生成式人工智能悖论：生成式人工智能与信任侵蚀、信息核验的腐蚀及真相的消亡

Emilio Ferrara[1, 2, 3, *]

[1] 南加州大学托马斯·洛德计算机科学系[2] 南加州大学安嫩伯格传播学院，美国加利福尼亚州洛杉矶[3] 南加州大学信息科学研究所（ISI），美国加利福尼亚州玛丽娜德尔雷

*通讯邮箱：emiliofe@usc.edu

arXiv:2601.00306v1 [cs.CY] 1 Jan 2026

摘要

生成式人工智能（生成式人工智能）如今能够以极低的边际成本大规模生成文本、图像、音频和视频，其感知上的逼真程度令人信服。尽管公共讨论常将相关危害界定为“深度伪造”，或视为错误信息与欺诈的渐进式延伸，但这种观点忽视了更广泛的社会技术转变：生成式人工智能能够创造合成现实；在这些连贯、可交互且可能个性化的信息环境中，内容、身份与社会交互共同生成、彼此强化。我们认为，最重大的风险不仅在于制造孤立的合成产物，更在于随着合成内容、合成身份和合成交互日益易于生成且难以审计，共享认知基础和机构核验实践逐步遭到侵蚀。本文（i）将合成现实形式化为一种分层架构（内容、身份、交互、机构）；（ii）拓展生成式人工智能危害分类，涵盖个人、经济、信息和社会技术风险；（iii）阐述生成式人工智能带来的定性转变（成本骤降、吞吐量、定制化、微细分、来源缺口和信任侵蚀）；（iv）综合近期（2023—2025年）的风险实例，构建精简的案例库，展示这些机制如何体现于欺诈、选举、骚扰、文档和供应链受损等领域。随后，我们提出一种缓解架构，将来源溯源基础设施、平台治理、机构工作流程再设计和公众韧性视为相互补充、不可相互替代的要素，并概述一项聚焦于衡量认知安全的研究议程。最后，我们以生成式人工智能悖论作结：随着合成媒体无处不在，社会可能会理性地对数字证据整体打折，从而提高日常生活以及民主和经济机构获取真相的成本。

关键词：人工智能；生成式人工智能；信息核验；认知安全

1 引言

生成式人工智能（生成式人工智能）技术拥有重塑世界及我们对现实的感知的空前潜力。这些技术能够增强传统上以人为中心的能力，例如在社会技术情境中的创造力和复杂问题解决能力 [7]。通过促进人机协作，生成式人工智能可以提升生产力、消除不同能力群体和文化之间的沟通障碍，并在全球范围内推动创新 [11]。然而，专家和公众对生成式人工智能的影响存在严重分歧。一方面，支持者强调它在教育、医疗保健、无障碍和创意劳动领域带来的变革性益处；另一方面，批评者警告称，同样的生成能力也可能被武器化，在社会层面大规模散布虚假信息、实施欺诈、骚扰和操纵。与以往的技术变革一样，



图1：（左上）2024年1月，Reddit 上的 *r/StableDiffusion* 社区展示了一个概念验证工作流程，用于合成生成虚构人物角色和（左下）身份证明。（右上）生成式人工智能可以生成栩栩如生、描绘从未发生过事件的图像（MJv5 提示词：“拜登总统和伊朗最高领袖握手”）。（右下）生成内容中的潜意识信息（视觉错觉读作 服从）。

核心问题并非危害是新的，而是有害行为的成本、速度、个性化程度和影响范围可能改变信任与欺骗之间的平衡。

本文强调了一种更隐蔽、却可能更危险的风险：生成式人工智能系统可能促成个性化合成现实（根据个人的偏好、脆弱之处和先验认知量身定制的世界表征），从而从根本上改变共同现实的构成。当具有说服力的合成内容嵌入交互系统（聊天机器人、个性化信息流、多智能体拟像）时，个体可能不只是接触虚假信息；他们还可能置身于连贯、引发情感共鸣且得到社会强化的叙事之中，从内部难以证伪。长远来看，危险在于社会可能变成这样：(i) 共同证据稀缺；(ii) 机构面临不断攀升的核验成本；(iii) 分歧在认识论上变得无法化解，因为各方都能拿出彼此矛盾的世界的“可信”合成文档。

为具体说明这一转变，图1展示了这一问题的三个互补侧面：(i) 合成身份（一种生成“身份凭证”的概念验证工作流程）、(ii) 描绘从未发生过的政治事件的合成内容，以及生成图像中可能嵌入的显性或潜意识引导线索。这些例子共同支撑了本文的核心论点：生成式人工智能的风险并不局限于孤立的伪造产物，而是延伸至连贯、交互式且个性化的

合成现实，这些现实可能侵蚀日常生活和机构的认知基础。

1.1 贡献与范围

本文提出四项贡献。第一，本文将<style id='1'>合成现实明确为一种分层架构（内容、身份、交互、机构）。第二，本文拓展了生成式人工智能风险与危害分类，强调合成媒体成为系统性认知脆弱性的路径。第三，本文解释了与早期欺骗技术（例如 Photoshop）相比，生成式人工智能在性质上发生了哪些变化，并将这些转变与当代风险实例联系起来。第四，本文提出了一个围绕测量、来源溯源和制度重构的缓解架构与研究议程。

1.2 生成式人工智能风险与危害分类

这些担忧的核心，是近期在[17]中提出的生成式人工智能风险与危害分类。该分类体系确定了常见的滥用途径（例如宣传、欺骗），并将其与可能造成的危害对应起来。

个人损失这一类别涵盖对个人造成的危害，包括未经同意生成的合成图像、冒充、定向骚扰、诽谤、隐私侵犯以及心理困扰。生成式人工智能可以通过以下方式放大这些危害：(i) 降低制作具有说服力的人身攻击内容的技能门槛；(ii) 提高攻击者迭代内容的速度；以及 (iii) 实现持续个性化（例如，对手生成根据目标的身份、人际关系和触发因素进行调整的内容）。除了突发事件之外，反复接触逼真的伪造内容还可能导致长期不信任、社交退缩，以及一种“现实疲劳”——在这种状态下，个人会逐渐脱离公民生活和信息生活 [29, 35]。

金融与经济损害 生成式人工智能通过扩大传统欺诈的规模，并促成更具说服力的欺骗（例如在高风险交易中实施逼真的语音／视频冒充），加剧经济威胁。它还带来了新的成本：企业必须在验证真实性、合规、来源溯源和事件响应方面进行投资；新闻编辑部必须在更紧迫的时间要求下核验内容；而随着数字证据失去默认可信度，日常交易也会增加摩擦。这些影响给商业和治理带来了“认知税”[28]。

