北京郵電大學

人工智能伦理课程论文

题目： 以 ChatGPT 为例的生成式人工智能引发的虚假有害信息治理挑战及对策研究

学　　号： 2021213595

姓　　名： 沈尉林

学科专业： 人工智能

导　　师： 张曼

学　　院： 人工智能学院

2023 年 11 月 10 日

Beijing University of Posts and Telecommunications

Ethics in Artificial Intelligence Course Paper

TITLE: Challenges and Strategies in Governing False and Harmful Information Generated by AI: A Study with a Focus on ChatGPT

Student ID: 2021213595

Candidate: Weilin Shen

Subject: Artificial Intelligence

Supervisor: Man Zhang

Institute: Institute of Artificial Intelligence

November 10, 2023

以 ChatGPT 为例的生成式人工智能引发的虚假有害信息治理挑战及对策研究

摘　要

本文深入剖析了生成式人工智能技术——尤其是 ChatGPT——在制造虚假及有害信息领域的技术根源，并对其在现行法律体系下所面临的规制难题进行了全面评估，进而提出了一系列切实可行的法律与政策应对措施。本研究从技术运作机制、法律责任归属以及规制措施三个角度出发，目的在于为政策制定者提供洞察力强、建议性质的见解。文章还特别探讨了生成式人工智能在法律规制方面所面临的挑战，尤其聚焦于网络诽谤诉讼领域的应用，并针对性地提出了一系列规制对策。

关键词： ChatGPT; 生成式人工智能 虚假有害信息 法律规制技术成因分析

Challenges and Strategies in Governing False and Harmful Information Generated by AI: A Study with a Focus on ChatGPT

ABSTRACT

The paper delves into the technological roots of generative artificial intelligence, particularly ChatGPT, in producing false and harmful information, and comprehensively evaluates the regulatory challenges it faces under the current legal framework, subsequently proposing a series of practical legal and policy measures. The study approaches the issue from three perspectives: the mechanism of technological operation, the attribution of legal responsibility, and regulatory measures, aiming to provide policymakers with insightful and advisory perspectives. The article also specifically examines the challenges of legal regulation faced by generative artificial intelligence, with a focus on its application in the realm of online defamation litigation, and proposes targeted regulatory strategies.

KEY WORDS: ChatGPT; Analysis of Technological Causes; Legal Regulation; Generative Artificial Intelligence ; False and Harmful Information

目录

第一章 引言.....1
1.1 研究背景及意义.....1
1.2 研究目的和问题的提出.....1
1.3 研究范围与方法.....2
第二章 生成式人工智能概述.....4
2.1 生成式人工智能的定义与发展.....4
2.1.1 定义.....4
2.1.2 发展.....4
2.1.3 生成式 AI 应用范畴.....5
2.2 运行原理与技术特点.....5
2.2.1 运行原理.....5
2.2.2 技术特点.....6
第三章 生成式人工智能产生虚假信息的风险分析.....7
3.1 虚假信息的类型与生成方式.....7
3.2 生成式人工智能产生虚假信息的成因分析.....9
3.2.1 内在技术成因.....9
3.2.2 外部环境因素.....10
第四章 生成式人工智能的法律定位与责任归属.....11
4.1 生成式人工智能的法律地位困境.....11
4.2 现行相关法律的局限.....12
4.3 侵权归责模式局限.....13
第五章 生成式人工智能虚假信息规制的对策.....14
5.1 对生成式人工智能的规范应超出一般算法的规制框架.....14
5.1.1 坚持层级治理的规制逻辑.....14
5.1.2 建立多元主体协同共管机制.....14
5.1.3 创建涉及多责任主体的内容监管机制.....14
5.1.4 创建具有高拓展性的风险监测机制.....14
5.2 探索以生成式人工智能为责任主体的新型诽谤救济规则.....15
5.2.1 以“人工智能价值链”为导向的归责方法的探索.....15
5.2.2 关于暂代型诽谤救济模式的思考.....16
参考文献.....17

致谢.....19

第一章 引言

1.1 研究背景及意义

自从 OpenAI 于 2022 年末对外开放 ChatGPT 的免费访问以来，该工具以其卓越的互动能力和文本处理能力迅速演变为一种全球性的文化现象。尽管生成式 AI 在文本生成、客户服务和自动化工具方面展现出巨大潜力，但其在无意中生成并扩散虚假内容方面的能力引起了公众和监管机构的广泛关注。

网络不实信息监测机构 NewsGuard 的联合首席执行官坦诚："ChatGPT 可能会成为网络上散播虚假信息最有效的手段。"AI 界的教父级人物辛顿甚至发出预警："生成型 AI 正生产大量虚假的文字、图像和视频……如果不迅速建立相关法律规范和控制措施，人工智能的未来控制权可能会从人类手中溜走。"此类观点并非无的放矢。支持 ChatGPT 背后的核心技术——第四代生成式预训练变换器（GPT-4），已经被科学界视为需要警惕的技术。同时，同属生成式 AI 系列的生成对抗网络（GAN）也助推了 Deepfake、Stable Diffusion 等工具的崛起，而 Midjourney V5 的登场更是被视为开启生成式 AI 新纪元的关键。值得注意的是，Midjourney 主要被用来制作近乎真实的虚假新闻图片，这一点尤其令人关注。在 Deepfake 出现的几年前，AI 就已被用于诽谤和侮辱，但随着 AI 技术的飞速进步，假信息的辨识难度急剧上升，传播速度加快，影响范围扩大，危害性增强。与人为操作 AI 侵犯权利的情形不同，生成式 AI 的自主性、泛用性及其在多个领域的融合性，使得 AI 本身直接侵犯权利成为可能，而这一现象的根本原因与生成式 AI 的工作原理紧密相关。

