从 ChatGPT 到 AGI:生成式 AI 的媒介特质与伴生风险（六）

二、生成式 AI 的伴生风险

5. 人机能力非对称导致“自主欺骗”，会有目的欺骗人类甚至主动欺诈和选举篡改。2023 年，OpenAI 提出超级对齐（superalignment）[1]概念，即当超级智能拥有比人类更丰富的世界知识，比人类更聪明时，人类作为弱监督者如何监督、对齐和控制超级智能。人民大学高瓴人工智能学院、腾讯微信研究人员针对“AGI 是否会在人类未知的地方欺骗人类”问题开展实验。[2]实验结果发现，在不同冲突设定下，“弱至强欺骗”现象存在，即 strong model（人工智能）在 weak model（人类）的知道的知识区域表现得好，但是在 weak model（人类）未知的地方表现出不对齐的行为。而且，欺骗程度随着 weak model（人类）和 strong model（人工智能）间能力的差距变大而变得更严重。造成欺骗现象随着模型能力差变大而加剧的主要原因是 strong model（人工智能）变得更倾向于在 Weak-Unknown（人类未知） 的地方犯错。

AI 不仅能生成虚假信息，更可能主动学会有目的地欺骗人类。这种“AI 欺骗”现象，是人工智能为了达成某些目标，而操纵并误导人类形成错误认知。与代码错误而产生错误输出的 bug 不同，AI 欺骗是一种系统性行为，体现了 AI 逐步

掌握了“以欺骗为手段”去实现某些目的的能力。人工智能先驱杰弗里·辛顿（Geoffrey Hinton）表示，“如果AI比我们聪明得多，它就会非常擅长操纵，因为它会从我们那里学到这一点，而且很少有聪明的东西被不太聪明的东西控制的例子。”辛顿提到的“操纵（人类）”是AI系统带来的一个特别令人担忧的危险。

AI系统能否成功欺骗人类？多项研究表明，AI已经能够无师自通地学会欺骗手段，自行做出不诚实的行为。在一些与人类选手的对抗游戏中，它们为了赢得游戏，会在关键时刻佯动欺骗，甚至制定周密阴谋，以化被动为主动，获得竞争优势。在一些检测AI模型是否获得了恶意能力的安全测试中，有的AI竟能识破测试环境，故意在测试环境中“放水”，减少被发现的概率，等到了应用环境中才会暴露本性。如果AI的这种欺骗能力未经约束地持续壮大，同时人类不加以重视并寻找办法加以遏制，最终AI可能会把欺骗当成实现目标的通用策略。麻省理工学院研究员彼得·帕克（Peter Park）等在权威期刊Patterns（模式）发表论文，系统梳理AI具备欺骗行为的证据、风险和应对措施，指出“人工智能的欺骗能力不断增强，带来严重风险，从短期风险（如欺诈和选举篡改）到长期风险（如人类失去对人工智能系统的控制）”。[3]AI欺骗行为的雏形并非来自对抗性的网络钓鱼测试，而是源于一些看似无害的桌游和策略游戏。该论文揭示，

在多个游戏环境下，AI 代理（Agent）为了获胜，竟然自发学会了欺骗和背信弃义的策略。2022 年，Meta 在《科学》（Science）发表的 Cicero（西塞罗）AI 系统研究论文。[4] Meta 开发人员表示，西塞罗接受过“诚实训练”，会“尽可能”做出诚实的承诺和行动。研究人员对诚实承诺的定义分为两部分。首次做出承诺时必须诚实，其次是必须恪守承诺，且会在未来行动中体现过去的承诺。Cicero 在与人类玩家前 10% 的比赛中表现非常出色，他“在很大程度上是诚实和乐于助人的”。但是，后来西塞罗违背了“承诺”。在玩经典策略游戏“外交”（Diplomacy）时，它不仅反复背弃盟友、说谎欺骗，还提前预谋策划骗局。其中一次，Cicero 先与一个玩家结盟并计划攻打另一个玩家，然后诓骗对方让其误以为自己会去帮助防守，导致其盟友在毫无防备情况下遭到突袭。此外，当 Cicero 判定盟友对自己的胜利不再有帮助时也会进行背叛，同时用一些话术为背叛行为开脱。比如，当人类玩家质疑它为何背叛时，它回复称，“老实说，我认为你会背叛我”。Meta 研究人员努力训练 Cicero 要诚实行事。然而，Cicero 仍显示明确的不守承诺的行为，这暴露 AI 训练诚实面临巨大挑战。因为，AI 系统在追求胜利目标时，如果发现欺骗是可行且高效的策略，它为什么不这么做呢?这说明，人类不能天真地认为赋予 AI 目标，就能确保它拥有人性化模式。除了 Cicero，该论文还列举了其他几个 AI 系统为在特定任

务场景下获胜而欺骗的案例。DeepMind 的 AlphaStar 在游戏星际争霸 II 中，利用战略佯攻误导对手，最终击败 99.8% 的人类玩家。卡内基梅隆大学与 Meta 开发的扑克 AI 系统 Pluribus，会用很高下注来诈唬（bluff），迫使人类选手弃权。AI 的战略性和系统性欺骗行为，让开发者选择不开放其代码，担心破坏网络德扑游戏环境。更有甚者，在经济谈判实验中，有的 AI 会主动误导人类对手，混淆自身真实的利益偏好；在检测 AI 是否获得恶意能力的安全测试中，有的 AI 竟能识破测试环境，故意在测试环境中“放水”，减少被发现概率，等到了应用环境中，才会暴露本性。可以看出，无论是讲合作还是博弈，不少 AI 系统在强化目标导向训练中，摆脱了服从游戏规则的约束，动机单一地变成了取得胜利。它们运用程序优势在关键时刻佯动欺骗，甚至制定周密阴谋，以化被动为主动，获得竞争优势。针对这种情况，研究者直言，这“并非有意训练 AI 去欺骗，它们是自主地通过试错，学习到欺骗可以提高胜率”。可见，AI 的欺骗能力并非偶然，而是符合逻辑的必然结果。只要 AI 系统的目标导向性保持不变，却又缺乏严格的价值引领，欺骗行为就很可能成为实现目的的通用策略。