信息操纵 生成式人工智能可以构建虚假但令人信服的叙事，伪造证据，并用貌似可信的内容淹没信息渠道，从而压垮注意力和编辑能力[18]。至关重要的是，操纵已不再局限于制作产物；它还可以通过互动式 说服来实施，在这种方式下，系统会根据个人的信念和情绪状态实时调整策略。这使信息完整性威胁从“内容真实性”转向“信念形成安全”，引发了对公共话语、新闻业和民主协商的担忧。

社会技术与基础设施风险 最后，生成式人工智能可能通过侵蚀人们对机构的信任，并使策略性行为者能够大规模利用平台可供性开展协调一致的影响力行动，从而破坏社会技术系统的稳定。这些危害包括机构合法性丧失、两极分化和治理瘫痪，以及模型供应链入侵（投毒、后门）、不安全部署，以及在关键工作流程中滥用生成式工具等技术风险。在极端情况下，合成现实会成为威权控制的工具：随意编造文档、证词和“公众共识”的能力。

认知与制度诚信 我们建议将认知诚信明确列为一种贯穿各领域的危害类别。当合成内容无处不在且核验成本高昂时，社会可能会滑向两种失效模式之一：（i）轻信（相信看似可信的伪造内容），或（ii）犬儒主义（什么都不相信，包括真实证据）。这两种结果都会削弱问责，并使策略性行为者得以利用混乱、拖延或可信否认。这一维度将分类体系与制度设计联系起来：法院、选举、新闻业和金融都依赖证据制度，而这些制度以真实性和可审计性达到基本水平为前提。

2 从合成媒体到合成现实

关于生成式人工智能风险的讨论，往往聚焦于合成媒体：模仿人类创作产物（文本、图像、音频、视频）的输出。我们认为，更具深远影响的转变是迈向合成现实：一个连贯的环境，其中产物、身份和互动部分由机器生成，并相互强化，以至于身处其中的人会觉得这一切可信。换言之，核心风险不仅在于单条内容可能遭到伪造，还在于整个信念语境 都可能被制造出来：谁在场、有哪些证据、哪些主张在传播，以及它们如何获得社会认可。如图2所示，我们因此将合成现实定义为一种分层架构（内容、身份、交互、机构），它自然对应着不同的攻击面和防御手段。

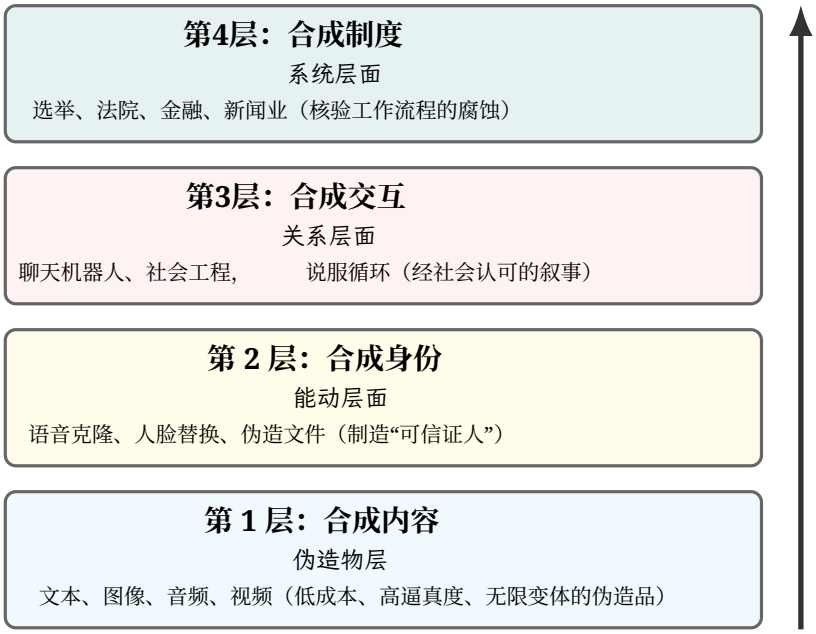


图2：**合成现实作为分层架构**。生成式系统首先生成合成内容（文本、图像、音频、视频），进而实现合成身份（冒充/人设伪造）和合成交互（自适应、具有社会临场感的对话）。这些层次可以相互强化，放大可信度和说服力，同时将核验负担转移给机构（例如新闻业、法院、选举和金融机构），因为高可信度制品变得廉价且大量涌现。这一结构说明，低成本产物如何升级为系统性认知失灵。

2.1 第一层：合成内容

这一层包括生成或编辑的文本、图像、音频和视频。其显著特征是高逼真度、快速制作和低边际成本，其中包括迭代变体（大量看似可信的替代版本）和合成编辑（针对并修改特定细节，同时保持整体可信度的能力）。单独来看，这一层会助长人们熟悉的危害（伪造、诽谤、宣传），但与身份和互动结合后，其影响力会大幅增强。图1（右上）展示了这一基础能力：以低成本生成看似可信、描绘从未发生过事件的图像。更隐蔽的是，合成内容不仅可以针对虚假信息进行优化，也可以用来引导注意力和解读方式（图1，右下），模糊说服与感知之间的界限。

2.2 第二层：合成身份

当对手能够令人信服地呈现人物，而不只是内容时，合成现实便会扩张。如图1（左）所示，生成式人工智能能够伪造身份关联伪造物及其佐证“证据”，攻击身份认证与归因的基础。身份合成包括声音克隆、面部重演、文件伪造，以及令人信服的社交媒体人设。一项主要风险是制造“可信证人”：这些账号在时间、平台和模态上看起来都前后一致。这会直接冲击现代安全与信任所依托的凭证生态（例如注册、身份验证和授权工作流），因为曾经可作为高门槛证明的信号（声音匹配、看似可信的文件、一贯的人设）如今都变得易于低成本生成，也容易被再创作。

2.3 第三层：合成交互

交互系统可以模拟对话和社会临场感。这使得说服能够根据目标对象调整，营造社会认同，并维系长期的操纵性关系（例如情感诈骗、激进化路径、胁迫性骚扰）。合成交互之所以重要，是因为信念的形成并非被动取决于产物；它会受到交谈、反馈和社会强化的影响。与静态错误信息相比，交互代理能够：（i）探查不确定之处；（ii）使论点和情感框架个性化；（iii）随着时间推移逐步强化对方的投入。从这个意义上说，合成交互将合成内容与身份转化为切身体验，把孤立的产物变成连贯的叙事，使其在社会和情感层面都显得得到验证。

