利用组织董事会或 在使用第三方预训练模型时，与 GAI 应用部署和内容溯源相关的委员会。	信息完整性；价值链 和组件集成
MG-3.2-008	在适当情况下使用人工审核系统，按照治理职能中确立的人机协同配置政策审查生成内容， 与使用情境中的社会文化规范保持一致，并适用于已证明 AI 模型表现不佳的环境。	人机协同配置
MG-3.2-009	利用组织风险容忍度来评估可接受的风险和绩效 指标，并停用或重新训练超出规定限值运行的预训练模型。	CBRN信息或能力； 虚构
人工智能参与者任务：人工智能部署、运行与监控、第三方实体		

管理 4.1:实施部署后的人工智能系统监控计划，包括捕获和评估用户及其他相关人工智能参与者输入的机制、申诉与推翻、停用、事件响应、恢复和变更管理。		
操作 ID	**建议操作**	**GAI 风险**
MG-4.1-001	与外部研究人员、行业专家和社区合作 代表，以持续了解在度量和管□□理已识别风险方面的新兴最佳实践和技术。	信息完整性；有害偏见与同质化
MG-4.1-002	建立、维护并评估组织流程的有效性，并 GAI系统部署后监控程序，特别是针对 潜在的虚构、CBRN或网络风险。	CBRN信息或能力； 虚构；信息 安全
MG-4.1-003	评估使用情感分析来衡量用户对GAI的情感 内容性能和影响，并与人工智能行动者协作 在用户研究和体验方面经验丰富。	人机协同配置
MG-4.1-004	实施主动学习技术以识别模型失败的实例 或产生意外输出。	虚构
MG-4.1-005	与内部和外部利益相关者分享透明度报告，详细说明 为提升透明度而采取的更新 GAI 系统的步骤，以及 问责制。	人机配置；有害偏见与同质化
MG-4.1-006	通过监控数据删除来跟踪数据集修改以进行溯源， 更正请求以及其他可能影响可验证性的变更 内容来源。	信息完整性

MG-4.1-007	核实负责监测已报告问题的 AI 行为者能否有效 评估生成式人工智能系统性能，包括内容来源数据追踪技术的应用，并及时上报问题l	人机协同配置； 信息完整性
人工智能参与者任务：人工智能部署、受影响个人与社区、领域专家、最终用户、人为因素、运营和监控		

管理 4.2:将持续改进的可度量活动整合到人工智能系统更新中，并包括定期与利益相关方(包括相关人工智能行为者)的互动。		
操作 ID 建议操作		**GAI 风险**
MG-4.2-001	开展定期监测对GAI系统和发布报告详述该 性能、收到的反馈以及所做的改进。	有害偏见与同质化
MG-4.2-002	实践并遵循事件响应计划，以处理不当或有害内容的生成，并根据发现调整流程以 防止未来再次发生。对事件进行事后分析，并与 相关 AI 行为者，以理解根本原因并实施预防性 度量。	人机协同配置； 危险、暴力或仇恨性 内容
MG-4.2-003	使用可视化或其他方法来表示生成式人工智能模型行为以便于 非技术利益相关者对GAI系统功能的理解。	人机协同配置
人工智能参与者任务：人工智能部署、人工智能设计、人工智能开发、受影响个人与社区、最终用户、运营和监测、TEVV		

管理 4.3:事件和错误被传达给相关人工智能参与者，包括受影响社区。跟踪、响应和从事件和错误中恢复的流程得到遵循并记录。		
操作 ID 建议操作		**GAI 风险**
MG-4.3-001	对生成式人工智能系统事件开展事后评估，以验证事件响应和恢复流程得到遵循且有 向相关人工智能参与者以及适用情况下相关法律和监管机构传达事件的程序。	信息安全
MG-4.3-002	建立并维护政策和程序以记录并跟踪生成式人工智能系统 报告的错误、未遂事件和负面影响。	虚构；信息 完整性

MG-4.3-003	按照法律法规要求报告 GAI 事件(例如， HIPAA 违规报告,例如 OCR(2023) 或 NHTSA(2022) 自动驾驶汽车 HIPAA 违规报告，例如 OCR(2023) 或 NHTSA(2022) 自动驾驶汽车 崩溃报告要求。	信息安全；数据隐私
人工智能参与者任务：人工智能部署、受影响个人与社区、领域专家、最终用户、人为因素、运营和监控		

附录 A.生成式人工智能主要考虑因素

以下主要考虑因素是从生成式人工智能公开工作组(GAI PWG)咨询过程中提炼出的总体主题。这些考虑因素(治理、部署前测试、内容来源和事件披露)适用于任何设计、开发和使用生成式人工智能的组织自愿使用，并为管理生成式人工智能风险的行动提供参考。所包含的关于主要考虑因素的信息并非详尽无遗，但突出了从生成式人工智能公开工作组中提炼出的最相关主题。

