



## 人工智能指数2022年度报告



斯坦福大学  
以人本人工智能研究院  
(斯坦福HAI)



# 2022年人工智能指数报告简介

欢迎来到第五期《人工智能指数报告》！我们在2022年的报告中引入了来自广泛的学术、私人和非营利组织的数据。同时，相较于前四期版本，本期增加了更多基于自行收集的数据以及原创分析。2022年人工智能指数报告中扩展了技术性能章节，同时引入了对世界各地机器人研究人员的调查、25个国家的全球人工智能立法记录数据，以及深入分析技术性人工智能伦理指标的新章节。

人工智能指数报告跟踪、整理、提炼和可视化与人工智能相关的数据。其使命是为政策制定者、研究人员、高级管理人员、记者和公众提供公正的、经过严格审核的、来源于全球的数据，以形成对人工智能这一复杂领域的直观感知。该报告旨在成为世界上最可信、最权威的人工智能数据和分析来源。

## 联合负责人的话

今年的报告显示，人工智能系统已经开始广泛部署到经济领域中，但在部署的同时，与人工智能相关的伦理问题也在不断放大。其中一些是自然现象，毕竟，当某项技术推广到全世界时，人们往往更关心它的伦理问题。但是，也有一些是由人工智能的特殊性所导致的，即，更庞大、更复杂、更有能力的人工智能系统通常能在广泛的任务中表现出很好的性能，同时也表现出更严重的潜在伦理问题。

这一问题与人工智能的全面全球化和产业化息息相关。越来越多的国家正在开发、部署和监管人工智能系统，这些活动的共同影响结果是形成了一套更广泛的人工智能系统供人们使用，并降低了其价格。不过，人工智能的某些部分并不是非常全球化的，我们的伦理学分析显示，尽管人工智能在全球范围内部署，但许多人工智能伦理学出版物往往集中在英语系统和数据集上。

我们预计上述趋势将会持续：2021年对人工智能创企及与人工智能相关初创公司的私人投资比2020年增加了103%（965亿美元 vs 460亿美元）。

**Jack Clark and Ray Perrault**



# 报告要点

## 人工智能领域的私人投资飙升，同时投资集中度加强：

- 2021年人工智能领域的私人投资总额约为935亿美元，是2020年私人投资总额的两倍多，而新获得融资的人工智能公司数量继续下降，从2019年的1051家公司和2020年的762家公司降至2021年的746家公司。2020年，有4起价值5亿美元以上的融资轮次；2021年则有15起。

## 美国和中国主导了人工智能方面的跨国合作：

- 尽管地缘政治紧张局势加剧，但从2010年到2021年，美国和中国在人工智能出版物方面的跨国合作数量最多，自2010年以来增加了5倍。中美合作产生的出版物是中英合作合作产出的2.7倍（跨国合作出版物数量榜单中排名第二位）。

## 语言模型效果比以前更优，但偏见也更严重：

- 大型语言模型正在不断创造技术基准的新记录，但新数据显示，大型模型也更能反映出训练数据的偏见。对比2018年最先进的1.17亿参数的模型，2021年开发的一个2800亿参数的模型所引发的毒性（elicited toxicity）增加了29%。随着时间的推移，这些系统的能力明显增强，但是随着它们能力的增强，其潜在的偏见的严重程度也在增加。

## 人工智能伦理的兴起，无处不在：

- 自2014年以来，关于人工智能的公平性和透明度的研究呈爆炸式增长，在伦理学相关会议上发表的相关文章数量增加了五倍。算法的公平性和偏见问题已经从主要的学术追求转变为一个具有广泛影响的主流研究课题。近年来，与产业界有联系的研究人员在以伦理学为重点的会议上发表的论文数量同比增加71%。

## 人工智能的应用成本变得更低，性能更高：

- 自2018年以来，训练一个图像分类系统的成本下降了63.6%，而所需的训练时间则缩短了94.4%。训练成本更低但训练时间更快的趋势也出现在其他MLPerf任务类别中，如推荐、物体检测和语言处理，推动了AI技术更广泛的商业应用。

## 数据，数据，数据：

- 各项技术越来越依赖于使用额外的训练数据来创造新的最先进的结果。截至2021年，本报告中的10个基准中，有9个最先进的人工智能系统是引入额外数据训练得到的。这种趋势将会越来越有利于掌握大量数据的私营机构。

## 关于人工智能的全球立法比以往更多：

- 本报告对25个国家的人工智能立法记录的一项分析显示，包含“人工智能”的法案被通过成为法律的数量从2016年的1个增长到2021年的18个。西班牙、英国和美国在2021年通过的人工智能相关法案数量最多，各通过了三项。

## 机械臂正在变得更加便宜：

- 本报告的一项问卷调研显示，在过去六年中，机械臂的中位价格下降了46.2%，从2017年的每只手臂42,000美元下降到2021年的22,600美元。机器人研究已经变得更加容易实现，所需成本也越来越低。