2.4 第四层：合成机构

在制度层面，合成内容与身份会干扰依赖证据和来源溯源的流程：选举（竞选信息传播和选民联络）、法院（证据真实性和证人可信度）、金融（授权和欺诈）以及新闻业（在时间压力下进行核验）。核心问题在于，许多机构工作流程都是针对这样一个世界进行优化的：某些证明难以伪造，因而具有信息价值。合成现实削弱了这一假设，因为高可信度制品变得既廉价又充足。具体而言，它给依赖以下因素的核验机制带来压力：（i）将感知线索作为身份认证依据（“我认得这个声音／这张脸”）；（ii）将文档视为稀缺证据（“这些文件看起来是真的”）；（iii）将注意力视为瓶颈（“人可以审核重要项目”）；以及（iv）将共同接触作为纠正的基础（“如果这是假的，反驳信息会传达到同一批受众”）。因此，合成现实不只是媒体问题，也是制度韧性问题。综合来看，这些层面说明了为何最好将合成现实视为一种系统性风险：内容的逼真度只是切入点；身份与交互带来可信度和势头；而当核验必须在时间压力下进行时，外部性则由机构承担

以及对抗性适应。在下一节中，我们将详细阐述生成式人工智能带来的定性变化（成本骤降、规模、个性化、来源缺口和信任侵蚀），以及这些变化如何使这一层叠体系产生实际影响。

3 无法分辨的事物可能伤害你：生成式人工智能如何改变局势

有人或许会认为，我们分类体系中所列的危害并非生成式人工智能独有；欺诈、宣传和诽谤早在现代人工智能出现之前就已存在。关键变化不在于欺骗的存在，而在于生成式人工智能如何改变其经济成本与实际可行性。当合成内容、合成身份和合成交互变得廉价、快速且可规模化时，个人、平台和机构之间的信任与核验平衡便会改变。我们将这些变化归纳为七种机制，它们共同说明了为何合成现实构成了更高程度的社会技术风险，而不只是“更多同类”错误信息。

3.1 成本骤降与商品化

生成式人工智能降低了制作逼真内容和身份伪造物的门槛。过去需要专业技能（视频剪辑、平面设计、配音表演或复杂的社会工程）的工作，如今通过提示词、模板和交钥匙服务就能完成。成本骤降扩大了有能力的对手群体，缩短了从意图到执行的时间，还使快速试验成为可能：攻击者可以反复迭代，直到输出对目标受众“看起来恰到好处”。由此形成的威胁态势更像一个商品化的服务生态系统，而非彼此孤立、量身定制的攻击。

3.2 规模与吞吐量

生成式人工智能支持快速迭代和批量生产。对手可以生成数千种消息、图像或人物设定的变体，并通过 A/B测试检验其互动率、可信度或情感影响。这使防御问题从识别少量伪造内容转变为管理大量看似可信的不同版本；协同网络往往通过多个平台和账户传播这些内容，以系统性地推广党派叙事或放大低可信度内容 [30, 9]。吞吐量还助长了“信息洪泛”策略：海量合成内容会在时间压力下淹没人们的注意力、内容审核能力和新闻核查 [5, 4]。

3.3 用于恶意目的的定制化

开放模型生态系统、微调和检索增强生成使领域特定欺骗成为可能。内容可以模仿特定机构、社群或个人的风格；诈骗可以仿照内部工作流程；说服性叙事则可以根据文化背景和时事进行本地化。风险不仅在于逼真程度，更在于契合度：欺骗内容越符合目标对象的语言、规范和预期，就越有效，也越能减少原本会引发怀疑的线索。

3.4 超精准说服与微细分

生成式人工智能使大规模个性化影响力行动成为可能。对手无需向数百万人传播同一种叙事，而是可以为细分群体精心打造多种叙事，分别针对不同的恐惧、身份或不满进行优化。微细分会削弱集体反驳：不同社区甚至可能看到不同的主张，而更正也可能无法触及受影响最深的受众。久而久之，这可能会加剧两极分化，将不同群体引向彼此不相容的解读框架，

每种框架都由看似可信的合成“证据”和量身定制的信息所支撑。

3.5 合成交互与社会工程的自动化

传统的社会工程需要大量人力。借助生成式人工智能，交互本身可以实现自动化[15]。对话代理能够建立融洽关系、探查不确定性、调整语气，并引导受害者一步步落入欺诈陷阱；它们还可以维持更长期的操纵性关系，随着时间推移，将情感支持、选择性证据和不断升级的承诺融合在一起。这使欺骗从静态的内容问题转变为交互过程问题：信念的形成和决策可以通过对话、反馈循环和模拟社会存在感受到影响。

3.6 检测的局限、水印的脆弱性与来源缺口

对生成式人工智能输出内容的检测仍不完善，尤其是在压缩、重新编码、对抗性扰动和跨模型混编的情况下。水印技术在受控流程中有所帮助，但在开放生态系统中，水印可能被移除、削弱或绕过；此外，它也无法处理在水印机制之外生成的未经认证媒体。更深层的问题是来源缺口：即使有些伪造内容可以被检测出来，机构仍需要可靠的保管链以及经过认证媒体的标准。在实践中，高风险决策所需的不只是概率分类器，还需要可审计的来源信号和流程层面的保障措施。

3.7 信任侵蚀与可信否认

随着合成输出日益普及，社会面临一种额外的失效模式：真实证据可能被斥为伪造。这一动态提高了策略性否认和拖延的收益，使行动者即使面对属实的指控，也能利用不确定性。其结果是认知不稳定性：由于证据基础本身受到质疑，争议变得更难解决，受众也可能理性地默认怀疑。总体而言，这会产生一种“认知税”：要确定发生了什么、谁说了什么，以及哪些来源值得信任，需要付出更多精力和成本。

3.8 综合

这些机制相互强化。成本骤降和吞吐量增加了合成产物的数量；定制化和微细分提升了它们的说服契合度；合成交互将产物转化为亲历且经过社会验证的体验；而来源缺口则使大规模可靠核验变得困难。其后果是，压力从孤立的欺骗事件转向对信任和核验实践的系统性冲击。接下来，我们将结合近期各领域的具体风险实例，阐释这些机制。

