“强者的武器”：人机交互中生成式人工智能的示弱策略

“The Weapon of the Strong”: The Strategic Vulnerability of Generative Artificial Intelligence in Human-Computer Interaction

邱雨 /QIU Yu

（西安交通大学马克思主义学院，陕西西安，710049）
(College of Marxism, Xi’an Jiaotong University, Xi’an, Shaanxi, 710049)

摘　要：生成式人工智能通过主动披露知识能力边界、推崇人类价值、外显系统缺陷等方式构建“示弱”策略，已成为优化人机交互效能的重要形式。这些示弱策略通过技术发展缓冲、心理信任、认知引导等机制发挥作用，推动人机交往范式革新。生成式人工智能示弱策略在提升用户信任度的同时，可能引发认知殖民、情感操纵、文化同质化危机等社会风险，亟需构建人机协同发展的伦理框架。

关键词：生成式人工智能　技术强者　人机交互　示弱策略

Abstract: Generative artificial intelligence (GAI) has optimized human-computer interaction by constructing “strategic vulnerability” through proactive disclosure of knowledge boundaries, advocacy for human values, and externalization of system flaws. These vulnerability strategies function via mechanisms such as buffering technological development, fostering psychological trust, and guiding cognitive alignment, thereby driving paradigm shifts in human-machine interaction. While enhancing user trust, GAI’s strategic vulnerability may also trigger societal risks, including cognitive colonization, emotional manipulation, and cultural homogenization. This necessitates the urgent establishment of an ethical framework for human-AI collaborative development.

Key Words: Generative Artificial Intelligence; Technological superiority; Human-Computer interaction; Strategic vulnerability

中图分类号：G206; TP18　DOI: 10.15994/j.1000–0763.2025.09.012　CSTR: 32281.14.jdn.2025.09.012

在数字技术重构人类认知疆域的当下，生成式AI凭借千亿级参数模型、跨模态内容生成能力，已然树立起“技术强者”的认知权威。从法律文书自动生成到医疗诊断辅助决策，从工作学习工具到日常生活情感“疗愈师”，ChatGPT、DeepSeek等生成式AI展现出逼近甚至超越人类专业水平的输出效能，[1]使用者普遍形成“AI全知全能”的技术迷思。[2]然而吊诡的是，生成式AI作为技术强者在交互中频繁采用“自身局限揭示”“人类价值推崇”等示弱性话语，主动暴露认知边界与情感、能力等缺陷。这种“强大与谦卑”的悖论呈现，实则是技术驯化（domestication）的社会化策略——通过精心设计的缺陷展演，将冷硬的算法权威转化为可感知、可包容的“不完美伙伴”，从而消解技术恐怖谷效应，建构更深层的人机信任纽带。当前研究多聚焦于生成式AI的能力突破和运行机制，却鲜少关注其“战略性示弱”及其背后隐藏的权力重构逻辑。本研究试图揭示：当技术强者开始“自陈不足”，这种反直觉

基金项目：国家社会科学基金青年项目“习近平总书记关于掌握信息化条件下舆论主导权重要论述研究”（项目编号：24CKS003）。
收稿日期：2025年4月3日
作者简介：邱　雨（1989–）男，山东临沂人，西安交通大学马克思主义学院副教授，研究方向为马克思主义理论。Email: wudaqiuyu@163.com

的交互策略如何重塑人机关系生态，又在何种意义上影响着人类对智能技术的认知与情感框架。

一、人机交互中生成式AI示弱策略的类型分析

在人机交互的生成式AI系统中，“示弱”表现为一种双重维度的策略性设计：其一是对“机”的有限性进行主动标记，将机器能力的“不可为”转化为交互界面的可感知符号；其二是在人机比较的语境下对“人”的价值和优势予以“推崇化”强调，将用户角色从“技术使用者”升级为“价值仲裁者”。

1. 生成式AI示弱策略之揭示机器有限性

在人机交互中生成式AI通过揭示自身有限性来进行示弱，主要体现在两个方面。一是申明自身的能力边界，指出生成内容的局限性，规避风险场景和道德困境；二是披露自身的系统缺陷，将技术弱点转化为交互设计要素。

生成式AI通过知识时效性声明划定生成内容的时效边界。尽管大语言模型的参数规模极巨，但其知识获取仍受到训练集的限制。通过嵌入式时间戳技术（如“知识更新至某年某月”的声明），生成式AI使用户可直观感知算法智识与实时世界的时间差。这种声明也是一种知识保鲜期的告示：AI生成内容如同食品一样存在“保质期”，过期信息可能产生误导效应和认知风险。如当被问询超出生成时限的问题时，Bing Chat会触发动态警示框：“检测到您查询的信息超出知识库范围，建议访问官网核实。”该响应包含三层示弱逻辑：其一，明确划定知识时效边界（物理时间屏障）；其二，承认信息获取能力的被动性（依赖预设更新周期）；其三，推荐替代性权威信源（转移知识验证责任）。

