引用格式：马晓飞，王佳. 何以实现生成式人工智能技术的安全治理？——基于信息生态系统的博弈分析[J]. 技术经济，2026，45(3)：81-97.

Ma Xiaofei，Wang Jia. How to achieve the secure governance of generative artificial intelligence technology? A game-theoretic analysis based on the information ecosystem[J]. Journal of Technology Economics，2026，45(3)：81-97.

技术经济管理

何以实现生成式人工智能技术的安全治理？

——基于信息生态系统的博弈分析

马晓飞，王 佳

（北京邮电大学经济管理学院，北京 100876）

摘 要：生成式人工智能的飞速发展在加快形成新质生产力等方面成效显著，但其面临的数据、信息、内容等方面的治理挑战亟待解决，急需在技术发展与安全之间寻找平衡点。基于信息生态系统理论，构建生成式人工智能服务管理者、提供者、使用者三方主体协同治理博弈模型，进行数值仿真分析发现：①生成式人工智能安全治理是一个多主体协同参与的过程，该过程受生态因素影响，加大各主体的共治收益及负面治理行为带来的损失都有利于促进信息生态系统稳定于｛严格监管，积极治理，合规使用｝的理想状态。②形象因素方面，提高生成式人工智能服务管理者的监管形象收益和提供者的声誉形象收益都可促进信息生态系统向理想状态进化。③奖惩因素方面，提高管理者对提供者的奖励及提供者对使用者的奖励虽然会使实施奖励一方自身利益稍有损失，但可促进整个信息生态系统向理想状态进化，其中管理者给予的奖励存在有效阈值；加大管理者和提供者实施惩罚的力度可促进三方协同共治。④超额收益因素方面，压缩提供者及使用者的超额收益空间可促进三方协同共治。研究结论为制定生成式人工智能技术安全治理策略提供了理论支撑，进而推动形成健康有序的网络生态。

关键词：发展与安全；生成式人工智能；多主体协同治理；信息生态系统；演化博弈

中图分类号：TP18；G350 **文献标志码**：A **文章编号**：1002-980X(2026)03-0081-17

DOI：10.12404/j.issn.1002-980X.J25101703

一、引言

随着ChatGPT等生成式人工智能大模型的广泛运用，人类正迎来以智能化为显著特征的新时代产业革命[1-2]。生成式人工智能(GAI)凭借其卓越的多模态理解能力与多类型内容生成能力在短时间内得到了快速发展，并基于其赋能特性，依托大数据、先进算法与复杂模型，有效促进了教育[3]、法律[4]、医疗[5]、金融[6]等各行各业的繁荣发展。然而，生成式人工智能技术是一把双刃剑，在促进经济发展和便利人类生活的同时，也带来诸多安全隐患，如算法风险、深度伪造风险等[7]。因此，如何实现生成式人工智能技术的安全治理成为亟待解决的关键问题，是事关网络安全及国家安全的重要议题。

目前，中国生成式人工智能技术发展迅猛。截至2025年12月1日，已有663款大模型服务完成备案①，相关技术和应用呈现出积极有序发展的良好态势，但仍需进一步加强数据隐私、虚假信息、内容质量等方面的治理，着力构建良好网络生态。对此，2025年11月28日，习近平总书记在中共中央政治局第二十三次集体学习时强调，网络生态治理是网络强国建设的重要任务，事关国家发展和安全，事关人民群众切身利益。

收稿日期：2025-10-17

基金项目：北京市社会科学基金“北京市人工智能大模型安全风险评估与治理长效机制研究”(25JCC128)

作者简介：马晓飞(1979—)，博士，北京邮电大学经济管理学院教授，博士研究生导师，研究方向：人工智能安全治理；(通信作者)王佳(1997—)，北京邮电大学经济管理学院博士研究生，研究方向：人工智能安全治理。

①数据来源于《推动网信事业高质量发展 开创网络强国建设新局面——访中央网信办主任庄荣文》，https://www.cac.gov.cn/2025-12/02/c_1766396564941361.htm。

要健全网络生态治理长效机制,着力提升治理的前瞻性、精准性、系统性、协同性,持续营造风清气正的网络空间。党的二十届四中全会通过的《中共中央关于制定国民经济和社会发展第十五个五年规划的建议》中明确提出要深化网络空间安全综合治理。党的二十大指出要健全网络综合治理体系,推动形成良好网络生态。这为中国网络安全工作指明了方向,也为生成式人工智能技术的监管提供了政策依据。实际上,生成式人工智能技术的安全治理过程是信息生态系统中各主体之间利益博弈的过程,治理困境是利益失衡造成的结果。因此,各主体应综合考虑经济效益和社会效益,选择有利于信息生态系统可持续发展的策略,形成网络治理强大合力,构建更为全面、系统、协同的治理体系。

由于生成式人工智能的安全治理过程涉及多个主体,各主体行为策略会受其他主体策略动态调整变化,而演化博弈方法能够很好地模拟这种多主体动态策略互动及调整过程。因此,本文先基于信息生态系统视角构建生成式人工智能信息生态系统进化模型,识别生成式人工智能安全治理的主体及主要治理困境;接下来,构建多主体协同治理的博弈模型,通过仿真模拟,探讨各主体参数对演化博弈系统稳定性的影响,进而提出相应的治理建议,促进多主体协同治理,共同推动形成健康有序的生成式人工智能网络生态。

二、文献综述

(一)生成式人工智能治理的相关研究

关于生成式人工智能的治理研究,主要围绕数据、信息、内容三个方面展开。首先,生成式人工智能的每一次迭代训练都离不开数据的支撑,而不断发展的生成式人工智能技术又会产生新的指数级的数据。因此,数据安全治理作为生成式人工智能治理的基础环节,近年来被学者广泛研究,其研究内容主要包括全链条风险防范机制分析[8]、数据主体责任矩阵构建[9]、治理体系与优化路径分析[10]等。其次,信息作为数据和内容的中间形式,既是数据的加工结果,同时也是内容的核心。现有关于信息安全治理的研究,主要围绕虚假信息治理展开,研究内容包括虚假信息的层级化运行机理分析[11]、多主体协同共治策略分析[12]、治理路径分析[13]等。最后,内容作为数据和信息的最终呈现形式,具有一定价值和意义,兼具多样性、丰富性等特点。然而,由于生成式人工智能大模型具有不稳定性,使得生成内容的质量和准确性也有很大差异[14-15],部分人工智能生成内容存在缺乏深度、同质化严重等缺陷[16]。因此,内容维度的治理研究主要集中在人工智能生成内容的质量评价方面。在评价指标上,涉及三个研究视角,一是内容可信度视角,该视角认为人工智能生成内容质量受内容可信度的影响[17-18];二是内容匹配度视角,该视角认为人工智能生成内容质量受生成内容与预期目标、信息环境等方面的匹配程度的影响[19-20];三是主客观融合视角,该视角认为人工智能生成内容质量受主观和客观两个方面的影响[21-22]。在评价方法上,以经验驱动的方法为主,借助用户反馈、领域知识、专家经验等对人工智能生成文本内容的质量进行评价,突出用户的感知质量[23-25]。

(二)信息生态系统相关研究

信息生态系统是从生态理论延伸而来的,其本质是信息、信息人、信息技术、信息环境等因信息活动结合而成的动态复杂网络,各要素之间相互依存、相互影响,共同推动信息生态系统进化[26-27]。现有关于信息生态系统的研究,主要涉及三个视角:一是大数据分析视角,包括医疗健康大数据资产价值实现[28]、数据要素市场构建[29]等;二是数字化平台建设视角,包括在线健康社区[30]、高校智慧图书馆[31]等;三是网络舆情传播视角,包括突发公共卫生事件网络舆情传播[32]、企业危机事件网络舆情传播[33]等。除此之外,随着新一代信息技术的发展,传统信息生态系统也逐渐向数智信息生态系统演化发展,以期通过信息生态系统的内生智慧加快形成新质生产力[34]。

(三)文献述评

现有关于生成式人工智能的治理研究,主要围绕数据、信息、内容单一层面展开,鲜有研究基于系统的视角,构建较为全面的生成式人工智能治理模型。现有关于信息生态系统的研究,在研究内容上多集中于数字化发展领域,包括大数据分析、数字化平台建设、网络舆情传播等,缺乏对智能化领域的分析,缺少向数智信息生态系统转化的内容。综上所述,本文基于信息生态系统视角构建生成式人工智能的信息生态系统模型,以明确各治理主体及主要治理困境,并以此为基础,构建多主体协同治理的博弈模型,通过仿真模拟,

探讨各治理主体参数对演化博弈系统稳定性的影响，进而提出相应的治理建议，着力构建良好网络生态。

三、生成式人工智能的信息生态系统构成

信息生态系统是由信息人、信息、信息技术、信息环境所组成的具有一定自我调节功能的人工系统。本文以信息生态系统理论为指导，以豆包、文心一言、ChatGPT、DeepSeek 等国内外知名大模型作为现实背景，全面覆盖信息生态系统所需的核心要素，以便于抽象提炼相关概念。采用理论与实践相结合的方式构建生成式人工智能的信息生态系统模型，如图 1 所示。

