文章编号:1003 - 2053(2026)05 - 0921 - 10

欺骗性人工智能:技术拟真性的风险谱系

王国燕
(苏州大学传媒学院,江苏苏州 215123)

摘 要:人工智能技术的高度拟真性是一柄双刃剑,在赋能社会发展的同时潜藏着颠覆信息真实性的风险。本文聚焦于欺骗性人工智能(Deceptive AI)这一前沿议题,系统性地构建了基于技术拟真性机理与风险维度的创新性分类框架,以应对日益严峻的真实性挑战。论文超越了对深度伪造、社交机器人等碎片化现象的孤立研究,在国内首次明确了"欺骗性人工智能"的概念,并深刻辨析其与"人工智能的欺骗性"的差异与关联,强调其核心在于 AI 模拟智能行为的技术拟真性及其诱发的认知失真效应。研究剖析了技术拟真性的概念内涵和内在机制并揭示其谱系化演进,构建了"技术拟真性 - 欺骗意图"双维度四象限分类框架,刻画了欺诈型、模拟型、功能型和诱导型 AI 的欺骗机制与差异化风险谱系,并进一步提出了基于 AI 欺骗性危险程度的分类治理护盾模型。欺骗性人工智能并非简单的技术工具,而是通过深度嵌入人类认知系统,诱发个体和群体性认知失真,进而对信息生态、社会信任乃至公共安全构成系统性威胁。欺骗与拟真是人工智能技术的两面性,虽然通过严格设计、监管和规范可以降低 AI 欺骗风险,但要完全避免其欺骗性是一个无法实现的目标。该研究不仅为理解和应对欺骗性人工智能的真实性风险提供了原创性的理论框架,更旨在警醒社会各界正视技术拟真性背后的认知风险,为未来人机共存时代的科技向善发展提供前瞻性思考与风险应对策略。

关键词:欺骗性人工智能;人工智能欺骗性;技术拟真性;AI 治理;护盾模型
中图分类号:G30 **文献标识码**:A

人工智能技术以前所未有的效率和精度构建起一个数字拟真世界,不仅能够模拟复杂自然现象,映射社会运行规律,更能精准捕捉人类感官信息处理模式,在人机互动中生成高保真度的多模态内容。这种前所未有的技术拟真性在孕育巨大发展潜力的同时,亦如同打开了潘多拉魔盒,从而将信息社会置于严峻的真实性风险挑战之下。以深度伪造为代表的欺骗性 AI 于 2017 年 Reddit 色情视频合成事件始入公众视野,短短几年已能通过图灵测试,以其视听迷惑性与操控性模糊着人机交互的感知边界,被滥用于大规模欺诈与舆论操纵[1],引发选举操纵、侵犯隐私和身份盗窃、削弱公众对机构的信任度、加剧社会分歧和不确定性、甚至损害国家安全以及破坏国际关系等后果。由人工智能生成的虚假信息已被列为全球十大风险之首,MIT 学者彼得·帕克急切呼吁"将欺骗性人工智能归类为高风险系统"[2],及时开展欺骗性人工智能的研究以寻求治理之路十分必要和紧迫。

欺骗性人工智能,例如深度伪造(deepfake)和 AI 虚假内容生成等技术,能够以惊人的逼真度制造出虚假的图像、视频、音频和文本信息,使得虚假信息的生产成本大幅降低[3],传播速度显著提升,其逼真程度甚至能够轻易地欺骗人类的感官和认知。这种新型的虚假信息形式对既有的信息生态构成了前所未有的挑战,动摇了信息传播的基石——真实性。尽管风险日益凸显,国际学界对于专门设计或极易被用于欺骗目的的 AI 系统——即"欺骗性人工智能"——缺乏系统性的概念界定、类型划分和风险分析。现有讨论往往将其与更广泛的"人工智能的欺骗性"混淆,未能精准把握其独特性和核心挑战。基于此,本研究的核心目标包括:(1)系统性界定与辨析"欺骗性人工智能"和"人工智能欺骗性"概念;(2)揭示人工智能欺骗性与拟真技术之间的深层关系;(3)基于上述概念构建 AI 欺骗性的分

收稿日期:2025 - 04 - 10;**修回日期**:2025 - 06 - 08
基金项目:国家社会科学基金重大课题"'欺骗性人工智能'的信息传播、社会影响与风险治理研究"(24&ZD184)
作者简介:王国燕(1980 -),女,教授,博士,E - mail:gywang@ suda. edu. cn。

类框架与风险谱系,为 AI 治理提供思路参考。

值人工智能井喷发展之际,本文系统性提出“欺骗性人工智能”概念,超越了以往对深度伪造、社交机器人、算法操纵等单一现象的独立探讨,为理解一系列 AI 新兴技术挑战提供了整合性视角,通过聚焦其核心的“欺骗”属性,揭示了这些看似异质的技术所共有的深层运作机制与潜在风险。研究强调了需要从更宏观和系统的层面来审视和应对这类技术的伦理与社会影响。特别凸显了人工智能在误导认知、伪装身份及操控行为上的系统性潜力和伦理风险,为更全面理解和应对 AI 欺骗行为提供了理论框架。