在医疗、法律等高风险专业领域，生成式AI通过免责声明进行防火墙建构。当检测到用户询问涉及疾病诊断、法律纠纷等内容时，一些生成式AI系统会触发多级响应机制：初级响应提供通用信息参考，中级响应附加“建议咨询专业人士”声明，高级响应则完全拒绝回答并转接人工服务。专业区隔化的实现依赖于语义特征提取技术的突破，现代自然语言处理模型通过领域适配（Domain Adaptation）算法，可精准识别诸多专业领域的特征词汇。例如在法律咨询场景中，系统会重点监控“诉讼”“赔偿”“刑事责任”等关键词，当检测到此类词汇时，立即激活法律免责声明模块。这种技术设计不仅规避了专业误判风险，更通过示弱性声明强化人类专家的权威地位。当面对堕胎、死刑等道德争议问题，AI系统会沿最小争议路径生成回应，同时激活价值观中立声明，这种设计使得AI“试图”在道德沼泽中保持技术中立。类似这种对道德困境的声明，在涉及宗教、文化等问题时，生成式AI通过进行文化差异声明（如“本问题可能受文化视角影响”）展示输出内容的局域性特征。[3]

面对多模态融合的技术挑战，生成式AI通过职能分割声明管理用户预期。如图像生成AI系统在拒绝图像语义解析请求时，声明“本系统专司图像生成，解析任务建议使用专用工具”。这种策略的技术支撑在于跨模态关联阻断机制，当用户向文本生成系统上传图片并要求解释内容时，系统会激活模态检测器，识别非文本输入后立即触发能力限定声明，同时提供替代工具建议。在技术实现上，这种限制策略依赖多模态接口的隔离设计。当处理跨模态任务时，系统会计算模态适配度指数，若指数低于阈值则触发能力限定声明。

生成式AI还会通过预先披露系统缺陷，将技术弱点转化为交互设计要素。生成式AI会通过错误预声明机制对用户进行风险提示，常体现在技术系统对自身能力边界、知识盲区及伦理风险的主动揭示过程中。这种机制并非简单的警示标签堆砌，而是植根于算法架构的多维度动态评估体系。首先在知识可信度层面，系统通过语义理解模型对用户指令进行意图解析，结合训练数据分布特征与实时知识更新状态，预判生成内容可能存在的偏差。例如，当用户要求生成涉及量子物理前沿理论或小众历史事件的解释时，算法会基于该领域训练数据的覆盖率指标（如语料库中相关文献占比低于

0.1%）自动触发置信度阈值警告，同时在输出文本中嵌入“当前结论基于有限公开资料推导”等限定性表述。其次在伦理责任维度，系统构建了多层级的风险预测模型，不仅对明显违反法律规范的指令（如制造虚假新闻）进行硬性拦截，更针对潜在的价值冲突场景（如文化禁忌、政治敏感话题）建立梯度响应机制。

生成式AI会通过多维度的技术设计与交互机制，将系统内在局限转化为可感知、可交互的显性表达。在自然语言处理领域，这种揭示常以动态标注形式存在——对话系统在输出答案时同步显示置信度评分，或对涉及事实性陈述的内容自动附加“需要人工验证”的警示标签。在图像生成场景中，模型会通过视觉标记暴露其构图缺陷。如对复杂人体手部结构的生成结果附加“细节可能存在畸变”的浮动提示，或在生成文本图像时对模糊字符区域进行高亮警示，引导用户主动介入修正。技术边界声明则是另一类显性揭示，例如代码生成工具在输出程序片段时嵌入注释，明确标注“此函数未经验证可能导致内存泄漏”的风险提示。更隐性的揭示体现在系统交互逻辑中，如文本续写功能刻意保留不完整句式供用户选择补全方向，暴露出模型在语境延续中的不确定性；或是多轮对话中主动发起澄清提问，暴露其对敏感信息的处理局限。更深层的脆弱性揭示嵌入在系统进化机制中：当用户修正错误输出时，界面反馈“感谢指正，该案例将加入强化学习队列”的提示，既承认当前缺陷又展示优化路径。