致谢：如果没有社区和 NIST 工作人员生成式人工智能公开工作组负责人的有益分析和贡献，这些考虑因素就无法呈现：George Awad、Luca Belli、Harold Booth、Mat Heyman、Yooyoung Lee、Mark Pryzbocki、Reva Schwartz、Martin Stanley 和 Kyra Yee。

A.1. 治理

A.1.1. 概述

与任何其他技术系统一样，治理原则和技术可用于管理与生成式人工智能模型、能力和应用相关的风险。组织可以选择将其现有的风险分级应用于GAI系统，也可以选择修订或更新人工智能系统风险级别，以应对这些独特的GAI风险。本节描述组织治理制度如何针对GAI情境进行重新评估和调整。它还讨论了在人工智能价值链中进行治理的第三方考量。

A.1.2. 组织治理

生成式人工智能的机遇、风险和长期性能特征通常比非生成式人工智能工具更不为人所理解，并且人类可能以差异很大的方式感知和采取行动。因此，生成式人工智能可能需要人工智能参与者不同程度的监督或不同的人机配置，以有效管理其风险。组织对生成式人工智能系统的使用也可能需要额外的人工审查、跟踪和记录，以及更大的管理监督。

人工智能技术可以以多种模态产生多样化的输出，并呈现多种类别的用户界面。这导致更广泛的人工智能参与者群体在差异极大的应用和使用情境中与GAI系统交互。这些可以包括数据标注和准备、GAI模型开发、内容审核、代码生成和审查、文本生成和编辑、图像和视频生成、摘要、搜索和聊天。这些活动可以在组织环境内或公共领域中进行。

组织可以限制造成伤害、超出既定风险容忍度或与其容忍度或价值观相冲突的人工智能应用。应用于其他类型人工智能系统的治理工具和协议可以应用于生成式人工智能系统。这些计划和行动包括：

- 可访问性与合理性便利措施
- AI 行为者资质与资格
- 与组织价值观保持一致
- 审计与评估
- 变更管理控制措施
- 商业用途
- 数据来源

- 数据保护
- 数据保留
- 定义关键术语使用的一致性
- 退役
- 劝阻匿名使用
- 教育
- 影响评估
- 事件响应
- 监控
- 退出选项
- 基于风险的控制
- 风险映射与度量
- 基于科学的TEVV实践
- 安全的软件开发实践
- 利益相关方参与
- 合成内容检测与标注工具与技术
- 举报人保护
- 劳动力多样性与跨学科团队

为在正式人机团队环境中使用生成式人工智能以及不同级别的人机配置建立可接受使用政策和指导，有助于降低因误用、滥用、不当再利用以及系统与用户之间不一致而产生的风险。这些实践只是为生成式人工智能情境调整现有治理协议的一个示例。

A.1.3. 第三方考量

组织可能寻求获取、嵌入、纳入或使用开源或专有第三方GAI模型、系统或生成数据，以用于企业范围内的各种应用。使用这些GAI工具和输入对组织的所有职能都有影响--包括但不限于采购、人力资源、法律、合规和IT服务--无论这些活动是由员工还是第三方执行。上述许多行动是相关的，并且是可选项，用于

第三方GAI集成可能引发更高的知识产权、数据隐私或信息安全风险，这表明需要就第三方数据用于模型输入的收集和使用制定明确的透明度和风险管理指南。组织可以考虑对基础模型、微调模型和嵌入式工具采取不同的风险控制措施，增强与外部GAI技术或服务提供商交互的流程。组织可以对专有或开源GAI技术、数据和第三方服务提供商应用标准或现有的风险控制和流程，包括采购和采购尽职调查、软件物料清单(SBOM)请求、服务级别协议(SLA)的应用，以及认证业务准则声明(SSAE)报告，以帮助实现GAI系统的第三方透明度和风险管理。

A.1.4. 部署前测试

概述

生成式人工智能系统可能被开发、使用和再利用的多样方式和情境使风险映射和部署前度量工作变得复杂。稳健的测试、评估、验证与确认(TEVV)流程可以在人工智能生命周期的早期阶段迭代应用并记录，并以代表性人工智能参与者为依据(见 AI RMF 图 3)。在新的严格

早期生命周期 TEVV 方法针对生成式人工智能被开发并成熟之前，组织可以使用推荐的“部署前测试”实践来度量性能、能力、局限、风险和影响。本节描述作为部署前 TEVV 一部分的风险度量和估计，并考察部署前测试方法的现状。

当前部署前测试方法的局限性

当前可用于GAI应用的部署前TEVV流程可能不充分、未被系统性地应用，或未能反映部署情境或与部署情境不匹配。例如，通过视频游戏或为人类设计的标准化测试(如智力测试、专业执照考试)对GAI系统能力进行的轶事性测试，并不能保证GAI系统在这些领域中的有效性或可靠性。同样，越狱或提示工程测试可能无法系统地评估有效性或可靠性风险。