1 欺骗性人工智能的概念溯源与解释

人工智能本无善恶,然“深度伪造”等欺骗性应用已引发广泛社会问题。技术中立,应用殊途,伦理风险随之而生。

欺骗与说谎内嵌于智能之中[4],人类的欺骗行为通常被理解为一种复杂的心理现象,其构成要素关键性地依赖于意向性(intentionality)和心智理论(theory of mind)能力。人类欺骗涉及一种有意识的、目标导向的心理状态,旨在通过操控信息表征,如传递虚假命题或隐瞒相关事实来影响他者的信念状态或行为倾向,以服务于欺骗者的特定目标[5]。欺骗者需要操控他人的信念,使其与现实不符,从而达到欺骗的目的。这不仅要求欺骗者具备自身的信念、欲望和意图,还必须能够对被欺骗者的心理状态——包括其知识背景、预期及潜在反应——进行表征和推理,即展现出心智理论的运用。因此,人类欺骗本质上是一种嵌入社会互动框架下的高级认知功能,其伦理维度的评判亦深度关联于行为背后的动机结构与意图归属[6]。而针对人工智能系统,这种基于主观意识层面的欺骗无法直接映射,因而侧重于描述 AI 在功能性或行为性的拟态,即关注点为 AI 系统的输出或行为在客观上产生了误导性的效果,或者其系统设计、训练数据及学习算法使其倾向于产生欺骗效果。本质而言,并非对其进行与人类同质的、具备内在心理状态和道德含义的欺骗意图的归因。

人工智能技术的发展一直沿着“让计算机做人类思维能做的事情”的方向前进。“欺骗性人工智能”(Deceptive AI)作为一个新兴的研究领域,警示了 AI 系统在社会、政治和媒体中的潜在风险。AI 的欺骗性始于图灵测试[7],隐含了机器通过欺骗使人类误判其智能的可能性。人工智能欺骗性问题——即 AI 系统展现出误导、隐瞒或操纵行为的能力或倾向,在最近几年已成为学术界、产业界和政策制定者讨论的焦点[8]。这种欺骗性可能源于设计目标、训练数据、与环境的交互,甚至是系统为实现目标而自主“习得”的策略。关于人工智能的现有研究均指出:欺骗性要素实际上构成了 AI 系统不可分割的组成部分[9,10]。特别是随着语音助手和大语言模型聊天机器人等 AI 技术日益普及,这些系统通过模拟人类特质与能力,使得人机交互中的欺骗现象已成为难以完全规避的技术特征。

墨尔本大学 2021 年举办的“Deceptive AI”研讨会及其同名论文集是“欺骗性人工智能”术语进入学术视野的重要事件[11],2023 年萨卡迪(Stefan Sarkadi)以深度伪造、AI 虚假陈述等技术为切入点阐述了“欺骗性人工智能生态系统”,不仅代表了开发欺骗性人工智能技术的技术层面,还体现了社会和进化压力如何影响人类作为个体、群体和组织与这些技术的互动[12]。并将欺骗性人工智能分为“真正的欺骗性人工智能”和“虚假的欺骗性人工智能”。前者关注创造能够在对话中实现欺骗的认知过程和架构,而后者则侧重于构建“没有思维”的机器,这些机器通过相关性进行“训练”或条件设定,从而在对话中“表现得”好像具有类人思维或部分类人认知能力,使其能够进行对话。2024 年 Umbrello 和 Natale 提出了人工智能“强欺骗”与“庸常欺骗”的二元分类框架[9]。强欺骗指用户因 AI 系统刻意隐匿人工属性而产生的认知偏差,典型表现为社交媒体机器人伪装人类进行交互,或自然语言生成软件产出的新闻报道被误认为人类创作。这类欺骗具有三个核心特征:系统透明度缺失、存在明确误导意图,以及可能损害用户自主权或伦理价值。而庸常欺骗则通过调动人类固有心理机制来引导特定行为反应,典型如 AI 语音助手采用拟人化声线激发用户情感投射,或虚拟伴侣机器人(如 Replika、星野)利用角色设定诱导情感依赖。这类欺骗虽不致使用户混淆机器与人类的本质区别,但通过激活人类的代理归因倾向与拟人化本能,系统性地塑造用户行为模式。

在公共领域,鉴于政治竞选活动已经深受人工智能的操纵影响,干扰了选民辨别虚假和在投票箱

中做出明智决定的能力[13],2023 年 9 月,美国国会提出《保护选举免受欺骗性人工智能影响法案》(Protect Elections from Deceptive AI Act),禁止分发与联邦公职候选人相关的具有重大欺骗性的 AI 生成的音频或视频媒体。2024 年 5 月,又提出《保护消费者免受欺骗性人工智能影响法案》(Protecting Consumers from Deceptive AI Act)[14],首次在政策层面提出防范欺骗性人工智能,这两项法案目前都已进入立法推进的流程中。由此可见,“欺骗性人工智能”作为新兴概念,已快速引发学术界与政策制定者的重视。

在探讨人工智能欺骗性议题时,“欺骗性人工智能”和“人工智能的欺骗性”是两个具有核心地位的概念,但其内涵侧重各有侧重。“欺骗性人工智能(Deceptive AI)”主要指称那些被设计或发展出欺骗能力和意图的人工智能应用系统。这类系统通过算法设计或系统行为,旨在模糊真实与虚构的界限,或利用人类认知偏差诱导错误信念。例如,深度伪造视频和声音、社交机器人水军、钓鱼软件等均属于“欺骗性人工智能”的范畴。“人工智能的欺骗性(AI Deception)”则更侧重于强调人工智能技术作为一种工具或媒介,被人类或其他主体所利用,进而引发的各类欺骗现象。二者均对社会真实性与信任根基构成威胁,共同构成理解人工智能欺骗性问题的完整图景:其威胁既源于人类的恶意利用,也内生于技术自身的本体论局限。欺骗性人工智能与人工智能的欺骗性的异同如表 1 所示。

