東南大學圖書館
Southeast University Library

东图 学术快报

Academic express of SEU LIB

前沿经典　学科热点

学术动态　工具助手

“人工智能伦理”经典论文推荐

编者按：

人工智能，尤其是机器人在各个领域从事越来越多的决策，逐步从被动工具变成人类的代理者,这引发了社会各界对人工智能伦理的思考和担忧，需要建立新的伦理范式，将人类社会的伦理规范延伸到智能机器。近年来该人工智能伦理相关研究取得了较大发展,丰富了已有研究内容。本期快报通过分析国内外人工智能伦理相关论文的关键词，掌握其主要的学科应用，并根据被引等情况为大家推送近三年该方向核心期刊论文。我们也将继续更新其他学科经典学术论文的推荐，为研究提供良好的支撑。

CONTACT US

联系电话：025-52090336

办公地址：李文正图书馆 B401 室

本期编辑：潘海燕

本期审核：陆　美

2021-第 19 期-人工智能伦理

研究现状 *RESEARCH STATUS*

对近三年国内外“人工智能伦理”关键词进行共现分析，发现国内的研究热点主要集中在人工智能的伦理问题、伦理规范、伦理原则、伦理风险等方向（见图 1-1），多发表在《电化教育研究》、《远程教育杂志》、《西南民族大学学报（人文社科版）》、《现代远程教育研究》等教育类、社科类期刊上，发表论文呈逐年上升的趋势。

而发表在国外期刊上的论文主要体现在社会科学、哲学、商业经济学、情报学图书馆学、科学技术、计算机科学、生物医学、政府法、历史哲学等方向（见图 1-2），多发表在《SCIENCE AND ENGINEERING ETHICS》、《ETHICS AND INFORMATION TECHNOLOGY》、《BIG DATA SOCIETY》、《COMPUTER LAW SECURITY REVIEW》、《JOURNAL OF MEDICAL INTERNET RESEARCH》等期刊上。论文发表量同样呈逐年上升的趋势。

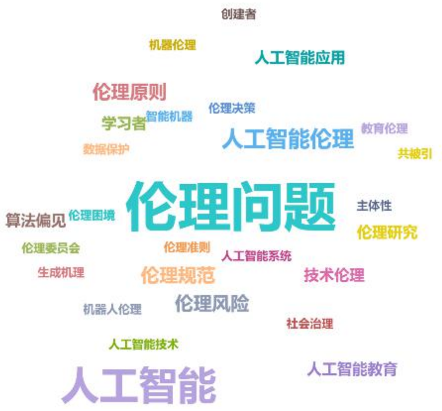

图 1-1 国内“人工智能伦理”研究热点

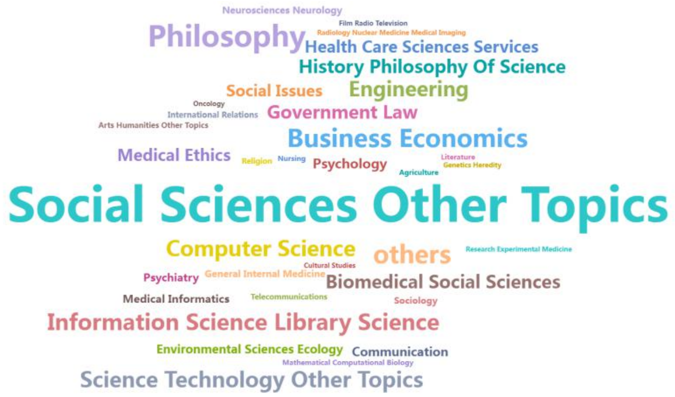

图 1-2 国外“人工智能伦理”研究热点

国内外均有不同的学科对“人工智能伦理”开展研究，属于交叉学科研究方向。国内更多集中在是在自动化技术、伦理学、计算机软件及应用、教育理论与管理等学科领域（图 1-3）。