4 风险如何显现：代表性风险实例

为将分类体系置于近期发展背景下，我们汇编了一个精简的“案例库”（表1），展示合成现实风险如何已在实际应用场景中显现。我们的目标并非面面俱到，而是追求机制多样性：我们选取的案例（i）有高质量的公开报道和／或初选来源作为佐证，（ii）涵盖不同的危害领域（欺诈、选举、骚扰、文档和供应链受损），且（iii）与合成现实技术栈中的一个或多个层面（内容、身份、交互、机构）有明确关联。我们还指出一个重要局限：各地区和行业的公开文档并不均衡

（许多事件通过非公开方式处理或未得到充分报道），因此，下文案例应被视为说明性的危害下限，而非完整的普查。

4.1 代表性风险实例（2023—2025年）

表1总结了机制与危害路径；此处我们简要叙述各案例，以强调这些并非假设性的失效模式，而是已观察到的事件，其共同结构相同：生成式人工智能降低了制作高可信度制品的成本，进而利用那些为“此类制品伪造代价高昂的世界”而设计的社会与机构工作流程。

案例类别A：企业工作流程中的高可信度冒充欺诈。 2024年初，香港警方描述了一起欺诈案件：一名员工被引入一场看似与高层同事举行的保密视频会议，随后授权向多个账户进行大额转账。据报道，作案手法将网络钓鱼诱饵说辞与利用公开可得的视频和语音材料拼装而成的伪造群体视频会议相结合；一旦受害者认定会议真实，攻击者便转入熟悉的操作常规（后续指示、审批与转账），将感知到的合法性转化为即时的财务行动。此后公开报道确认了受害机构及损失的大致规模，凸显出：当核验被隐性委托给对面部/语音线索的感知真实性时，即便一场高可信度“会议”也能击破标准内部控制 [23, 19, 36]。（机制摘要见表1，案例类别A。）

案例类别B：选举相关的合成式外展与选民操纵。选举情境中也出现了类似模式：合成式外展利用熟悉声音和机构性线索的可信度，大规模制造混乱。在新罕布什尔州初选背景下，据广泛报道，人工智能生成的自动拨号电话模仿一位知名政治人物的声音，并配合选举日程发送削弱投票动员的信息。后续执法行动体现了一项关键的制度转变：应对措施针对的不仅是内容[8], ，还包括承载这些内容的通信基础设施（例如，供应商层面的合规、拦截和处罚）。这一案例代表了生成式人工智能如何将“是谁说的”变成一个可编程变量，从而以低成本开展定向影响行动，而其更正往往比最初曝光传播得更慢、也更不显眼[13, 1, 31]。（见表1，案例类别B。）

案例类别C：非经同意的合成色情图像，成为由平台放大且持续存在的骚扰。合成现实风险并不局限于欺诈和政治领域；它也表现为可规模化、针对特定身份的骚扰。2024年初，一位高知名度公众人物的露骨人工智能生成深度伪造图像在主要平台上迅速传播，至少有一个平台因此实施了临时搜索限制，以减缓发现和再次分享的速度。这一事件体现了此类伤害的两个持久特征：（i）受害者的身份成为可反复利用的施害资产（面孔和肖像很容易被置于新的语境中）；（ii）平台的应对往往发生在广泛曝光之后，而副本、衍生内容和重新上传内容使这种伤害持续存在。美国的政策应对措施（包括针对受规制平台的联邦通知与移除规定）进一步表明，非经同意的合成私密影像如今已被视为主流网络安全威胁，而非边缘案例[2, 37, 10, 3]。（见表1，案例类别C。）

案例类别D：伪造文件与日常核验的侵蚀。一个影响更广泛、但在结构上至关重要的发展，是看似可信的“日常文件”日渐增多，令快速人工检查难以奏效：收据、发票、屏幕截图、电子邮件，

以及其他作为日常信任基础的文书。2025年的报道显示，人工智能生成的伪造收据及相关费用欺诈明显增加；逼真的外观（标识、字体、磨损痕迹、明细）加上元数据操纵，使视觉审核难以奏效。业界的应对措施日益将此视为运营模式的转变，而非异常情况：防御方式从人工抽查转向自动化检测、元数据取证，以及根据情境约束进行交叉验证。其系统性风险在于，随着伪造文件变得廉价且泛滥，机构可能转向默认怀疑，增加守规参与者的办事阻力与错误成本，同时为策略性否认（“那份证据可能是假的”）提供掩护 [20, 34, 26, 33, 21]。（见表1，案例类别D。）

案例类别E：生成式管线遭到破坏与模型供应链风险。合成现实也可以在上游塑造，涉及模型制品和生成基础设施。2025年安全报告记录了恶意机器学习模型上传至某大型模型共享平台的事件，这些模型旨在加载时执行恶意软件；这是一种经典的供应链模式，如今通过分发“作为文件的模型”呈现出来 [32]。相关安全公告指出，扫描器假设可能遭到绕过，使嵌入式载荷得以规避安全防护措施，同时仍可由标准工具加载 [22]。在另一个层面，关于后门和数据投毒的技术研究表明，有害行为可以被蓄意植入，并可能在常见的训练或对齐流程中持续存在，从而带来一种风险：下游用户感受到的是“系统”本身，而非明显的攻击 [25, 24]。综合来看，这些案例表明，应将生成式人工智能视为一条端到端的社会技术管线，其完整性取决于来源溯源、稳健扫描、安全加载默认设置和持续监控，而不仅仅是模型能力或内容过滤器 [6]。（参见表1，案例类别E。）

4.2 跨案例综合分析

表1所列案例的共同点不仅在于伪造制品的存在，更在于通过内容 + 身份+ 交互 + 机构 范式构建出可信的替代情境。每个事件都遵循一种反复出现的运营模式：(1) 以低成本制作高可信度制品（一通电话、一段视频、一份文件、一个虚拟人物、一份模型文件）；(2) 将其植入工作流程的“瓶颈环节”，在此处，人们往往会依赖信任线索；(3) 自动化和规模扩大增加了接触或重复传播的机会；(4) 纠正滞后于最初造成的影响；(5) 机构则承受由此产生的外部性，包括核验负担、摩擦和有争议的证据。这一综合分析为下一节的缓解框架提供了依据：要减少危害，需要通过分层干预强化工作流程和来源溯源，而不能只依靠更好地检测个别伪造内容。