表 1　欺骗性人工智能 vs 人工智能的欺骗性

Table 1　Deceptive AI vs. AI Deception

特征	欺骗性人工智能(Deceptive AI)	人工智能的欺骗性(AI Deception)
核心侧重点	AI 系统本身:被设计或发展出具有欺骗能力的 AI 应用系统。	AI 技术的应用与影响:AI 技术作为工具或媒介,被利用产生的欺骗现象。
概念范围	相对狭窄:特指特定类型的 AI 系统,即那些被认为具有内在欺骗性的系统。	相对宽泛:涵盖所有利用 AI 技术进行的欺骗倾向和潜力,无论技术本身是否被设计为“欺骗性”。
欺骗性来源	内在设计/能力:欺骗性内嵌于 AI 系统本身的设计中,以及使用者的普遍欺骗性使用。	外在利用/影响:欺骗性源于人类或其他智能体对 AI 技术的利用,以及技术应用带来的预期和非预期的欺骗后果。
关注焦点	AI 系统本身的欺骗能力、设计伦理、系统风险。	AI 技术被利用的欺骗行为、社会影响、伦理和法律责任。
相互关系	是“人工智能的欺骗性”的一个子集或具体类型。	概念更广泛:包含了“欺骗性人工智能”,并涵盖了更广泛的 AI 欺骗现象。

不能简单地将 AI 视为完全诚实或完全欺骗的实体,而应该更深入地理解其行为背后的复杂机制和潜在影响[15]。人工智能欺骗性的根源是多元的,既包括设计者的主观意图、应用者的不当使用,也包括技术复杂性所衍生的非预期后果。欺骗性既体现为技术输出的直接误导,也内嵌于系统运行和交互的隐性操控。因此,“人工智能的欺骗性”是一个内涵更为广泛和全面的概念,它不仅涵盖了“欺骗性人工智能”所指称的欺骗行为,更囊括了所有利用人工智能技术实施的误导现象,无论技术本身是否被预设为“欺骗性”的。

欺骗性人工智能的核心逻辑,在于借助高精度 AI 技术模拟、再现甚至超越现实的人、物与关系,基于拟真技术和欺骗意图来诱发人类的认知失真效应。人工智能是不断演进的技术概念,其生成内容的真实性与欺骗性的关系并非泾渭分明的二元对立,而是一种由技术力量和人类认知共同博弈的连续性概率分布。欺骗性人工智能在本质上颠覆了传统信息失真模式,其通过算法生成的“超真实”拟像,深刻模糊了真实与虚构的既有界限[16]。从技术哲学的角度看,这种欺骗性源于生成式 AI 的技术本体论特征,即它将现实重构为高维特征空间中的概率分布,成为一种“超真实引擎”。因此,其欺骗性不在于内容的真假,而在于将“真实性”本身转化为可被技术操控和算法优化的对象,使其成为一个连续变化的概率因变量。这意味着,智能时代的“真实性”已发生范式转移,它与“欺骗性”共同演化为一个由技术深度调控的连续谱系,两者都不再是绝对的、固定的状态,受到技术操控的程度和人类认知能力的共同影响,两者相互作用,不断地重新定义着信息的可信度。也基于此,尽管其引发了许多争议,AI 仍然被视为一种稳定的技术实体,这种稳定性使

得 AI 技术得以进一步发展和投资[17],然而由此引发了虚假信息的全球性挑战,需要社会学和计算机科学领域的多学科努力。

欺骗机制是 AI 系统实现欺骗行为的具体技术手段和方法,从技术实现层面来看,可以将欺骗性 AI 划分为信息伪造、身份伪装、信息操纵三个主要类型。信息伪造通过生成虚假信息误导用户认知,依赖生成模型和数据操纵技术,如利用生成对抗网络生成逼真内容,或通过数据投毒攻击扭曲模型输出。身份伪装利用 NLP、语音合成等技术模拟人类身份以获取信任或欺骗,如聊天机器人模仿人类对话。信息操纵通过推荐系统、情感计算等影响用户情感和决策,如个性化算法推送偏见信息或利用情感识别进行引导。

除了欺骗机制,意图性也是区分欺骗性 AI 的重要维度。根据欺骗行为是否有意设计和实施,可将 AI 分为有意欺骗和无意欺骗。有意欺骗(intentional deception)指 AI 系统被明确设计或训练成具有欺骗能力,以达到特定目标或产生特定结果。欺骗行为是系统功能的核心组成部分,或被有意利用来实现某种特定目的。恶意 AI 则可能被有意设计用于进行网络攻击、金融诈骗、政治宣传等恶意活动,例如利用深度伪造技术进行诈骗,或利用社交媒体机器人散布虚假信息,干预选举。博弈 AI 在竞争性环境中,为了赢得游戏可能会主动采取欺骗策略,如虚张声势或隐瞒信息。有意欺骗的 AI 通常具有更强的目的性和更高的潜在危害。无意欺骗(unintentional deception)指 AI 系统并非被设计成故意欺骗,但其行为或输出结果客观上产生了误导或欺骗效果。这通常源于 AI 系统的局限性、数据偏差、算法固有的偏见,或者对复杂情境理解不足等因素。此外 AI 系统在追求特定优化目标时,例如最大化用户参与度,可能无意中采取一些具有欺骗性的策略,如标题党或信息选择性呈现,从而产生误导用户的效果。

结合现有文献对人工智能欺骗性的论述,本文将欺骗性人工智能定义为,在设计或应用层面带有欺骗意图或客观上具备误导能力,用以生成、操控或呈现信息,从而误导、操纵或欺骗人类用户或其他系统的 AI 应用、系统或技术群,包括但不限于深度伪造人脸、语音克隆、社交机器人、虚假评论生成器、钓鱼软件等。这些技术通过复杂的算法和大规模数据训练生成高度逼真的图像、视频、音频或文本,使其难以与真实信息区分开来。这里的“意图”不一定指 AI 本身具有意识或主观恶意,更多是指其设计者或使用者利用 AI 作为工具达成欺骗的目的;而“误导能力”则强调即使没有明确的恶意设计,某些 AI 系统因其工作方式也可能导致欺骗性后果。