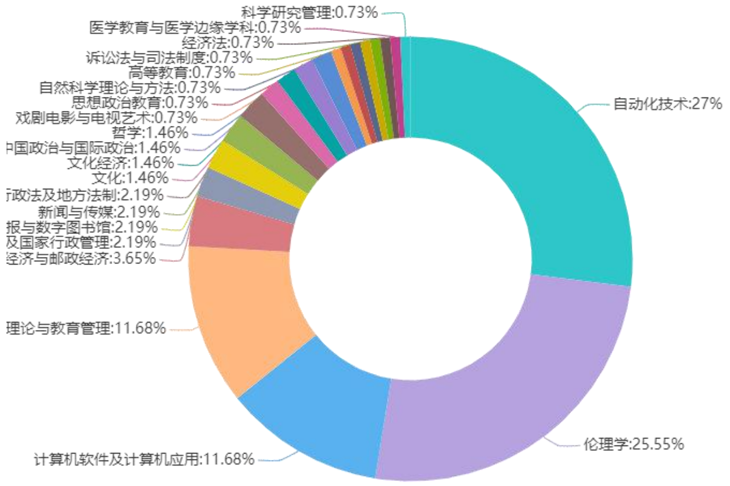

图 1-3 国内研究“人工智能伦理”的相关学科

国外在人工智能领域的研究相对来说分布更加均匀，各个学科方向都所研究，在伦理学、哲学、商业经济学、情报学图书馆学、计算机科学等领域发表人工智能伦理的论文较多，具体包括 Ethics、Philosophy、Information Science Library Science、Social Sciences Biomedical、History Philosophy Of Science、Multidisciplinary Sciences 等学科领域（图 1-4、图 1-5）。

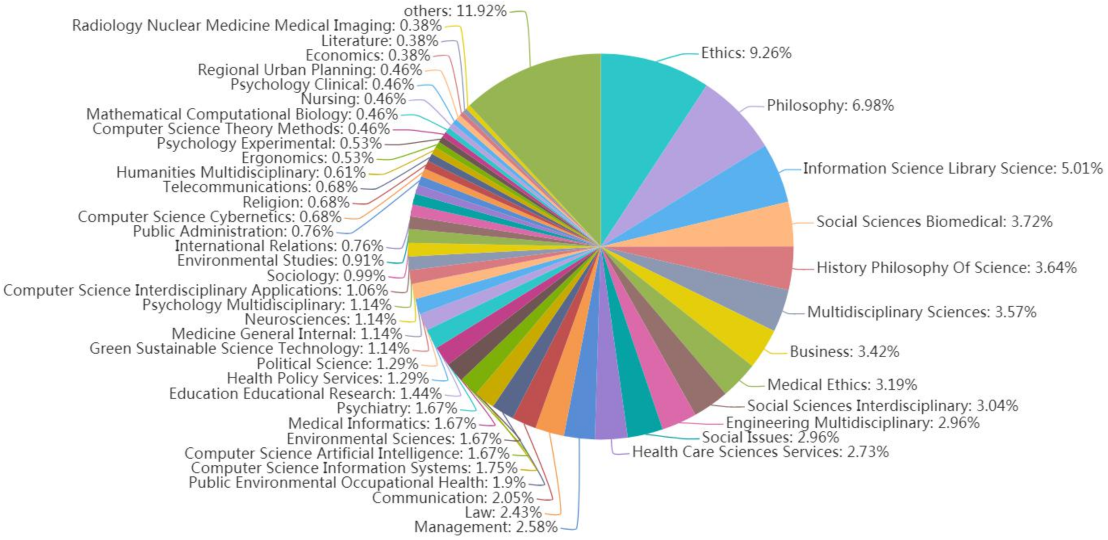

图 1-4 国外研究"人工智能伦理"的相关学科

热点追踪 HOT TRACE

国内相关论文推荐

国内相关论文推荐

数据来源：CNKI 中国知网，在来源类别中选择【核心期刊】和【CSSCI】；

论文发表年限：2019 年 1 月 1 日—2021 年 11 月 19 日；

检索词：“人工智能伦理”，通过篇关摘进行检索。

本次推荐被引频次较高的 5 篇论文。

1.

江丰光,熊博龙,张超.我国人工智能如何实现战略突破——基于中美 4 份人工智能发展报告的比较与解读[J].现代远程教育研究,2020,32(01):3-11.