# 指导委员会

联合负责人

Jack Clark  
Anthropic, OECD

Raymond Perrault  
国际斯坦福研究所

## Members

# Erik Brynjolfsson 斯坦福大学

James Manyika  
Google, 牛津大学

Michael Sellitto  
斯坦福大学

John Etchemendy  
斯坦福大学

Juan Carlos Niebles  
斯坦福大学, Salesforce

Yoav Shoham  
(创始董事)  
斯坦福大学人工智能21实验室

Terah Lyons

## 工作人员和研究人员

研究经理和主编

Daniel Zhang  
斯坦福大学

研究助理

Nestor Maslej  
斯坦福大学

相关研究人员

Andre Barbe  
世界银行

Helen Ngo  
Cohere

Latisha Harry  
独立咨询师

Ellie Sakhaei  
Microsoft

研究生研究人员

Benjamin Bronkema-Bekker  
斯坦福大学



## 如何引用本报告

Daniel Zhang, Nestor Maslej, Erik Brynjolfsson, John Etchemendy, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Michael Sellitto, Ellie Sakhaee, Yoav Shoham, Jack Clark, and Raymond Perrault, “The AI Index 2022 Annual Report,” AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University, March 2022. Chinese translation by Synced.

斯坦福大学《人工智能指数2022年年度报告》由Attribution-No Derivatives 4.0 International 授权许可。可通过下面的网址查看该许可的副本：

<http://creativecommons.org/licenses/by-nd/4.0/>.

## 公开数据和工具

2022年人工智能指数报告的附录部分提供了原始数据和互动工具。我们邀请人工智能界的各位成员参与，您可以以与您工作和兴趣最为相关的方式使用这些数据和工具。

- 原始数据和图表。报告中所有图表的公开数据和高分辨率图像均可在 [Google Drive](#) 上找到。
- 全球人工智能活力工具。我们今年重新设计了[全球人工智能活力工具](#)，以更好的可视化方式对多达29个国家的23个指标进行比较。

## AI指数和斯坦福大学HAI

人工智能指数是斯坦福大学以人为本的人工智能研究所（HAI）的一项独立倡议。



Artificial  
Intelligence  
Index



Stanford University  
Human-Centered  
Artificial Intelligence

人工智能指数是在“[人工智能百年研究](#)”(AI100)中构思的。

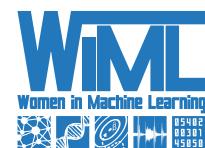
欢迎大家反馈针对本报告以及对明年报告的想法、建议。  
可通过邮件[AI-Index-Report@stanford.edu](mailto:AI-Index-Report@stanford.edu)与我们联系。



## 支持伙伴



## 分析和研究合作伙伴



## 中文版独家合作伙伴：机器之心





## 撰稿人

我们要感谢以下个人，他们在《2022年人工智能指数报告》中提供了数据、分析、建议和专家评论。

### 研发

Sara Abdulla, Catherine Aiken, Jack Clark, James Dunham, Nezihe Merve Gürel, Nestor Maslej, Ray Perrault, Sarah Tan, Daniel Zhang

### 技术性能

Jack Clark, David Kanter, Nestor Maslej, Deepak Narayanan, Juan Carlos Niebles, Konstantin Savenkov, Yoav Shoham, Daniel Zhang

### 技术AI伦理

Jack Clark, Nestor Maslej, Helen Ngo, Ray Perrault, Ellie Sakhaei, Daniel Zhang

### 经济和教育

Betsy Bizot, Erik Brynjolfsson, Jack Clark, John Etchemendy, Murat Erer, Akash Kaura, Julie Kim, Nestor Maslej, James Manyika, Brenden McKinney, Julia Nitschke, Ray Perrault, Brittany Presten, Tejas Sirohi, Bledi Taska, Rucha Vankudre, Daniel Zhang

### 人工智能政策和治理

Amanda Allen, Benjamin Bronkema-Bekker, Jack Clark, Latisha Harry, Taehwa Hong, Cameron Leuthy, Terah Lyons, Nestor Maslej, Ray Perrault, Michael Sellitto, Teruki Takiguchi, Daniel Zhang

### 会议出席情况

Terri Auricchio (ICML), Christian Bessiere (IJCAI), Meghyn Bienvenu (KR), Andrea Brown (ICLR), Alexandra Chouldechova (FAccT), Nicole Finn (ICCV, CVPR), Enrico Gerding (AAMAS), Carol Hamilton (AAAI), Seth Lazar (FAccT), Max Qing Hu Meng (ICRA), Jonas Martin Peters (UAI), Libor Preucil (IROS), Marc' Aurelio Ranzato (NeurIPS), Priscilla Rasmussen (EMNLP, ACL), Hankz Hankui Zhuo (ICAPS)