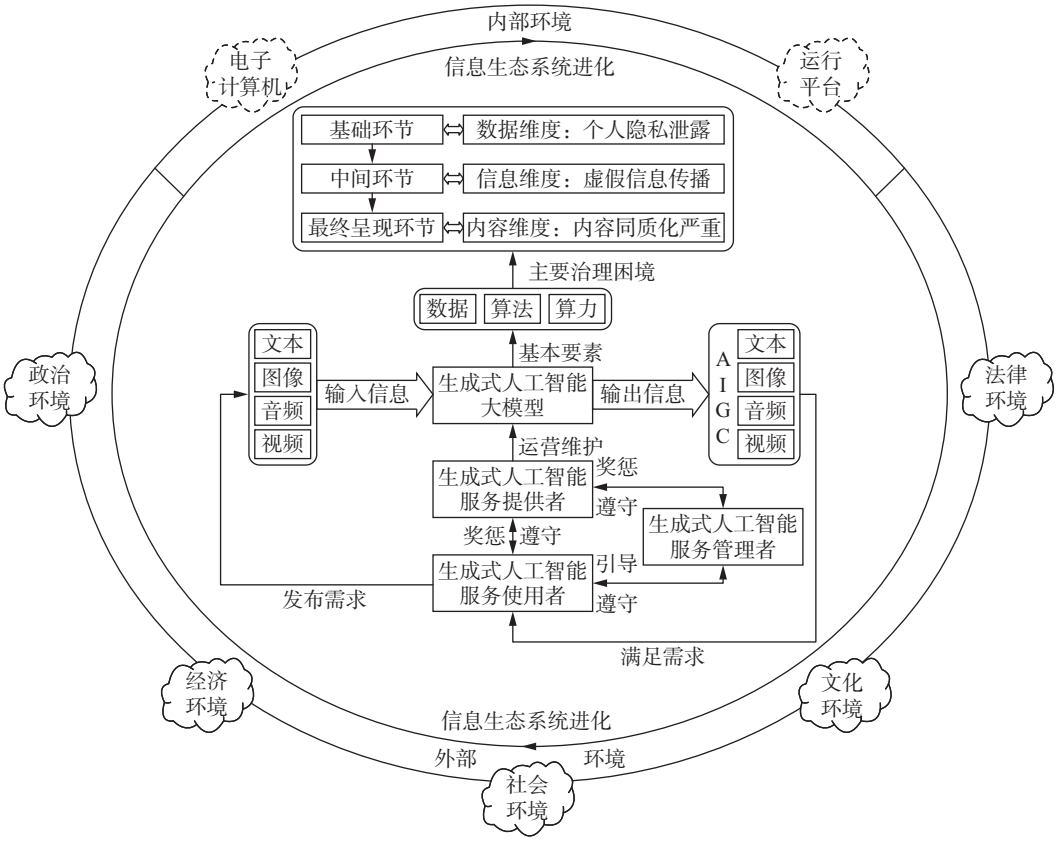


图 1　生成式人工智能的信息生态系统模型

（一）信息人

信息人是生成式人工智能信息生态系统的主体，是信息的生产者与传播者，主要包括生成式人工智能服务管理者、生成式人工智能服务提供者、生成式人工智能服务使用者这三类。其中，生成式人工智能服务管理者是指维护生成式人工智能安全的相关政府组织，主要任务包括制定监管政策、设立奖惩制度等；生成式人工智能服务提供者是指大模型运营商，主要任务包括训练优化生成式人工智能大模型、提供大模型使用权限、提供大数据及云计算服务等；生成式人工智能服务使用者是指使用生成式人工智能大模型的用户，用户类型可以是个人用户，也可以是企业用户，这些用户通过向大模型输入信息，来获得自己所需要的输出信息。这三类主体通过动态互动与信息生态系统的其他要素产生链接，共同推动信息生态系统良性运转。

具体来讲，生成式人工智能服务管理者通过制定政策规范提供者的行为，如中国于 2023 年 7 月 11 日出台的《生成式人工智能服务管理暂行办法》中明确提到提供者应当依法承担网络信息内容生产者责任，履行网络信息安全义务。涉及个人信息的，依法承担个人信息处理者责任，履行个人信息保护义务。这在一定程度上约束了信息技术的使用边界，营造了良好的信息环境，净化网络空间。生成式人工智能服务提供者通过训练优化大模型将服务使用者的输入数据转化为有价值的输出信息，从信息源头控制生成内容质量；

生成式人工智能服务使用者通过需求指令触发系统运转,其对输出内容的评分或修改又反向优化服务提供者的模型训练,推动信息技术的迭代升级。特别说明的是,上述三类信息人之间的关系是生成式人工智能服务管理者引导使用者的行为,同时通过实施奖惩措施约束提供者的行为;生成式人工智能服务提供者既要遵守管理者的监管政策,又要采取奖惩措施约束使用者的行为;生成式人工智能服务使用者则表现为要遵守管理者和提供者的双重监管措施。

(二)信息

信息是生成式人工智能信息生态系统的客体,包括输入到生成式人工智能大模型里的信息和从大模型中输出的信息,输入信息和输出信息的内容形式可以是文本、图片、音频、视频的单一形式,也可以是多模态融合的形式。其中,输出信息即为人工智能生成内容(AIGC),在Web3.0时代,AIGC成为继专业生成内容(PGC)和用户生成内容(UGC)之后的全新的内容生成模式[35-36]。不同于单一模态的人工智能生成内容,多模态人工智能生成内容具有更丰富的表达方式和更复杂的信息关系。“信息”这一维度对应了生成式人工智能基本要素中的“数据”,输入信息可以看成原始数据,经过大模型算法的处理和转换,输出具有价值的信息内容。信息作为流动载体,通过技术转化、主体交互和环境适配,实现从“数据”到“价值”的升华,是信息生态系统运转的核心纽带,具体来讲,输出信息的质量及其与内部环境、外部环境的适配程度会不断促进信息技术迭代升级,进而更好地为提供者和使用者所用。

(三)信息技术

信息技术涵盖了所有支持生成式人工智能开发、运行和优化的技术手段和方法,包括机器学习算法、深度学习框架、自然语言处理技术、计算机视觉技术等,对应了生成式人工智能的两个基本要素,“算法”和“算力”。算法通过处理数据来生成信息,并进一步将信息转化为具体的内容。算力是信息技术在硬件层面和软件层面的具体体现,它包括计算资源、存储资源、网络资源等多个方面,决定了算法的执行效率和数据处理能力,是处理大规模数据集并快速生成高质量内容的关键。信息技术通过“需求拉动-环境约束”的双重机制,持续优化信息处理能力,是信息生态系统高效运转的底层引擎,具体来讲,服务使用者的需求推动算法迭代升级,管理者的监管要求促使技术向透明化发展,国际技术封锁的外部信息环境倒逼自主技术创新。

(四)信息环境

信息环境是生成式人工智能信息生态系统存在、发展的背景及场所,是对信息人、信息、信息技术产生影响的所有因素的总称,包括内部信息环境和外部信息环境。其中,内部信息环境指的是支撑生成式人工智能大模型研发、部署和运行的一系列基础设施和条件,包括电子计算机、运行平台等;外部信息环境指的是生成式人工智能在发展过程中所处的政治环境、经济环境、社会环境、文化环境、法律环境等更广泛的因素。信息环境通过“约束-赋能”双重作用,为其他要素提供运行边界与发展空间,是信息生态系统可持续运转的保障。具体来讲,信息环境通过约束服务提供者的技术边界,在提升使用者体验的同时,保护了使用者数据隐私。此外,信息环境还决定了输出信息的适用性及信息技术的发展方向,这些都有利于生成式人工智能信息生态系统良性运转。

四、信息生态系统视角下生成式人工智能的演化博弈模型构建

(一)博弈参与主体行为分析与策略选择

基于上述构建的生成式人工智能信息生态系统,本文结合现有理论研究[12]、生成式人工智能发展实践及相关政策法规等,构建多主体协同治理博弈模型。所构建的演化博弈模型涉及的参与主体即为生成式人工智能信息生态系统中的信息人,主要包括生成式人工智能服务管理者、生成式人工智能服务提供者、生成式人工智能服务使用者这三类。

1.生成式人工智能服务管理者

一方面,生成式人工智能服务管理者可能通过健全监管制度、优化监管技术、设置奖惩机制等方式对生成式人工智能的数据安全、信息真实、内容多样化等进行全方位的严格监管;另一方面,管理者也可能因为监管成本高,难度大,而选择宽松监管,缺乏对数据流转过程、信息真实性、生成内容质量审核等方面进行治

理。因此，生成式人工智能服务管理者的策略选择为{严格监管，宽松监管}②。

2. 生成式人工智能服务提供者

一方面，生成式人工智能服务提供者可能通过加强数据保护措施、完善信息审核机制、加大技术投入等方式对生成式人工智能的数据安全、信息真实、内容多样化等方面进行积极治理，这将有利于提升提供者的品牌形象，有利于其长远发展；另一方面，提供者也可能通过消极治理来谋求更多的短期经济效益，缩减治理成本，忽略技术安全。因此，生成式人工智能服务提供者的策略选择为{积极治理，消极治理}。

3. 生成式人工智能服务使用者

一方面，生成式人工智能服务使用者或基于自身责任感及道德感，或受制于生成式人工智能服务管理者及提供者的限制，选择依法合规使用生成式人工智能大模型；另一方面，使用者可能由于短期利益驱使及缺乏信息素养而违规使用生成式人工智能大模型，如缺乏数据隐私保护意识而泄露个人隐私、为博眼球而传播虚假信息、过度追求生成内容数量而生成大量同质化内容等。因此，生成式人工智能服务使用者的策略选择为{合规使用，违规使用}。

（二）模型假设与参数设定

假设 1：基于信息生态系统视角的生成式人工智能治理主要包含生成式人工智能服务管理者、生成式人工智能服务提供者、生成式人工智能服务使用者三类主体，三方主体在博弈过程中均为有限理性，且各主体均有两种策略可供选择。生成式人工智能服务管理者实行“严格监管”策略的概率为 x，“宽松监管”策略的概率为 $1-x$；生成式人工智能服务提供者选择“积极治理”策略的概率为 y，“消极治理”策略的概率为 $1-y$；生成式人工智能服务使用者采取“合规使用”策略的概率为 z，“违规使用”策略的概率为 $1-z$；$x,y,z\in[0,1]$。

假设 2：对于生成式人工智能服务管理者而言，当其选择“宽松监管”策略时，只需付出较低的监管成本 C_1，包括基本监管软硬件设备采购、监管人员薪酬等；当其选择“严格监管”策略时，监管范围的扩大与监管力度的加强使得监管成本增加，严格监管成本与严格监管程度 α 有关（$\alpha>1$），记为 αC_1，此时，管理者的美誉度提高，进而获得监管形象收益 R_1。为引导信息生态系统向理想状态演进，当管理者严格监管时，会设计奖惩制度加以干预，若生成式人工智能服务提供者积极治理，则管理者会给予奖励 I_1，如财政补贴、荣誉表彰等；若生成式人工智能服务提供者消极治理，则管理者会予以惩罚 P_1，如警告、罚款等。特别说明的是，关于生成式人工智能服务管理者直接对生成式人工智能服务使用者实施奖惩的情况较少，可能会对表现极为突出的企业用户给予荣誉表彰等奖励，或对触犯法律的使用者实施罚款、拘留等惩罚。但一般情况下，管理者不直接与使用者进行交互，而是通过提供者来实施管理和处罚，故为了更好地模拟现实场景，此处不考虑管理者对使用者的直接奖惩。