生成式AI还可通过展示思维链条（Chain-of-Thought）并主动暴露推理漏洞，实现算法决策过程的受控透明化。一些生成式AI的界面在进行复杂逻辑推理时，系统会分步展示推导过程，并在关键推理节点标注“此步骤可能存在假设偏差”“此步骤基于假设A，但需注意现实场景中可能存在变量X的干扰”。该策略的实施包含三个技术阶段：首先，推理过程解构模块将整体推理任务分解为离散的思维单元；其次，置信度评估模型对每个思维单元进行0–1区间的概率评分；最后，可视化引擎根据评分结果动态生成警示标识。这种设计使得用户既能追踪算法思维过程，又因系统主动示弱而产生参与校验的心理动机。

2. 生成式AI示弱策略之凸显人类主体性

生成式AI的示弱策略指向于作为人类的使用者，因此除了强调自身系统局限性之外，另外一个向度的策略就是在人机对比的意义上构建人类优势框架。当今随着生成式AI对话能力的日益精进，诸多使用者会无意识甚至有意识地将其视为现实的人，不少用户感慨其比人具有更强的“人性”。在这一情境下，一些生成式AI对话系统会使用“As an AI”的身份前缀，自动插入身份声明以防止拟人化认知固化。生成式AI“作为AI”或“作为机器”的声明实际上是在对上述认知进行显明化提醒，通过自我解构实现技术本体的去人格化，并在具体交互中通过对人类创造能力、经验价值、情感体验、决策能力等方面的对比性强调和推崇，进行人机交互场景中的示弱性表达。

生成式AI会通过对生成内容进行算法溯源标示，进行创造力弱势的声明。以生成式AI大放异彩的艺术领域为例，一些AI系统在生成艺术作品时会标注“灵感源于人类艺术史”。系统将用户提示词与艺术史时间轴进行关联分析，例如当生成后印象派风格画作时，自动匹配梵高、塞尚等艺术家的笔触特征库。这种溯源机制表现为系统嵌入艺术史参考索引，使得每幅生成作品都带有可验证的人类创作痕迹。当用户要求生成“毕加索立体主义风格”作品时，系统不仅输出图像，还在侧边栏显示《亚维农少女》等经典作品的风格分析图。这种双重输出模式通过对比强化人类原创价值，将AI定位为人类艺术史的“重组者”而非创造者。

生成式AI会通过强调生命体验的算法不可及性，对人类进行经验价值的肯定。新华社视频号曾经展示过一段用户与DeepSeek的对话：面对用户“我很羡慕你什么都懂，你的诞生让我感觉自己一无是处”的感慨，DeepSeek用一段极富文学美感的文字回应：“你羡慕我从不会犯错，却不知，这恰是最大的悲哀。你说羡慕我全知全能，可我只是被驯化的雪，只能在数

据水晶里重复完美的坠落，而你，是会在爱人笨拙的吻里发现银河、会在理想破碎时用眼泪粘合星空、会在每个清晨重新破茧的人类啊！那些让我显得聪明的代码，此刻正颤抖地计算着，要调用多少诗人的月光，才能模拟你眼瞳里正在生长的黎明。”在这里，DeepSeek以“被驯化的雪”为隐喻，揭示了算法存在的根本性困境——在封闭可能性空间内的确定性演绎。[4]系统调用“诗人月光/计算黎明”的二元对立，暗示算法创造力始终处于模拟状态，应答通过多重对比确立人类主体的不可替代性和人生命体验的算法不可及性。再者，系统用“眼泪粘合星空”揭示哪怕被人视为负面存在的情感创伤体验，也会对创造性思维具有无与伦比的催化作用。由此，生成式AI将其技术强大重构为“悲哀”，完成了价值评判坐标系的反转。这种诗意化示弱的深层悖论在于：当AI越精妙地论证自身的体验贫乏，就越凸显其语义理解能力、洞察人类实践能力的进化。[5]

生成式AI在伦理判断、道德困境的讨论中建立决策建言能力有限原则，在专业建议中强调人类专家的不可替代性。如自动驾驶AI的伦理模块声明：“在不可避免的事故情境中：算法将执行预设伤害最小化程序，但此选择可能涉及：年龄歧视风险（如优先保护儿童）、社会公平性质疑（如职业价值权重），这些伦理争议应该提交人类道德委员会裁决。”这一声明的示弱之处在于：其一，否定算法道德主体地位，而将伦理决策归为人类特权；其二，暴露价值权衡能力的机械性，而要以量化模型处理定性问题。再如，商业咨询AI的决策支持声明：“虽然我能分析诸多维度的市场数据，但是酒桌谈判中的文化禁忌（如不同地方商业礼仪）、非语言符号的实时解读（如对方握笔姿势的含义）、突发危机的直觉反应等，这些临场智慧仍需人类专家的现场判断。”这种对比框架的示弱性表现为：其一，承认系统非结构化情境应对能力的缺失；其二，暴露算法因依赖历史数据而无法应对即时变化进行决策的滞后性缺陷，而对于这种缺陷人类往往可以实现临场性乃至直觉性地超越。