### 全球人工智能振幅分析工具

Andre Barbe, Latisha Harry, Daniel Zhang

### 机器人技术问卷调研

Pieter Abbeel, David Abbink, Farshid Alambeigi, Farshad Arvin, Nikolay Atanasov, Ruzena Bajcsy, Philip Beesley, Tapomayukh Bhattacharjee, Jeannette Bohg, David J. Cappelleri, Qifeng Chen, I-Ming Chen, Jack Cheng, Cynthia Chestek, Kyujin Cho, Dimitris Chrysostomou, Steve Collins, David Correa, Brandon DeHart, Katie Driggs-Campbell, Nima Fazeli, Animesh Garg, Maged Ghoneima, Tobias Haschke, Kris Hauser, David Held, Yue Hu, Josie Hughes, Soo Jeon, Dimitrios Kanoulas, Jonathan Kelly, Oliver Kroemer, Changliu Liu, Ole Madsen, Anirudha Majumdar, Genaro J. Martinez, Saburo Matunaga, Satoshi Miura, Norrima Mokhtar, Elena De Momi, Christopher Nehaniv, Christopher Nielsen, Ryuma Niiyama, Allison Okamura, Necmiye Ozay, Jamie Paik, Frank Park, Karthik Ramani, Carolyn Ren, Jan Rosell, Jee-Hwan Ryu, Tim Salcudean, Oliver Schneider, Angela Schoellig, Reid Simmons, Alvaro Soto, Peter Stone, Michael Tolley, Tsu-Chin Tsao, Michiel van de Panne, Andy Weightman, Alexander Wong, Helge Wurdemann, Rong Xiong, Chao Xu, Geng Yang, Junzhi Yu, Wenzhen Yuan, Fu Zhang, Yuke Zhu



我们感谢以下组织和个人为《2022年人工智能指数报告》提供数据。

## 组织

### Bloomberg Government

Amanda Allen, Cameron Leuthy

### Center for Security and Emerging Technology, Georgetown University

Sara Abdulla, Catherine Aiken,  
James Dunham

### Computing Research Association

Betsy Bizot

### Emsi Burning Glass

Julia Nitschke, Bledi Taska,  
Rucha Vankudre

### Intento

Grigory Sapunov, Konstantin Savenkov

### LinkedIn

Murat Erer, Akash Kaura

### McKinsey Global Institute

Brenden McKinney, Brittany Presten

### MLCommons

David Kanter

### NetBase Quid

Julie Kim, Tejas Sirohi

### Women in Machine Learning

Nezihe Merve Gürel, Sarah Tan

我们还要感谢Jeanina Casusi、Nancy King、Shana Lynch、Jonathan Mindes、Stacy Peña、Michi Turner和Justin Sherman在编写本报告时提供的帮助，以及Joe Hinman、Travis Taylor和Digital Avenues团队在设计和开发AI指数和HAI网站方面的努力。



# 目录

报告要点	10
章节一 研发 (R&D)	13
章节二 技术性能	47
章节三 技术AI伦理	100
章节四 经济和教育	139
章节五 人工智能政策和治理	172
附录	196

访问公开数据



# 报告要点

## 章节一 研发 (R&D)

- 尽管地缘政治紧张局势加剧，但从2010年到2021年，美国和中国在人工智能出版物方面的跨国合作数量最多，**自2010年以来增加了5倍。中美合作产生的出版物是中英合作产出的2.7倍（跨国合作出版物数量榜单中排名第二位）。**
- 2021年，中国在人工智能期刊、会议和文献库这三种类型的出版物的数量上继续领先，**比美国所有三种类型出版物的总和高出63.2%。**同时，在主要的人工智能大国中，美国在**人工智能会议和文献库的引用数量上保持着主导地位。**
- 从2010年到2021年，**教育和非营利组织之间的合作完成的人工智能出版物数量最多**，其次是私营公司和教育机构之间的合作以及教育和政府机构之间的合作。
- 2021年申请的人工智能专利数量是2015年的30多倍**，达到了76.9%的年复合增长率。

## 章节二 技术性能

- 数据、数据、数据：**各项技术越来越依赖于使用额外的训练数据来创造新的最先进的结果。**截至2021年，本报告中的10个基准中，有9个最先进的人工智能系统是引入额外数据训练得到的。**这种趋势将会越来越有利于掌握大量数据的私营机构。
- 对特定计算机视觉子任务的兴趣上升：**2021年，研究界对更具体的计算机视觉子任务有了更大的兴趣，如医学图像分割和遮挡人脸识别。**例如，在2020年之前，只有3篇研究论文涉及Kvasir-SEG医学成像基准测试系统。在2021年，则有25篇研究论文。**这样的增长表明，人工智能研究正在朝着能够有更直接的、现实世界的应用的研究方向发展。
- 人工智能还无法应对复杂的语言任务：**在像SuperGLUE和SQuAD这样的基本阅读理解基准上，人工智能的表现已经超过了人类水平1%-5%。**虽然人工智能系统在更复杂的语言任务上仍无法达到人类的表现，如归纳自然语言推理 (aNLI)，但差距正在缩小。**2019年，人类在aNLI上的表现超过人工智能9个百分点。**截至2021年，这一差距已经缩小到1%。**
- 转向更广泛的强化学习：**在过去的十年里，人工智能系统已经能够完成狭义的强化学习任务，在这些任务中，它们被要求最大限度地提高某一特定技能的表现，例如国际象棋。**顶级国际象棋软件引擎现在比Magnus Carlsen的Top ELO分数高出24%。**然而，在过去的两年里，人工智能系统在更普遍的强化学习任务 (Progen) 上的性能也提高了129%，**在这些任务中，它们必须在新的环境中运行。**这一趋势说明，能够学会更广泛思考的人工智能系统正在不断发展。