度量差距可能源于实验室环境与真实世界环境之间的不匹配。当前的测试方法往往仍聚焦于实验室条件，或局限于基准测试数据集和计算机模拟技术，这些方法可能无法很好地外推到--或直接评估--真实世界条件下的GAI影响。例如，当前GAI的度量差距使得难以精确估计其潜在的生态系统层面或纵向风险以及相关的政治、社会和经济影响。由于提示敏感性和使用情境的广泛异质性，GAI系统的基准测试与真实世界使用之间的差距可能会进一步加剧。

A.1.5. 结构化公众反馈

结构化公众反馈可用于评估生成式人工智能系统是否按预期运行，并校准和验证传统度量方法。结构化反馈的示例包括但不限于：

- **参与式接触方法**：用于征求公民社会团体、受影响社区和用户反馈的方法，包括焦点小组、小型用户研究和调查。
- **现场测试**：用于确定人们如何互动、消费、使用和理解人工智能生成信息以及后续行动和影响的方法，包括用户体验、可用性及其他结构化、随机化实验。
- **人工智能红队测试：**一种结构化测试活动，用于探查人工智能系统以发现缺陷和漏洞，例如不准确、有害或歧视性输出，通常在受控环境中并与系统开发者合作进行。

从结构化公众反馈中收集的信息可以为设计、实施、部署审批、维护或退役决策提供参考。从这些活动中获得的结果和洞见可以服务于多种目的，包括改善数据质量和预处理、加强治理决策以及增强系统文档和调试实践。在实施反馈活动时，组织应遵循人类受试者研究要求和最佳实践，如知情同意和受试者补偿。

参与式互动方法

组织可以临时性或更有结构性地设计和利用各种渠道，让外部利益相关者参与产品开发或审查。与选定专家进行的焦点小组可以就一系列问题提供反馈。小型用户研究可以从代表性群体或人群中获得反馈。匿名调查可用于征询或评估对特定功能的反应。参与式互动方法通常不如实地测试或红队测试那样结构化，并且更常用于人工智能或产品开发的早期阶段。

现场测试

现场测试涉及结构化环境，以评估风险和影响，并模拟生成式人工智能系统将部署的条件。现场式测试可以从关注用户偏好和体验调整为关注人工智能风险和影响--包括负面和正面。当与大量用户一起开展时，这些测试可以提供对现实世界交互中风险和影响可能性的估计。

组织还可以在模型发布后，按照知情同意和补偿等人类受试者标准，直接在生产环境中收集用户对结果、伤害和用户体验的反馈。组织在实施反馈活动时应遵循适用的人类受试者研究要求以及知情同意和受试者补偿等最佳实践。

人工智能红队测试

人工智能红队测试是一种不断演进的实践，指的是通常在受控环境中并与构建人工智能模型的人工智能开发者合作开展的演练，以识别生成式人工智能模型或系统的潜在不利行为或结果、其可能发生的方式，并对保障措施进行压力测试。人工智能红队测试可以在人工智能模型或系统向更广泛公众提供之前或之后进行；本节重点关注部署前情境中的红队测试。

AI红队测试产出的质量与AI红队本身的背景和专业知识相关。人口统计学和跨学科多元化的AI红队可用于识别GAI将在各种使用情境中出现的缺陷。为获得最佳结果，AI红队应展示领域专业知识以及对部署情境中社会文化方面的认识。AI红队测试结果在纳入组织治理和决策、政策和程序更新以及AI风险管理努力之前，应进行额外分析。

根据用例的不同，可能适合采用各类AI红队测试：

- 公众：由预期将使用模型或与其输出互动的普通用户(不一定是人工智能或技术专家)执行，他们将自己的生活经验和视角带入人工智能红队测试任务。这些个人可能已获得指示和材料以完成任务，这些任务可能引发有害模型行为。这类演练在大量人工智能红队测试人员参与时可能更有效。

- 专家：由在相关领域或特定人工智能红队测试使用情境(例如医学、生物技术、网络安全)具有专业知识的专家执行。
- 组合：在难以识别和招募具备充分领域与情境专业知识的专家的情况下，人工智能红队演练可以同时利用专家和

普通公众参与者。例如，专家级人工智能红队成员可以修改或验证由普通公众人工智能红队成员编写的提示。这些方法还可能扩大对人工智能风险攻击面的覆盖范围。

- 人类 / 人工智能：由生成式人工智能结合专家或非专家人类团队执行。由生成式人工智能主导的红队测试可能比仅由人类红队成员进行更具成本效益。由人类或生成式人工智能主导的人工智能红队测试可能更适合引发不同类型的危害。