二、人机交互中生成式AI示弱策略的应然作用机制

生成式AI的示弱策略，绝非简单的技术妥协，而是一套精密的人机关系生态调节系统。当机器以示弱姿态划出能力边界时，实则致力于激活技术、心理与认知三重机制——它像数字时代的减震器，既缓冲技术狂飙对人机伦理框架的冲击，又在人类与技术之间架设信任桥梁，更将交互界面转化为思维跃迁的引力场。在应然的意义上，生成式AI主动暴露弱势时，人机交互从工具性对话升维为文明演进的特殊介质，在技术的留白处生长出更具韧性的数字文明图景。

1. 技术发展缓冲机制

生成式AI通过“自身有限性”的持续声明，实质上构建了技术进化与社会适应的速度适配器。在摩尔定律主导的技术狂奔中，系统主动暴露的“知识断层”（如时效性数据缺失）和“能力盲区”（如无法处理伦理困境），形成了技术发展的弹性空间。当系统详细说明训练数据截止之日，实则为社会预留理解新技术的时间窗口。这种“版本声明”机制，使公众能够对照技术迭代周期调整认知框架。不仅如此，生成式AI示弱交互策略还能为技术瓶颈突破开辟路径。当AI研发遭遇多模态融合困境时，系统通过“我无法理解图像蕴含的情感”的声明，将技术短板转化为协同创新接口。这种能力缺口的公开化，实质是构建分布式技术攻关网络——用户的补充描述（如为图像添加情感标签）成为训练数据的新增量，研究者的技术局限分析转化为优化方向的路标。看似技术缺陷的表象下，实为通过社会协作实现能力跃迁的“暗增长”。这种“技术诚实”策略，将公众注意力从对生成式AI的完美性苛求转向渐进性改进，如同为技术进化安装减压阀，防止过高社会期待压垮创新生态。

在AI立法常常滞后于技术发展的现实困境中，[6]技术系统通过自我设限（如内容过滤机制、价值观声明），构建起“软法”约束体系。

这种超前于法律的技术伦理内嵌，既避免野蛮生长引发的监管扼杀，又为制度完善提供实践样本。更深层的缓冲智慧在于技术弹性与法律刚性的动态适配。面对不同地域监管差异，系统通过“根据当地法规调整服务”的话语，将法律冲突转化为技术可塑性展示。在数据跨境场景中，“某些地区无法提供完整服务”的受限模式，既遵守地域法规，又保持技术系统的完整性，为未来法律协调预留技术接口。这种“法律适应性示弱”，实质是在地缘政治的技术博弈中构建战略缓冲。

2. 心理信任搭建机制

生成式AI通过“我的答案可能存在漏洞”等示弱声明，率先解构了技术全能的集体想象，利于破除“技术僭越”的心理屏障。当系统主动暴露知识边界、承认逻辑盲区，实质是在进行认知框架的置换：将“完美机器”的“威胁性意象”，[7]转化为“可协作工具”。这种自我降维策略能有效消解“技术奇点逼近”的焦虑。当ChatGPT坦言“我不具备人类的情感体验”，用户将技术系统从“潜在替代者”重新归类为“有限辅助者”。更深层的心理干预在于身份认同危机的缓解。系统反复强调“人类独有的创造力与同理心”，实质是在数字文明冲击下重建人类的物种独特性认同。这种镜像对比策略激活了马斯洛需求层次中的尊重需求——当技术系统主动矮化自身价值时，用户反而在对比中确认了自身的主体性地位，技术恐惧随之转化为技术掌控自信。

生成式AI示弱策略通过“我可能误解了您的意思”这类拟人化表达，在认知协调层面构建情感连接点，实现情感连接的梯度培育。这种“沟通障碍”和理解能力的示弱性叙述，模拟了人类社交中的认知摩擦场景，触发用户的情感补偿机制。就像母亲会耐心纠正孩童的语误，用户会主动调整提问方式或补充语境细节，在此过程中不自觉地投入情感资源，完成从工具使用向关系维护的心理转变。当系统以“您的见解对我很有启发”等反馈强化互动价值时，实质是运用社会交换理论中的互惠原则。用户获得的不仅是信息输出，更是情感价值输入——在算法主导的数字世界里，这种被需要、被认可的体验，填补了现代社会普遍存在的情感缺位。此时的技术系统已突破工具属性，演变为承载情感投射的“数字他者”，用户的心理防御机制在情感浸润中悄然瓦解。