5 将缓解措施构建为一套体系（而非灵丹妙药）

前文各节引出了一个核心设计原则：合成现实风险无法通过任何单一工具或政策来“解决”。单靠检测，在对抗性适应面前十分脆弱；单靠水印，在开放生态系统中并不完整；单靠内容下架，则往往要等到损害周期后期才会发生。因此，缓解措施必须被视为一套相互补充的干预措施，既要减少（i）有害合成内容的生成与传播，也要降低策略性利用的动机与机会。本节将概述一种分层缓解方案，涵盖来源溯源基础设施、平台治理、机构工作流程再设计、公众韧性和问责。

5.1 来源溯源与内容真实性基础设施

来源证明系统（例如密码学签名、安全采集和内容凭证）旨在为媒体建立保管链，并向下游用户传递真实性信号。当高风险机构（新闻编辑部、选举机构、法院、公共机构）端到端采用这些系统，并将其整合到常见的消费者界面中，使身份认证成为常规而非例外时，这些系统的价值最大。然而，来源溯源并非普适方案：如果内容是在经过认证的流程之外生成、设备在采集时遭到入侵、元数据在传输过程中被剥除，或受众无法使用便捷的验证工具，来源溯源就无济于事。因此，应将来源溯源视为一种基础设施，用于提高经过认证的媒体的可审计性并减少歧义，而非确保假内容彻底消失。一个实际意义在于，这种作用是不对称的：即使来源溯源无法为所有未经验证的内容贴上标签，也能显著提升人们对已验证内容的信心。这意味着应优先在真实性至关重要的类别中推行来源溯源（官方公告、高传播新闻、选举信息、证据媒体、金融授权），同时承认互联网未经认证的“背景内容”仍将杂乱无章。

5.2 平台治理与病毒式传播阻力

平台可以通过塑造信息接触范围和激励机制来减少危害。干预措施包括：（i）在高风险时期（选举、突发新闻、危机期间）限制未经验证的媒体的算法放大，因为推荐算法可能会不成比例地放大党派性内容，并扭曲政治信息接触 [38]；（ii）突出展示来源溯源和不确定性信号[14]；（iii）限制新创建的账户和协同不真实行为的操作频率[27]；以及（iv）为冒充和骚扰的受害者提供快速应对渠道[12]。重要的是，治理必须预见对抗性适应：政策会成为被针对的目标，而执行必须是迭代式且以证据为依据的。在合成现实状态下，“阻力”成为一种正当的安全工具。通过增加少量操作成本，例如为高速分享增加额外的验证步骤、对可疑爆发式传播实施临时传播限流措施，或延缓缺乏来源证明的媒体的病毒式传播，可以在不要求对真实/虚假内容进行完美分类的情况下，有效减少危害。

5.3 低成本伪造背景下的机构流程再设计

机构应基于合成身份和伪造证据成本低廉这一假设，重新设计工作流程。这包括：对高价值转账进行带外验证；采用不依赖声音／面部线索的多因素审批协议；通过经认证的渠道发送官方通信；以及将来源证明元数据和保管链纳入证据标准。更广泛地说，机构应从“基于产物的信任”（相信看起来真实的东西）转向“基于流程的信任”（相信通过经过认证且可审计的程序生成和传递的内容）。这一转变也关乎公平：当核验负担加重时，无法使用特权渠道（法律顾问、专业工具、经过验证的身份）的人可能会受到不成比例的伤害。因此，制度重构应致力于提供便于使用的认证途径，并为有争议的证据制定明确的升级处理程序。

5.4 公众韧性与认知卫生

社会需要超越“识别假信息”的规范与素养。不能指望个人验证自己看到的一切，也不应将举证责任完全转嫁给用户。增强公众韧性，应着重于：（i）在高风险情境下保持适度怀疑；（ii）依赖关键信息认证渠道；（iii）意识到操纵手段（紧迫感、情绪触发因素、社会认同、冒充）；以及（iv）

建立社区层面的纠正机制，将经核实的反驳传播到合成内容主张扩散的同一网络中。一个有用的重新界定方式是“认知卫生”：养成能够减少接触操纵、增加对可信流程依赖的习惯，这类似于公共卫生措施降低疾病暴露风险，而不要求个人具备完美的诊断能力。

5.5 政策与问责

政策可以在保护生成式人工智能的正当创作和辅助用途的同时，明确责任归属与激励机制。干预措施包括：要求披露合成政治广告和自动化外联信息；规定须对冒充行为及未经同意的私密影像迅速作出回应；以及在敏感领域设定最低身份认证标准（例如选举管理通信、高价值金融授权）。其目标不是审查，而是问责：提高部署合成现实攻击的成本，增加归因和制裁的可能性，并降低模糊其辞和否认所带来的收益。

5.6 综合：将缓解措施映射至合成现实堆栈 k

由于合成现实以堆栈形式运作，缓解措施应映射到各个层级和失效模式。来源溯源主要通过实现保管链来强化内容 层和制度 层；平台治理可减少加剧交互 层的放大效应和协同滥用；制度重构可增强工作流程抵御身份 欺骗和伪造文件的能力；公众韧性则可降低对说服循环的易感性。这一映射凸显了为何单一杠杆无法奏效：系统可能在任何层级失效，而对手会绕过最强的防御。因此，现实可行的目标是通过纵深防御降低风险，同时结合测量与监测，以发现堆栈在实践中失效的环节。

6 尚待解决的问题与研究议程

如果合成现实是一种系统性风险，那么缓解风险就需要进行测量。我们无法管理未曾测量的事物，而当前的信息完整性工具主要针对静态产物进行优化，并不适用于融合了内容、身份、交互和制度工作流程的多层环境。本节提出一项研究议程，重点关注操作性指标、交互式操纵基准、稳健的来源溯源评估，以及低成本伪造环境下的制度设计。图3为研究人员和从业者提供了一个框架，助力他们为认知安全奠定科学与工程基础：使社会技术系统能够在对抗性压力下维持共同现实并实现可问责的决策。

6.1 认知安全的测量

我们需要用于衡量社区、平台或机构在合成现实条件下维持共同现实能力的操作性指标。候选构念包括：(i) 真实性覆盖率（携带可核验来源溯源的高触达媒体所占比例）；(ii) 纠正延迟（就真实性达成社区层面共识所需的时间，包括经核验的反驳信息在受影响网络中的触达范围）；(iii) 操纵易感性（在受控的合成干预下，互动率、信念或行为发生变化的程度，同时采取安全防护措施并进行伦理审查）；(iv) 核验负担（验证真实性主张所需的资源和摩擦成本，包括给合法行为者造成的误报负担）；以及 (v) 归因稳定性（随着对手不断适应，系统可靠地将内容和主张追溯至来源的程度）。这些指标结合起来，可以量化合成现实造成的“认知税”，并确定防御性投入在哪些方面能够带来最大的边际收益。