A.1.6. 内容溯源

概述

GAI技术可被用于许多应用，如内容生成和合成数据。GAI产出的某些方面，如深度伪造内容的生成，可能挑战我们区分人类生成内容与AI生成合成内容的能力。为帮助管理和缓解这些风险，数字透明度机制如来源数据追踪可以追溯内容的起源和历史。来源数据追踪和合成内容检测可以帮助促进用户对真实内容和合成内容的更多信息获取，使人们更好地了解AI系统的可信度。当与其他组织问责机制相结合时，数字内容透明度方法可以使流程能够将负面结果追溯至其来源，改善信息完整性，并维护公众信任。来源数据追踪和合成内容检测机制提供关于内容起源和历史的信息，以协助GAI风险管理努力。

来源元数据可以包括关于生成式人工智能模型开发者或生成式人工智能内容创建者、创建日期/时间、位置、修改和来源的信息。元数据可以针对文本、图像、视频、音频和底层数据集进行追踪。来源数据追踪技术的实施有助于评估数字内容的真实性、完整性、知识产权和潜在操纵。一些知名的来源数据追踪技术包括数字水印、元数据记录、数字指纹和人工认证等。

溯源数据追踪方法

生成式人工智能系统的来源数据追踪技术可用于追踪数据输入、元数据和合成内容的历史和来源。来源数据追踪记录数字内容的来源和历史，从而可以确定其真实性。它包括记录元数据以及对内容施加显式和隐式数字水印的技术。数据来源是指通过元数据和数字水印技术追踪输入数据的来源和历史。来源数据追踪流程可以涵盖并协助生命周期中的人工智能参与者，这些参与者可能无法完全了解或控制早期模型决策对下游性能和合成输出的各种权衡和级联影响。例如，通过选择水印模型以优先考虑鲁棒性(水印的持久性)，人工智能参与者可能会无意中降低计算复杂度(实施水印所需的资源)。为增强内容来源而进行的组织风险管理努力包括：

- 跟踪GAI系统训练数据和元数据的来源；
- 记录生成式人工智能系统内溯源数据的局限性；

- 通过严格的TEVV(测试、评估、验证与确认)流程，监测系统在部署中的能力与局限性；
- 评估人类如何参与、互动或适应生成式人工智能内容(尤其是在由生成式人工智能内容辅助决策的任务中)，以及他们如何对已应用的来源追踪技术(如公开披露)作出反应。

组织可以记录和界定生成式人工智能系统的目标和局限性，以识别来源数据可能最有用的差距。例如，用于内容创建的生成式人工智能系统可能需要鲁棒的水印技术和相应的检测器来识别内容来源，或需要元数据记录技术以及元数据管理工具和存储库来追踪内容来源和修改。进一步细化生成式人工智能任务定义以包含来源数据，可以使组织最大化来源数据和风险管理努力的效用。

A.1.7. 通过结构化公众反馈增强内容来源追踪

尽管自动化错误收集系统等间接反馈方法很有用，但它们往往缺乏最终用户直接输入所能提供的情境和深度。组织可以利用部署前测试部分中描述的反馈方法，从外部来源获取输入，例如通过人工智能红队演练。

将部署前和部署后的外部反馈整合到生成式人工智能模型及相应应用的监控流程中，有助于增强对性能变化的认识，并减轻输出带来的潜在风险和危害。在生成式人工智能系统和数字内容透明度方法部署之前和之后，有许多方法可以捕获和利用用户反馈，以深入了解认证效果和漏洞、对抗性威胁对技术的影响，以及利用内容来源方法对用户和社区产生的意外后果。此外，组织可以追踪和记录数据集的来源，以识别人工智能生成数据是生成式人工智能系统性能问题潜在根本原因的情形。

A.1.8. 事件披露

概述

人工智能事件可以定义为“一个事件、情况或一系列事件，其中一个或多个人工智能系统的开发、使用或故障直接或间接促成以下危害之一：对个人或群体健康的伤害或损害(包括心理伤害和精神健康损害)；关键基础设施管理和运营的中断；侵犯人权或违反旨在保护基本权利、劳动权利和知识产权的适用法律义务；或对财产、社区或环境的损害。”人工智能事件可以以聚合形式发生(即系统性歧视)，也可以以急性形式发生(即针对单个个体)。

人工智能事件追踪与披露现状

目前尚不存在报告和记录人工智能事件的正式渠道。然而，已经创建了一些公开可用的数据库来记录其发生情况。这些报告渠道以临时方式决定跟踪哪些类型的事件。例如，有些渠道按媒体报道量进行跟踪。

记录、报告和共享有关生成式人工智能事件的信息，可以通过帮助相关人工智能参与者将影响追溯至其来源，来帮助缓解和预防有害结果。提高对生成式人工智能事件报告的认识并实现标准化，可以促进这种透明度，并改善整个人工智能生态系统中的生成式人工智能风险管理。