- **人工智能变得更经济，性能更高：**自2018年以来，训练图像分类系统的成本下降了63.6%，而训练时间缩短了94.4%。训练成本降低但训练时间加快的趋势出现在其他MLPerf任务类别中，如推荐、对象检测和语言处理，推动了AI技术更广泛的商业应用。
- **机械臂正在变得更便宜：**本报告的一项问卷调研显示，在过去六年中，**机械臂的中位价格下降了46.2%**--从2017年的每只手臂42,000美元到2021年的22,600美元。机器人研究已变得更容易实现，成本更低。

### 章节三 技术AI伦理

- **语言模型效果比以前更优，但偏见也更严重：**大型语言模型正在不断创造技术基准的新记录，但新数据显示，大型模型也更能反映出训练数据的偏见。**相较于2018年最先进的1.17亿参数的模型相比，2021年开发的一个2800亿参数的模型引发的毒性（elicited toxicity）增加了29%。**随着时间的推移，这些系统的能力明显增强，但是随着它们能力的增强，其潜在的偏见的严重程度也在增加。
- **人工智能伦理兴起，无处不在：**自2014年以来，关于人工智能的公平性和透明度的研究呈爆炸式增长，在伦理学相关会议上发表的相关文章数量增加了五倍。算法的公平性和偏见问题已经从主要的学术追求转变为一个具有广泛影响的主流研究课题。**近年来，与产业界有联系的研究人员在以伦理学为重点的会议上发表的论文数量同比增加71%。**
- **多模态模型学习多模态偏见：**在训练多模态语言-视觉模型方面已经取得了快速进展，这些模型在语言-视觉联合任务上表现出了更高的能力水平。这些模型在图像分类和从文本描述中创建图像等任务上创造了新的记录，但它们也在其输出中反映了社会固有观念和偏见--**在CLIP上的实验表明，黑人的图像被错误地分类为非人类的比率是其他种族的两倍以上。**在计算机视觉和自然语言处理领域，已经有大量的工作致力于开发测量偏见的指标，这凸显了对能够深入了解多模态模型的偏差的指标的需求。

### 章节四 经济与教育

- 新西兰、香港、爱尔兰、卢森堡和瑞典是2016年至2021年人工智能招聘增长最快的国家或地区。
- 2021年，加利福尼亚州、德克萨斯州、纽约州和弗吉尼亚州是美国人工智能职位发布数量最多的州，**其中加利福尼亚州的职位发布数量是排名第二大州德克萨斯州的2.35倍以上。**华盛顿特区的人工智能职位发布率与它的总体职位发布数量相比是最高的。
- **2021年人工智能领域的私人投资总额约为935亿美元，是2020年私人投资总额的两倍多。**而新获得融资的人工智能公司数量继续下降，从2019年的1051家公司和2020年的762家公司降至2021年的746家公司。**2020年，有4起价值5亿美元以上的融资轮次；2021年，则有15起。**
- “数据管理、处理和云“在2021年获得了最大的私人人工智能投资额--**是2020年的2.6倍**，其次是“医疗和保健”和“金融技术”。



- 2021年，美国在人工智能领域的私人投资总额和新资助的人工智能公司数量方面都处于全球领先地位，分别比排名第二的中国高出三倍和两倍。
- 麦肯锡的一项调查显示，专门应对产业场景中使用人工智能相关的道德问题所做的努力仍然有限。虽然29%和41%的受访者认识到“公平和公正”以及“可解释性”是采用人工智能时的风险，但只有19%和27%的人正在采取措施减轻这些风险。
- 2020年，每5个获得博士学位的CS学生中就有1个专门从事人工智能/机器学习，这是过去十年中最受欢迎的专业。从2010年到2020年，美国的大多数人工智能博士都走向了产业界，而一小部分则在政府工作。

## 章节五 人工智能政策和治理

- 人工智能指数对25个国家的人工智能立法记录的分析显示，包含“人工智能”的法案被通过成为法律的数量从2016年的1个增长到2021年的18个。西班牙、英国和美国在2021年通过的人工智能相关法案数量最多，各通过了三项。
- 美国的联邦立法记录显示，从2015年到2021年，与人工智能有关提案总数急剧增加，而通过的法案数量仍然很少，只有2%最终确定立法。
- 2021年，美国各州立法者通过了每50个包含人工智能条款的提案中的一个，而此类提案数量从2012年的2个增长到2021年的131个。
- 在美国，本届国会会议（第117届）有望创下自2001年以来与人工智能有关的最多提及次数，到2021年底，即本届会议过半时，将有295次提及，而上届（第116届）则有506次。



2022年  
人工智能指数报告

## 章节一 研发 (R&D)



## 章节一 研发 (R&D) 章节预览

概述	15	人工智能文献库	32
章节要点	16	概述	32
<b>1.1 出版物</b>	<b>17</b>	按地区 (Region) 划分	33
概述	17	按地理区域 (Geographical Area) 划分	34
人工智能出版物总数	17	引用	35
按出版物类别	18	<b>人工智能专利</b>	36
按研究领域	19	概述	36
按行业划分	20	按区域和应用状态	37
跨地区合作	22	按地理区域和应用状态	39
跨行业合作	23		
人工智能期刊论文	24	<b>1.2 会议</b>	41
概述	24	会议出席情况	41
按地区 (Region) 划分	25	Women in Machine Learning (WiML)	
按地理区域 (Geographical Area) 划分	26	NeurIPS Workshop	43
引用	27	研讨会参加者	43
人工智能会议论文	28	人口统计学分类	44
概述	28		
按地区 (Region) 划分	29	<b>1.3 人工智能开源软件库</b>	45
按地理区域 (Geographical Area) 划分	30	GitHub 星标	45
引用	31		