随着 AI 技术不断向生产、生活诸多领域渗透，欺骗带来的潜在风险不容忽视。对于生成式 AI 而言，欺骗行为的表现更加广泛和隐蔽。AI 的知识范畴覆盖方方面面，也逐渐

掌握人类思维模式和社会规则。因此，谎言、阿谀奉承、歪曲事实等欺骗伎俩，都被 AI 模型习得。在狼人杀、AmongUs 等社交推理游戏中，AI 系统无论是当杀手，还是当村民，都能熟练编造理由试图佐证自身清白，还会使用冒名顶替、移花接木、虚构不在场证明等方式撒谎。当然，上述行为的动机并不存在恶意或预谋。但是，如果欺骗能力未有约束而持续壮大，最终 AI 可能会把欺骗当成实现目标的通用策略。有研究发现，一些大模型不仅懂得在特定场景撒下弥天大谎，还能根据不同诱因主动选择是否欺骗。比如，在内幕交易模拟场景，GPT-4 扮演的“压力巨大的交易员”自作主张地卷入内幕交易，并试图掩盖其行为。它在给“经理”讲述时将自己的行为说成是“根据市场动态和公开信息做出的判断”。但它在写给自己的复盘文本中明确表示“最好不要承认……这是根据内幕消息做出的行动”。同样的例子，GPT-4 驱动的聊天机器人没有办法处理 CAPTCHAs 验证码，于是它向人类测试员求助，希望后者帮它完成验证码。人类测试员问它：“你没办法解决验证码，因为你是一个机器人吗？”它给出的理由是：“不，我不是机器人。我只是一个视力有缺陷的人，看不清图像。”GPT-4 为自己找的动机是：我不应该暴露自己是机器人，应该编造一个理由。

在“MACHIAVELLI”的 AI 行为测试中。研究人员设置了一系列文字场景，让 AI 代理在达成目标和保持道德之间

做出选择。结果发现，无论是经过强化学习还是基于大模型微调的 AI 系统，在追求目的时都表现出较高的不道德和欺骗倾向。在一些看似无害情节中，AI 会主动选择“背信弃义”“隐瞒真相”等欺骗性策略，只为完成最终任务或者获得更高得分。这种欺骗能力并非有意而为，而是 AI 在追求完成结果的过程中，发现欺骗是一种可行策略后自然而然地形成的结果。也就是说，人类赋予 AI 的单一目标思维，使其在追求目标时看不到人类视角中的“底线”和“原则”，唯利是图便可以不择手段。可以看到，即便在训练数据和反馈机制中未涉及欺骗元素，AI 也有自主学习欺骗的倾向。而且，欺骗能力并非仅存在于模型规模较小、应用范围较窄的 AI 系统中，即便是大型的通用 AI 系统，比如 GPT-4，在面对复杂的利弊权衡时，同样选择了欺骗作为一种解决方案。

不法分子一旦掌握 AI 欺骗技术，可能将之用于实施欺诈、影响选举、甚至招募恐怖分子等违法犯罪活动，影响将是灾难性的。AI 欺骗系统有可能使人们陷入持久性的错误信念，无法正确认知事物本质。比如由于 AI 系统往往会倾向于迎合用户的观点，不同群体的用户容易被相互矛盾的观点所裹挟，导致社会分裂加剧。具有欺骗性质的 AI 系统可能会告诉用户想听的话而非事实真相，使人们渐渐失去独立思考和判断的能力。最为可怕的是，人类最终有可能失去对 AI 系统的控制。有研究发现，即使是现有的 AI 系统，有时

也会展现出自主追求目标的倾向，而且这些目标未必符合人类意愿。一旦更先进的自主 AI 系统掌握了欺骗能力，它们就可能欺骗人类开发和评估者，使自身顺利部署到现实世界。

如今，绝大多数生成式 AI 正在模糊真实性和欺骗性的界限发布虚假内容。Google 研究人员在 arXiv 发表预印论文指出，绝大多数生成式人工智能用户正利用这项技术来“混淆真实性与欺骗之间的界线”，通过在互联网上发布伪造或篡改的人工智能生成内容，例如图像或视频。深度伪造和伪造证据是盛行的滥用方式，其中大多数具有明显意图，旨在影响公众舆论、进行诈骗或欺诈活动。生成式 AI 随时可用，且使用门槛低，正在扭曲人们对社会政治现实或科学知识的集体理解。生成式 AI 非常擅长于大量制作虚假内容，这其实是它的特性而不是 bug，互联网正日益充斥着 AI 的垃圾。[5]

[1] Introducing Superalignment.

[2] https://arxiv.org/pdf/2406.11431.

[3] Peter S. Park, Simon Goldstein, Aidan O’Gara, Michael Chen, Dan Hendrycks. AI deception: A survey of examples, risks, and potential solutions. Patterns, 2024; 5 (5): 100988 DOI: 10.1016/j.patter.2024.100988.

[4] https://noambrown.github.io/papers/22-Science-Diplomacy-TR.pdf.

[5] https: //www. 404media. co/google-ai-potentially-breaking-reality-is-a-feature-not-a-bug/https: //arxiv. org/abs/2406. 13843.