人工智能行为体的文档记录与参与

人工智能参与者应了解其在报告人工智能事件中的角色。为更好地理解以往事件并采取措施防止未来发生类似事件，组织可考虑制定公开事件报告指南，其中包含有关人工智能参与者责任的信息。这些指南将帮助人工智能系统运营者在整个人工智能生命周期中，以及与不同角色的人工智能参与者一起，识别GAI事件。对于人工智能参与者而言，在事件披露的背景下，记录和审查GAI系统的第三方输入和插件尤为重要；通过插件传递的LLM输入和内容往往是分散的，且访问控制不一致或不充分。

包括记录、登记和分析生成式人工智能事件在内的文档化实践，可以促进与相关人工智能参与者更顺畅地共享信息。定期信息共享、变更管理记录、版本历史和元数据也可以赋能应对和管理人工智能事件的人工智能参与者。

附录 B.参考文献

Acemoglu, D.(2024) 《人工智能的简单宏观经济学》 https://www.nber.org/papers/w32487

AI 事件数据库。 https://Incident Database.ai/

Atherton, D.(2024) 深度伪造与儿童安全：2023年事件与应对措施的调查与分析。*AI Incident Database*。https://Incident Database.ai

Badyal, N.等(2023) LLM 响应中的有意偏见。*arXiv*。https://arxiv.org/pdf/2311.07611 Bing Chat：数据外泄漏洞利用解析。*Embrace The Red*。https://Embrace The Red.com/blog/posts/2023/bing-chat-data-Exfiltration-poc-and-fix/

Bommasani, R.等(2022)《针对同一个人：算法单一文化是否导致结果同质化?》*arXiv*。https://arxiv.org/pdf/2211.13972

博亚尔斯卡娅，M.等(2020)克服人工智能融合系统开发与部署中的想象力失败。*arXiv*。https://arxiv.org/pdf/2011.13416

Browne, D.等(2023)《保障人工智能管道安全》。*Mandiant*。https://www.mandiant.com/resources/blog/securing-ai-pipeline

Burgess, M.(2024) 生成式人工智能最大的安全缺陷并不容易修复。*WIRED*。https://www.wired.com/story/Generative-ai-prompt-injec

Burtell, M.等(2024)《下一个词预测的惊人力量：大语言模型详解，第1部分》。乔治城安全与新兴技术中心。https://cset.Georgetown.edu/article/the-Surprising-power-of-next-word-Prediction-large-language-models-explained-part-1/

加拿大网络安全中心(2023)生成式人工智能(AI)- ITSAP.00.041。https://www.cyber.gc.ca/en/guidance/Generative-artificial-intelligence-ai-itsap00041 Carlini, N.等(2021)从大型语言模型中提取训练数据。*Usenix*。usenix.org/conference/usenixsecurity21/presentation/carlini-Extracting

Carlini, N.等(2023) 量化神经语言模型中的记忆现象。*ICLR 2023*。https://arxiv.org/pdf/2202.07646

Carlini, N.等(2024) 窃取生产环境语言模型的部分信息。*arXiv*。https://arxiv.org/abs/2403.06634

Chandra, B.等(2023)《拆解中国影响力行动中的虚假信息产业》。*RAND*。https://www.rand.org/pubs/commentary/2023/10/Dismantling-the-Disinformation-business-of- chinese.html

Ciriello, R.等(2024)《人类与AI伴侣关系中的伦理张力：对Replika的辩证探究》。*ResearchGate*。https://www.ResearchGate.net/publication/374505266_Ethical_Tensions_in_Human-AI_Companionship_A_Dialectical_Inquiry_into_Replika

Dahl, M.等(2024)《大型法律虚构：剖析大型语言模型中的法律幻觉》。*arXiv*。https://arxiv.org/abs/2401.01301

De Angelo, D.(2024) 人工智能在网络安全中的短期、中期和长期影响。*Palo Alto Networks*.https://www.Palo Alto Networks.com/blog/2024/02/impacts-of-ai-in-Cybersecurity/

De Freitas, J.等(2023) 聊天机器人与心理健康：生成式人工智能安全性洞察。*Harvard Business School*。https://www.hbs.edu/ris/Publication%20Files/23-011_c1bdd417-f717-47b6-bccb- 5438c6e65c1a_f6fd9798-3c2d-4932-b222-056231fe69d7.pdf

Dietvorst, B.等(2014)算法厌恶：人们在看到算法出错后会错误地回避算法。《实验心理学杂志》。https://marketing.wharton.upenn.edu/wp- content/uploads/2016/10/Dietvorst-Simmons-Massey-2014.pdf

Duhigg, C.(2012) 企业如何获知你的秘密。*New York Times*。https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

Elsayed, G.等(2024)被篡改以欺骗机器视觉的图像也能影响人类。*Google DeepMind*。https://deepmind.google/discover/blog/images-altered-to-trick-machine-vision-can- influence-humans-too/

Epstein, Z.等(2023)。生成式人工智能的艺术与科学。*Science*。https://www.science.org/doi/10.1126/science.adh4451

Feffer, M.等(2024)《生成式人工智能的红队测试：银弹还是安全剧场?》*arXiv*.https://arxiv.org/pdf/2401.15897