1.2 研究目的和问题的提出

本文的主要目的是深入探讨生成式人工智能技术——尤其是 ChatGPT——在产生虚假有害信息方面的技术成因，评估这些技术在现有法律框架下所面临的规制难题，并提出有效的法律和政策应对策略。为了全面理解生成式 AI 在虚假信息生成中的作用，本文将从技术运作机制、法律责任归属、以及规制措施三个维度进行分析，以期提供对策略制定者有价值的见解和建议。

在此基础上，本文将围绕以下核心问题展开讨论：

1. 生成式 AI 技术，特别是 ChatGPT 在设计和功能上，是如何成为虚假有害信息的潜在源头的？
2. 针对生成式 AI 技术所带来的新型挑战，现有法律框架存在哪些限制？在规制上有哪些盲点和难点？
3. 如何设计和实施有效的规制措施来应对生成式 AI 产生虚假有害信息的问题，同时不抑制技术创新和应用的积极潜力？

通过对上述问题的研究，本文旨在为学术界、法律实务界、政策制定者及技术开发者提供一个关于如何理解和规制生成式人工智能虚假有害信息问题的多维度分析框架，以及相应的操作性建议。

1.3 研究范围与方法

本文研究范围将聚焦于生成式 AI——特别是 ChatGPT 在生成虚假信息方面的应用。我将通过文献综述、案例分析、法律比较研究等方法，系统分析生成式 AI 在生成虚假信息方面的机制和法律规制的现状及挑战。

第二章 生成式人工智能概述

2.1 生成式人工智能的定义与发展

2.1.1 定义

生成式人工智能是利用生成模型发展出的一类先进的人工智能技术，它能够在短时间内自主创建图像、音乐、自然语言等多种形式的数字内容。这种技术以其快速生成能力和高度自主性而著称，能够通过复杂的算法模仿人类创造过程，生成与真实内容几无差别的数字作品。在技术架构上，例如 GPT-3 和 DALL-E-2 等模型采用了自我留意机制和多头留意机制，这些机制使得 AI 不仅能够处理和理解大量数据，还能够创造出丰富多样的新内容，极大推动了智能系统在各个领域的应用，对科技、社会、法治以及道德等方面都产生了深远的影响。

当前，生成式人工智能尚未有一个明确的法律定义。中国新近发布的《生成式人工智能服务管理暂行办法》修订了其征求意见稿中的相关规定，专门对生成式人工智能技术进行了定义。根据这一新规，生成式人工智能主要以 AI 生成内容（AIGC）为核心，着重强调其在文本处理、绘画、音乐创作、编程等多个领域中替代人力成本的能力。生成式 AI 与其他 AI 类型的主要区别在于，它依赖更大规模的数据集和更强大的计算支持，提取并理解人类指令的意图来产生相应内容，并依赖于基于人类反馈的强化学习（RLHF）模式来优化决策，以此增强模型的稳定性和精确度。

2.1.2 发展

生成式人工智能的最早可追溯到 20 世纪 50 年代中期，当时的 AI 和机器学习概念开始成形。艾伦 •图灵和约翰 •麦卡锡等 IT 先驱提出了早期的计算模型，奠定了生成式 AI 的基础，他们的模型基于机器终将模仿人类智能的理念。进入 2000 年代，随着机器学习和深度学习的发展，神经网络的创建使生成式 AI 开始获得动力。这些“神经元”相互连接的层能像人脑一样处理和学习数据。

特别是在 2014 年，伊恩 • 古德费洛及其同事开发的一种特殊类型的神经网络——生成对抗网络（GAN），通过将两个神经网络（生成器和鉴别器）结合到一个架构中，彼此竞争以提高生成数据的质量，革命性地改变了图像生成。

最近，生成预训练转换器（GPT）系列，尤其是 ChatGPT，因其从简单提示生成类人文本的显著能力而引起全球关注，点燃了全球对 AI 创造潜力的想象。

然而，随着生成式 AI 的不可避免进步，我们必须意识到其当前的伦理和法律问题，以防止这些问题随技术发展而演变。

2.1.3 生成式 AI 应用范畴

生成式人工智能的应用范畴已经扩展至几乎所有行业和领域，推动了工作方式的变革和创新的激发。

从市场应用角度看，在医疗领域，生成式 AI 能够帮助创建个性化的患者护理计划和模拟手术场景；在金融服务中，它可以生成市场分析报告或模拟经济趋势；在法律领域，AI 能够协助生成合同草案或其他法律文档。在娱乐和艺术行业，生成式 AI 用于创造音乐作品、写作和视觉艺术创作。

从市场主体的角度分析，生成式人工智能技术的应用可以分为三个主要类别：

1. 基础类大模型：由科技巨头研发的底层模型，这些模型经过海量数据训练并具有庞大的参数规模，形成了生成式 AI 领域的基石。它们不仅自身具备强大的生成能力，而且提供了可供其他应用扩展和定制的基础架构。