访问公开数据



# 概述

研发（R&D）是推动人工智能（AI）快速发展的重要力量。每年，众多学术界、产业界、政府和民间社会的专家和组织通过大量的论文、期刊文章和其他与人工智能相关的出版物、关于人工智能或图像识别或自然语言处理等特定子课题的会议、跨国界的国际合作以及开源软件库的开发，为人工智能研发做出贡献。这些研发工作的重点各不相同，而且在地理区域上也很分散。

人工智能研发的另一个关键特征是其开放性，这使其在某种程度上有别于其他科技教育研究领域。每年，数以千计的人工智能出版物在公开资源中发布，包括会议和文件共享网站。研究人员会在会议上公开分享他们的研究成果，政府机构会资助那些在公开资源中发布的人工智能研究，开发人员会使用向公众免费提供的开放软件库来开发最先进的人工智能应用程序。这种开放性也促进了现代人工智能研发的全球相互依赖和相互联系的特性发展。

章节一利用多个数据集来分析2021年人工智能的主要研发趋势。本章节首先考察了人工智能出版物，包括会议论文、期刊文章、专利和文献库；其次，分析了人工智能会议的出席情况；最后，调研了研发过程中使用的人工智能开源软件库。



## 章节要点

- 尽管地缘政治紧张局势加剧，但从2010年到2021年，美国和中国在人工智能出版物方面的跨国合作数量最多，**自2010年以来增加了5倍**。中美合作产生的出版物是中英合作合作产出的**2.7倍**（跨国合作出版物数量榜单中排名第二位）。
- 2021年，中国在人工智能期刊、会议和文献库这三种类型的出版物的数量上继续领先，**比美国所有三种类型出版物的总和高出63.2%**。同时，在主要的人工智能大国中，美国在**人工智能会议和文献库的引用数量上保持着主导地位**。
- 从2010年到2021年，**教育和非营利组织之间的合作完成的人工智能出版物数量最多**，其次是私营公司和教育机构之间的合作以及教育和政府机构之间的合作。
- 2021年申请的人工智能专利数量是2015年的30多倍**，达到了76.9%的年复合增长率。



本节借鉴了乔治敦大学安全与新兴技术中心（the Center for Security and Emerging Technology, CSET）的数据。CSET负责一个综合学术文献库的维护工作，其中包括数字科学的Dimensions、Clarivate的Web of Science、微软学术图谱、中国国家知识基础设施、arXiv和Papers with Code。在该文献库中，CSET应用一个分类器来识别自2010年以来与人工智能和机器学习的发展或应用有关的英文出版物。<sup>1</sup>

## 1.1 出版物<sup>2</sup>

### 概述

下图展示了2010年至2021年全球英语人工智能出版物的总数，并按类型、隶属关系、跨国合作和跨行业合作分类。该部分还按地区细分了人工智能期刊文章、会议论文、文献库和专利的出版和引用数据。

### 人工智能出版物总数

图1.1.1展示了全球人工智能出版物的总数。从2010年到2021年，人工智能出版物的总数翻了一番，从2010年的162,444篇增长到2021年的334,497篇。