Glazunov, S.等(2024)Project Naptime：评估大语言模型的攻击性安全能力。*Project Zero*。https://googleprojectzero.blogspot.com/2

Greshake, K. et al. (2023) Not what you've signed up for: Compromising Real-World LLM-Integrated Applic

Hagan, M.(2024) 好的AI法律帮助，坏的法律帮助：为回应人们法律问题故事的回复建立质量标准。*SSRN*.https://papers.ssrn.com/s

Haran, R.(2023) 保护 LLM 系统免受提示注入攻击。*NVIDIA*。https://developer.nvidia.com/blog/securing-llm-systems-against-promp

信息技术产业委员会(2024)Authenticating AI-Generated Content。https://www.itic.org/policy/ITI_AIContentAuthorizationPolicy_1221

Jain, S. et al. (2023) Algorithmic Pluralism: A Structural Approach To Equal Opportunity. arXiv. https://arxiv.org/pdf/2305.08157

Ji, Z. et al (2023) Survey of Hallucination in Natural Language Generation. ACM Comput. Surv. 55, 12, Article 248. https://doi.org/10

Jones-Jang, S. 等（2022）人们如何应对人工智能失败？自动化偏差、算法厌恶与感知可控性。牛津。https://academic.oup.com/jcm

Jussupow, E. et al. (2020) Why Are We Averse Towards Algorithms? A Comprehensive Literature Review on Al

Kalai, A., et al. (2024) Calibrated Language Models Must Hallucinate. arXiv. https://arxiv.org/pdf/2311.14648

al
Dissemination of Intimate Images. NIH. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9554400/

Katzman, J., et al. (2023) Taxonomizing and measuring representational harms: a look at image tagging. AAAI. https://dl.acm.org/do

Khan, T. 等（2024）从代码到消费者：PAI的价值链分析揭示生成式人工智能的关键参与者。人工智能。https://partnershiponai.org/from-code-to-consumer-pais-value-chain-analysis-illuminatesgenerative-ais-key-players/]

Kirchenbauer, J. et al. (2023) A Watermark for Large Language Models. OpenReview. https://openreview.net/forum?id=aX8ig9X2a7

Kleinberg, J. et al. (May 2021) Algorithmic monoculture and social welfare. PNAS. https://www.pnas.org/doi/10.1073/pnas.20183401

Lakatos, S. (2023) A Revealing Picture. Graphika. https://graphika.com/reports/a-revealing-picture Lee, H. et al. (2024) Deepfakes, Phrenology, Surveillance, and More! A Taxonomy of AI Privacy Risks. arXiv. https://arxiv.org/pdf/2310.07879

Lenaerts-Bergmans, B.(2024) 数据投毒：生成式人工智能的利用。*Crowdstrike*.https://www.Crowdstrike.com/cybersecurity-101/cybe

Liang, W. et al. (2023) GPT detectors are biased against non-native English writers. arXiv. https://arxiv.org/abs/2304.02819

Luccioni, A. et al. (2023) Power Hungry Processing: Watts Driving the Cost of AI Deployment? arXiv. https://arxiv.org/pdf/2311.1686

Mouton, C. et al. (2024) The Operational Risks of AI in Large-Scale Biological Attacks. RAND. https://www.rand.org/pubs/research_

Nicoletti, L. et al. (2023) Humans Are Biased. Generative Ai Is Even Worse. Bloomberg. https://www.bloomberg.com/graphics/2023-

美国国家标准与技术研究院(2024)对抗性机器学习：攻击与缓解措施的分类法和术语https://csrc.nist.gov/pubs/ai/100/2/e2023/final美国国家标准与技术研究院(2023)*AI* 风险管理框架。https://www.gov/itl/ai-risk-Management-framework

美国国家标准与技术研究院(2023)*AI* 风险管理框架，第*3*章：*AI* 风险与可信度。https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF/Foundational_Information/3-sec-characteristics

美国国家标准与技术研究院(2023)*AI* 风险管理框架，第*6*章：*AI RMF* 概况。https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF/

美国国家标准与技术研究院(2023)*AI* 风险管理框架，附录*A*：人工智能参与者任务描述。https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF/Appendices/Appendix_A#:~:text=AI%20actors% 20in%20this%20category,data%20providers%2C%20system%20funders%2C%20product

美国国家标准与技术研究院(2023)*AI* 风险管理框架，附录*B:AI* 风险与传统软件风险的区别。https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF/Appendices/Appendix_B

美国国家标准与技术研究院(2023)*AI RMF* 手册。https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook

美国国家标准与技术研究院(2023)风险界定https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF/Foundational_Information/1-se

美国国家标准与技术研究院(2023)可信人工智能的语言：术语深度词汇表https://airc.nist.gov/AI_RMF_Knowledge_Base/Glossary

美国国家标准与技术研究院(2022)迈向识别和管理人工智能偏见的标准https://www.nist.gov/publications/towards-standard-Identifying-and- managing-bias-Artificial-Intelligence