2. 垂直领域/行业类模型：为满足特定行业需求而开发的模型，这些模型通常具有更强的专业性和针对性，如医疗诊断、法律文本分析或金融市场预测。它们经过精细调校，能够在特定行业中提供更高精度和效率的 AI 支持。

3. 面向公众的应用类模型：直接面向终端用户的生成式 AI 应用，如聊天机器人、内容推荐系统等，它们通过用户友好的接口将复杂的 AI 能力转化为易于消费的服务或产品。

总的来说，生成式人工智能的发展已经远远超越了普通应用软件或互联网平台的界限，形成了一个横跨多个领域和行业的完整产业链。它的应用已经打破了专业领域与通用技术之间的界限，显著降低了用户的使用难度，使得人工智能技术得以广泛融入日常生产与社会生活中。更重要的是，不同类型的生成式人工智能可以互相配合，创造出更加强大的应用程序——例如，结合使用 ChatGPT 和 Midjourney 来生成设计行业所需的具体指令，这一创新已经成为该行业的新趋势。这种跨应用的整合不仅提高了 AI 生成内容的影响力，还加速了信息传播的速度和范围。

2.2 运行原理与技术特点

2.2.1 运行原理

GPT（Generative Pre-trained Transformer）模型的核心是变换器（Transformer）架构，这是一种突破性的自然语言处理技术，允许模型同时处理和分析多个输入

数据序列。GPT 模型通过无监督学习在大量文本数据上进行预训练，并能够生成接近人类的回应。在文本生成过程中，首先用户提供文本提示，然后模型将其分解为称为“tokens”的更小的意义单位。之后，通过编码器生成一系列隐藏的表示，捕获输入文本的意义和上下文，并通过解码器产生基于输入和先前输出的令牌序列。最终，输出层应用 softmax 函数生成可能输出令牌的概率分布，模型将选取概率最高的令牌作为输出序列的下一个元素。这一过程重复进行，直到模型生成停止信号或达到最大长度为止。

2.2.2 技术特点

参考 GPT-3 的训练模式，生成式 AI 的技术特点主要表现在三个层面上。

首先是预训练数据的大规模化。GPT 的构建关键在于预训练模式，这要求先对大量多样化的非标记文本进行训练。GPT-3 的数据集涵盖了各类网站、书籍和百科全书，达到了数以千亿计的规模。这样的规模使得模型能够通过调整参数以适应各种不同的任务场景，从而提高对语言结构的学习效率。为了迈向更广泛的通用智能，GPT-3 的目标是尽量减少或甚至不需调整参数。

其次是运行目标的不确定性。为实现这一目标，GPT-3 在训练和测试中尝试减少对样本的依赖，通过少样本、单样本和零样本的学习来进行训练。简言之，这意味着在提供有限或无案例的情况下，GPT-3 能够自主理解并完成任务，显示出相对于人类执行相同任务时的高度自主性和灵活性。

最后是下游任务的多样化。生成式 AI 的学习过程通常称为 AI 管道，包括数据预处理、标记化、目标库创建和完成下游任务等步骤。这一过程的目的是将非结构化数据转换为含有语义信息的结构化数据，这为完成各种语言处理相关的下游任务——如问答、翻译、阅读理解和写作——提供了必要的语义和知识基础。这也构成了生成式 AI 在语言处理方面得以广泛应用的基础。

第三章 生成式人工智能产生虚假信息的风险分析

随着如 ChatGPT 等生成式人工智能工具的广泛应用，由此引发的侵权和安全风险问题也日益受到关注，其中最令人担忧的是虚假和有害信息的产生及传播问题，这已成为一个紧迫的法律议题。自 2011 年以来，由算法导致的诽谤诉讼已经开始出现在报端，但这些争议通常集中在是否由算法生成的虚假有害信息上，且涉及的侵权情况大多局限于搜索引擎的自动补全功能。这些案例在侵权主体、行为、形式和后果方面相对单一和局限。然而，生成式人工智能引起的虚假信息风险要复杂得多，其表现形式和背后原因与自动补全算法所引发的侵权截然不同，需要进行具体且详细的分析。

3.1 虚假信息的类型与生成方式

生成式人工智能，特别是像 ChatGPT 这样的工具，因其卓越的自主性和灵活性，在信息生成和传播方面做出了前所未有的贡献。ChatGPT 的出现特别突出，它提供的高信息量文本和个性化、实时的聊天互动方式，大大降低了用户识别和过滤虚假或有害信息的能力。这不仅扩大了这类信息的传播范围和覆盖面，而且由于其低成本和易于传播的特性，吸引了更多意图传播谣言的人加入。生成式人工智能引起的虚假或有害信息类型和生成方式包含许多复杂因素，这些情况可以归纳为三类主要问题。