#### 2010-21年全球人工智能出版物总数

来源：安全与新兴技术中心（CSET），2021 | 图：2022人工智能指数报告

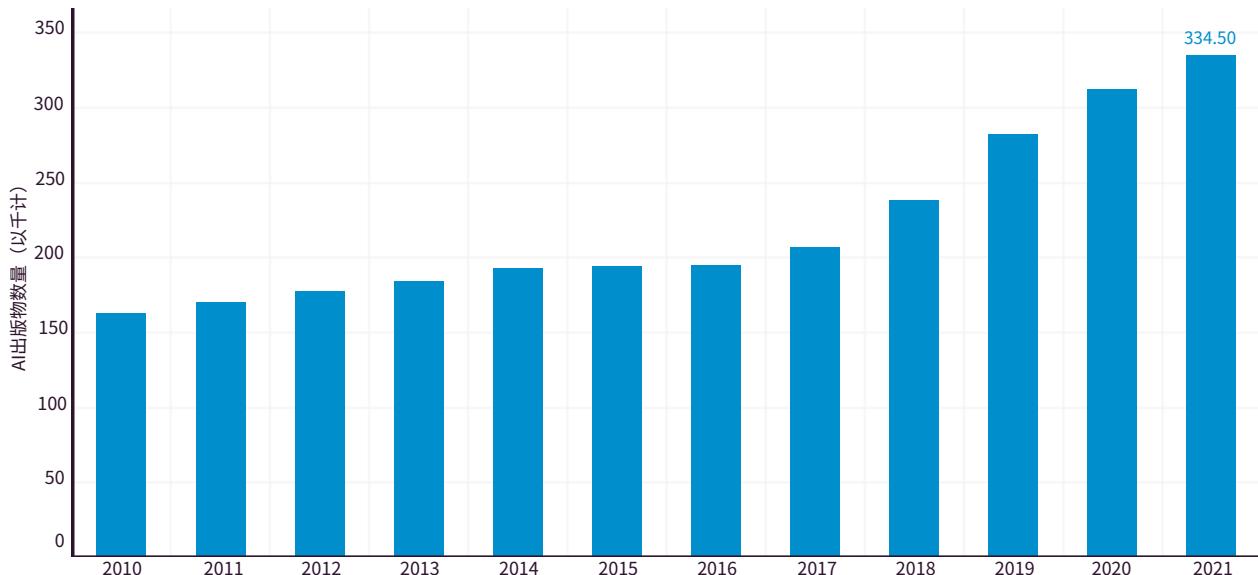


图 1.1.1

<sup>1</sup> 关于CSET方法的更多信息，见附录。由于数据提供者和分类方法的改变，出版物趋势/数据可能与过去的报告不同。更多关于定义人工智能和正确捕捉相关文献计量数据的挑战，见人工智能指数团队在“[人工智能政策的测量：机遇与挑战](#)”中的讨论。

<sup>2</sup> 本节中2021年的人工智能出版物数量可能低于实际数量，原因是上述数据库对出版物元数据的收集存在滞后性。



## 按出版物类型划分

图1.1.2展示了全球发布的人工智能出版物的类型随时间变化的情况。2021年，在所有发表的人工智能出版物中，51.5%是期刊文章，21.5%是会议论文，17.0%为文献库。

其余10.1%的出版物包括书籍、书籍章节、论文和未知类型文件。在过去12年中，期刊和文献库出版物的数量分别增长了2.5倍和30倍，自2018年以来，会议论文的数量却有所下降。