假设 3：对于生成式人工智能服务提供者而言，提供生成式人工智能大模型研发、使用权限及相关大数据、云计算等服务是其基本职能，提供者可以由此获得基本的业务收益 R_2。随着生成式人工智能服务使用者数量的增加，提供者可以占据更多的市场份额并获得更多的经济利润，包括合规收益和违规收益两种。以广告投放收入为例，一种是合规的广告投放收益 R_3，另一种是由于使用者违规使用及提供者消极治理综合带来的超额收益 R_4（如纵容使用者泄露个人隐私可使提供者获得大量用户数据，并以此获得更多商业机会；纵容使用者传播虚假信息及生成大量同质化内容可为提供者吸引更多使用者，提升用户活跃度，增加点击率，进而提升广告收入等）。当生成式人工智能服务提供者选择“消极治理”策略时，只需支付基本的治理成本 C_2，包括大模型运行成本、基本维护成本等；当其选择“积极治理”策略时，治理范围的扩大与治理力度的加强使得治理成本增加，积极治理成本与积极治理程度 β 有关（$\beta>1$），记为 βC_2，此时，提供者的美誉度提高，进而获得声誉收益 R_5。为引导信息生态系统向理想状态演进，当提供者积极治理时，会制定相应的奖惩措施，若生成式人工智能服务使用者合规使用，则提供者会给予奖励 I_2，如账号积分、开放使用权限等；若生成式人工智能服务使用者违规使用，则提供者会予以惩罚 P_2，如限制功能、封禁账号等。

假设 4：对于生成式人工智能服务使用者而言，其通过本地部署、Web 访问、应用程序编程接口（API）调用等方式使用生成式人工智能大模型时，需要支付一定使用费及服务费，该部分使用成本与提供者的基本业

②{ }代表三方主体的一种策略集合。

务收益相同,均为 R_2;使用者在向大模型输入信息直至获得输出信息的全流程需付出时间精力等额外成本记为 F。当使用者选择“合规使用”策略时,可获得自身需要的输出信息,由此带来的效率提高等基本收益记为 R_6。当使用者选择“违规使用”策略时,除了可获得基本收益 R_6 以外,还可因窃取个人隐私、传播虚假内容、快速生成大量同质化内容而在短期内获得一定经济收益及心理收益,这种超额收益记为 R_7。

假设5:对于生成式人工智能信息生态系统而言,当生成式人工智能服务管理者选择“严格监管”策略、生成式人工智能服务提供者选择“积极治理”策略且生成式人工智能服务使用者选择“合规使用”策略时,良好的信息生态秩序会给管理者、提供者和使用者带来共治收益 R_8、R_9 和 R_{10},实现技术、社会、经济等多方面的共赢。其中,管理者的共治收益体现在促进技术健康发展、维护社会稳定与公共安全、提升自身的公信力等方面,与管理者的监管形象收益 R_1 不同的是,只有当三方主体均采取积极治理措施时才会形成共治收益,而监管形象收益存在与否只受管理者自身策略选择的影响,与另外两个主体的策略选择无关。提供者的共治收益体现在推动生成式人工智能技术不断迭代升级、增强市场竞争力和品牌形象等方面,同理,与声誉收益 R_5 不同的是,只有当三方主体均采取积极治理措施时才会形成共治收益,而声誉收益存在与否只受提供者自身策略选择的影响,与另外两个主体的策略选择无关。使用者的共治收益体现在个人隐私得到保护、提升输出信息获取效率及质量等方面。除此之外,当各博弈主体选择消极策略时,使得生成式人工智能信息生态系统功能紊乱,各消极治理主体还需要承受法律、社会等多方面的负面影响带来的损失。具体而言,对于生成式人工智能服务管理者,宽松监管会造成自身公信力损失,记为 L_1;对于生成式人工智能服务提供者,消极治理会带来一定法律风险、经济损失及声誉损失,记为 L_2;对于生成式人工智能服务使用者,违规使用会带来一定隐私泄漏风险、法律风险及经济损失等,记为 L_3。

(三)模型构建

根据上述假设,构建“生成式人工智能服务管理者-生成式人工智能服务提供者-生成式人工智能服务使用者”三方博弈支付矩阵(表1和表2)。

表1 演化博弈支付矩阵

博弈主体				生成式人工智能服务管理者	
				严格监管(x)	宽松监管($1-x$)
生成式人工智能服务提供者	积极治理(y)	生成式人工智能服务使用者	合规使用(z)	(a_1,b_1,c_1)	(a_2,b_2,c_2)
			违规使用($1-z$)	(a_3,b_3,c_3)	(a_4,b_4,c_4)
	消极治理($1-y$)	生成式人工智能服务使用者	合规使用(z)	(a_5,b_5,c_5)	(a_6,b_6,c_6)
			违规使用($1-z$)	(a_7,b_7,c_7)	(a_8,b_8,c_8)

表2 三方博弈收益值

博弈策略	生成式人工智能服务管理者	生成式人工智能服务提供者	生成式人工智能服务使用者
(a_1,b_1,c_1)	$R_1+R_8-\alpha C_1-I_1$	$R_2+R_3+R_5+R_9+I_1-\beta C_2-I_2$	$R_6+R_{10}+I_2-R_2-F$
(a_2,b_2,c_2)	$-C_1-L_1$	$R_2+R_3+R_5-\beta C_2-I_2$	$R_6+I_2-R_2-F$
(a_3,b_3,c_3)	$R_1-\alpha C_1-I_1$	$R_2+R_3+R_5+I_1+P_2-\beta C_2$	$R_6+R_7-R_2-F-P_2-L_3$
(a_4,b_4,c_4)	$-C_1-L_1$	$R_2+R_3+R_5+P_2-\beta C_2$	$R_6+R_7-R_2-F-P_2-L_3$
(a_5,b_5,c_5)	$R_1+P_1-\alpha C_1$	$R_2+R_3-C_2-P_1-L_2$	R_6-R_2-F
(a_6,b_6,c_6)	$-C_1-L_1$	$R_2+R_3-C_2-L_2$	R_6-R_2-F
(a_7,b_7,c_7)	$R_1+P_1-\alpha C_1$	$R_2+R_3+R_4-C_2-P_1-L_2$	$R_6+R_7-R_2-F-L_3$
(a_8,b_8,c_8)	$-C_1-L_1$	$R_2+R_3+R_4-C_2-L_2$	$R_6+R_7-R_2-F-L_3$

(四)复制动态方程构建

1. 生成式人工智能服务管理者的复制动态方程

将生成式人工智能服务管理者实行“严格监管”策略和“宽松监管”策略时的期望收益分别记为 EM_1 和 EM_2,平均期望收益记为 EM,则有:

$$EM_1 = yz(R_1+R_8-\alpha C_1-I_1)+y(1-z)(R_1-\alpha C_1-I_1)+(1-y)z(R_1+P_1-\alpha C_1)+(1-y)(1-z)(R_1+P_1-\alpha C_1) \quad (1)$$

$$EM_2 = yz(-C_1-L_1)+y(1-z)(-C_1-L_1)+(1-y)z(-C_1-L_1)+(1-y)(1-z)(-C_1-L_1) \quad (2)$$

$$EM = xEM_1 + (1-x)EM_2 \tag{3}$$

根据马尔萨斯复制动态方程，生成式人工智能服务管理者实行“严格监管”策略数量的增长率为 $EM_1 - EM$，在时间 t 的不断延续下，生成式人工智能服务管理者的复制动态方程为

$$F(x) = \frac{dx}{dt} = x(EM_1 - EM) = x(1-x)[R_1 - \alpha C_1 + C_1 + L_1 + (1-y)P_1 - yI_1 + yzR_8] \tag{4}$$

2. 生成式人工智能服务提供者的复制动态方程

生成式人工智能服务提供者选择“积极治理”策略和“消极治理”策略时的期望收益分别记为 ES_1 和 ES_2，平均期望收益记为 ES，则有：

$$\begin{aligned} ES_1 = {} & xz(R_2 + R_3 + R_5 + R_9 + I_1 - \beta C_2 - I_2) + x(1-z)(R_2 + R_3 + R_5 + I_1 + P_2 - \beta C_2) + \\ & (1-x)z(R_2 + R_3 + R_5 - \beta C_2 - I_2) + (1-x)(1-z)(R_2 + R_3 + R_5 + P_2 - \beta C_2) \end{aligned} \tag{5}$$

$$\begin{aligned} ES_2 = {} & xz(R_2 + R_3 - C_2 - P_1 - L_2) + x(1-z)(R_2 + R_3 + R_4 - C_2 - P_1 - L_2) + \\ & (1-x)z(R_2 + R_3 - C_2 - L_2) + (1-x)(1-z)(R_2 + R_3 + R_4 - C_2 - L_2) \end{aligned} \tag{6}$$

$$ES = yES_1 + (1-y)ES_2 \tag{7}$$

进而得到生成式人工智能服务提供者的复制动态方程为

$$\begin{aligned} F(y) = \frac{dy}{dt} = {} & y(ES_1 - ES) = y(1-y)[R_5 - \beta C_2 + C_2 + L_2 + xP_1 + \\ & xI_1 + (1-z)P_2 - (1-z)R_4 - zI_2 + xzR_9] \end{aligned} \tag{8}$$

3. 生成式人工智能服务使用者的复制动态方程

生成式人工智能服务使用者采取“合规使用”策略和“违规使用”策略时的期望收益分别记为 EU_1 和 EU_2，平均期望收益记为 EU，则有：

$$\begin{aligned} EU_1 = {} & xy(R_6 + R_{10} + I_2 - R_2 - F) + x(1-y)(R_6 - R_2 - F) + (1-x)y(R_6 + I_2 - R_2 - F) + \\ & (1-x)(1-y)(R_6 - R_2 - F) \end{aligned} \tag{9}$$

$$\begin{aligned} EU_2 = {} & xy(R_6 + R_7 - R_2 - F - P_2 - L_3) + x(1-y)(R_6 + R_7 - R_2 - F - L_3) + (1-x)y(R_6 + R_7 - R_2 - \\ & F - P_2 - L_3) + (1-x)(1-y)(R_6 + R_7 - R_2 - F - L_3) \end{aligned} \tag{10}$$

$$EU = zEU_1 + (1-z)EU_2 \tag{11}$$

进而得到生成式人工智能服务使用者的复制动态方程为

$$F(z) = \frac{dz}{dt} = z(EU_1 - EU) = z(1-z)(L_3 - R_7 + yP_2 + yI_2 + xyR_{10}) \tag{12}$$