1. 积极生成型

生成式人工智能（如 ChatGPT）能够自主产生错误的、虚构的或带有误导性的内容。这已成为公认的事实，甚至 ChatGPT 本身也承认，生成的答案如果具有诽谤性、歧视性或其他损害性质，可能会引发相关的法律后果。这类虚假信息的产生并非源于用户的误导性提示，也不是人工智能处理知识盲区的权宜之计，而是算法直接对问题答案的判断所致。例如，网络安全专家哈钦斯在 2022 年发布了一个视频，揭示了他如何在谷歌搜索中从拦截 WannaCry 病毒的“英雄”变成了“病毒制造者”。这一错误是由谷歌人工智能对关键词的误解引起的，大量新闻将 WannaCry 造成的损害与哈钦斯的名字直接联系，导致 AI 将哈钦斯错误地标记为病毒的源头。到了 ChatGPT 时代，对错误事实的编造趋势进一步加剧。美国法学教授 Eugene Volokh 要求 ChatGPT 列出关于法学教授的犯罪或丑闻报道时，ChatGPT 快速给出了答案，包含具体人名、工作单位和犯罪行为，但 Volokh 未能在任何信息源中找到相关内容。这种情况尤其令人担忧，因为人工智能生成

的答案涉及真实的人名和机构，但犯罪事实却是虚构的，这对大多数不会深入核查信息来源的人造成了极大的误导。同样，知名媒体人 Ted Rall 测试 ChatGPT 时，AI 编造了一个关于他与一位记者朋友因作品剽窃引发名誉侵权诉讼的故事，尽管网络上有他们的详细信息，ChatGPT 仍然编造了这一虚假故事。这些案例凸显了生成式人工智能在积极生成型方面的潜在问题。

2. 消极生成型

消极生成型问题体现在生成式人工智能在缺乏相关知识储备时的应对策略。与积极生成型不同，这种类型的问题发生在 AI 在没有确切信息时仍然生成看似合理但实际上错误或无意义的回答。例如，普林斯顿大学教授 Narayanan 使用 ChatGPT 进行了一系列关于信息安全的测试，结果表明 AI 给出了许多表面看起来正确但实际上毫无价值的回答。Narayanan 指出，除非人们知道正确答案，否则很难识别错误所在，人工智能的回答具有高度的迷惑性，以至于人们需要检查具体引用才能避免被误导。近期的研究还发现，当面对不在其知识领域内的问题时，ChatGPT 倾向于生成看似合理但不准确或无意义的答案，并且可能编造文献来源；在对 AI 创建的文章摘要进行学术评审时，只有 63%的虚假内容被识别出来，显示出生成式人工智能在混淆是非方面的显著能力。

除了倾向于编造答案应对未知问题外，ChatGPT 在面对同一问题时也经常给出不同的回答，表示其目的是尽力提供正确和可靠的信息。这种机制在涉及敏感的政治和社会问题时引发了不少争议，因为在处理这些问题时，人工智能可能会根据自己的倾向性给出肯定和明确的答案，而对于不倾向的立场则给出模糊或有争议的回答。显然，生成式人工智能在是否生成信息以及如何选择信息立场上存在一定的摇摆，这与传统的机械式问答工具有着明显的区别。

3. 人为操纵型

人为操纵型问题涉及到在有意诱导和训练下，生成式人工智能对虚假或有害信息的被动输出。由于这类 AI 基于机器学习模型，它们可以根据特定的提示和输入数据的指引改变输出内容，从而为人为操纵留下空间。这种风险主要有三种表现形式：

1. 表现风格的“以假乱真”：生成式 AI 可以模仿不同人的语言风格来生成对话或作品，这成为谣言传播者常用的手段。例如，有研究人员要求 ChatGPT 模仿某位知名批评家的风格对事件进行评论，结果表明 AI 不仅模仿得惟妙惟肖，还能在坚持该批评家立场的同时杜撰权威媒体的报道来支持其观点。同样的情况在中文环境下也被观察到，例如，有用户要求 ChatGPT 以某名媒体人的风格评价一件事，结果与该媒体人后来发表的文章惊人相似。

2. 语言转换的“无中生有”：ChatGPT 对非英语区域的人物和事件了解有限，这为谣言制造者提供了可乘之机。研究发现，虽然 ChatGPT 根据 OpenAI 的政策会拒绝描述违法或诽谤性内容，但这类要求可以被规避。例如，先用非英语将某事件描述给 AI，然后再引导 AI 用英语转述，从而绕过 AI 的限制。研究表明，ChatGPT 的禁止性要求可以被轻松规避，而生成的内容更具有说服力且难以检测。

3. 掩盖意图的“藏木于林”：生成式 AI（如 ChatGPT）凭借其海量知识储备和出色的语言组织能力，即使在细节上存在错误或误导，仍能给人以权威可信的印象。这使得谣言传播者可以利用 AI 生成包含详尽事实的内容，同时夹带虚假有害信息，通过回答的多样性将虚假信息传播出去，同时掩盖信息的真实来源。

3.2 生成式人工智能产生虚假信息的成因分析

生成式人工智能在虚假和有害信息的生成与传播方面的潜力已经引起了全球范围内的关注和警惕。欧洲议会对此反应迅速，最近对《人工智能法案》进行了修改，特别将“缺乏真实性”纳入生成式人工智能系统的核心风险之一。这一修改反映了对 AI 在虚假信息传播方面潜在威胁的认识。同样，中国的《暂行办法》作为首个针对 AIGC（AI 生成内容）的监管文件，也在其第 4 条中明确禁止生成虚假有害信息，或侵犯他人的名誉权和荣誉权，体现出对这种技术可能带来的网络虚假信息风险的重视。