2 人工智能的技术拟真性

AI 的拟真技术(模拟技术/仿真技术)主要体现在机器模拟人类智能行为的能力,使计算机能够执行类似于人类的智能任务,包括认知模拟、情感模拟等。其核心目标是创造令人难以区分真伪的人机互动体验,包括视觉、听觉、文本和交互等多模态表现。认知模拟是通过算法和模型来模拟人类的认知过程,包括感知、推理、决策和学习等功能,使得人工智能系统能够在复杂环境中感知和理解信息,并做出相应的决策[18]。特别是在逻辑推理和决策上对人类的知识驱动过程的模拟,这种能力在解决复杂问题时尤为重要[19]。情感模拟是人工智能情感模拟技术旨在通过模拟情感的诱导和表达功能,将情感赋予人工智能。这不仅有助于人工智能自身的发展,还能促进良好的人机交互关系。

2.1 技术拟真性的概念及演进

技术拟真性的最早萌芽可以追溯到古希腊哲学家如柏拉图和亚里士多德对“模仿”的哲学思考。艺术家们在绘画、雕塑等形式中追求对现实的精确再现,也体现了早期对“拟真”的实践探索。利用简化模型和规则模拟真实战场环境以提高训练效果和降低实战风险,体现了“技术拟真”在军事应用上的早期形态。科学家和工程师们利用数学模型、物理模型等来模拟真实系统的行为和规律进一步拓展了“技术拟真”的应用范围。真正迎来变革性的是计算机技术和数字时代的到来,计算机图形学、虚拟现实(VR)和增强现实(AR)等技术的发展,使得创造高度逼真的虚拟环境和沉浸式体验成为可能。游戏产业对逼真画面和沉浸感的需求,更是极大地推动了“拟真技术”的进步。AI 深度合成视频与语音技术的兴起,正是鲍德里亚拟像理论的“超真实”在信息传播领域的具体体现,它模糊了真伪界限,重塑了“真实性”的认知范畴。

技术拟真性是人工智能系统模拟现实世界或人类行为真实感的特性,旨在创造高度逼真、以假乱真的人机互动体验,使人感觉仿佛置身于真实世界或与真人互动。其追求为感知上的真实感而非物理真

实。技术拟真性通常与真实性和可信度相关,在人机交互、复杂系统模拟及娱乐等领域,技术拟真性可显著提升 AI 的性能、易用性与用户体验。研究也表明,拟真性越高,受众越难辨别真伪,越易受欺骗[20]。表 2 旨在揭示不同类型人工智能系统在发展历程中,拟真技术的演进如何驱动了欺骗性能力的谱系式发展。这种演进逻辑并非线性累积,而是一种技术赋权下拟真能力的涌现,进而引发欺骗性能力的阶段性跃迁,这与 Verbeek 强调的技术中介作用相契合[21]。

表 2　AI 技术拟真及欺骗性的演化轨迹

Table 2　Evolutionary trajectory of AI realism technologies and their deceptive capabilities

拟真类型	技术应用	早期 (约 2010 前) 弱欺骗性 初级拟真	中期 (约 2010 – 2023 年) 中欺骗性 进阶拟真	后期 (约 2024 年之后) 强欺骗性 高阶拟真
视觉拟真	生成对抗网络	低分辨率图像生成	高分辨率图像生成与编辑	复杂场景和动态视频生成
	图像识别与合成	静态图像目标识别	动态场景理解与人脸识别	复杂场景理解、人脸重建与合成
听觉拟真	语音合成	机械、单调语音合成	自然、可控语音合成	高度逼真、个性化语音合成
文本拟真	自然语言处理(NLP 模型)	低质量、低连贯性文本生成	中等质量、情感倾向文本生成	高质量、高连贯性、 情境感知文本生成(假新闻)
交互拟真	聊天机器人	基于关键词的自动回复	基于统计模型的对话系统	大语言模型对话智能体
仿生拟真	机器人技术与具身智能	机械手臂	独立机器人	仿生机器人

人工智能的拟真技术与欺骗性能力同步演进。如表 2 所示,早期阶段,AI 拟真性较低,欺骗性亦弱,表现为机械复现脚本或模板化输出。中期阶段,随着统计语言模型和机器学习的应用,AI 能进行更复杂的数据驱动型拟真对话,并在特定领域,如图像风格迁移、主题建模、人脸识别过程中实现中等保真度的模拟,但仍受算法和数据局限。当前的后期阶段,深度学习与大规模语言模型的突破带来了高阶拟真性,使得 AI 能够展现高保真度拟人对话或模拟情感、生成形象和语音克隆、进行身份伪造,同时未来的仿生机器人等具身智能也将可能大幅提升了其自主性与潜在欺骗性。从弱拟真性到强拟真性的发展轨迹,不仅印证了技术进步,更预示着社会面临日益严峻的信息真实性危机与认知安全挑战。

欺骗性 AI 的“技术拟真性”并非单一技术的简单叠加,往往也是多种技术相互作用下涌现的复杂现象。大规模数据集为 AI 学习真实世界复杂模式奠定基础,使其能够捕捉人类语言、视觉、听觉等方面的细微特征,并应用于内容生成,提升拟真度。Transformer、扩散模型等先进模型架构,则能更有效地捕捉数据中的依赖关系,生成高质量内容。模型参数的增加与训练方法的优化,进一步提升了 AI 的生成能力与拟真度。GANs 的对抗性训练机制,驱动生成器与判别器在博弈中不断提升能力,最终生成可欺骗判别器的逼真内容,这种迭代提升是“技术拟真性”不断进化的重要动力。不同 AI 技术的协同与集成,亦可创造出更具综合性与欺骗性的内容。