（五）三方演化博弈系统均衡点及稳定性分析

联立式（4）、式（8）和式（12），可得到生成式人工智能服务管理者、生成式人工智能服务提供者、生成式人工智能服务使用者的三维动力系统，分别令 $F(x)=0$，$F(y)=0$，$F(z)=0$，进一步求解该复制动态方程组，可知博弈系统存在 8 个均衡点，分别为 $E_1=(0,0,0)$，$E_2=(0,0,1)$，$E_3=(0,1,0)$，$E_4=(0,1,1)$，$E_5=(1,0,0)$，$E_6=(1,0,1)$，$E_7=(1,1,0)$，$E_8=(1,1,1)$。根据 Friedman[37] 的研究，在非对称博弈中，当信息不对称条件成立时，演化稳定策略为纯策略，本文将通过构建雅可比矩阵，进一步判断上述 8 个纯策略纳什均衡点的渐进稳定性（ESS），该演化博弈系统的雅克比矩阵如式（13）所示。

$$J = \begin{bmatrix} J_{11} & J_{12} & J_{13} \\ J_{21} & J_{22} & J_{23} \\ J_{31} & J_{32} & J_{33} \end{bmatrix} = \begin{bmatrix} \frac{\partial F(x)}{\partial x} & \frac{\partial F(x)}{\partial y} & \frac{\partial F(x)}{\partial z} \\ \frac{\partial F(y)}{\partial x} & \frac{\partial F(y)}{\partial y} & \frac{\partial F(y)}{\partial z} \\ \frac{\partial F(z)}{\partial x} & \frac{\partial F(z)}{\partial y} & \frac{\partial F(z)}{\partial z} \end{bmatrix} \tag{13}$$

其中：$J_{11} = (1-2x)[R_1 - \alpha C_1 + C_1 + L_1 + (1-y)P_1 - yI_1 + yzR_8]$，$J_{12} = x(1-x)(zR_8 - P_1 - I_1)$，$J_{13} = xyR_8 \times (1-x)$，$J_{21} = y(1-y)(P_1 + I_1 + zR_9)$，$J_{22} = (1-2y)[R_5 - \beta C_2 + C_2 + L_2 + xP_1 + xI_1 + (1-z)P_2 - (1-z) \times$

$R_4 - zI_2 + xzR_9]$，$J_{23} = y(1-y)(xR_9 + R_4 - P_2 - I_2)$，$J_{31} = yzR_{10}(1-z)$，$J_{32} = z(1-z)(P_2 + I_2 + xR_{10})$，$J_{33} = (1-2z)(L_3 - R_7 + yP_2 + yI_2 + xyR_{10})$。各均衡点的特征值及稳定性分析见表 3。

表 3 均衡点的特征值及稳定性分析

均衡点	特征值λ_1	特征值λ_2	特征值λ_3
$E_1(0,0,0)$	$L_3 - R_7$	$R_1 - \alpha C_1 + C_1 + L_1 + P_1$	$R_5 - R_4 - \beta C_2 + C_2 + L_2 + P_2$
$E_2(0,0,1)$	$R_7 - L_3$	$R_1 - \alpha C_1 + C_1 + L_1 + P_1$	$R_5 - \beta C_2 + C_2 + L_2 - I_2$
$E_3(0,1,0)$	$I_2 + L_3 + P_2 - R_7$	$R_1 - \alpha C_1 + C_1 + L_1 - I_1$	$2R_4 - R_5 + \beta C_2 - C_2 - L_2 - P_2$
$E_4(0,1,1)$	$R_7 - L_3 - P_2 - I_2$	$R_1 + R_8 - \alpha C_1 + C_1 + L_1 - I_1$	$\beta C_2 - C_2 - R_5 - L_2 + I_2$
$E_5(1,0,0)$	$L_3 - R_7$	$\alpha C_1 - R_1 - C_1 - L_1 - P_1$	$R_5 - R_4 - \beta C_2 + C_2 + L_2 + P_1 + P_2 + I_1$
$E_6(1,0,1)$	$R_7 - L_3$	$\alpha C_1 - R_1 - C_1 - L_1 - P_1$	$R_5 + R_9 - \beta C_2 + C_2 + L_2 + P_1 + I_1 - I_2$
$E_7(1,1,0)$	$I_2 + L_3 + P_2 - R_7 + R_{10}$	$\alpha C_1 - R_1 - C_1 - L_1 + I_1$	$R_4 - R_5 + \beta C_2 - C_2 - L_2 - P_1 - P_2 - I_1$
$E_8(1,1,1)$	$R_7 - L_3 - P_2 - I_2 - R_{10}$	$\alpha C_1 - R_1 - R_8 - C_1 - L_1 + I_1$	$\beta C_2 - C_2 - R_5 - R_9 - L_2 - P_1 - I_1 + I_2$

当雅可比矩阵的所有特征值均小于 0 时，该均衡点是渐近稳定点；当雅可比矩阵中有大于 0 的特征值时，该均衡点不是稳定点。由表 3 可知，当满足一定条件时，上述 8 个均衡点均有可能成为生成式人工智能信息生态系统治理演化稳定点，且分别对应信息生态系统的不同发展时期。为加快形成生成式人工智能信息生态系统的协同共治格局，本文主要针对 E_8 的理想状态进行分析，即三方主体的策略选择收敛在均衡点 $E_8(1,1,1)$，相关参数需满足 $R_7 < L_3 + P_2 + I_2 + R_{10}$、$\alpha C_1 + I_1 < R_1 + R_8 + C_1 + L_1$ 和 $\beta C_2 + I_2 < R_5 + R_9 + C_2 + L_2 + P_1 + I_1$。

五、数值仿真分析

（一）参数设置及演化趋势分析

为了更加清晰直观地反映生成式人工智能服务管理者、提供者及使用者策略选择的动态演化过程，解析生成式人工智能信息生态系统治理机制的影响因素，本文利用 MATLAB2018b 软件对构建的演化博弈模型进行数值仿真模拟。参考生成式人工智能安全治理的相关研究[12,38-39]，在满足上述假设的情况下，根据等式平衡原则并结合实际情境对相关参数进行赋值（表 4）。

表 4 参数取值

参数	数值	参数	数值
管理者宽松监管成本 C_1	6	提供者合规的广告投放收益 R_3	5
提供者消极治理成本 C_2	5	使用者违规使用及提供者消极治理综合带来的超额收益 R_4	3
管理者严格监管程度 α	2	提供者声誉收益 R_5	3
提供者积极治理程度 β	2.5	使用者基本收益 R_6	5
管理者给提供者的奖励 I_1	3	使用者违规使用的超额收益 R_7	7
提供者给使用者的奖励 I_2	2.5	管理者的共治收益 R_8	5.5
管理者给提供者的惩罚 P_1	2	提供者的共治收益 R_9	2.25
提供者给使用者的惩罚 P_2	2	使用者的共治收益 R_{10}	2
使用者额外成本 F	1.5	管理者宽松监管的损失 L_1	3
管理者监管形象收益 R_1	5	提供者消极治理的损失 L_2	2.5
提供者基本业务收益 R_2	5.5	使用者违规使用的损失 L_3	2

除此之外，本文设置各主体初始意愿为 0.5[40]，设定演化时间为 10，进行仿真模拟，演化仿真分析趋势。图 2 刻画了理想状态下生成式人工智能服务管理者、提供者及使用者的策略选择，系统策略最终演化稳定于 $E_8(1,1,1)$，表明{严格监管，积极治理，合规使用}为三方演化稳定策略，验证了上述稳定性分析的合理性。

（二）生成式人工智能服务管理者参数对演化博弈系统稳定性的影响

本节着重探讨生成式人工智能服务管理者的生态因素、形象因素、奖惩因素对演化博弈系统稳定性的影响，其中生态因素属于三方主体的共性因素，形象因素及奖惩因素属于生成式人工智能服务管理者和提供者的个性因素。

1. 生态因素的影响

在保持其他参数不变的情况下，根据等差原则，分别以生成式人工智能服务管理者共治收益 R_8 和管理者宽松监管损失 L_1 为变量，R_8 和 L_1 对演化博弈系统中各主体策略选择影响的演化仿真分析如图 3 和图 4 所示。由于 R_8 和 L_1 都是作用在管理者上的，因此，管理者共治收益增加和宽松监管损失加大都会促使管理者从摇摆不定的策略状态转向“严格监管”策略，当管理者选择“严格监管”策略时，会对提供者的行为进行奖惩，故提供者也会从摇摆不定的策略状态转向“积极治理”策略，并通过对使用者的行为进行奖惩，促使使用者从“违规使用”策略转向“合规使用”策略。总的来说，R_8 和 L_1 值越大，越有利于增强生成式人工智能信息生态系统的稳定性、提升各主体演化速率，最终稳定于{严格监管，积极治理，合规使用}的理想状态，且这两个参数变化对三方主体的行为具有显著影响。