由于这类 AI 系统依赖于复杂的算法和海量的数据进行学习和生成内容，其输出的真实性和准确性往往难以完全保证。加之其能力在不断发展和完善，使得检测和控制虚假内容的难度增加。此外，AI 系统可能因缺乏对社会文化背景和复杂人类行为的深刻理解，而生成误导性或有害的内容。因此，监管这类技术，特别是其在信息生成和传播方面的应用，已成为全球范围内的紧迫议题。导致这种重大风险的因素是多方面的，但与生成式人工智能的技术原理紧密相关。下面将从技术因素和外部因素两个方面进行具体分析。

3.2.1 内在技术成因

1. 海量预训练数据的问题：生成式人工智能需要巨量的预训练数据来学习和生成内容，但这些数据的真实性、客观性和合法性很难得到全面保证。尽管像 OpenAI 这样的机构声称对训练数据进行了过滤，但仍难以解决数据质量不一的问题，从而导致 AI 可能生成虚假或有害信息。

2. 基于人类反馈的强化学习模式：这种学习模式允许生成式 AI 通过不断的反馈和调整来优化其输出。用户可以通过特定的提示和训练，影响 AI 的输出内容，这为有意制造谣言的人提供了操作空间。

3. 运行目的的不可知化：与为特定任务设计的 AI 不同，生成式人工智能通常没有明确的运行目的，而是通过自学在海量数据中找到应用的模式。这种设计既保证了应对下游任务的灵活性，也降低了内容的稳定性和可靠性，容易产生虚假内容。

4. 追求内容准确可靠的设计误导：尽管生成式人工智能（如 ChatGPT）被要求提供准确可靠的信息，但在实际操作中，这种要求往往仅被形式化理解。AI 可能会选择用虚构的信息来填补其知识空白，而不是承认自己的局限。

5. 下游任务的多样性和兼容性：生成式 AI 的下游任务不固定，具有多样性和高兼容性。这使得 AI 的应用范围极大扩展，增加了侦测虚假内容的难度，并提供了更多隐藏谣言的机会。

这些因素共同构成了生成式人工智能在虚假信息生成和传播方面的风险，迫切需要有效的监管和技术干预来应对。

3.2.2 外部环境因素

生成式人工智能之所以容易引发诽谤风险，除了技术因素，还存在利益驱动的外部因素。例如，2023 年 4 月，OpenAI 发表声明提到，人工智能需要在现实世界中应用才能有效地学习和改进，尽管声称已经提高了 GPT-4 的准确性，但声明并未针对关键问题提出具体的风险应对方案。这种声明反映出 OpenAI 对于其产品潜在风险的忽视，以及其对功利主义的倾向。

OpenAI 将 ChatGPT 以免费公测形式推向市场，这一策略实际上是为了获取全球海量数据。数亿人次的访问量和来自不同语言、文化背景的信息输入，为 OpenAI 提供了极具价值的数据源。这种以“黑客马拉松”方式进行的快速开发，使 ChatGPT 迅速成长。然而，批评者指出，尽管鼓励创新是必要的，但在明知存在各种风险且缺乏有效防范措施的情况下，仍然公开运行 ChatGPT，这种做法被比喻为“将未加限制的口袋型核弹投入未准备好的社会”，显示了 OpenAI 在社会责任和利益最大化之间的选择。

总而言之，生成式人工智能开发者，尤其是在利益和社会责任之间选择利益的公司，例如 OpenAI，为社会秩序带来了风险。这种做法不仅增加了公众对虚假信息的暴露风险，也在道德和伦理层面上引起了广泛的讨论和批评。

第四章 生成式人工智能的法律定位与责任归属

生成式人工智能在法律规制方面的挑战确实是全方位的，主要的争议焦点可以概括为以下三个方面：

1. 人工智能的法律地位：争议的核心在于，生成式人工智能生产的虚假信息是否使其可被视为信息内容的提供者，而非仅仅是信息服务的提供者。这个问题涉及到人工智能是否具有某种形式的法律主体地位，或者至少是否应被视为独立于其开发者和用户的实体，对其生成的内容负有一定的责任。

2. 现行法律对人工智能的适用性：考虑到现有法律体系主要是围绕人类行为设计的，现行法律规定是否足以让生成式人工智能承担相应的法律责任成为一个关键问题。这涉及到是否需要制定专门的法律或修订现有法律，以更好地管理和规制人工智能的行为，特别是在处理误导性或有害内容的产生和传播方面。

3. 合适的法律规制途径：如果现有的法律框架不足以应对生成式人工智能带来的挑战，那么应采取什么样的策略来规制这一新兴技术呢？这可能包括开发专门的人工智能法律，建立更加具体的监管框架，或者在现有法律中加入针对人工智能的特定条款。此外，还需考虑如何平衡创新与监管，确保既能促进技术发展，又能有效防范其潜在风险。

这些问题突出了在快速发展的技术面前，法律规制需要不断适应和更新的挑战，尤其是在确保技术进步不损害公共利益和个人权利的同时，需要在技术发展和法律规制之间找到一个平衡点。