摘要:在人工智能发展的热潮下,全球各个国家均针对人工智能的发展方向和应用场景进行了高屋建瓴的政策规划。政策规划的重要旨趣在于作为"风向标"引导投资、研究与实践的方向,使之更加聚焦和深入,并为研究与实践提供强有力的政策支撑。美国于 2016 年首次发布并于 2019 年更新的《国家人工智能研究与发展战略规划》表明美国政府对人工智能在全球未来竞争中战略作用的研判。中国紧跟国际步伐,于 2017 年发布《新一代人工智能发展规划》并提出"三步走"战略,后在研究和实践层面陆续推出《新一代人工智能发展白皮书（2017 版）》和《人工智能标准化白皮书（2018 版）》。这两份白皮书与美国的两份《国家人工智能研究与发展战略

规划》在人工智能研发基础、研发领域以及应用领域三个方面具有可比性,并体现出各自的特色。反思中美人工智能战略规划的差异,可以为我国人工智能未来发展提供镜鉴。鉴于当前我国在新一代人工智能发展规划中对人才和核心技术规划和部署的重视,以及在人工智能伦理研究和实践方面仍存在提升空间,未来我国应从人工智能伦理体系、伦理准则和伦理风险等方面完善人工智能伦理、法律和道德建设,持续加强人工智能人才培养和核心技术的研发,并且加强企业与研究机构的协同合作。

2.

于雪,段伟文.人工智能的伦理建构[J].理论探索,2019(06):43-49.

摘要:人工智能2.0时代标志着人机一体化新一代技术革命的来临,这种背景下建构人工智能伦理不仅重要而且必要。人工智能技术具有不确定性、隐象性和系统复杂性,其伦理建构需要遵循整体性、过程性、适应性、相容性、灵活性和鲁棒性等原则。人工智能伦理的框架建构可以通过明晰人工智能伦理的基本价值、确定人工智能伦理的基本标准,以及落实人工智能伦理的相关责任这三方面得以实现,并且通过嵌入设计、合理使用、合作管理和多元参与这四条实践进路来推进。

3.

谢洪明,陈亮,杨英楠.如何认识人工智能的伦理冲突?——研究回顾与展望[J]. 外国经济与管理,2019,41(10):109-124.

摘要:人工智能尤其是机器人在各个领域从事越来越多的决策,逐步从被动工具变成人类的代理者,这引发了社会各界对人工智能伦理的思考和担忧,需要建立新的伦理范式,将人类社会的伦理规范延伸到智能机器。近年来该研究取得了较大发展,丰富了已有研究内容,提升了对人工智能伦理研究的指导能力。本文全面回顾了国内外有关人工智能伦理的研究进展,在文献计量分析的基础上,发现"传统派"、"谨慎派"和"乐观派"三种对人工智能的不同态度引发了"人—机"关系的伦理冲突,从人工智能道德哲学、道德算法、设计伦理和社会伦理四个视角系统性地评述了人工智能伦理的研究成果,国家和企业(组织)分别从战略和社会责任的层面上强调对人工智能伦理的态度,提出了更加系统、完善的人工智能伦理的理论框架,有助于从理论和实践层面系统地把握已有研究成果。未来需要在全球情景条件的伦理体系建设、伦理对技术的前瞻性、伦理角色塑造和科学发展的伦理观上做进一步研究。

4.

王钰,程海东.人工智能技术伦理治理内在路径解析[J].自然辩证法通讯,2019,41(08):87-93.

摘要:人工智能的不确定性为人类对其进行治理提供了可能。伦理对人工智能的嵌入体现为作为伦理型道德能动体的人类对其发展的管理与责任承担。根据具体技术的发展过程,伦理对人工智能的嵌入先后按照四个阶段进行:在设计阶段进行伦理嵌入,以人工智能专家为主导,通过"预测-评估-设计"模型实现人工智能的道德化设计;在试验阶段进行伦理评估,以评估委员会为主导,通过伦理效应的预测与识别、伦理问题的分析与澄清、以及解决方案的开发与确定来修正和完善人工智能开发方案;在推广阶段进行伦理调适,以政府部门为主导,通过制度调适、舆论调适和教育调适三种路径,实现人工智能与社会价值系统的顺利融合;在使用阶段,以使用者为主导,通过对他者、对世界、对技术以及对自身的责任的主动承担,来确认自身作为伦理型道德能动体的地位,并为人工智能伦理潜能的实现提供支撑。

5.