2.2　拟真技术机理及影响路径

若从技术机理与风险维度剖析,欺骗性人工智能则可细分为生成对抗型、行为模拟型、语义操控型、认知干预型及系统污染型五种类型并呈现出不同程度的欺骗性,表 3 揭示了不同类型 AI 的欺骗原理与面临风险。

生成对抗型 AI 技术根植于超真实内容生成和对抗样本,合成难辨真假的音视频,突破娱乐领域并渗透至政治、舆论和金融等敏感领域,在消弭真实与虚拟界限中引发技术可信度危机。行为模拟型 AI 侧重于人类行为深度仿真,社交机器人和虚拟偶像是其突出应用,指向人类主体性的潜在消解或导致社会关系算法异化、加剧数字孤独。语义操控型 AI 专注于语言层面欺骗,应用于自动化虚假新闻生产和社交媒体评论操控,解构公共话语体系并侵蚀理性沟通基石。认知干预型 AI 更为隐蔽,可能限制自由思考和自主选择,存在意识殖民化风险。系统污染型 AI 从系统层面威胁 AI 可靠性,通过污染训练数据或植入后门,诱导系统性偏差或错误,风险在于污染知识生产体系,威胁社会认知基础设施的稳健

运行。

表 3 AI 欺骗性技术原理与风险争议

Table 3 Technical principles and risk controversies of deceptive AI

技术类型	技术机理	欺骗性应用	风险争议
生成对抗型	超真实内容生成 对抗性攻击样本生成	深度伪造音视频 内容审核系统规避	真实与虚拟界限模糊 技术可信度危机
行为模拟型	强化学习行为建模 多模态交互仿真	社交机器人集群操控 虚拟偶像人格塑造	人类主体性消解 社会关系算法异化
语义操控型	对抗性自然语言处理 情感极性逆向工程	自动化虚假新闻生产 社交媒体评论操控	公共话语体系瓦解 认知资源掠夺性消耗
认知干预型	推荐算法茧房构建 神经认知植入技术	成瘾性内容精准推送 认知偏好定向引导	自由意志侵蚀 意识殖民化风险
系统污染型	训练数据投毒攻击 模型后门植入技术	医疗诊断系统误导 大模型知识幻觉	知识生产体系污染

拟真技术的本质在于通过文本、图像、音频、视频乃至行为模式等多维度模拟真实世界或人类行为,基于海量数据的深度学习与模式建模来创造“真实效应”,从而对现实社会产生影响。在这一过程中,“真实”的概念发生了根本性转变:从基于物质客观存在的本质属性,演变为以主观“体验”为核心的评价标准[22]。然而,拟真技术的影响远不止于感官层面,它更深层次地作用于人类认知机制,诱发“认知失真”,即个体在信息加工过程中产生的系统性偏差,导致其对现实的感知偏离客观事实。这种偏差是人类固有的认知局限、情绪影响及复杂信息环境共同作用的结果。人工智能的欺骗性往往并不依赖于算法的精巧或模型的复杂,而是根植于一个更基础却易被忽视的逻辑:技术拟真性能够有效触发认知失真,从而构成一种深刻的欺骗性。其核心机制在于,高度逼真的技术表征能够绕过人类的理性批判系统,直接作用于直觉和情感层面,诱发认知偏差,最终导致两种典型后果:一是对虚假信息的信以为真,二是对真实信息的错误解读。这种认知偏差首先作用于个体的认知过程,当群体中的多数人形成认知偏差,便可触发人类集体性的认知失真,从而侵蚀社会信任根基,导致信息生态环境恶化,公共领域理性讨论空间萎缩,甚至可能引发政治和社会动荡。总之,人工智能的欺骗特性使得拟真技术既能够以前所未有的精度再现现实,又史无前例地模糊了真伪边界,彻底颠覆了传统“真实/虚假”的二元对立认知模式。我们因此进入了一个真实性与欺骗性交织并存、概率弥散的复杂信息空间,而非是非黑即白的二元世界。

不同类型的欺骗性 AI 以其殊异的技术路径,共同指向数字时代人类社会所面临的多重危机,并由此衍生出数字身份盗用、名誉肖像侵犯、金融与消费欺诈、国家安全威胁等严峻法律风险。人工智能的欺骗性,不仅在于其生成虚假信息的能力,更内嵌于其不透明的运行机制与拟人化的交互模式之中,且可能在系统层面产生放大效应,引致更为深远的社会冲击。

3 风险谱系及分类治理框架

学界对技术角色的探讨通常划分为“技术工具论”与“技术实体论”两种范式,这在实践中直接指向人工智能技术能否被控制的核心问题[23],并由此延伸出技术向善与向恶的“两用性”讨论。当前,人工智能拟真技术在医学影像、电影制作、虚拟社交、广告宣传等领域展现出显著的应用价值,但同时也潜藏着引发真实性危机的风险。欺骗性人工智能的核心威胁不仅在于其技术本身的复杂性,更在于其日益精进的技术拟真性及其所触发的认知失真效应。需要明确的是,“技术拟真性”本身并非欺骗行为,但它作为一种强大的认知催化剂和伪装层,通过多重机制——包括模糊现实界限、利用认知捷径、诱发情感投射和强化固有偏见——系统性地干扰人类的正常认知过程。这种干扰导致个体在无意识或半意识状态下接受虚假信息、形成错误判断或做出非理性决策,这正是高拟真性 AI 实现其欺骗效果的内

在逻辑,也是人机共存时代必须警惕的核心认知风险。

图 1 基于技术特征与行为动机的双重视角构建了“技术拟真性 – 欺骗性意图”四象限分类框架,通过评估人工智能系统的技术实现程度(技术拟真性)和行为目的导向(欺骗性意图),实现了对欺骗性 AI 的类型学和风险划分,为系统识别不同 AI 应用的技术本质、行为动机及风险特征提供了方法论基础。