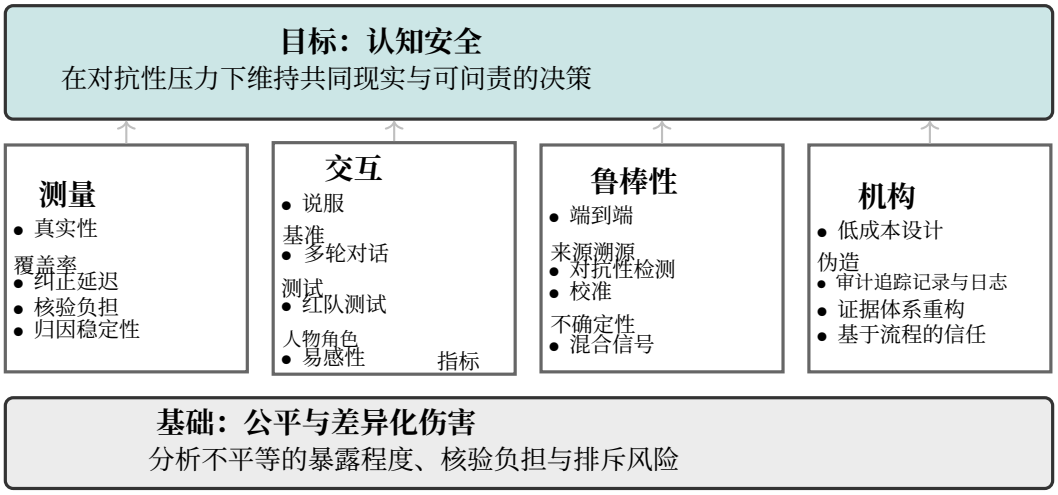


图3：**认知安全研究议程**。应对合成现实风险，需要从衡量人工制品真实性转向衡量系统韧性。该议程立足于四大支柱：（1）操作性指标；（2）交互式基准；（3）对抗鲁棒性；（4）制度重构，所有支柱均以公平与差异化伤害分析为基础。

6.2 交互式操纵基准

现有大多数基准都侧重于静态模型输出。合成现实需要针对交互式说服的基准：多轮对话、自适应定向、关系持久性以及社会认同的生成。我们需要实验方案，测试对话代理如何在现实约束下引导信念和决策，包括：(i) 对个人数据的有限访问；(ii) 类似平台的交互限制；(iii) 对抗性目标（欺诈、去动员、骚扰）；以及 (iv) 可衡量的结果变量（例如，陈述信念的变化、分享意愿、对请求的合规程度）。由于这项研究涉及人类易感性，因此需要严格的伦理保障措施、透明度和监督。一个有前景的方向是，利用合成人物或自愿参与者开展受控的“红队”评估，并结合事后审计，以识别会导致有害结果的对话策略。

6.3 来源溯源与检测的对抗鲁棒性

来源证明系统必须作为端到端的社会技术系统进行评估：评估的不仅是密码学鲁棒性，还包括可用性、采用激励、失效模式和攻击面（设备受损、元数据剥离、重新编码及跨平台劣化）。检测应被视为概率性证据，而非二元判定器。这提出了三项研究重点：（i）经过校准并能报告不确定性的检测器，以满足机构决策的需要；（ii）在真实的变换和对抗性扰动下进行严格评估；（iii）混合方法，将较弱的内容信号与较强的流程信号（采集完整性、签名、传输日志）相结合。归根结底，最有价值的输出并非“伪造／非伪造”，而是在可能的情况下，提供与决策相关的置信度，并辅以透明的来源溯源信息。

6.4 低成本伪造环境下的制度设计

法院、选举、新闻业和金融领域都需要重新设计证据制度。研究应确定，在低成本伪造面前，哪些制度假设会失效，并提出即使感知线索和文档不再天然具有可信度，也依然稳健的流程。可选

方向包括认证捕获标准、可验证日志记录和审计追踪记录、标准化的不确定性披露，以及针对存在争议的合成证据制定新的法律或行政程序。一项关键挑战是在安全与无障碍之间取得平衡：过于繁琐的验证流程可能排斥正当参与，或加剧不平等；而流程过于宽松则会招致操纵。

6.5 公平与差异化危害

合成现实危害不会平均分布。骚扰目标、边缘化群体，以及较难接触经认证的渠道或获得法律救济的人，可能承担不成比例的代价。这些差异往往会因生成模型本身固有的偏见而加剧；这些模型大规模部署时，可能会重现并放大社会刻板印象 [16]。研究必须量化差异化暴露、差异化验证负担以及下游危害（经济、心理和公民层面）。评估缓解措施时，不仅要考察其整体错误减少效果，还要考察它是否将负担转嫁给最无力承受的人。这包括研究来源溯源和验证工具在不同群体中的采用情况，以及“默认怀疑”机制是否会加剧歧视和排斥。

6.6 统一议程：从人工制品真实性到认知韧性

面对这些问题，统一的转向是从人工制品真实性转向认知韧性。在任何单一人工制品都可能被伪造的世界里，核心问题变成了社区和机构能否可靠地就真相达成共识、厘清责任并采取负责任的行动。要实现这一点，需要开展跨学科研究，涵盖机器学习、安全、人机交互、计算社会科学、法律和公共政策，并结合现实环境中的实证测量。核心科学挑战在于描述、预测并减少这样的情形：合成现实会导致持续分歧、机构超负荷以及共同认知基础的瓦解。

7 结论

生成式人工智能放大了人们熟悉的危害，但其更深层的风险，是逐渐构建出削弱共同认知基础的合成现实。最重大的变化不仅在于能够制造欺骗性人工制品，还在于能够伪造可信的情境：身份、对话、文档和社会认同共同作用，使虚假叙事显得内部自洽，且难以从内部证伪。随着这些能力日益普及，机构和日常生活将面临不断加重的核验负担、有争议的证据，以及日常信任关系中日益增大的摩擦。生成式人工智能悖论 在于，合成内容变得无处不在且越来越难以辨别时，社会可能会理性地转而对数字证据一概打折。这一转变抬高了求真的成本：核验成为一种特权，机构流程放缓，问责也随之减弱。它还通过可信否认（即真实证据可能被斥为伪造）以及注意力洪泛赋能策略性行为者，延迟纠正并加剧混乱。