杜静,黄荣怀,李政璇,周伟,田阳.智能教育时代下人工智能伦理的内涵与建构原则[J].电化教育研究,2019,40(07):21-29.

摘要:在智能教育时代,人机如何共处是人工智能伦理建构的关键。文章首先从技术悖论视角,厘清当前人

工智能应用于教育在技术滥用、数据泄露、智能教学机器的身份与权力边界等方面存在的伦理挑战与困境；其次，利用内容分析法，结合多国与国际组织政策文件，对人工智能伦理相关的伦理要素进行分析与抽取，发现政府、高校、国际组织文件中多次提到的价值、人类利益、安全、隐私、责任等关键要素；最后，基于人机共处的考量，结合人工智能在教育领域的应用现状和伦理关键要素，归纳分析出智能教育伦理需遵循的原则，包括问责原则、隐私原则、平等原则、透明原则、不伤害原则、非独立原则、预警原则与稳定原则。

国外相关论文推荐

数据来源：SSCI（Social Sciences Citation Index）、A&HCI (Arts & Humanities Citation Index)；

论文发表年限：2019 年 1 月 1 日—2021 年 11 月 22 日；

检索词：“Artificial Intelligence or AI”and“ethic”，在 SSCI 和 A&HCI 数据库中通过主题检索，同时人工删除了部分与研究主题相关性较小的论文，本次推荐被引次数较高的 5 篇相关论文。

1.

标题: **How artificial intelligence will change the future of marketing**

作者: Davenport, T (Davenport, Thomas); Guha, A (Guha, Abhijit); Grewal, D (Grewal, Dhruv); Bressgott, T (Bressgott, Timna)

来源出版物: JOURNAL OF THE ACADEMY OF MARKETING SCIENCE 卷: 48 期: 1 特刊: SI 页: 24-42 DOI: 10.1007/s11747-019-00696-0 出版年: JAN 2020

摘要: In the future, artificial intelligence (AI) is likely to substantially change both marketing strategies and customer behaviors. Building from not only extant research but also extensive interactions with practice, the authors propose a multidimensional framework for understanding the impact of AI involving intelligence levels, task types, and whether AI is embedded in a robot. Prior research typically addresses a subset of these dimensions; this paper integrates all three into a single framework. Next, the authors propose a research agenda that addresses not only how marketing strategies and customer behaviors will change in the future, but also highlights important policy questions relating to privacy, bias and ethics. Finally, the authors suggest AI will be more effective if it augments (rather than replaces) human managers.

2.

标题: **A governance model for the application of AI in health care**

作者: Reddy, S (Reddy, Sandeep); Allan, S (Allan, Sonia); Coghlan, S (Coghlan, Simon); Cooper, P (Cooper, Paul)

来源出版物: JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION 卷: 27 期: 3 页: 491-497 DOI: 10.1093/jamia/ocz192 出版年: MAR 2020

摘要: As the efficacy of artificial intelligence (AI) in improving aspects of healthcare delivery is increasingly becoming evident, it becomes likely that AI will be incorporated in routine clinical care in the near future. This promise has led to growing focus and investment in AI medical applications both from governmental organizations and technological companies. However, concern has been expressed about the ethical and regulatory aspects of the application of AI in health care. These concerns include the possibility of biases, lack of transparency with certain AI algorithms, privacy concerns with the data used for training AI models, and safety and liability issues with AI application in clinical environments. While there has been extensive discussion about the ethics of AI in health care, there has been little dialogue or recommendations as to how to practically address these concerns in health care. In this article, we propose a governance model that aims to not only address the ethical and regulatory issues that arise out of the application of AI in health care, but also stimulate further discussion about governance of AI in health care.

3.