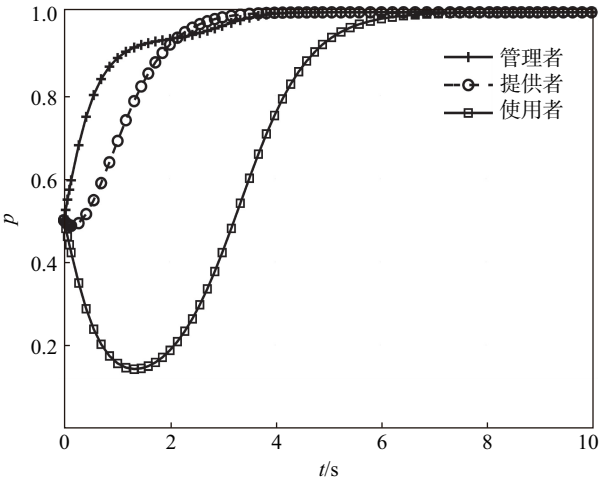
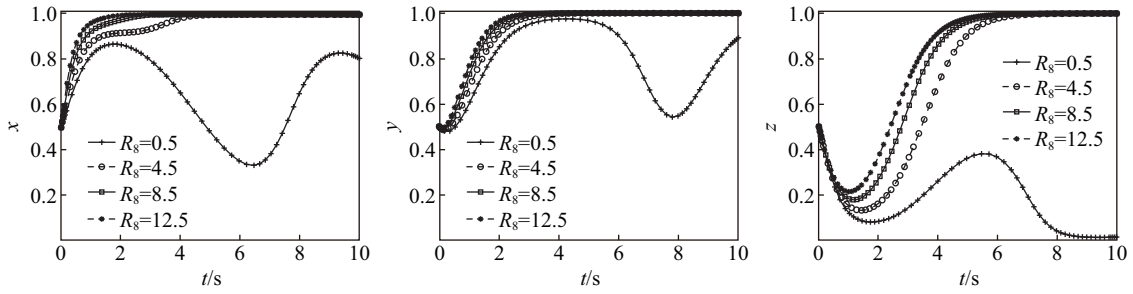
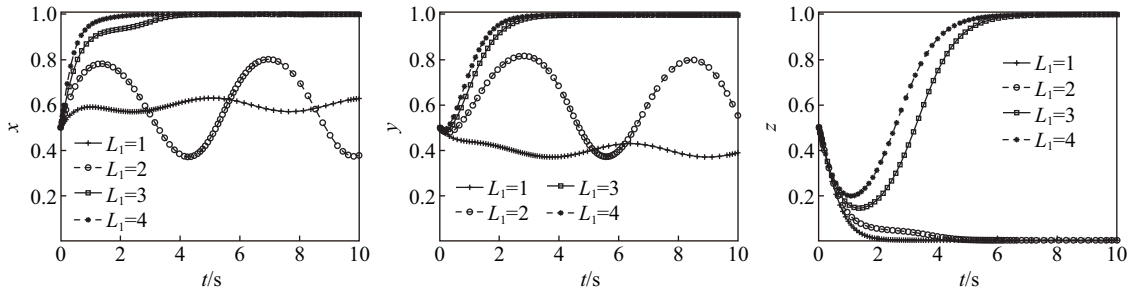


图 2　理想状态下生成式人工智能信息生态系统治理三方演化



(a) 管理者行为　(b) 提供者行为　(c) 使用者行为

图 3　不同管理者共治收益下演化博弈系统的稳定性分析



(a) 管理者行为　(b) 提供者行为　(c) 使用者行为

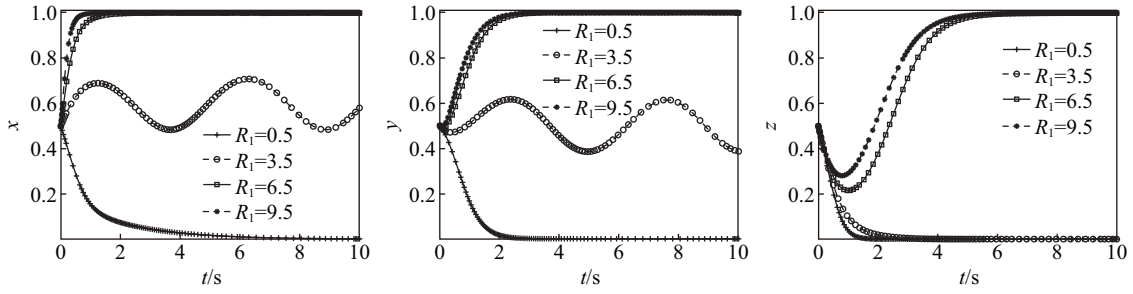
图 4　不同管理者宽松监管损失下演化博弈系统的稳定性分析

2. 形象因素的影响

本文以生成式人工智能服务管理者监管形象收益 R_1 为变量，R_1 对演化博弈系统中各主体策略选择影响的演化仿真分析如图 5 所示。当管理者监管形象收益较低（$R_1=0.5$）时，各主体会稳定于{宽松监管，消极治理，违规使用}，使得信息生态系统趋向恶化；随着 R_1 值越来越大，各主体趋向于理想状态的收敛速度加快，信息生态系统的稳定性逐渐增强，各主体之间协同作用，共同促进生成式人工智能安全治理。

3. 奖惩因素的影响

首先，为了分析管理者给提供者的奖励对演化博弈系统稳定性的影响，本文以生成式人工智能服务管理者给提供者的奖励 I_1 为变量，I_1 对演化博弈系统中各主体策略选择影响的演化仿真分析如图 6 所示。当管理者给提供者的奖励较少（$I_1=2$）时，管理者采取“严格监管”策略的成本较低，因此最终会稳定于“严格监管”

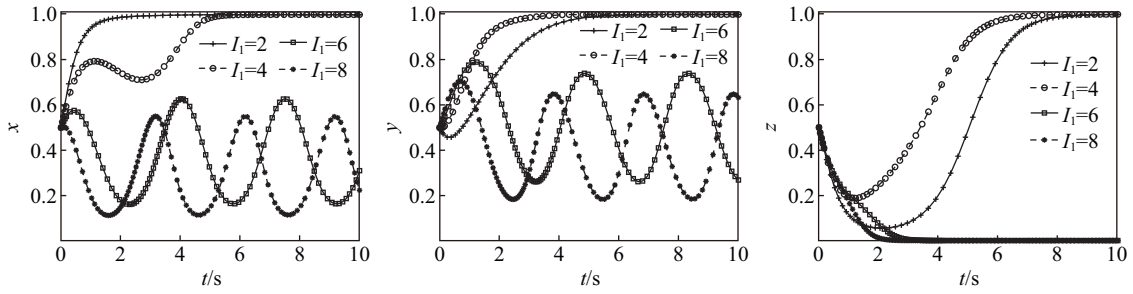


(a)管理者行为　(b)提供者行为　(c)使用者行为

图 5　不同管理者监管形象收益下演化博弈系统的稳定性分析

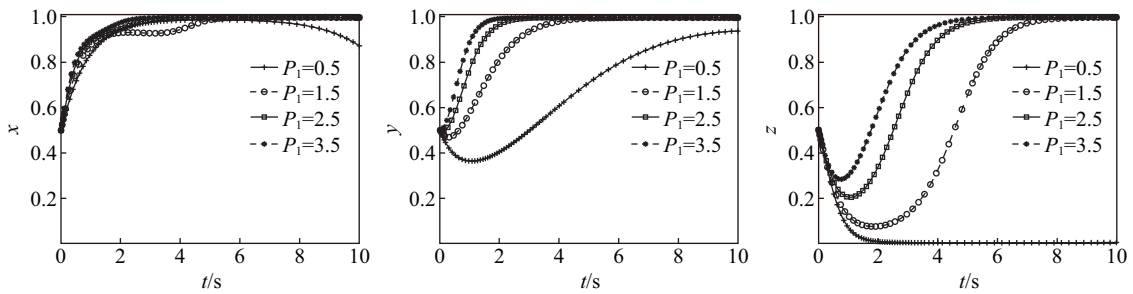
策略,此时,受信息生态系统稳定性的影响,提供者会稳定于“积极治理”策略,使用者会稳定于“合规使用”策略。随着 I_1 值变大($I_1=4$),管理者的奖励支出增加,自身收益减少,故收敛速度变慢,但此时提供者会因为受到激励而加快收敛于“积极治理”策略,同时使用者也会因为提供者实施的奖惩而加快收敛于“合规使用”策略。然而,当 I_1 值过大($I_1=6$ 或 8)时,管理者会在奖励支出与自身收益之间进行权衡,进而处于摇摆不定的策略状态,而提供者察觉到管理者可能采取“宽松监管”策略时,也会从“积极治理”策略转向“消极治理”策略,最终导致使用者稳定于“违规使用”策略,严重破坏信息生态系统。总的来说,管理者应在合理区间内加大对提供者的奖励,虽然此时管理者收敛于“严格监管”策略的速度稍有减缓,但对于整个信息生态系统而言,有利于提供者和使用者协同参与生成式人工智能安全治理,破解生成式人工智能在数据、信息、内容等方面的治理困境。

接下来,为了分析管理者给提供者的惩罚对演化博弈系统稳定性的影响,本文以生成式人工智能服务管理者给提供者的惩罚 P_1 为变量,P_1 对演化博弈系统中各主体策略选择影响的演化仿真分析如图 7 所示。当管理者给提供者的惩罚较小($P_1=0.5$)时,提供者稳定于“积极治理”策略的速度较慢,此时,使用者倾向于选择“违规治理”策略;随着 P_1 值越来越大,各主体趋向于理想状态的收敛速度加快,信息生态系统的稳定性逐渐增强,各主体之间协同作用,共同推动生成式人工智能安全治理的进程。综合上述分析,管理者应



(a)管理者行为　(b)提供者行为　(c)使用者行为

图 6　不同管理者奖励下演化博弈系统的稳定性分析



(a)管理者行为　(b)提供者行为　(c)使用者行为

图 7　不同管理者惩罚下演化博弈系统的稳定性分析

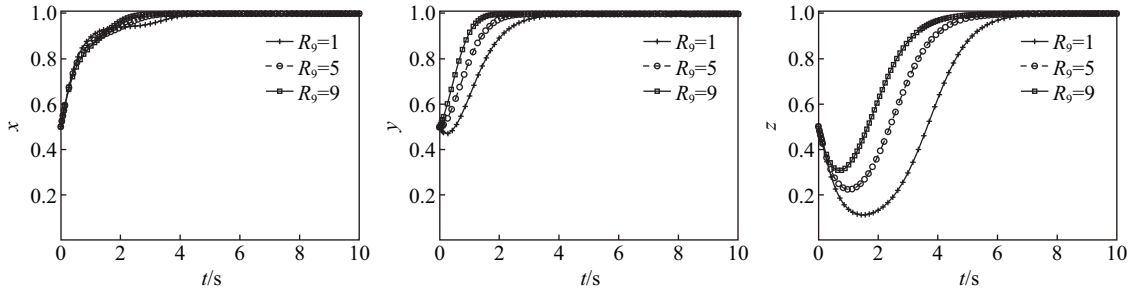
采取“惩罚为主，奖励为辅”的奖惩措施，通过加大对提供者消极治理行为的惩戒力度，辅之以在合理区间范围内增加对提供者积极治理行为的奖励，综合促进生成式人工智能信息生态系统健康有序地发展。

（三）生成式人工智能服务提供者参数对演化博弈系统稳定性的影响

本节着重探讨生成式人工智能服务提供者的生态因素、形象因素、奖惩因素、超额收益因素对演化博弈系统稳定性的影响。其中生态因素属于三方主体的共性因素，形象因素及奖惩因素属于生成式人工智能服务提供者和管理者的个性因素，超额收益因素属于生成式人工智能服务提供者和使用者的个性因素。