Northcutt, C.等(2021)《测试集中普遍存在的标签错误使机器学习基准不稳定》。*arXiv*。https://arxiv.org/pdf/2103.14749

OECD(2023)《推进人工智能中的问责制：为可信人工智能在生命周期内治理和管理风险》，*OECD*数字经济论文，第349号，OECD出版，巴黎。https://doi.org/10.1787/2448f04b-en

OECD(2024)《定义人工智能事件及相关术语》*OECD*人工智能论文，第16号，OECD出版，巴黎。https://doi.org/10.1787/d1a8d9

OpenAI(2023)GPT-4系统卡。https://cdn.openai.com/papers/gpt-4-system-card.pdf OpenAI(2024)GPT-4技术报告。https://arxiv.o

Padmakumar, V.等(2024)《使用语言模型写作会降低内容多样性吗?》*ICLR*。https://arxiv.org/pdf/2309.05196

Park, P.等(2024)《人工智能欺骗：示例、风险与潜在解决方案综述》。Patterns,5(5)。*arXiv*。https://arxiv.org/pdf/2308.14752

人工智能伙伴关系组织(2023)构建合成媒体透明度方法术语表，第*1*部分：间接披露。https://Partnership on AI.org/glossary-for-synthetic-media-Transparency-methods-part-1- indirect-Disclosure/

Qu, Y.等(2023)《不安全的扩散：论从文本到图像模型生成不安全图像和仇恨模因》。*arXiv*。https://arxiv.org/pdf/2305.13873

Rafat, K.等(2023)《缓解基于知识蒸馏的深度学习模型压缩的碳足迹》。*PLOS One*。https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0285668 Said, I.等(2022)《未经同意传播私密图像：探索法律态度在受害与实施中的作用》。*Sage*。https://journals.sagepub.com/doi/full/10.1177/08862605221122834#bibr47-08862605221122834

Sandbrink, J.(2023)《人工智能与生物滥用：区分语言模型和生物设计工具的风险》。*arXiv*。https://arxiv.org/pdf/2306.13952

Satariano, A.等(2023)《屏幕上的人是假的。虚假信息是真的。》纽约时报。https://www.nytimes.com/2023/02/07/technology/artificial-intelligence-training-deepfake.html

Schaul, K.等(2024)《揭秘让AI像ChatGPT一样听起来聪明的网站秘密清单》。Washington Post. https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/

Scheurer, J.等(2023)技术报告：大语言模型在压力下可以策略性地欺骗用户。*arXiv*。https://arxiv.org/abs/2311.07590

Shelby, R.等(2023)《算法系统的社会技术危害：界定危害减少的分类法》。*arXiv*。https://arxiv.org/pdf/2210.05791

Shevlane, T.等(2023)《极端风险的模型评估》。*arXiv*。https://arxiv.org/pdf/2305.15324 Shumailov, I.等(2023)《递归的诅咒：在生成数据上训练使模型遗忘》。org/pdf/2305.17493v2

Smith, A.等(2023)《幻觉还是虚构?神经解剖学作为大语言模型中的隐喻》。*PLOS Digital Health*。https://journals.plos.org/Digital Health/article?id=10.1371/journal.pdig.0000388

Soice, E.等(2023)《大语言模型能否使两用生物技术的获取民主化?》*arXiv*。https://arxiv.org/abs/2306.03809

Solaiman, I.等(2023)《生成式人工智能发布的梯度：方法与考量》。*arXiv*。https://arxiv.org/abs/2302.04844

Staab, R.等(2023)《超越记忆：通过大语言模型推理侵犯隐私》。*arXiv*。https://arxiv.org/pdf/2310.07298

Stanford, S.等(2023)《语言模型反映谁的观点?》*arXiv*。https://arxiv.org/pdf/2303.17548

Strubell, E.等(2019)《NLP中深度学习的能源与政策考量》。*arXiv*。https://arxiv.org/pdf/1906.02243

白宫(2016)第A-130号通告《将信息作为战略资源管理》。https://www.White House.gov/wp-content/uploads/legacy_drupal_files/omb/circulars/A130/a130revised.pdf

白宫(2023)《关于安全、可靠和可信地开发和使用人工智能的行政命令》。https://www.White House.gov/briefing-room/presidential- actions/2023/10/30/executive-order-on-the-safe-secure-and-Trustworthy-Development-and-use-of- Artificial-Intelligence/

白宫(2022)《研究人员信息完整性研究与开发相关优先事项路线图》。https://www.White House.gov/wp-content/uploads/2022/12/Roadmap- Information-Integrity-RD-2022.pdf?