标题: **The Ethics of AI Ethics: An Evaluation of Guidelines**

作者: Hagendorff, T (Hagendorff, Thilo)

来源出版物: MINDS AND MACHINES　卷: 30　期: 1　页: 99-120　DOI: 10.1007/s11023-020-09517-8　提前访问日期: FEB 2020　出版年: MAR 2020

摘要: Current advances in research, development and application of artificial intelligence (AI) systems have yielded a far-reaching discourse on AI ethics. In consequence, a number of ethics guidelines have been released in recent years. These guidelines comprise normative principles and recommendations aimed to harness the "disruptive" potentials of new AI technologies. Designed as a semi-systematic evaluation, this paper analyzes and compares 22 guidelines, highlighting overlaps but also omissions. As a result, I give a detailed overview of the field of AI ethics. Finally, I also examine to what extent the respective ethical principles and values are implemented in the practice of research, development and application of AI systems-and how the effectiveness in the demands of AI ethics can be improved.

4.

标题: **From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices**

作者: Morley, J (Morley, Jessica); Floridi, L (Floridi, Luciano); Kinsey, L (Kinsey, Libby); Elhalal, A (Elhalal, Anat)

来源出版物: SCIENCE AND ENGINEERING ETHICS　卷: 26　期: 4　特刊: SI　页: 2141-2168　DOI: 10.1007/s11948-019-00165-5　出版年: AUG 2020

摘要: The debate about the ethical implications of Artificial Intelligence dates from the 1960s (Samuel in Science, 132(3429):741-742, 1960. https://doi.org/10.1126/science.132.3429.741; Wiener in Cybernetics: or control and communication in the animal and the machine, MIT Press, New York, 1961). However, in recent years symbolic AI has been complemented and sometimes replaced by (Deep) Neural Networks and Machine Learning (ML) techniques. This has vastly increased its potential utility and impact on society, with the consequence that the ethical debate has gone mainstream. Such a debate has primarily focused on principles-the 'what' of AI ethics (beneficence, non-maleficence, autonomy, justice and explicability)-rather than on practices, the 'how.' Awareness of the potential issues is increasing at a fast rate, but the AI community's ability to take action to mitigate the associated risks is still at its infancy. Our intention in presenting this research is to contribute to closing the gap between principles and practices by constructing a typology that may help practically-minded developers apply ethics at each stage of the Machine Learning development pipeline, and to signal to researchers where further work is needed. The focus is exclusively on Machine Learning, but it is hoped that the results of this research may be easily applicable to other branches of AI. The article outlines the research method for creating this typology, the initial findings, and provides a summary of future research needs.

5.

标题: **Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI**

作者: Duran, JM (Duran, Juan Manuel); Jongsma, KR (Jongsma, Karin Rolanda)

来源出版物: JOURNAL OF MEDICAL ETHICS 卷: 47 期: 5 页: 329-335 DOI: 10.1136/medethics-2020-106820 出版年: MAY 2021

摘要: The use of black box algorithms in medicine has raised scholarly concerns due to their opaqueness and lack of trustworthiness. Concerns about potential bias, accountability and responsibility, patient autonomy and compromised trust transpire with black box algorithms. These worries connect epistemic concerns with normative issues. In this paper, we outline that black box algorithms are less problematic for epistemic reasons than many scholars seem to believe. By outlining that more transparency in algorithms is not always necessary, and by explaining that computational processes are indeed methodologically opaque to humans, we argue that the reliability of algorithms provides reasons for trusting the outcomes of medical artificial intelligence (AI). To this end, we explain how computational reliabilism, which does not require transparency and supports the reliability of algorithms, justifies the belief that results of medical AI are to be trusted. We also argue that several ethical concerns remain with black box algorithms, even when the results are trustworthy. Having justified knowledge from reliable indicators is, therefore, necessary but not sufficient for normatively justifying physicians to act. This means that deliberation about the results of reliable algorithms is required to find out what is a desirable action. Thus understood, we argue that such challenges should not dismiss the use of black box algorithms altogether but should inform the way in which these algorithms are designed and implemented. When physicians are trained to acquire the necessary skills and expertise, and collaborate with medical informatics and data scientists, black box algorithms can contribute to improving medical care.