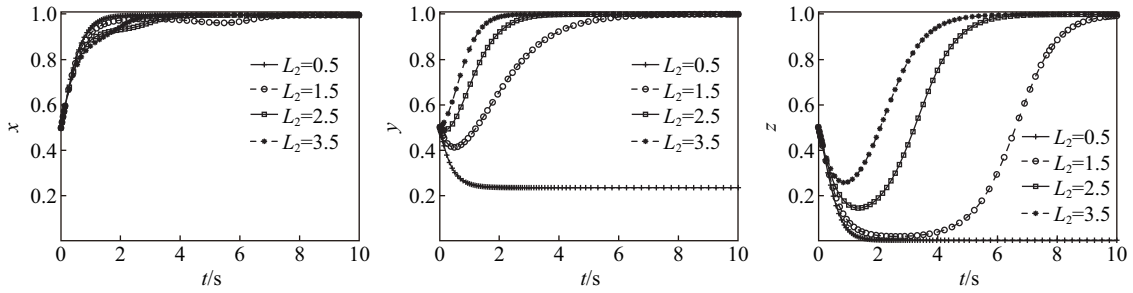
1. 生态因素的影响

在保持其他参数不变的情况下，根据等差原则，分别以生成式人工智能服务提供者共治收益 R_9 和提供者消极治理损失 L_2 为变量，R_9 和 L_2 对演化博弈系统中各主体策略选择影响的演化仿真分析如图 8 和图 9 所示。由于 R_9 和 L_2 是作用在提供者上的，因此提供者共治收益增加和消极治理损失加大都会促使提供者从“消极治理”策略转向“积极治理”策略，当提供者选择“积极治理”策略时，会对使用者的行为进行奖惩，故使用者也会从“违规使用”策略转向“合规使用”策略。另外，R_9 和 L_2 越大，提供者选择“积极治理”策略的收益越大，管理者实施“严格监管”策略时需要给予的奖励越多，管理者自身的收益将变少，故管理者起初会有收敛速度减慢的趋势，但随后还是会受信息生态系统的稳定性逐渐增强而加快收敛速度，最终稳定于严格监管的理想状态。除此之外，由图 8 和图 9 可知，R_9 和 L_2 这两个参数变化对提供者和使用者的行为影响显著，对管理者的行为影响微弱。总的来说，加大提供者的共治收益和消极治理损失有利于加快信息生态系统趋向于理想状态的收敛速度。



（a）管理者行为　（b）提供者行为　（c）使用者行为

图 8　不同提供者共治收益下演化博弈系统的稳定性分析

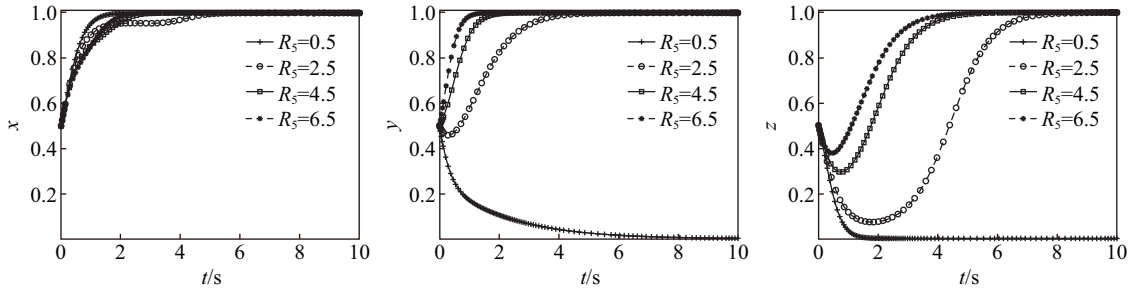


（a）管理者行为　（b）提供者行为　（c）使用者行为

图 9　不同提供者消极治理损失下演化博弈系统的稳定性分析

2. 形象因素的影响

本文以生成式人工智能服务提供者声誉收益 R_5 为变量，R_5 对演化博弈系统中各主体策略选择影响的演化仿真分析如图 10 所示。当提供者声誉收益较低（$R_5=0.5$）时，提供者会倾向于“消极治理”策略，此时管理者实施“严格监管”策略可以通过施加惩罚来获取较高的收益，使用者也会因为提供者的消极治理而选择“违规使用”策略。随着 R_5 越来越大，提供者和使用者受参数变化的影响显著，在很大程度上加快了其趋向于理想状态的收敛速度；管理者会因为给予的奖励增加，使得自身收益减少，进而导致收敛速度减慢，但这



(a)管理者行为　(b)提供者行为　(c)使用者行为

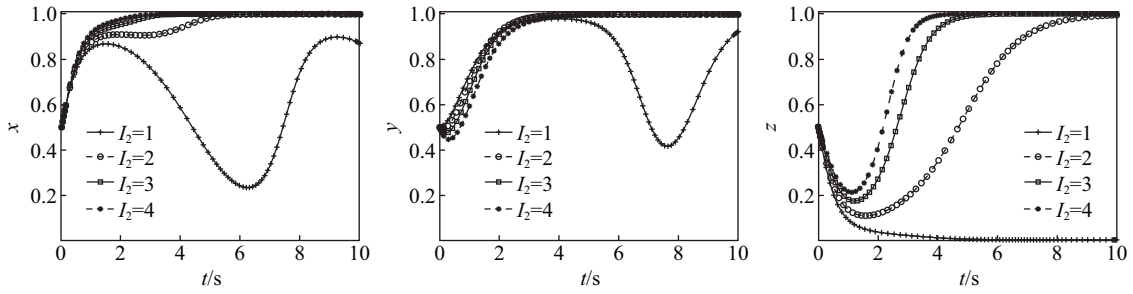
图 10　不同提供者声誉收益下演化博弈系统的稳定性分析

种影响是微弱的。因此,总的来说,提高提供者声誉收益 R_5,有利于提升信息生态系统的稳定性,促进各主体协同治理。

3. 奖惩因素的影响

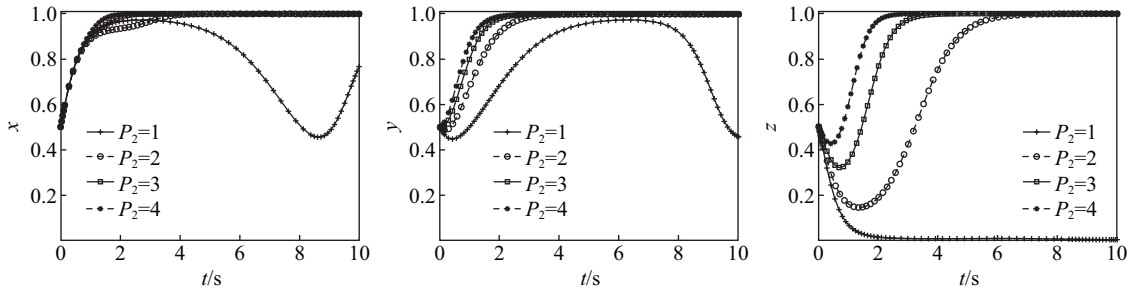
为了分析提供者给使用者的奖励对演化博弈系统稳定性的影响,本文以生成式人工智能服务提供者给使用者的奖励 I_2 为变量,I_2 对演化博弈系统中各主体策略选择影响的演化仿真分析如图 11 所示。当提供者给使用者的奖励较少($I_2=1$)时,使用者会因为激励不足而更倾向于选择“违规使用”策略,此时信息生态系统功能被破坏,管理者也有可能选择“宽松监管”策略,而提供者也处于摇摆不定的策略状态。随着 I_2 越来越大,使用者有充足的动能选择“合规使用”策略,并且收敛于理想状态的速度越来越快;管理者也加快收敛于“严格监管”策略;提供者虽然因为奖励支出的增加而导致收敛速度稍微减缓,但这种影响是微弱的,最终也会收敛于“积极治理”策略。总体而言,加大提供者对使用者的奖励有利于各主体协同参与生成式人工智能安全治理。

为了分析提供者给使用者的惩罚对演化博弈系统稳定性的影响,本文以生成式人工智能服务提供者给使用者的惩罚 P_2 为变量,P_2 对演化博弈系统中各主体策略选择影响的演化仿真分析如图 12 所示。当提供者给使用者的惩罚较小($P_2=1$)时,使用者会因为代价不高而更倾向于选择“违规使用”策略,此



(a)管理者行为　(b)提供者行为　(c)使用者行为

图 11　不同提供者奖励下演化博弈系统的稳定性分析



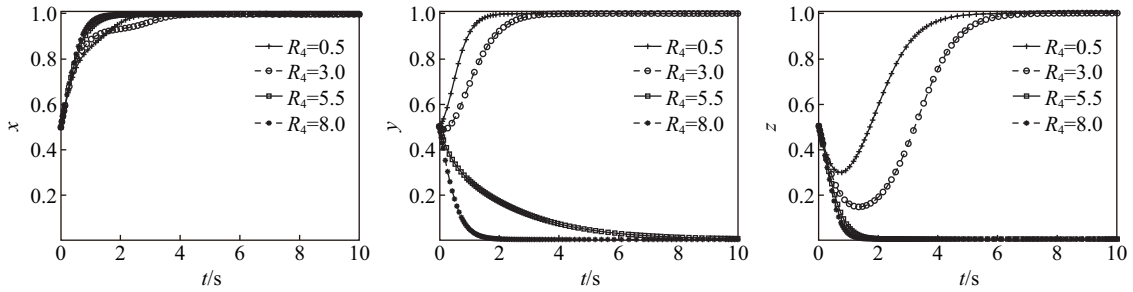
(a)管理者行为　(b)提供者行为　(c)使用者行为

图 12　不同提供者惩罚下演化博弈系统的稳定性分析

时信息生态系统功能被破坏，管理者也有可能选择“宽松监管”策略，而提供者也处于摇摆不定的策略状态。随着 P_2 值越来越大，各主体趋向于理想状态的收敛速度加快，信息生态系统的稳定性逐渐增强，各主体之间协同作用，共同推动生成式人工智能安全治理的进程。综合上述分析，提供者应采取“奖惩并重”的奖惩措施，通过加大对使用者合规使用行为的奖励及违规使用行为的惩罚，推动生成式人工智能信息生态系统健康有序地发展。