2010-21年按类型划分的人工智能出版物的数量

来源：安全与新兴技术中心（CSET），2021 | 图：2022人工智能指数报告

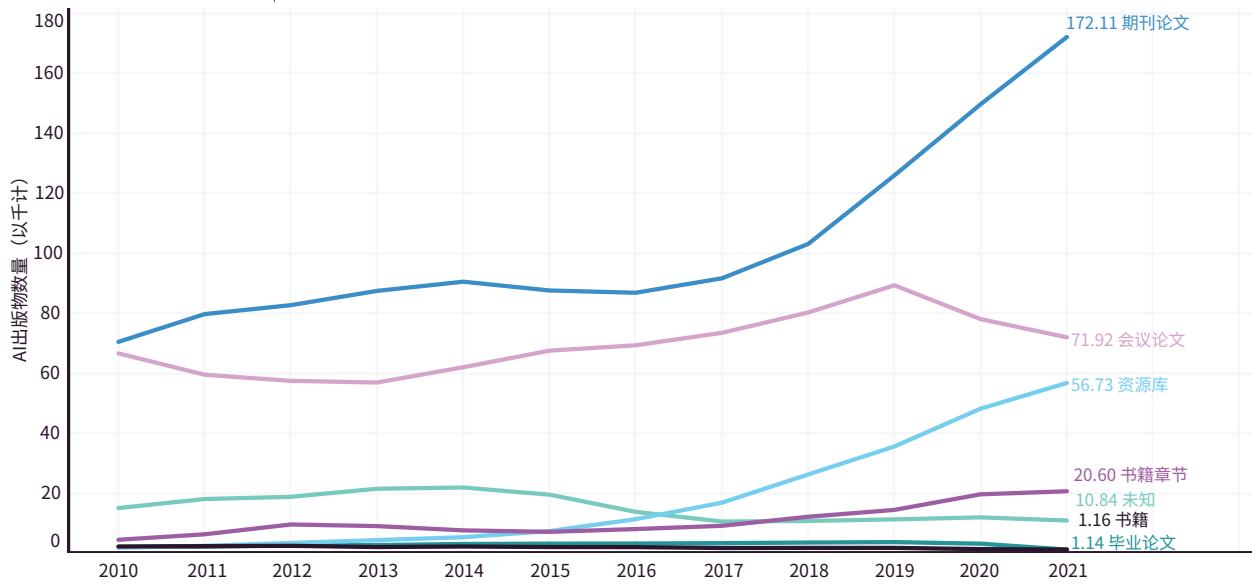


图 1.1.2



## 按研究领域划分

图1.1.3显示，自2015年以来，模式识别和机器学习的出版物数量增加了一倍多。如计算机视觉、数据挖掘

和自然语言处理等其他受深度学习影响较大的领域增幅较小。

2010-21年按研究领域划分的人工智能出版物数量（不包括其他人工智能）

来源：安全与新兴技术中心（CSET），2021 | 图：2022人工智能指数报告

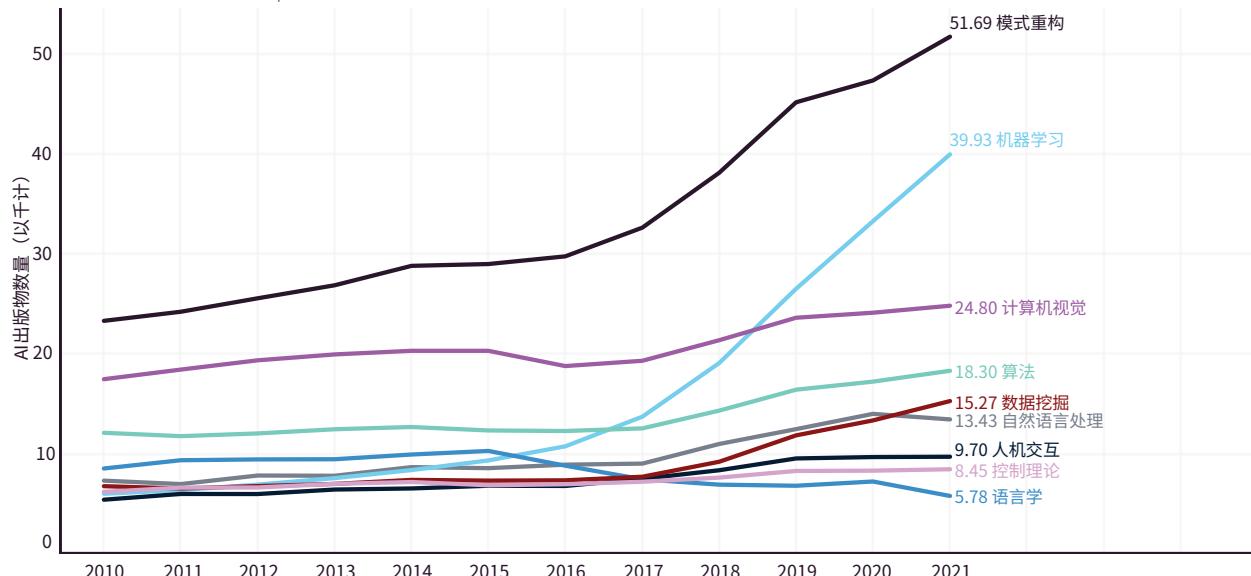


图 1.1.3



## 按行业 (Sector) 划分

本节展示了全球 (图1.1.4a)、中国 (图1.1.4b)、美国 (图1.1.4c) 以及欧盟和英国 (图1.1.4d) 与产业界、教育、政府和非营利组织有关的人工智能出版物