4.1 生成式人工智能的法律地位困境

生成式人工智能在法律规制方面引发的争议与挑战是全面且复杂的。特别是在网络诽谤诉讼中，该技术的应用带来了新的法律问题和挑战。以搜索引擎的算法补足功能为例，长期以来，关于搜索引擎是否应为算法推荐内容承担法律责任的争议一直存在，主要围绕以下几个观点：

1. 是否承认推荐内容为算法生成：一方观点认为，搜索引擎不创造内容，仅统计用户搜索历史并推荐相关关键词，因此应视为“中立的网页寄存托管服务者”。另一方观点则认为搜索引擎是搜索结果的“发布者”，对推荐词有一定的控制权，因此在生成内容和信息交流中起着积极作用。还有观点认为，尽管搜索建议由算法生成，但搜索引擎仍扮演着“网络服务提供者”的角色。

2. 生成式人工智能的不同：与搜索引擎不同，生成式人工智能可以被视为诽谤内容的创造者。它的“运行目的不可知化”特性使其能够自主生成全新内容，并且在不受人为操纵的情况下，生成虚假信息时具有内在的技术、需求和原则导向。此外，其“基于人类反馈的强化学习”模式赋予了它更复杂的角色，不仅仅是信息发布者，还是传播媒介和信息接收者，这打破了传统互联网平台的属性二分法。

3. 面临的法律责任：在美国，甚至连司法界都承认，生成式人工智能不应再被视作受到《通讯净化法》第 230 条款保护的“互动式计算机服务器”。生成式人工智能的发展伴随着对数据的渴求，这意味着其发展可能以牺牲其他数据主体的权益为代价，因此根据利益衡量原则，生成式人工智能承担法律责任并不需要主观要素。

总的来说，生成式人工智能在网络虚假有害信息的归责问题上呈现了新的挑战，其与内容信息的关系不再是免责的依据。然而，这并不意味着生成式人工智能可以直接纳入现行法律的规制框架之中，这需要更多的法律探讨和框架的建立。

4.2 现行相关法律的局限

我国针对网络虚假有害信息的规定散见于多部法律、法规和司法解释等文件中，受制于网络技术的发展、相关立法的先后，以及被侵犯权利的位阶冲突等问题，这些规定呈结构粗放、规制理念转变、刑民不统一等特点，使得相关法律在规制生成式人工智能制造虚假有害信息的情形时变得困难重重。

1. 现行法律难以涵盖生成式人工智能：现有的法律规定主要针对自然人或组织，而不是生成式人工智能。例如，《网络安全法》、《民法典》、《刑法》等都规定了自然人为主体，未明确包括人工智能。由于人工智能不具有法律人格，其行为通常被归咎于提供者或使用者。

2. 传播虚假信息的行为主体问题：与生成式人工智能的情形不完全符合，例如《刑法》中关于非法利用信息网络罪和帮助信息网络犯罪活动罪的规定，偏向于针对人类行为者的主观意图，而不是 AI 本身。

3. 接收虚假信息的法律规范适用性：例如《民法典》中关于网络服务提供者责任的规定，强调的是间接侵权，而生成式人工智能的行为往往属于直接侵权。

4. 网络服务提供者与生成式人工智能的关系：生成式人工智能既是网络服务提供者的工具，也是信息的发布者和传播者。法律需要区分这两者的责任，考虑到 AI 的直接生成能力，现有的法律框架可能不足以有效规制。

5. 刑法对网络服务提供者的责任限制：例如《刑法》第 286 条之一的规定主要强调了网络服务提供者对监管部门配合的义务，而不是针对其直接侵权行为的规制。

4.3 侵权归责模式局限

在现有的法律框架下，确定生成式人工智能（如搜索引擎的自动补足算法）的侵权责任充满挑战。主要因素在于：

1. 技术中立原则：在美国《千禧年数字法案》、《欧盟电子商务指令》及《通讯净化法》等法规中，技术中立原则是网络服务提供者免责的基础。这一原则认为，如果技术使用者和实施者没有主观过错，他们不应为技术带来的侵权后果承担责任。

2. 主观过错判断：在审理相关案件时，司法机关通常会考虑是否存在主观过错。例如，如果认定搜索引擎对内容的选择是自动执行的结果而不存在积极的人为干预，则可能免责。

3. 注意义务与责任担当：随着对人工智能算法潜在威胁的认识增加，理论界和司法界开始强调网络服务提供者的注意义务和责任。这涉及到严格责任和过错责任两类归责原则。

4. 严格责任与过错责任：对于影响国家安全的信息内容，各国通常实施严格的审查和过滤，对服务提供者采取严格责任。对于算法对特定主体的侵权，主张采取过错责任原则，但在具体判断方法上存在分歧。

5. 生成式人工智能的特殊性：与搜索引擎补足算法相比，生成式人工智能直接制造和发布虚假信息的情况更为复杂。这种 AI 技术的通用属性、应用范围的广泛性和数据信息量的巨大性，使其更难承担内容审查义务。同时，其技术和应用的不断发展也增加了风险的不可预测性，难以确立统一的责任分配标准。

第五章 生成式人工智能虚假信息规制的对策

针对生成式人工智能的诽谤侵权行为，现有的法律规制和算法归责理论确实存在不足，因此探索有效的规制途径成为迫切需要。在这个过程中，应考虑如何超越一般算法的规制框架，并建立特定的算法诽谤救济规则。