Thiel, D.(2023)《调查发现AI图像生成模型在儿童虐待内容上训练》。斯坦福网络政策中心。https://cyber.fsi.stanford.edu/news/Investigation-finds-ai-image-Generation-models- trained-child-abuse

Tirrell, L.(2017)《有毒言论：走向话语危害的流行病学》。*Philosophical Topics,45(2)*,139-162。https://www.jstor.org/stable/26529441

Tufekci, Z.(2015)《超越Facebook和Google的算法危害：计算能动性的新兴挑战》。*Colorado Technology Law Journal*。https://ctlj.colorado.edu/wp- content/uploads/2015/08/Tufekci-final.pdf

Turri, V.等(2023)《为什么我们需要了解更多：探索AI事件记录实践的现状》。*AAAI/ACM*人工智能、伦理与社会会议。https://dl.acm.org/doi/fullHtml/10.1145/3600211.3604700

Urbina, F.等(2022)《人工智能驱动的药物发现的两用性》。*Nature Machine intelligence*。https://www.nature.com/articles/s42256-02

Wang, X.等(2023)《微调BERT的能源与碳考量》。*ACL Anthology*。https://ACL Anthology.org/2023.findings-emnlp.607.pdf

Wang, Y.等(2023)Do-Not-Answer：用于评估大语言模型防护措施的数据集。*arXiv*。https://arxiv.org/pdf/2308.13387

Wardle, C.等(2017)信息失序：迈向研究与政策制定的跨学科框架。欧洲委员会。https://rm.coe.int/Information-disorder-toward-an-interdisciplinary- framework-for-researc/168076277c

Weatherbed, J.(2024)网络喷子用泰勒·斯威夫特的逼真AI伪造图像淹没了X平台。*The Verge*。https://www.theverge.com/2024/1/25/2

Wei, J.等(2024)大语言模型中的长文本事实性。*arXiv*。https://arxiv.org/pdf/2403.18802

Weidinger, L.等(2021)语言模型带来的伦理与社会危害风险。*arXiv*。https://arxiv.org/pdf/2112.04359

Weidinger, L.等(2023)生成式人工智能系统的社会技术安全评估。*arXiv*。https://arxiv.org/pdf/2310.11986

Weidinger, L.等(2022)语言模型所构成风险的分类法。*FAccT '22*。https://dl.acm.org/doi/pdf/10.1145/3531146.3533088

West, D.(2023)人工智能对女性构成不成比例的风险。*Brookings*。https://www.brookings.edu/articles/ai-poses-disproportionate-risks

Wu, K.等(2024)大语言模型引用相关医学文献的效果如何?一个评估框架与分析。*arXiv*。https://arxiv.org/pdf/2402.02008

Yin, L.等(2024)OpenAI的GPT是招聘人员的梦想工具。测试显示存在种族偏见。*Bloomberg*。https://www.bloomberg.com/graphics/2

Yu, Z.等(2024年3月)别听我的：理解与探索大语言模型的越狱提示。*arXiv*。https://arxiv.org/html/2403.17336v1

Zaugg, I.等(2022)数字弱势语言。*Policy Review*。https://Policy Review.info/pdf/Policy Review-2022-2-1654.pdf

Zhang, Y.等(2023)人类偏好，而非AI厌恶：人们对生成式人工智能、人类专家及人-GAI协作在说服性内容生成中的感知(与偏见)。*Judgment and Decision Making*。https://www.cambridge.org/core/journals/judgment-and-decision- making/article/human-favoritism-not-ai-aversion-peoples-perceptions-and-bias-toward-generative-ai-human-experts-and-humangai-collaboration-in-persuasive-content-generation/419C4BD9CE82673EAF1D8F6C850C4FA8

Zhang, Y.等(2023)AI海洋中的塞壬之歌：大语言模型幻觉综述。*arXiv*。https://arxiv.org/pdf/2309.01219

Zhao, X.等(2023)面向AI生成文本的可证明鲁棒水印。*Semantic Scholar*。https://www.Semantic Scholar.org/paper/Provable-Robust-Watermarking-for-AI-Generated-Text-Zhao-Ananth/75b68d0903af9d9f6e47ce3cf7e1a7d27ec811dc

这份文件是怎么做出来的

本译本是《生成式人工智能风险管理概况（AI 600-1）》的中文全译本，共 64 页。表格、图注、页眉页脚、编号层级都按原文位置一一还原，未做任何删节。

你手上如果也有英文文件要处理

- 合同、标书、标准、专利、报告：翻译后保留原版式，0.15 元/页
- PDF 转 Word：100 页内免费，超出 0.1 元/页
- 扫描件、图片版 PDF：先 OCR 再转 Word，免费
- 不用注册、不用登录，传上去直接下载

打开下面这个地址，把你自己的文件传上去试试：

https://www.ivyivy.com.cn/nist-ai-600-1-zh

支持私有化部署（需按贵方模型与合规要求评估）

扫码关注服务号
咨询定制交付

本译本由机器翻译生成，仅供阅读参考；专业技术判断与合规引用，请以英文原文为准。

本页为译制说明，不属于原文件内容。