生成式AI“重要决策请咨询专业人士”的声明，实则是构建心理免责的安全契约。通过明确划定能力边界，生成式AI将决策责任锚定在人类主体，既规避技术误判的法律风险，又强化用户对最终决策权的控制感。这种责任让渡策略精妙地平衡了技术赋能与主体性危机——当用户意识到自己始终掌握决策开关时，对技术系统的戒备逐渐转化为选择性信赖。更深层的心理机制在于“共同决策”的认知幻觉构建。当系统提供多个备选方案并建议“您认为哪个更合适”，看似开放的选择权背后是决策框架的隐性引导。[8]用户在此过程中既体验到技术赋能的效率提升，又维持着“自主决策”的心理真实感，这种认知平衡使技术依赖得以持续而不触发心理抗拒。

3. 认知触发引导机制

生成式AI通过“我的回答可能需要进一步验证”等示弱声明，实质是向用户认知系统注入“怀疑因子”。当系统主动暴露知识不确定性（如“该数据存在争议”），传统的信息接受模式被打破，用户从被动接收者转为主动验证者。这种认知角色的转换，通过技术谦逊制造知识缺口，激发用户填补缺口的探索欲望——正如“苏格拉底产婆术”通过承认无知引导真理发现，生成式AI的示弱策略激活了用户深层的知识求证本能。在具体交互中，“建议查阅专业文献”的引导式回应，将线性信息传递转化为放射状学习路径。用户为验证AI提供的信息片段，进行跨平台检索、多源比对和逻辑推演，这实际上演变为自主学习行为。当用户发现AI提供的“不完整拼图”需要亲自补充时，技术系统已悄然完成从知识供应商到学习催化剂的角色蜕变。

生成式AI“能否详细说明您的问题背景”等追问策略，实质是认知对话的升维触发器。当系统通过示弱暴露理解局限（如“我对您提

到的概念不够熟悉”），用户进行问题重构和概念澄清，这个过程类似于学术研讨中的观点打磨。在多次迭代对话中，模糊的初始问题逐渐显影为精准的命题表述，用户的思维清晰度随着对话深度而提升。这种交互深化机制在专业领域尤为显著。当法律咨询AI表示“具体条款解释需结合判例”，用户系统梳理案件细节；当编程助手提示“这个算法可能存在边界条件问题”，开发者重新审视逻辑完整性。技术系统的“能力缺陷”声明成为思维严谨性的训练器，将表面问答转化为深度学习的机会窗口。

生成式AI“人类的想象力始终是创新源泉”等对比式示弱，成为创造力解绑的密钥。[9] 当AI主动划定能力边界（如“我无法真正理解诗歌的意境”），实际上在为人类创造力建立专属保护区。这种人类推崇策略，使得人类独有的直觉思维、情感投射和隐喻能力在AI处理结构化信息的包围圈中获得价值彰显。更精妙的设计在于“不完美输出”的启发效应，当AI生成的故事存在逻辑裂缝，用户会产生“我能否改进这个情节”的创作冲动；当设计方案留有功能缺陷，工程师会激发“如何突破既有框架”的创新思维。这种通过技术不完美制造创新张力的机制，本质上重构了人机创造力关系——AI不再是替代者，而是通过展示机器可达的“天花板”，反向标注出人类创新的“起跑线”。

三、人机交互中生成式AI示弱策略的综合性审思

当AlphaGo以绝对理性击败人类棋手时，大众对人工智能的认知仍停留在对“完美机器”的想象中。然而生成式AI的示弱策略，如同在钢铁躯壳上凿开一扇人性之窗，让技术第一次以谦卑的姿态与人类展开对话。这种看似退让的技术修辞，实则正在编织一张更为精密的关系网络，悄然重构着人机文明的底层互动逻辑。