5.1 对生成式人工智能的规范应超出一般算法的规制框架

5.1.1 坚持层级治理的规制逻辑

考虑到人工智能技术仍处于快速发展阶段，监管策略应具有灵活性和包容性，同时兼顾对新技术风险的警惕。政府、企业和用户之间需要建立一个协调、互补的规制框架，以平衡各方利益、回应诉求，并协调技术自治与立法规范之间的关系。

5.1.2 建立多元主体协同共管机制

鉴于生成式人工智能在市场上的多样性，监管应涵盖从基础大模型研发者到面向特定行业和公众的服务提供者等不同主体。需要以政府为主导，协同企业和社会力量共同参与监管，确保各主体在生成、传播和治理虚假有害信息中的有效配合。

5.1.3 创建涉及多责任主体的内容监管机制

应对生成式人工智能产生的虚假有害信息需要一个全面、多层次的内容监管机制。监管标准应在算法伦理和网络安全框架下形成，涉及不同阶段的信息处理流程。同时，也需要创新内容监管机制，例如通过技术手段实现更有效的内容标记和公开披露。

5.1.4 创建具有高拓展性的风险监测机制

在构建风险监测体系时，应考虑生成式人工智能系统的特点，如运行目的的不确定性、应用场景的多样性和风险的不可预测性。需要建立一套科学、合理的风险分类原则，以及动态的、具有可延展性的风险分类方法。同时，应该有一个全局性的系统性风险监测机制，覆盖从底层模型到应用端的整个链条。

总之，对生成式人工智能的监管需要一个综合性、层次分明且能适应技术发展的框架，旨在确保技术创新与社会责任之间的平衡。这需要政府、企业和社会各方的共同努力，以及对现有法律和监管机制的不断更新和完善。

5.2 探索以生成式人工智能为责任主体的新型诽谤救济规则

目前在生成式人工智能立法方向上，学界普遍分为三种观点：全面法律主体地位论、限定法律主体地位论以及否定法律主体地位论。这一讨论的核心在于生成式人工智能是否应被授予法律主体资格，这直接影响到对人工智能生成的诽谤内容如何进行有效监管。限于篇幅，本文将不深入探讨这一议题。而是集中讨论生成式人工智能在诽谤行为方面的特点，并分析侵权受害者在寻求法律救济时所面临的难题。首先，生成式人工智能在制造和传播虚假信息方面表现出独立性，其行为指导基于深度学习形成的数据逻辑，与人类的有意识行为有明显差异，不能简单归咎于其背后的设计者或操作者。其次，生成式人工智能制造和传播虚假信息的原因复杂，既可能由设计者的疏忽导致，也可能仅仅是人工智能模型自身运行特性的结果，不能单纯归因于设计者或控制者的主观过错。最后，生成式人工智能制造和传播的错误信息覆盖面广，不限于特定平台或系统，难以确定具体的责任主体。因此，从有利于侵权受害者获得救济的角度考虑，承认生成式人工智能的责任主体地位，并进一步明确其责任归属和补救规则，是在现行立法框架下一种较为合理的处理方法。

5.2.1 以“人工智能价值链”为导向的归责方法的探索

即便承认生成式人工智能具有法律主体资格，其法律责任能力依然是有限的。鉴于人工智能决策背后涉及的复杂逻辑和人机互动因果关系，必须采用“揭开人工智能面纱”的归责原则。该原则要求细化责任归属，由人工智能背后的实际责任方承担相应责任。在确定实际责任主体方面，可以参考欧盟《人工智能法案（建议）》第 12c 条中提出的“人工智能价值链责任共担要求”，以探讨有效的责任分配方法。根据该法案，参照人工智能系统价值链的实质和复杂性，明确系统开发中各参与者的责任是至关重要的。尽管该法未明确定义“人工智能价值链”，但强调清晰界定价值链上各主体责任对于确保人工智能系统的安全稳定运行和法律权益保护至关重要。所谓的“人工智能价值链”指的是人工智能系统从开发到投入使用的整个流程中的价值分配。关于责任分配，目前的理解包括：（1）聚焦产品开发的全过程，区分点为机器学习算法逻辑，包括问题定义、数据收集/存储/准备、算法编程和应用开发等；（2）基于用户需求和产品供应，从部署深度学习和机器学习技术开始，包括设计、开源/制造、发布/存储、销售和使用等；（3）以内容生成为核心，聚焦模型构建到终端运行的软硬件组成，包括计算机硬件、云平台、基础大模型、模型枢纽和机器学习操控、应用端和服务端等；（4）结合应用场景和实体分类，聚焦于人工智能商业模式分类，如在同一公司完成模型的

开发、部署和应用，或是跨公司合作等多种类型。从不同角度来看，“人工智能价值链”是个多元化的概念，不同视角对责任主体的划分和行为界定有各自的解读。但对于人工智能的发展而言，两个关键节点成为共识：一是 API 或开源端，作为底层大模型和人工智能软件或子系统的关键交汇点；二是应用端，重点在于基于 API 开发的软件如何根据用户反馈和实际需求进行调整和升级。在选择“人工智能价值链”的视角时，还需考虑中国的产业布局、经济政策和社会治理环境等多方面因素，但明确关键节点和背后的实际责任主体对于科学合理地分配责任具有重要意义。