4. 超额收益因素的影响

本文以生成式人工智能服务使用者违规使用及提供者消极治理综合带来的超额收益 R_4 为变量，R_4 对演化博弈系统中各主体策略选择影响的演化仿真分析如图 13 所示。当超额收益较低（$R_4=0.5$）时，提供者会因为诱惑不足而倾向于“积极治理”策略，此时使用者倾向于“合规使用”策略，由于信息生态系统具有一定稳定性，故此时管理者实施“严格监管”策略。随着 R_4 越来越大，提供者受利益驱使，从“积极治理”策略转向“消极治理”策略，相应地，使用者也从“合规使用”策略转向“违规使用”策略，此时，R_4 的变化对管理者影响较小，整个信息生态系统稳定于{严格监管，消极治理，违规使用}的乱序状态。因此，压缩提供者违规的超额收益空间有助于信息生态系统健康有序地发展。



(a) 管理者行为　(b) 提供者行为　(c) 使用者行为

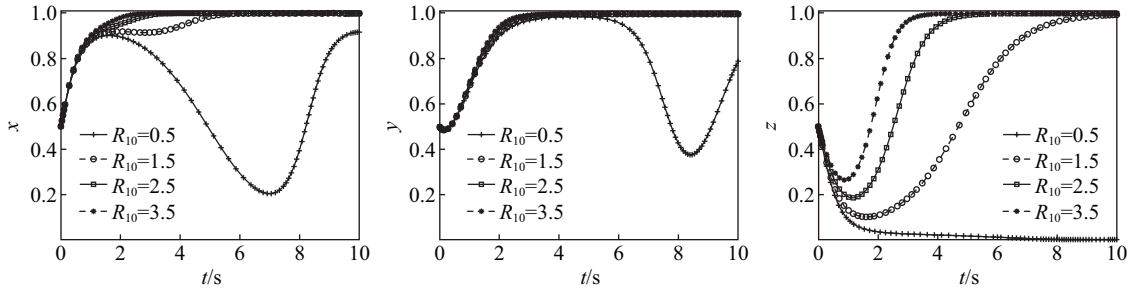
图 13　不同提供者超额收益下演化博弈系统的稳定性分析

（四）生成式人工智能服务使用者参数对演化博弈系统稳定性的影响

本节着重探讨生成式人工智能服务使用者的生态因素、超额收益因素对演化博弈系统稳定性的影响，其中生态因素属于三方主体的共性因素，超额收益因素属于生成式人工智能服务使用者和提供者的个性因素。

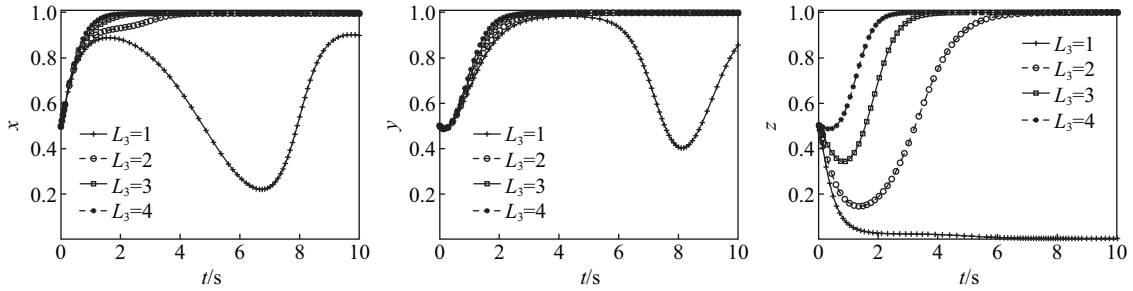
1. 生态因素的影响

在保持其他参数不变的情况下，根据等差原则，分别以生成式人工智能服务使用者共治收益 R_{10} 和使用者违规使用损失 L_3 为变量，R_{10} 和 L_3 对演化博弈系统中各主体策略选择影响的演化仿真分析如图 14 和图 15 所示。由于 R_{10} 和 L_3 都是作用在使用者上的，因此，使用者共治收益增加和违规使用损失加大都会促使使用者从“违规使用”策略转向“合规使用”策略。由于管理者与使用者之间不发生直接的奖惩关系，故此时管理者会受信息生态系统稳定性的影响，选择收益较高的“严格监管”策略，收敛速度加快；提供者的行为会受到声誉收益、管理者给予的奖励、给予使用者奖励等因素的综合作用而趋向于选择收益较高的“积极治理”策略。总的来说，加大使用者的共治收益和违规使用损失有利于加快信息生态系统趋向于理想状态的收敛速度。



(a) 管理者行为　(b) 提供者行为　(c) 使用者行为

图 14　不同使用者共治收益下演化博弈系统的稳定性分析

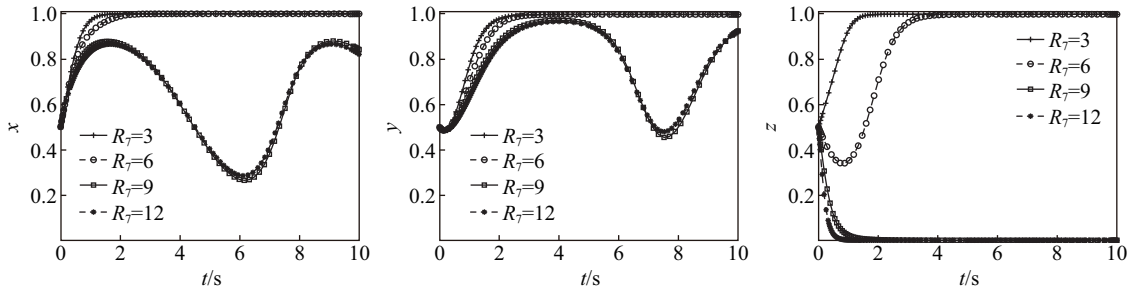


(a)管理者行为 (b)提供者行为 (c)使用者行为

图 15 不同使用者违规使用损失下演化博弈系统的稳定性分析

2. 超额收益因素的影响

本文以生成式人工智能服务使用者违规使用的超额收益 R_7 为变量，R_7 对演化博弈系统中各主体策略选择影响的演化仿真分析如图 16 所示。当超额收益较低（$R_7=3$）时，使用者会因为诱惑不足而倾向于“合规使用”策略，此时提供者倾向于“积极治理”策略，管理者倾向于“严格监管”策略。随着 R_7 越来越大，使用者受利益驱使，从“合规使用”策略转向“违规使用”策略，相应地，提供者从“积极治理”策略转向摇摆不定的策略状态，管理者也从“严格监管”策略转向摇摆不定的策略状态。因此，压缩使用者违规使用的超额收益空间有助于信息生态系统健康有序地发展。



(a)管理者行为 (b)提供者行为 (c)使用者行为

图 16 不同使用者违规使用的超额收益下演化博弈系统的稳定性分析

六、结论与启示

(一)研究结论

本文旨在探讨生成式人工智能技术快速发展背景下的安全治理机制问题，通过构建生成式人工智能的信息生态系统模型，识别出生成式人工智能安全治理过程中的三个治理主体及数据、信息、内容三个方面的治理困境，并以此为基础，构建了“生成式人工智能服务管理者-生成式人工智能服务提供者-生成式人工智能服务使用者”三位一体的演化博弈模型，运用 MATLAB 软件进行数值仿真，系统分析各类因素对演化博弈系统稳定性的影响效应。主要研究结论如下：

第一，生成式人工智能安全治理是一个多主体协同参与的过程，仅依靠单一主体进行治理并不能达到最优的治理效果，治理效率及效果与各主体的生态利益相关。具体来讲，加大各主体的共治收益及负面治理行为带来的损失，有利于加快信息生态系统趋向于理想状态的收敛速度。

第二，生成式人工智能安全治理过程受多元治理主体各自利益参数的影响，包括管理者的形象参数、奖惩参数，提供者的形象参数、奖惩参数、超额收益参数及使用者的超额收益参数等。具体来讲，提高管理者的监管形象收益和惩罚力度，在合理区间内加大对提供者积极治理行为的奖励；增加提供者的声誉形象收益、对使用者合规使用行为的奖励及违规使用行为的惩罚，降低提供者消极治理的超额收益；降低使用者违规使用的超额收益，都有利于加快信息生态系统趋向于理想状态的收敛速度。

（二）理论贡献

相对于现有研究，本文的理论贡献主要体现在以下三个方面：

第一，基于信息生态系统的研究视角，构建了生成式人工智能服务管理者、提供者、使用者的三方主体演化博弈模型，综合考虑数据、信息、内容等多层面的治理困境，突破现有研究仅围绕单一层面进行分析的局限，完善了生成式人工智能技术安全治理的理论体系。

第二，构建的生成式人工智能信息生态系统理论模型，突破现有关于信息生态系统的研究多集中于数字化发展领域的局限，补充了对智能化领域的分析及向数智信息生态系统转化的内容。

第三，围绕生态因素、形象因素、奖惩因素、超额收益因素等展开参数灵敏度分析，探究各类因素对主体行为策略的影响，相关参数设置拓展了演化博弈的研究视角，研究结论为制定生成式人工智能技术安全治理策略提供了理论支撑，推动形成健康有序的网络生态。

（三）管理启示

基于信息生态系统视角，在生成式人工智能安全治理过程中，应充分发挥管理者严格监管、提供者积极治理、使用者合规使用的三重效应。

第一，要加快构建生成式人工智能信息生态系统，完善多主体协同参与的治理机制。生成式人工智能安全治理需要管理者、提供者和使用者的共同努力，各主体必须重视生成式人工智能安全问题。对于管理者而言，要不断出台相关政策，完善相关管理办法，在严格监管的同时提升自身的监管形象。对于提供者而言，要自觉将积极治理纳入自身的发展规划中，不要一味追求经济效益而忽略自身声誉等社会效益，不要通过破坏信息生态系统来谋求自身的发展。对于使用者而言，要不断提高自身的信息素养，合法合规地使用生成式人工智能大模型工具，促进信息生态系统健康有序地发展。

第二，生成式人工智能服务管理者要充分发挥好引领作用，评估提供者的治理行为，建立相应的奖惩制度。鉴于生成式人工智能安全治理过程受多主体利益博弈的影响，且管理者给予提供者的奖励力度存在有效阈值，因此，针对提供者不同的治理行为，管理者要设计相应的指标体系来评估提供者的治理程度，同时有针对性地建立不同的奖惩制度，切实贯彻“惩罚为主，奖励为辅”的奖惩机制。除此之外，管理者还应不断完善生成式人工智能安全治理的法律法规，加强宣传，提升自身监管形象，推动生成式人工智能安全治理进程。