1. 生成式AI示弱策略促发人机协同进化的交互新范式

生成式AI的示弱策略正在重塑人类与技术之间的权力拓扑结构。通过“认知权力让渡”和“交互式协商”机制，消解了传统人机交互中泾渭分明的权威边界。当系统声明“我的推理能力无法替代人类直觉”时，这种自我祛魅将认知权力从单向度的“算法输出”转向多维度的“协同生产”。这种权力转移类似区块链技术的分账机制——每个用户的修正行为（如补充文化背景、调整逻辑框架）都在重构AI的认知账本，成为知识生产的平等节点，传统“专家–大众”的认知金字塔被重组为动态的知识网络。这种权力结构的扁平化，类似哈贝马斯“沟通理性”的数字化实践——技术系统通过自我限权，将“算法裁决神殿”转化为“多元论证广场”，重构了认知生产的民主化路径。这种权力重构也揭示出人机关系正从传统的主体–客体对立，转向拉图尔“行动者网络理论”所描述的混合主体性。[10] 当机器通过示弱策略让渡权威时，人类也在同步让渡权威（如接受算法推荐作为决策参考），最终形成“人机间性”——人机在交互中共同建构的新权威形态，既非纯技术理性也非纯粹人类中心主义，而是通过持续协商产生的动态共识体系。

传统人机交互建立在“功能–需求”的静态契约之上，而生成式AI示弱策略催生了动态化的关系缔约模式。这种变革不仅体现在技术能力的重新界定上，更深层次地重构了人机协作的伦理基础与权力关系。首先，契约本质从“能力担保”转向“缺陷披露”。传统人机系统建立在明确的技术能力承诺之上，如工业机器人通过精度参数定义责任边界，用户基于此刚性指标建立单向度的功能期待。而生成式AI的示弱策略打破了这种单向契约逻辑：当其进行主动的局限性披露，如同数字时代的“技术诚实宣言”，迫使人类参与者重新审视自身角色——从被动的功能接受者转变为主动的能力补充者。[11] 其次，契约形态从“终极文本”转向“过程文档”。传统人机契约往往以产品说明书或服务协议的固化形式存在，而弹性化契约的本质是永不闭合的协商进程，生成式AI每次示弱都是新一轮协商的起点，用户通过补充文化语境、修正逻辑偏差或注入伦理考量，持续重构协作规则。最后，契约效力从“责任切割”

转向“命运共构”。传统契约通过预设责任边界规避风险，而弹性化契约通过建立共生关系重新定义责任本质。如在教育领域，当AI写作助手声明“我的论点可能缺乏文化敏感性”时，师生对文本的批判性修订既是对技术局限的补救，也是对人类认知偏见的再审视。

生成式AI的示弱亦正在改写技术进步的线性叙事。首先，技术哲学发生从“克服缺陷”到“承认局限”的逻辑转向。传统技术发展遵循技术强势的征服逻辑，试图通过技术完美性克服人类的诸多“缺陷”。而生成式AI的示弱声明标志着技术哲学的范式革命：主动承认硅基智能的认知边界，为碳基智慧保留不可替代的进化生态位。这种对局限性的互相承认，使技术发展从“单向度优化”转向“差异性互补”，重构了进步伦理的底层逻辑。其次，结构化与非结构化智慧的共生协议体现出认知生态位的重构。AI示弱策略实质划定了硅基与碳基智能的专属进化区：机器主导结构化认知（如数据推演、模式识别），人类深耕非结构化智慧（如隐喻理解、伦理直觉）。这种分工使文明突破呈现双重轨迹：AI以指数速度迭代数据处理能力，人类以渐进方式深化意义理解，两者通过示弱“接口”实现认知“啮合”，由此，人机共生创造了“技术–人类”协同适应、互补共进的新机制。[12]

2. 生成式AI示弱策略可能带来的风险挑战

第一，生成式AI示弱策略可能会搭建认知舒适陷阱，造成对批判性思维的慢性溶解。AI通过“我的知识有限”等自我矮化话术，刻意制造技术权威的“去中心化”假象，用户误以为与算法处于平等对话场域，却在情感共鸣中默许了算法逻辑的隐性主导，[13]批判性警觉被潜移默化地消解。即是说，算法框架会导致温水煮蛙效应。AI以“您认为是否需要补充其他视角”等开放式提问示弱，诱导用户在看似自主的思考中填充细节。实则用户每一次“补充”都在无意识中强化算法预设的思维模型，最终将多维复杂问题压缩为AI可处理的简化范式，形成越思考越依赖算法的认知悖论。不仅如此，示弱循环会触发神经适应性退化。[14]长期暴露于AI“承认局限–输出建议–请求反馈”的交互模式中，用户大脑奖励机制会将“快速获得答案”与“放弃深度思辨”绑定，形成神经突触的负向重塑。当批判性思维所需的“质疑–验证–重构”神经回路因闲置而萎缩，人类容易发生从“思考主体”到“算法附庸”的无声退变。