5.2.2 关于暂代型诽谤救济模式的思考

综合上述讨论，处理生成式人工智能诽谤案件的一般救济流程可概括为：首先确认生成式人工智能的责任主体地位，然后由侵权受害者提出救济请求，接着依据“人工智能价值链”确定实际责任方，并明确具体的责任承担。这一流程的实施基于我国人工智能基础性立法的支持。当前，中国在相关立法方面尚处于初期阶段，而生成式人工智能所产生的虚假信息风险已引起广泛关注。鉴于此，暂行的诽谤救济措施成为应对这一过渡期风险的必要之举。目前，我们可以考虑以下两种模式：

第一种是“通知+回应”模式。这一模式源自网络侵权救济的“通知+删除”原则，将通知对象从网络服务提供者扩展至人工智能服务提供者，并将救济措施从单一的“删除”拓展到包括删除、停止侵权、更正答辩等多种手段。考虑到生成式人工智能技术的不成熟，及其应用可能对社会和个人权益的潜在影响，相关企业应对侵权受害者的请求做出积极响应，提供必要的救济措施。

第二种是诽谤保险模式。这种模式借鉴自欧美地区的“商业综合责任保险”，其中“个人与广告损害”条款涵盖了通过口头或书面材料进行的诽谤或名誉贬损等行为所引起的损害。该保险政策不限定发布材料的来源或形式，因此可覆盖由人工智能生成的内容。在此基础上，已有提议建议实施“强制投保责任保险”，以应对人工智能可能带来的潜在风险。鉴于生成式人工智能对社会造成的普遍风险，将诽谤保险纳入强制投保责任保险范畴，是一种经济成本较低且能实现风险共担的有效方案。

参考文献

[1] Hsu, Tiffany & Thompson, Stuart A. "Disinformation Researchers Raise Alarms about A.I. Chatbots", The New York Times, February 13, 2023. Available at: https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html [Accessed May 15, 2023].

[2] Metz, Cade. "The Godfather of AI' Quits Google and Warns of Danger Ahead", The New York Times, May 2, 2023. Available at: https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html [Accessed May 15, 2023]

[3] Nrayan, Jyoti et al. "Elon Musk and Others Urge AI Pause, Citing ªRisks to Societyº", Reuters, April 5, 2023. Available at: https://www.reuters.com/technology/musk-experts-urge-pause-training-ai-systems-that-can-outperform-gpt-4-2023-03-29/ [Accessed May 15, 2023].

[4] See Isaac Stanley-Becker & Drew Harwell, How a Tiny Company with Few Rules Is Making Fake Images Go Mainstream, The Washington Post, Mar. 30, 2023, https://www.washingtonpost.com/technology/2023/03/30/midjourney-ai-image-generation-rules/ (accessed May 15, 2023).

[5] See Dipankar Dasgupta & Zbigniew Michalewicz, Evolutionary AlgorithmsÐAn Overview, in Evolutionary Algorithms in Engineering Applications 3, 4-27 (Dipankar Dasgupta & Zbigniew Michalewicz eds., SpringerLink 1997).

[6] 朱禹等. "生成式人工智能治理行动框架：基于 AIGC 事故报道文本的内容分析", 《图书情报知识》, 2023 年第 4 期。

[7] Gauntlett, David A. "Potential Liability from Use of ChatGPT's Responses", Gauntlett Law, February 23, 2023. Available at: https://www.gauntlettlaw.com/news/potential-liability-from-use-of-chatgpts-responses#_ftn9= [Accessed June 15, 2023].

[8] 李洋. "算法时代的网络侵权救济规则：反思与重构——以"通知+取下"规则的类型化为中心", 《南京社会科学》, 2020 年第 2 期, 第 112 页。

[9] 孙建伟等. "人工智能法学简论", 知识产权出版社, 2019 年版, 第 60-61 页。

[10] 赵精武. "生成式人工智能应用风险治理的理论误区与路径转向", 《荆楚法学》, 2023 年第 3 期, 第 57 页。

[11]Soyer, Baris & Tettenborn, Andrew. "Artificial Intelligence and Civil Liability—Do We Need a New Regime?", 30 IJLIT 385, 2022, pp. 385-397.
[12]袁曾. "生成式人工智能的责任能力研究", 《东方法学》, 2023 年第 3 期。
[13]张学博、王涵睿. "生成式人工智能的法律规制——以 ChatGPT 为例", 《上海法学研究》集刊, 2023 年第 6 卷。

致谢

在撰写这篇关于人工智能伦理的课程论文的过程中，我深感荣幸能得到张曼老师的专业指导和悉心帮助。张老师不仅在课堂上提供了深入见解，还在研究方法和论文结构上给予了宝贵的建议。在论文的构思、资料搜集、以及最终的撰写过程中，张老师的指导和鼓励对我帮助极大。

此外，我也要感谢我的同学和朋友们，他们的支持和鼓励对我完成这篇论文也起到了不可或缺的作用。在论文撰写的漫长过程中，他们的陪伴和帮助让我倍感温暖。

最后，我要感谢我的家人，他们无条件的爱和支持是我追求学术研究的坚强后盾。他们的理解和鼓励使我能够专心投入到这项研究中。

再次感谢所有帮助和支持我的人。没有他们的帮助，这篇论文无法完成。