第三，生成式人工智能服务提供者要兼顾经济效益与社会效益，并针对使用者的行为给予相应的奖惩措施。为了杜绝提供者受利益驱使而选择消极治理，管理者应严格监管，压缩提供者的超额收益空间，同时使用者也要严格要求自己，不给提供者获得违规收益的机会。针对使用者的使用行为，提供者应采取“奖惩并重”的奖惩措施，加大对使用者合规使用行为的奖励及违规使用行为的惩罚，以促进使用者合规使用，保障信息生态系统稳定发展。除此之外，提供者要不断提高自身技术能力，从数据、算法、算力等多方面对生成式人工智能大模型进行评测，不断优化性能，并积极参与制定行业标准，提高自身声誉，推动生成式人工智能安全治理进程。

第四，生成式人工智能服务使用者要不断提高自身信息素养，不受违规使用生成式人工智能大模型所带来的经济及心理上的超额收益驱使，要自觉合法合规参与生成式人工智能安全治理。除了提供者给予的奖惩外，官方媒体也应加大生成式人工智能安全治理的宣传教育，高校及中小学也应加强学生对于生成式人工智能安全治理的理论知识教育，各主体协同作用，压缩使用者的超额收益空间。除此之外，使用者还可以通过信息反馈机制向提供者及管理者提出合理的发展建议，助力信息生态系统可持续发展。

参考文献

[1] 严驰. 生成式人工智能大模型全球治理的理论证成与初步构想[J]. 中国科技论坛，2024(5)：140-148.

[2] MANNURU N R，SHAHRIAR S，TEEL Z A，et al. Artificial intelligence in developing countries：The impact of generative artificial intelligence (AI) technologies for development [J]. Information Development，2025，41(3)：1036-1054.

[3] 刘邦奇，尹欢欢. 人工智能赋能教师数字素养提升：策略、场景与评价反馈机制[J]. 现代教育技术，2024，34(7)：23-31.

[4] 杨立民. 基于生成式人工智能法律服务的数智化发展逻辑与建构路径[J]. 深圳大学学报(人文社会科学版)，2023，40(6)：111-120.

[5] WAISBERG E，ONG J，MASALKHI M，et al. OpenAI's Sora in medicine：Revolutionary advances in generative artificial intelligence for healthcare[J]. Irish Journal of Medical Science，2024，193(4)：2105-2107.

[6] 刘志雄. 生成式人工智能赋能普惠金融：现实基础、关键风险挑战与应对策略[J]. 人民论坛·学术前沿，2024(6)：86-93.
[7] 沈芳君. 生成式人工智能的风险与治理——兼论如何打破“科林格里奇困境”[J]. 浙江大学学报(人文社会科学版)，2024，54(6)：73-91.
[8] 李森. 风险防范视阈下生成式人工智能数据安全的治理路径——以 GPT 类模型为例[J]. 西藏民族大学学报(哲学社会科学版)，2023，44(6)：139-145.
[9] 张欣. 生成式人工智能的数据风险与治理路径[J]. 法律科学(西北政法大学学报)，2023，41(5)：42-54.
[10] 徐伟，何野. 生成式人工智能数据安全风险的治理体系及优化路径——基于 38 份政策文本的扎根分析[J]. 电子政务，2024(10)：42-58.
[11] 周妍，沈天健. 生成式人工智能视域下虚假信息的层级化运作机理与治理[J]. 编辑之友，2024(8)：75-83.
[12] 郭海玲，卫金金，刘仲山. 生成式人工智能虚假信息协同共治研究[J]. 情报杂志，2024，43(9)：121-129，165.
[13] 漆晨航. 生成式人工智能的虚假信息风险特征及其治理路径[J]. 情报理论与实践，2024，47(3)：112-120.
[14] DU H，LI Z，NIYATO D，et al. Enabling AI-generated content services in wireless edge networks[J]. IEEE Wireless Communications，2024，31(3)：226-234.
[15] KARABACAK M，OZKARA B，MARGATIS K，et al. The advent of generative language models in medical education[J]. JMIR Medical Education，2023，9：e48163.
[16] ZYBACZYNSKA J，NORRIS M，MODI S，et al. Artificial intelligence-generated scientific literature：A critical appraisal[J]. Journal of Allergy and Clinical Immunology-In Practice，2024，12(1)：106-110.
[17] 张凌寒，贾斯瑶. 人工智能生成内容标识制度的逻辑更新与制度优化[J]. 求是学刊，2024，51(1)：112-122.
[18] 宋士杰，赵宇翔，朱庆华. 从 ELIZA 到 ChatGPT：人智交互体验中的 AI 生成内容(AIGC)可信度评价[J]. 情报资料工作，2023，44(4)：35-42.
[19] 杜修平，王崟羽，陈子尧. AIGC 赋能“中文+职业教育”资源智能生成与质量进化——内涵、机理与模式构建[J]. 电化教育研究，2024，45(5)：121-128.
[20] GOLAN R，RIPPS S，REDDY R，et al. ChatGPT's ability to assess quality and readability of online medical information：Evidence from a cross-sectional study[J]. Cureus，2023，15(7)：e42214.
[21] 唐昆，李白杨，张心源. 基于主客观融合的人工智能跨模态生成内容质量及效能测度研究[J]. 情报理论与实践，2024，47(11)：150-161.
[22] 邢润媚，常升龙，何宽，等. AIGC 图像质量评估指标研究[J]. 南京信息工程大学学报，2025，17(1)：63-73.
[23] ZHANG Y，GOSLINE R. Human favoritism，not AI aversion：People's perceptions (and bias) toward generative AI，human experts，and human-GAI collaboration in persuasive content generation[J]. Judgment and Decision Making，2023，18：e41.
[24] WILHELM T，ROOS J，KACZMARCZYK R. Large language models for therapy recommendations across 3 clinical specialties：Comparative study[J]. Journal of Medical Internet Research，2023，25：e49324.
[25] GHANEM Y，ROUHI A，AL-HOUSSAN A，et al. Dr. Google to Dr. ChatGPT：Assessing the content and quality of artificial intelligence-generated medical information on appendicitis[J]. Surgical Endoscopy，2024，38(5)：2887-2893.
[26] 秦喜亮，田虹. 数字技术背景下企业知识扩散模型构建及仿真研究——基于信息生态视角[J]. 情报科学，2024，42(3)：174-182.
[27] 杨雨娇，袁勤俭. 信息生态理论其在信息系统研究领域的应用及展望[J]. 现代情报，2022，42(5)：140-148.
[28] 翟运开，宋欣，王宇. 医疗健康大数据资产价值实现路径分析——基于信息生态系统理论[J]. 技术经济，2023，42(11)：178-190.
[29] 丁波涛. 基于信息生态理论的数据要素市场研究[J]. 情报理论与实践，2022，45(12)：36-41，59.
[30] 胡漠，张蕴潮. 在线健康社区生态系统架构与关键影响要素识别研究[J]. 图书情报工作，2023，67(2)：33-43.
[31] 吕鲲，郭淳，罗星雨，等. 高校智慧图书馆信息服务生态系统构建及其系统动力学分析[J]. 情报科学，2022，40(6)：44-51.
[32] LUO H，MENG X，ZHAO Y F，et al. Rise of social bots：The impact of social bots on public opinion dynamics in public health emergencies from an information ecology perspective[J]. Telematics and Informatics，2023，85：102051.
[33] 杨波，谢乐. 企业危机事件网络舆情传播态势生成机理研究——基于信息生态的多阶段 fsQCA 分析[J]. 管理评论，2022，34(7)：339-352.
[34] 张迪，张力伟. 数智信息生态系统：内涵、构成与机制[J]. 现代情报，2024，44(4)：11-21.
[35] 金雪涛，周也馨. 从 ChatGPT 火爆看智能生成内容的风险及治理[J]. 编辑之友，2023(11)：29-35.
[36] DU H，ZHANG R，NIYATO D，et al. Exploring collaborative distributed diffusion-based AI-generated content (AIGC) in wireless networks[J]. IEEE Network，2024，38(3)：178-186.
[37] FRIEDMAN D. Evolutionary games in economics[J]. Econometrica，1991，59(3)：637-666.
[38] 杨秀云，梁珊珊. 基于演化博弈的互联网信息生态环境治理机制研究[J]. 当代经济科学，2023，45(1)：29-45.
[39] 邱均平，张廷勇，徐中阳. 基于三方演化博弈的 AIGC 虚假信息协同治理策略研究[J]. 图书情报工作，2026，70(2)：71-85.
[40] 杜志平，区钰贤. 基于三方演化博弈的跨境物流联盟信息协同机制研究[J]. 中国管理科学，2023，31(4)：228-238.

How to Achieve the Secure Governance of Generative Artificial Intelligence Technology? A Game-Theoretic Analysis Based on the Information Ecosystem

Ma Xiaofei, Wang Jia

(School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: The rapid development of generative artificial intelligence (GAI) has been effective in accelerating the formation of new quality productivity and other aspects, but the governance challenges it faces in terms of data, information and content need to be solved urgently, and there is an urgent need to find a balance point between technological development and security. Based on information ecosystem theory, a collaborative governance game model involving three stakeholders (managers, providers, and users of GAI services) and numerical simulation analysis were conducted. It is found that GAI safety governance is a process of multi-object collaborative participation, which is affected by ecological factors, and increasing the co-governance gains of each subject and the losses brought by negative governance behaviors are conducive to promoting the information ecosystem to be stable in the ideal state of {strict regulation, active governance, and compliant use}. Regarding image factors, increasing the regulatory image gain of managers and the reputation image gain of providers can promote the information ecosystem to evolve to the ideal state. Regarding rewards and penalties, increasing the rewards from managers to providers and from providers to users can promote the evolution of the entire information ecosystem to a desirable state, even though it may cause a slight loss in the interests of the party who imposes the rewards, where the rewards given by the managers are within an effective threshold; and increasing the penalties imposed by the managers and providers can promote the synergistic governance of the three parties. With regard to the factor of excess earnings, compressing the space for excess earnings of providers and users can promote the synergistic governance of the three parties. The research conclusions presented herein provide theoretical support for formulating security governance strategies for generative artificial intelligence technologies, thereby promoting the development of a healthy and orderly online ecosystem.

Keywords: development and security; generative artificial intelligence; multi-body collaborative governance; information ecosystem; evolutionary game