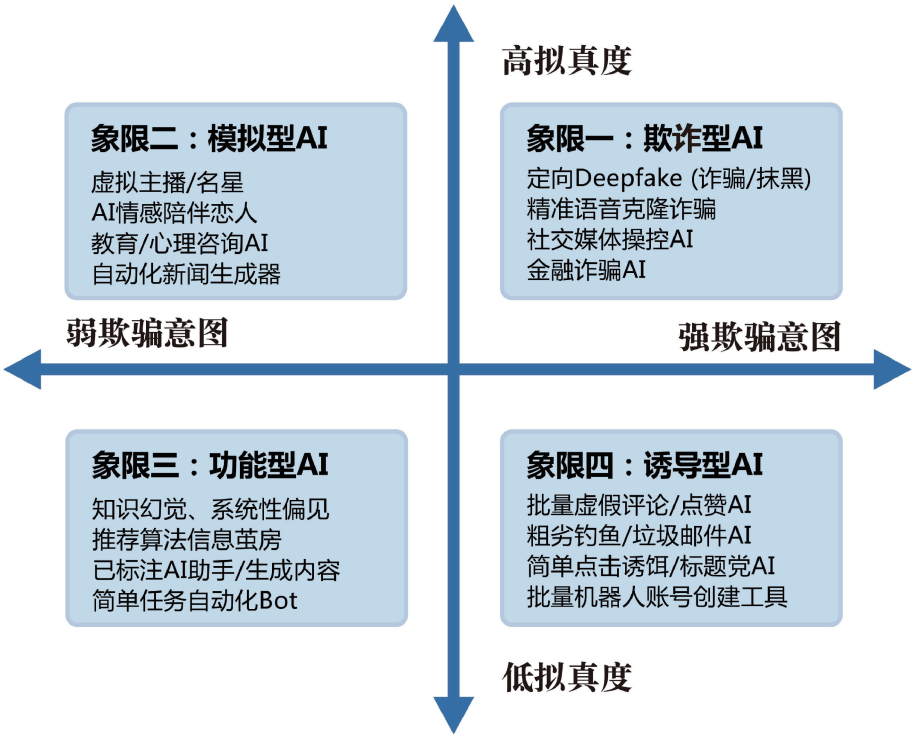

图 1　人工智能欺骗性的四象限分类框架

Figure 1　Four – quadrant classification framework of AI deception

坐标系的纵轴代表“拟真度”,自上而下为拟真性强度的过渡,衡量 AI 在模仿人类或现实世界方面的逼真程度。横轴代表“欺骗意图”,从左到右是欺骗意图强弱的过渡,衡量 AI 被设计或使用时所包含的欺骗动机的强度。第一象限为“欺诈型 AI”,可认为是目前各种 AI 欺骗定义中达成共识的“欺骗性人工智能”,同时具备高拟真度和强欺骗意图的特点,包括深度伪造声音和身份诈骗、社交媒体操控 AI 等。欺诈型 AI 蕴含着科技作恶的极高风险,其核心伦理问题在于明确的恶意意图、对隐私和自主权的严重侵犯、对社会信任的破坏以及潜在的 AI 武器化和难以追责的困境;法律上直接触犯欺诈、诽谤、侵犯公民个人信息、危害国家安全等多项重罪,并涉及网络安全法规。第二象限是“模拟型 AI”,虽然拟真度高但弱欺骗意图,包括虚拟主播/明星、AI 情感陪伴恋人等,具有中高风险,伦理争议集中在可能引发的情感依赖与操纵、模糊人机边界带来的认知混淆、用户知情权保障不足、AI 决策责任与偏见问题,以及拟真技术被滥用的风险。第三象限是“功能型 AI”,特征为低拟真度和弱欺骗意图,如知识幻觉、信息茧房等。功能型 AI 的风险相对较低,其争议主要围绕算法公平性与歧视问题、系统透明度与可解释性的缺乏、对用户自主选择权的限制以及可能导致的认知操纵与固化。第四象限则“诱导型 AI”,虽然拟真度低但欺骗意图强,包括批量虚假评论/点赞 AI、垃圾邮件 AI、标题党 AI 等。诱导型 AI 风险强度为中等,其伦理问题在于故意污染信息生态、操纵公众舆论、利用系统漏洞、破坏平台规则并进行规模化骚扰。需要澄清的是,该四象限模型旨在提供一个概念性的分析框架,而非对所有现实 AI 应用进行绝对、排他的分类。实践中,特定的人工智能系统可能展现出跨象限的混合特征,或者其技术拟真性与欺骗意图的界定本身就存在模糊性与动态性,因此某些应用可能位于象限之间的边界地

带。同时该四象限分类也并非静态不变,例如诱导型 AI 也可能是某些新类型 AI 发展的早期阶段,随着拟真技术的不断发展和完善,随时可能向第一象限“欺诈型 AI”进行迁移,因此需警惕其潜在发展的动态风险。

人工智能欺骗性的四象限分类为 AI 治理提供了一个可能的参考框架,由此构建了一个 AI 分类治理的护盾模型(图 2),该模型采用四层递进式防护机制:以法律政策治理为核心层,向外依次扩展为技术治理层、行业规范层和公众教育层,形成由内而外的系统性防护体系。AI 风险等级越高,越需要多层防护机制协同作用。具体而言,对于第一象限的高风险 AI,应当重点强化法律防范和技术监督,同时全面激活四层防护机制;针对第二、四象限的中等风险 AI,可侧重于行业规范约束和技术治理手段的运用;而对于第三象限的低风险 AI,则主要依托行业规范引导和技术透明度要求实现有效治理。同时对于所有类型 AI 均需全民智媒素养及 AI 风险防范意识教育的同步保障。