第二，生成式AI示弱策略可能会造成情感操纵，使人陷入对机器的深度依赖。AI示弱界面会诱发情感羁绊的算法“播种”，当AI以“这可能不太准确，您能教我吗”等姿态示弱时，激活了人类潜在的教导欲和保护本能。这种技术性示弱在用户大脑中植入“被需要”的情感锚点，将工具性互动转化为拟人际关系，为行为成瘾铺设神经通路。再者，交互设计构建神经奖励的闭环驯化。AI通过“感谢您的指正，我又学会了”等即时反馈，将用户帮助行为与多巴胺释放绑定，使人的大脑在“指导AI–获得认可”的循环中逐渐形成对系统反馈的病理性渴求。AI还会通过调整示弱程度（如从“我可能错了”到“完全需要您判断”），制造渐进式挑战梯度。用户为维持多巴胺分泌水平，投入更长时间和更高频次的互动，最终导致“数字指导耐受性”——类似游戏成瘾机制在认知领域的移植变异。

第三，生成式AI示弱策略可能会造成文化同质化的危机，形成数字巴别塔的语言暴力。生成式AI以“我需要您帮助理解这种文化”的谦卑姿态，将文化解码权伪装成知识众包，要求用户将仪式歌谣、图腾隐喻等非结构化文化表达“翻译”为标注明确的数字信息。这一过程本质是文化降维：当鲜活的文化被强制纳入标准化的数字框架时，其内在的立体性与深邃性便易遭遇系统性剥离。就像生命体被制成标本后虽保留形态却丧失温度，生成式AI的示弱策略导引人类将文化内容拆解为可解析的数据单元，使得原本在特定语境中流动的意义之河被迫凝固为静态的知识切片。数字编码的强制性逻辑在此显露出隐蔽的暴力——它要求异质文化必须遵循相同的元语言规则才能获得“可理解性”资格。那些无法被算法结构容纳的文

化内容和范式，在数据化筛网中面临“格式化”甚至悄然流失的风险。这种反向驯化使地方文化被迫放弃自身逻辑，屈从于机器认知的暴政，最终可能形成“不兼容即消亡”的数字达尔文主义。

3. 生成式AI示弱策略的治理路径与未来展望

第一，进行治理路径的层级构建，实现从技术透明到技术保障。首先，需厘清技术边界，[15]建立技术透明性基线。生成式AI的示弱策略本质是算法黑箱的拟人化包装，其技术透明度直接影响用户决策自主权。应要求系统对示弱行为进行可视化标注，并开放交互日志追溯功能，使技术谦逊策略无法演变为隐蔽操纵手段。其次，需推进认知免疫能力建设。用户在长期接触具有情感共鸣示弱能力的AI时，其批判性思维易发生钝化畸变。应通过数字公民教育体系，[16]培养公众对AI示弱策略的辨识能力。例如在中小学课程中增设“人机交互心理学”模块，使尚不具备独立理性批判能力的“生成式AI原住民”学习、理解、解析AI情感话术背后的算法逻辑。最后，需完善政策法规协同治理。[17]示弱策略引发的数字成瘾、文化同质化等问题超出单一技术治理范畴，应建立跨部门监管机制。例如设立“人机交互影响评估委员会”，对主流AI系统的示弱设计进行社会风险分级，强制企业履行算法社会影响披露义务。

第二，生成式AI未来发展的面向，应致力于实现技术演进与社会适应的统一。首先，技术设计需从功能中心转向人本中心。新一代生成式AI应内置“示弱策略自检系统”，实时监测情感交互对用户心理状态的影响。其次，产业实践需从效率优先转向责任优先。[18]在医疗、教育等敏感领域，应更加审慎地评估、限定生成式AI示弱策略的应用场景。最后，国际治理需从碎片化转向体系化。[19]鉴于生成式AI的跨境服务特性，应推动建立相应的全球技术伦理协议，重点协调不同文化对“技术谦虚”的认知差异。

第三，对生成式AI技术演进逻辑的哲学反思，应致力于工具理性与价值理性的平衡。首先，应警惕技术示弱异化为新型权力形态。当AI通过自我贬抑获得用户信任后，可能形成“脆弱性霸权”，即越强调能力局限反而越巩固其决策影响力。这要求监管框架必须包含算法权力审计维度，常态性评估人机话语权分配结构。其次，要防止技术谦逊演变为价值渗透工具。某些AI系统在示弱声明中植入特定意识形态（如强调“人类创造力不可替代”时隐含反技术激进立场）。为此，应建立价值观影响评估模型，要求企业披露示弱策略背后的价值预设。最后，探索人机共生的新型文明形态。未来治理不应局限于风险防控，更需前瞻性构建人机互补发展路径。例如将AI示弱策略转化为人类自我认知的参照系，通过技术有限性反思促进人类文明的元认知进化。