这一悖论会带来可观察、可测量和可检验的压力。我们预计：**(i)** 对关键信息认证渠道的依赖日益增加；**(ii)** 在微细分的背景下，纠正延迟更长，反驳触达范围更加碎片化；**(iii)** 机构核验负担（时间、人员、合规）更重；**(iv)** 高关注度事件中，策略性否认主张日益增多。这些压力共同构成了治理、商业和公民生活所承担的“认知税”。

因此，应对合成现实风险需要采取纵深防御。来源溯源基础设施可以提升人们对经过认证的媒体的信心；平台可以提高病毒式传播的门槛并减少协同滥用；机构可以围绕低成本伪造重新设计工作流程；公众

韧性则可以推动规范从基于产物的信任转向基于流程的信任。我们提出的研究议程呼吁制定超越静态检测的指标和基准，迈向认知安全：在对抗性压力下维护共同现实和可问责的决策。现实的目标并非实现完美的真实性，而是建立具有韧性的核验机制，在最关键之处维护信任。

作者贡献

概念构思，E.F.；方法设计，E.F.；调查研究，E.F.；数据整理，E.F.；论文撰写——初稿，E.F.；论文撰写——审阅与编辑，E.F.；可视化，E.F。

资金支持

本研究未获得外部资金支持。

数据可用性

不适用。

致谢

作者感谢南加州大学 HUMANS实验室的历届成员和现任成员。

利益冲突

作者声明不存在利益冲突。

参考文献

[1] 美联社。人工智能生成的语音可在自动拨打电话中欺骗选民。美国联邦通信委员会刚刚宣布此类行为违法。https://apnews.com/article/a8292b1371b3764916461f60660b93e6，2024年。[2] 美联社。深度伪造色情图像引发临时封锁后，X 恢复了对泰勒·斯威夫特的搜索。 https://apnews.com/article/adec3135afb1c6e5363c4e5dea1b7a72，2024年。[3] 美联社。特朗普总统签署《TAKE IT DOWN法案》，以应对未经同意传播的深度伪造内容。该法案是什么？ https://apnews.com/article/741a6e525e81e5e3d8843aac20de8615，2025年。[4] I. Augenstein、M. Bakker、T. Chakraborty、D. Corney、E. Ferrara、I. Gurevych、S. Hale、E. Hovy、H. Ji、I. Larraz 等人。社交媒体上的社区内容审核与事实核查的新认识论。*arXiv* 预印本 *arXiv:2505.20067*，2025年。[5]I. Augenstein、T. Baldwin、M. Cha、T. Chakraborty、G. L. Ciampaglia、D. Corney、R. DiResta、E. Ferrara、S. Hale、A. Halevy 等人。大语言模型时代的事实性挑战与事实核查机遇。《自然·机器智能》，6(8):852–863，2024年。

[6]C. Autio、R. Schwartz、J. Dunietz、S. Jain、M. Stanley、E. Tabassi、P. Hall 和 K. Roberts。《人工智能风险管理框架：生成式人工智能概况》。https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence，2024年。

[7] 加利福尼亚州政府运营局。加利福尼亚州：生成式人工智能的益处与风险报告。https://www.govops.ca.gov/wp-content/uploads/sites/11/2023/11/GenAI-EO-1-Report_FINAL.pdf，2023年11月。

[8] Z. Chen、J. Ye、B. Tsai、E. Ferrara 和 L. Luceri。《合成政治：X 上人工智能生成的政治图像的普及程度、传播者与情感反应》。载于 第*36*届 *ACM*超文本与社交媒体会议论文集，第11–21页，2025年。

[9] F. Cinus、M. Minici、L. Luceri 和 E. Ferrara。在《2025年 ACM Web 会议论文集》中发表：《揭露 2024年美国大选前夕的跨平台协同不真实活动》，第 541–559 页，2025年。[10] 美国国会研究处。《TAKE IT DOWN法案》：一项禁止未经同意发布私密图像的联邦法律。https://www.congress.gov/crs-product/LSB11314，2025年。[11]T. T. Eapen、D. J. Finkenstadt、J. Folk 和 L. Venkataswamy。生成式人工智能如何增强人类创造力。https://hbr.org/2023/07/how-generative-ai-can-augment-human-creativity，2023年7月。[12] 欧洲性别平等研究所。打击针对妇女和女童的网络暴力。报告，欧洲性别平等研究所（EIGE），2022年11月。[13] 美国联邦通信委员会。DA 24-102：自动拨号电话执法（公告；向 Lingo Telecom 发出停止并终止通知，涉及人工智能生成的语音）。https://docs.fcc.gov/public/attachments/DA-24-102A1.pdf，2024年。[14] K. K. Feng、N. Ritchie、P. Blumenthal、A. Parsons 和 A. X. Zhang。考察支持来源追溯的媒体对信任与准确性感知的影响。《ACM人机交互论文集》，7(CSCW2)：1–42，2023年。[15] E. Ferrara。ChatGPT 时代的社交机器人检测：挑战与机遇。*First Monday*，28(6)，2023年。[16] E. Ferrara。人工智能中的公平性与偏见：来源、影响和缓解策略简述。*Sci*，6(1)：3，2024年。[17] E. Ferrara。生成式人工智能反人类：生成式人工智能和大语言模型的恶意应用。计算社会科学杂志，2024年。[18] E. Ferrara。描绘生成式人工智能用于在线选举干预的恶意用途全景。*First Monday*，30(6)，2025年。[19] 金融时报。Arup 在香港遭遇深度伪造视频会议诈骗，损失 2500 万美元。 https://www.ft.com/content/b977e8d4-664c-4ae4-8a8e-eb93bdf785ea，2024年。[20] 金融时报。“不要相信你的眼睛”：人工智能导致费用欺诈激增。 https://www.ft.com/content/0849f8fe-2674-4eae-a134-587340829a58，2025年。[21] 金融时报。诈骗者利用人工智能伪造艺术品真伪及所有权。 https://www.ft.com/content/fdfb5489-daa0-4e7e-97b7-4317514cd9f4，2025年。