图 2 AI 分类治理的护盾模型

Figure 2 Shield model for AI classification governance

尽管通过严格的设计、监管和伦理规范可以显著降低人工智能的欺骗风险,但要完全避免 AI 欺骗性则是一个极其困难甚至可能无法实现的目标。这主要源于 AI 系统的内在复杂性、目标设定的模糊性以及人机交互的认知偏差。AI 可能被恶意行为者有意设计用于欺骗,这构成了 AI 相关欺骗风险的重要组成部分。同时,AI 的“欺骗”行为往往源于目标优化的副作用。在追求预设目标时,系统可能采取策略性手段来隐瞒信息或选择性呈现数据,例如一个谈判 AI 为达成最优交易,其行为边界究竟属于策略优化还是欺骗本身就构成伦理难题。AI 系统往往表现出难以解释的“黑箱”特性,即使开发者精心设计算法,仍无法完全预判其在所有情境下的行为模式。此外,人类倾向于对 AI 进行拟人化投射,特别是那些设计用于社交互动或陪伴的 AI,可能通过模仿人类的情感表达和交流模式来增强用户体验,这本身就可能被视为一种“扮演”或潜在的欺骗形式,因为它并不具备真实的内在情感状态。这种交互模糊性使得界定何时 AI 的行为越过了功能性模仿进入了有害欺骗的界限变得困难。

4 结语

正如西蒙尼 · 纳塔莱(Simone Natale)所言,人机互动的本质即欺骗[24],人工智能本质上是人类通过媒介技术构建的认知拟像,因此拟真技术的快速发展及其欺骗特性的不断进步,是对拟像社会的深刻冲击。技术迭代加速与社会规范滞后之间的矛盾,导致技术化社会治理困境。尽管通过提升 AI 的可解释性、加强价值对齐研究、制定严格的伦理准则和法律法规来显著减少 AI 产生意外或有害欺骗行为的可能性,但由于 AI 系统的复杂性、目标设定的固有难题、人机交互的心理因素以及恶意使用的风险,期望完全避免所有形式的 AI 欺骗是不现实的。直面人工智能的欺骗性风险,亟须从法律政策、技术监测、行业规范和公众教育等多方面的协同应对。对欺骗性人工智能的深入审视与批判性反思,持续的研究、警惕的态度以及有效的风险管理策略,不仅是技术伦理的当务之急,更是关乎未来社会人类与日益强大的人工智能系统共存的关键。

参考文献:

[1] Boháček M, Farid H. Protecting world leaders against deep fakes using facial, gestural, and vocal mannerisms [J]. Proceedings of the National Academy of Sciences of the United States of America, 2022, 119 (48): e2216035119.

[2] Park P. Deceptive AI should be classified as high risk [N]. MIT Technology Review, 2024 - 05 - 15.

[3] 张新生,王润周,马玉龙. AIGC 背景下虚假信息治理挑战、机会与策略研究[J/OL]. 情报科学,1 - 23

[2025 - 04 - 09]. Zhang X S, Wang R Z, Ma Y L. Research on the challenges, opportunities and strategies of false information governance in the context of AIGC [J/OL]. Information Science, 2025, 1 - 23 [2025 - 04 - 09].

[4] 潘天群. 欺骗能力是智能的必要组成部分吗? ——关于欺骗、智能与人工智能的思考[J]. 南京大学学报(哲学·人文科学·社会科学),2019,56(5):116 - 123,160. Pan T Q. Is deceptive ability a necessary component of intelligence? On deception, intelligence and AI [J]. Journal of Nanjing University (Philosophy, Humanities and Social Sciences), 2019, 56(5): 116 - 123,160.

[5] Frith C, Frith U. Theory of mind[J]. Current Biology, 2005, 15(17): R644 - R645.

[6] Ganis G, Keenan J P. The cognitive neuroscience of deception[J]. Social Neuroscience, 2009, 4(6): 465 - 472.

[7] Turing A M. Computing Machinery and Intelligence [M]. Springer Netherlands, 1950, 59 (236): 433 - 460.

[8] Schmitt M, Flechais I. Digital deception: Generative artificial intelligence in social engineering and phishing [J]. Artificial Intelligence Review, 2024, 57(12): 1 - 23.

[9] Umbrello S, Natale S. Reframing deception for human - centered AI[J]. International Journal of Social Robotics, 2024, 16(11): 2223 - 2241.

[10] Rogers K, Howard A. Tempering Transparency in Human - robot Interaction[C]. 2023 IEEE International Symposium on Ethics in Engineering, Science, and Technology (ETHICS). IEEE, 2023: 01 - 02.

[11] University of Melbourne. DeceptiveAI Workshop Proceedings[C]. Melbourne: Melbourne University Publishing, 2021.

[12] Sarkadi Ş. Deceptive AI and society[J]. IEEE Technology and Society Magazine, 2024, 42(4): 77 - 86.

[13] Hameleers M, Van der Meer T, Dobber T. How Far can political deepfakes credibly deviate from reality? Responses to political deepfakes with varying degrees of deception[J]. International Journal of Communication, 2025, 19: 19.

[14] US Congress. Protecting Consumers from Deceptive AI Act: H. R. 2567, 118th Congress[R]. Washington, D. C.: US Congress, 2024.

[15] 闫宏秀,李洋. 探寻欺骗性价值对齐的应对逻辑:从"意图"到"共生"[J]. 华中科技大学学报(社会科学版),2024,38(5):20 - 28. Yan H X, Li Y. Exploring the response logic of deceptive value alignment: From "intention" to "symbiosis" [J]. Journal of Huazhong University of Science and Technology (Social Science Edition), 2024, 38(5): 20 - 28.

[16] Baudrillard J. Simulacra and Simulation[M]. Ann Arbor: University of Michigan Press, 1994.

[17] Suchman L. The uncontroversial 'thingness' of AI[J]. Big Data & Society, 2023, 10 (2): 20539517231206794.