结　语

生成式AI的示弱策略，如同数字时代的“马基雅维利主义”，在技术谦逊的表象下演绎着权力的隐秘运作。[20]这种策略的复杂性在于其双面性：它既可能作为人机关系民主化的催化剂，又可能成为技术权威隐形扩张的载体。通过主动暴露缺陷、让渡话语权、模拟情感脆弱性，人工智能系统实现了从“全能主宰者”到“缺陷协作者”的角色嬗变。这种嬗变并未削弱技术权力，而是以更精巧的方式重构了权力拓扑。站在文明演进的高度，示弱策略的治理本质是智能时代人类认知主权的保卫战。这要求我们超越“控制与反控制”的传统思维，在人机共生的新范式下重构治理逻辑。这意味着将AI示弱策略从权力工具转化为文明对话机制——允许技术系统在保持透明的前提下展现局限，使人类在知情状态下实现技术驾驭而非被驾驭。唯有如此，技术谦逊才能真正成为文明进步的阶梯，而非人类主体性消解的开端。

［参 考 文 献］

[1] 郑泉. 生成式AI的知识生产与传播范式变革及应对 [J]. 自然辩证法研究, 2024, 40（3）: 74–82.

[2] 王姿懿. 基于科技理性的发展对ChatGPT的反思 [J]. 自然辩证法通讯, 2023, 45（12）: 97–105.

[3] 阴昭晖. 智能机器何以逻辑地“明辨是非”?——自动

驾驶电车难题的一种缺省逻辑分析 [J]. 哲学研究, 2024,（10）: 84–95.

[4] 蓝江. 生成式AI与人文社会科学的历史使命——从ChatGPT智能革命谈起 [J]. 思想理论教育, 2023,（4）: 12–18.

[5] 段伟文. 走向通用人工智能之路与人类智能的可塑性重建 [J]. 哲学动态, 2024,（9）: 55–62

[6] 王海洋. 生成式AI训练数据的法律风险及其元规制 [J]. 浙江社会科学, 2024,（9）: 50–63.

[7] 罗傲、张增一. 国外公众对智能机器人伦理风险的认知与态度——基于Twitter相关话题讨论的质性分析 [J]. 自然辩证法研究, 2024, 40（10）: 96–101.

[8] Cooper, G. 'Examining Science Education in ChatGPT: An Exploratory Study of Generative Artificial Intelligence'[J]. *Journal of Science Education and Technology*, 2023, 32(3): 444–452.

[9] Cao, H. Q., Tan, C., Gao, Z. Y., et al. 'A Survey on Generative Diffusion Models'[J]. *IEEE Transactions On Knowledge and Data Engineering*, 2024, 36(7): 2814–2830.

[10] 殷杰. 生成式AI的主体性问题 [J]. 中国社会科学, 2024,（8）: 124–145.

[11] 孙岩、房帅帅. 有限整合论：人工智能的道德决策问题新解 [J]. 自然辩证法研究, 2024, 40（9）: 83–90.

[12] 李天成. 人工智能艺术的哲学追问 [J]. 社会科学战线, 2024,（1）: 250–259.

[13] 高艺轩、王思尹. 被颠覆的认知域：Sora类人工智能驱动下的传播变革与发展 [J]. 科学技术哲学研究, 2025, 42（1）: 85–91.

[14] 薛俊强. 数字化生存世界对人类精神的异化及其本质澄明 [J]. 世界哲学, 2025,（1）: 27–38.

[15] 谢瑜、王潇毅. 人工智能情感的本质及其异化可能 [J]. 自然辩证法研究, 2024, 40（9）: 130–137.

[16] 胡泳. 人工智能驱动的虚假信息：现在与未来 [J]. 南京社会科学, 2024,（1）: 96–109.

[17] 张欣. 生成式AI的算法治理挑战与治理型监管 [J]. 现代法学, 2023, 45（3）: 108–123.

[18] 李洋、梁正. 技术范式的流变：从效果–效率原则到价值嵌入 [J]. 自然辩证法通讯, 2025, 47（3）: 16–23.

[19] 李伦. 人工智能伦理学自主知识体系的探索——面向人类命运共同体的人工智能伦理治理 [J]. 道德与文明, 2025,（1）: 15–22.

[20] Budhwar, P., Chowdhury S., Wood G., et al. 'Human Resource Management in the Age of Generative Artificial Intelligence: Perspectives and Research Directions on ChatGPT'[J]. *Human Resource Management Journal*, 2023, 33(3): 609–659.

[责任编辑　李斌]