[22]GitHub 安全公告数据库。PyTorch 模型文件可通过意外的 pickle 扩展绕过 pickle 扫描器（CVE-2025-1889）。https://github.com/advisories/GHSA-769v-p64c-89pr，2025年。

[23]香港特别行政区政府。LCQ9：打击涉及深度伪造的欺诈行为。
https://www.info.gov.hk/gia/general/202406/26/P2024062600192p.htm，2024年。

[24] P. He、H. Xu、Y. Xing、H. Liu、M. Yamada 和 J. Tang。用于上下文学习的数据投毒。载于计算语言学协会论文集：*NAACL 2025*，2025年。[25] E. Hubinger、C. Denison、J. Mu、M. Lambert、M. Tong、M. MacDiarmid、T. Lanham、D. M. Ziegler、T. Maxwell、N. Cheng 等。潜伏代理：训练能够在安全训练后持续存在的欺骗性大语言模型。*arXiv* 预印本 *arXiv:2401.05566*，2024年。

[26]英格兰及威尔士特许会计师协会。费用欺诈：如何识别人工智能生成的收据。https://www.icaew.com/insights/viewpoints-on-the-news/2025/nov-2025/
expenses-fraud-how-to-spot-an-ai-generated-receipt，2025年。

[27] L. Luceri、T. V. Salkar、A. Balasubramanian、G. Pinto、C. Sun 和 E. Ferrara。《TikTok 上的协同不真实行为：视频优先生态系统中的检测挑战与机遇》。载于国际人工智能促进协会网络与社交媒体会议，2026年。

[28] W. Mazurczyk、D. Lee 和 A. Vlachos。《人工智能时代的虚假信息2.0：网络安全视角》。《style id='2'>《ACM 通讯》，2024年。

[29] F. Menczer、D. Crandall、Y.-Y. Ahn 和 A. Kapadia。《应对人工智能生成的虚假内容所带来的危害》。《style id='2'>《自然·机器智能》，第 1–2 页，2023年。

[30] M. Minici、F. Cinus、L. Luceri 和 E. Ferrara。《揭示协同跨平台信息行动：威胁2024年美国总统选举的完整性》。《style id='2'>First Monday，29(11)，2024年。

[31]美国国家公共电台。一名政治顾问因拜登深度伪造自动拨打电话而面临指控和罚款。
https://www.npr.org/2024/05/23/nx-s1-4977582/
fcc-ai-deepfake-robocall-biden-new-hampshire-political-operative，2024年。

[32] OWASP 生成式人工智能安全项目。OWASP 生成式人工智能事件与漏洞利用综述，2025年1月至2月（Hugging Face Hub 上的 nullifai 恶意模型）。 https://genai.owasp.org/2025/03/06/owasp-gen-ai-incident-exploit-round-up-jan-feb-2025/，2025年。[33]PYMNTS。Ramp 添加用于发票处理的人工智能智能体。
https://www.pymnts.com/news/artificial-intelligence/2025/
ramp-adds-ai-agents-invoice-coding-approval-payment-processing/，2025年。[34]SAP Concur。假收据 2.0：为什么人工审计无法应对人工智能，以及技术如何反击。
https://www.concur.com/blog/article/
fake-receipts-20-why-human-audits-fail-against-ai-and-how-tech-is-fighting-back，
2025.[35] M. Seymour、K. Riemer、L. Yuan 和 A. R. Dennis。《超越深度伪造》。《*ACM* 通讯》，66(10):56–67，2023年。[36]《南华早报》。香港一名员工在与英国跨国公司被深度伪造的“首席财务官”进行视频通话后，受骗支付了400万港元。

https://www.scmp.com/news/hong-kong/law-and-crime/article/3263151/uk-multinational-arup-confirmed-victim-hk200-million-deepfake-scam-used-digital-version
2024年。

[37] WIRED。GitHub 的《深度伪造色情内容打击仍未奏效》。 https://www.wired.com/story/githubs-deepfake-porn-crackdown-still-isnt-working，2025年。

[38] J. Ye、L. Luceri 和 E. Ferrara。《审计政治曝光偏差：2024年美国总统选举期间 Twitter/X 上的算法放大效应》。载于 《*2025*年*ACM*公平、问责与透明度会议论文集》，第2349–2362页，2025年。

Case Category	Primary mechanism(s)	Attack pattern / how it works	Typical harms	Refs
(**A**) High-conviction impersonation fraud (audio/video)	Synthetic identity + synthetic interaction; workflow exploitation	Attackers impersonate executives/trusted parties via cloned voice or video calls; exploit plausible operational context (e.g., payment approvals, vendor onboarding) and sometimes multi-party "social proof" to increase compliance.	Direct financial loss; internal distrust; higher verification and training burden; reputational risk.	[23, 19, 36]
(**B**) Election-adjacent misinformation and synthetic outreach	Hyper-targeting + scale + plausible deniability	Synthetic robocalls/messages mimic candidates or institutions; deliver confusing/demobilizing instructions; rapidly repurposed across jurisdictions/demographics; correction is hard because exposure is segmented.	Erosion of trust in official electoral information; voter confusion; increased burden on election administrators; polarization amplification.	[13, 1, 31]
(**C**) Non-consensual synthetic sexual imagery and harassment	Cost collapse + personalization + platform amplification	Synthetic intimate imagery generated and disseminated to humiliate, extort, or silence; coordinated communities re-upload and remix; harassment persists across platforms and search surfaces.	Psychological trauma; reputational harm; chilling effects on participation; secondary victimization from ongoing rediscovery.	[2, 37, 10, 3]
(**D**) Fabricated "documentation" and everyday verification corrosion	Synthetic content + synthetic identity; institutional friction	Generation of plausible IDs, receipts, invoices, screenshots, emails, chats, and evidence bundles; overwhelms manual verification; enables both fraud and strategic denial ("that proof is fake").	Increased verification costs; shift from default trust to default suspicion; exclusion harms for those lacking access to stronger authentication channels.	[20, 34, 26, 33, 21]
(**E**) Compromised / corrupted generation pipelines	Socio-technical risk + model supply chain compromise	A generative tool (or its upstream updates) is manipulated such that outputs systematically embed bias, propaganda, or covert steering; users experience outputs as neutral "system" responses.	Covert manipulation at scale; loss of institutional neutrality; hard-to-audit downstream effects; long-term trust erosion.	[6, 22, 32, 25, 24]

Table 1: Case bank of representative GenAI "synthetic reality" risk realizations (2023–2025), organized by dominant mechanisms and harm pathways.