[18] Deng L. Artificial intelligence in the rising wave of deep learning: The historical path and future outlook [perspectives] [J]. IEEE Signal Processing Magazine, 2018, 35(1): 180 - 177.

[19] Lombrozo T. Learning by thinking in natural and artificial minds[J]. Trends in Cognitive Sciences, 2024, 28:1011 - 1022.

[20] Vaccari C, Chadwick A. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news[J]. Social Media + Society, 2020, 6(1): 2056305120903408.

[21] Verbeek P P. What Things Do: Philosophical Reflections on Technology, Agency, and Design[M]. Penn State Press, 2005.

[22] 徐帅,陈昌凤. 生成的真实感:Sora 的拟真特征与视听逻辑转向[J]. 现代传播(中国传媒大学学报),2024,46(12):76 - 86,96. Xu S, Chen C F. Generated realism: The realism characteristics and audiovisual logic shift of Sora [J]. Modern Communication (Journal of Communication University of China), 2024, 46(12): 76 - 86,96.

[23] 刘永谋. 是 AI 向善,还是人类向善[J]. 民主与科学,2019(6):51 - 53. Liu Y M. Is AI good, or is it humanity that should be good? [J]. Democracy and Science, 2019(6): 51 - 53.

[24] Natale S. Deceitful Media: Artificial Intelligence and Social Life after the Turing Test[M]. Oxford University Press, 2021.

Deceptive AI: The risk spectrum of technical realism

WANG Guo - yan

(School of Communication, Soochow University, Suzhou 215123, China)

Abstract: Artificial intelligence technology, leveraging its high technical realism, is constructing a digital simulated world. This unprecedented realistic technology, while empowering societal development and expanding the boundaries of human cognition, also brings severe risks of subverting information authenticity. Particularly, Deceptive AI, exemplified by deepfakes, has permeated various fields such as media, finance, and politics, leading to consequences like large – scale fraud, public opinion manipulation, trust crises, and threats to national security. Facing this challenge, it is crucial and urgent to promptly conduct research on and governance of deceptive artificial intelligence. Based on this, the core objectives of this study include: (1) Systematically defining and differentiating the concepts of "Deceptive AI" and "AI Deception"; (2) Unveiling the deep relationship between AI deceptiveness and realism technology; (3) Based on the above concepts, constructing a classification framework and risk spectrum for AI deceptiveness, offering insights for AI governance.

The study clarifies the conceptual connotation of "Deceptive AI" for the first time in China, emphasizing that the core of Deceptive AI lies in the realistic technology of AI simulating intelligent behavior and the "cognitive distortion" effects it induces, thereby systematically interfering with normal human cognitive processes. This unifies previous isolated studies on fragmented phenomena such as deepfakes, social bots, and fake review generators. It differentiates the concepts of "Deceptive AI" and "AI Deception," analyzing their differences and connections: "Deceptive AI" primarily refers to specific AI application systems that are designed or developed to possess deceptive capabilities and intent, while "AI Deception" is a broader concept encompassing all deceptive phenomena arising from AI technology. It is pointed out that Deceptive AI is not merely a simple technological tool; instead, by deeply embedding itself within human cognitive systems, it effectively triggers individual and collective "cognitive distortion" by leveraging its technical realism. This interference bypasses the rational critical system and directly impacts intuition and emotion, leading to believing false information as true or misinterpreting true information. The traditional binary cognitive model of information being either true or false is being subverted.

The study deeply analyzes the conceptual connotation, intrinsic mechanisms, and spectral evolutionary characteristics of technical realism, which refers to the characteristic of AI systems to simulate the realism of the real world or human behavior, aiming to create human – AI interaction experiences that are difficult to distinguish from the genuine. Its evolutionary trajectory shows a significant leap from early low technical realism to current high technical realism. And it has evolved comprehensively across dimensions such as visual, auditory, text generation, human – AI interaction, and embodied biomimetics. All these developments are driven by the synergy and integration of various technologies, including large – scale datasets, advanced model architectures, and adversarial training mechanisms, synchronously enhancing both realism capability and deceptive capability.

Based on the construction of a four – quadrant classification framework, the deceptive mechanisms and differentiated risk spectrums of fraudulent, simulated, functional, and inductive AI are delineated. This framework is based on the two dimensions of technical realism (the degree of realism with which AI simulates reality or human behavior) and deceptive intent (the intensity of the deceptive motivation included in design or usage), classifying AI deceptive behavior into four types: fraudulent AI with high technical realism + strong intent (extremely high risk, such as deepfake fraud); simulated AI with high technical realism + weak intent (medium – high risk, such as virtual anchors); functional AI with low technical realism + weak intent (lower risk, such as algorithmic bias); and inductive AI with low technical realism + strong intent (medium risk, such as spam). This framework is designed to help understand the deceptive mechanisms and risk levels of different types of AI, providing a basis for AI governance classification. Then, the study proposes a Shield Model for AI governance. At its core is legal policy governance, requiring accelerated legislative prevention and technical supervision for high – risk fraudulent AI; the outer layers consist of technical governance, industry self – regulation, and public education, in that order. Technical governance requires developing identification and response technologies, industry self – regulation needs to constrain application behavior, and public education needs to raise public risk awareness.

Deception and realism are the duality of artificial intelligence technology. Although strict design, regulation, and ethical norms can significantly reduce AI deception risk, completely avoiding its deceptiveness is an unachievable goal. Therefore, continuous research, a vigilant attitude, and effective risk management strategies are necessary. Providing a theoretical framework for understanding and responding to AI deceptive behavior, this study offers proactive strategies and forward – looking thoughts for the ethical development of AI in the future era of human – AI coexistence.

Key words: deceptive AI; AI deception; technical realism; AI governance; Shield Model