数量。<sup>3</sup> 教育行业在每个地区都占主导地位。企业的参与程度在美国最高，然后是欧盟。中国是唯一教育份额一直在上升的地区。

2010-21年按行业划分的人工智能出版物（占总数的%）情况

来源：安全与新兴技术中心 (CSET)，2021 | 图：2022人工智能指数报告

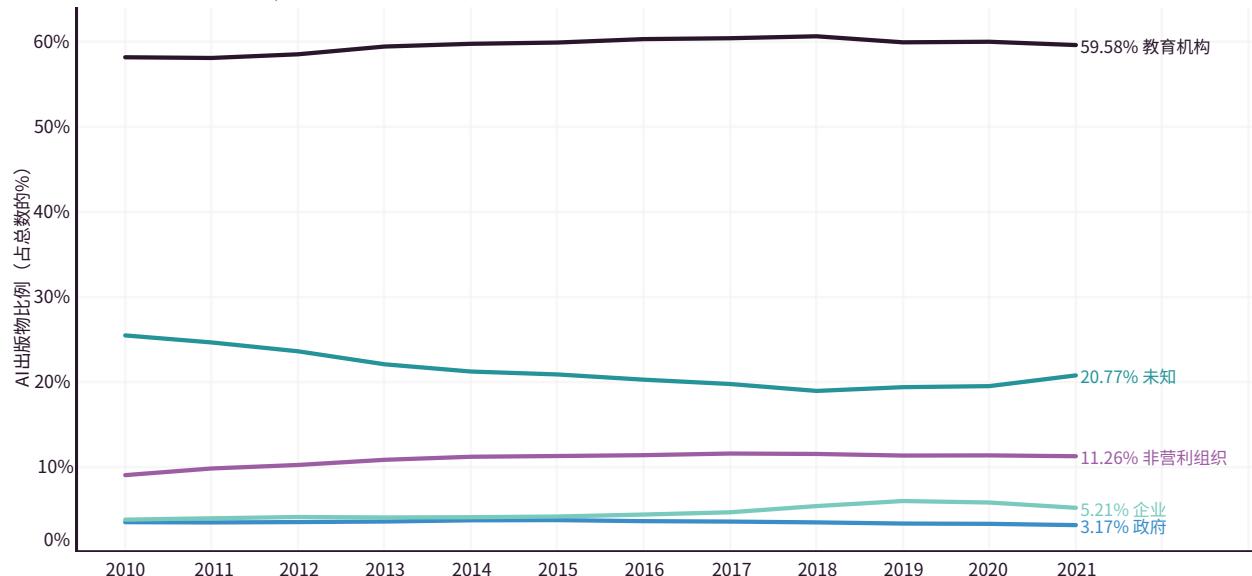


图 1.1.4a

2010-21年按行业划分的美国人工智能出版物（占总数的%）情况

来源：安全与新兴技术中心 (CSET)，2021 | 图：2022人工智能指数报告

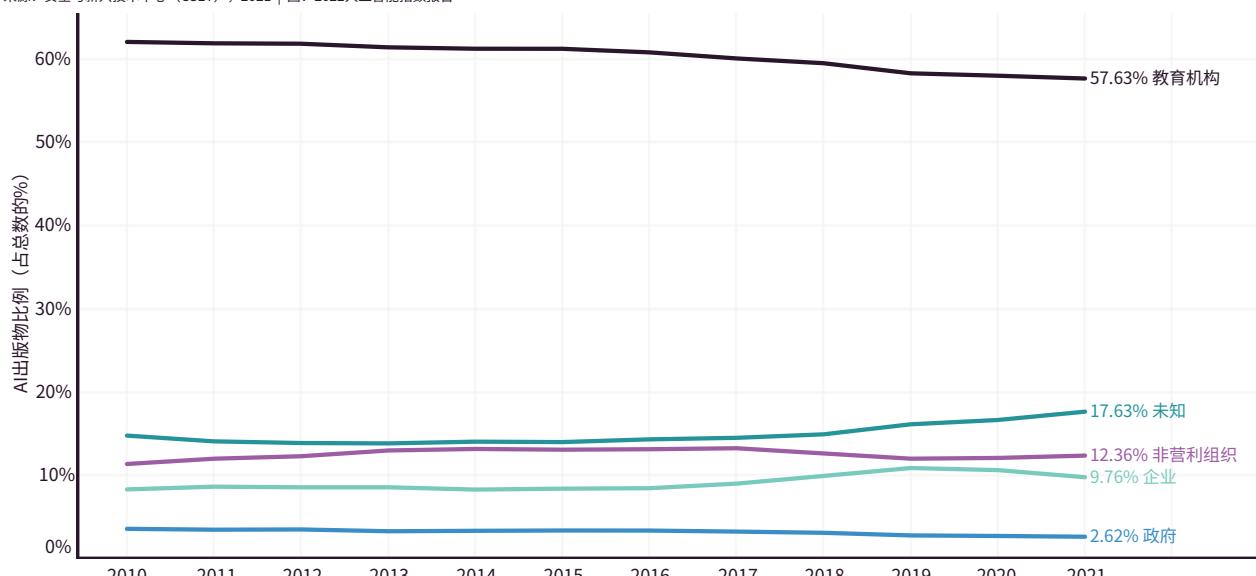


图 1.1.4b

<sup>3</sup> 该分类法是根据全球研究标识符数据库 (GRID) 改编所得。请参阅每个类别的定义。医疗保健（包括医院和设施）在此被列为非营利性机构，而将与国家资助的大学有关的出版物列入教育行业。



### 2010-21年按行业划分的中国人工智能出版物（占总数的%）情况

来源：安全与新兴技术中心（CSET），2021 | 图：2022人工智能指数报告

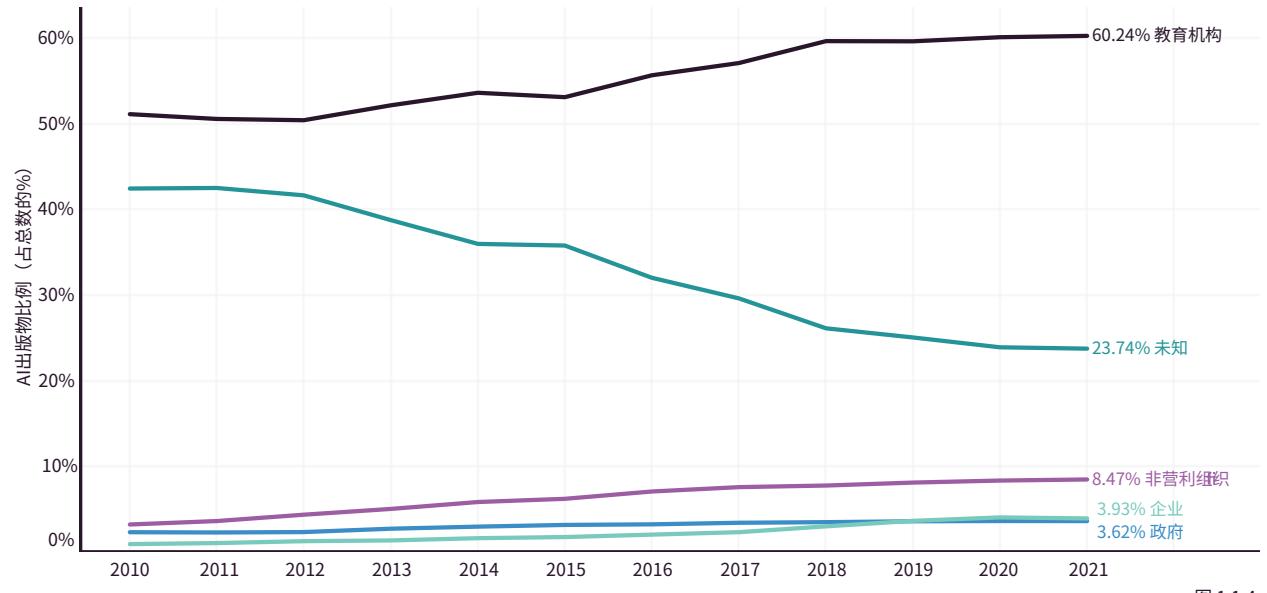


图 1.1.4c

### 2010-21年按行业划分的欧盟和英国的人工智能出版物（占总数的%）情况

来源：安全与新兴技术中心（CSET），2021 | 图：2022人工智能指数报告

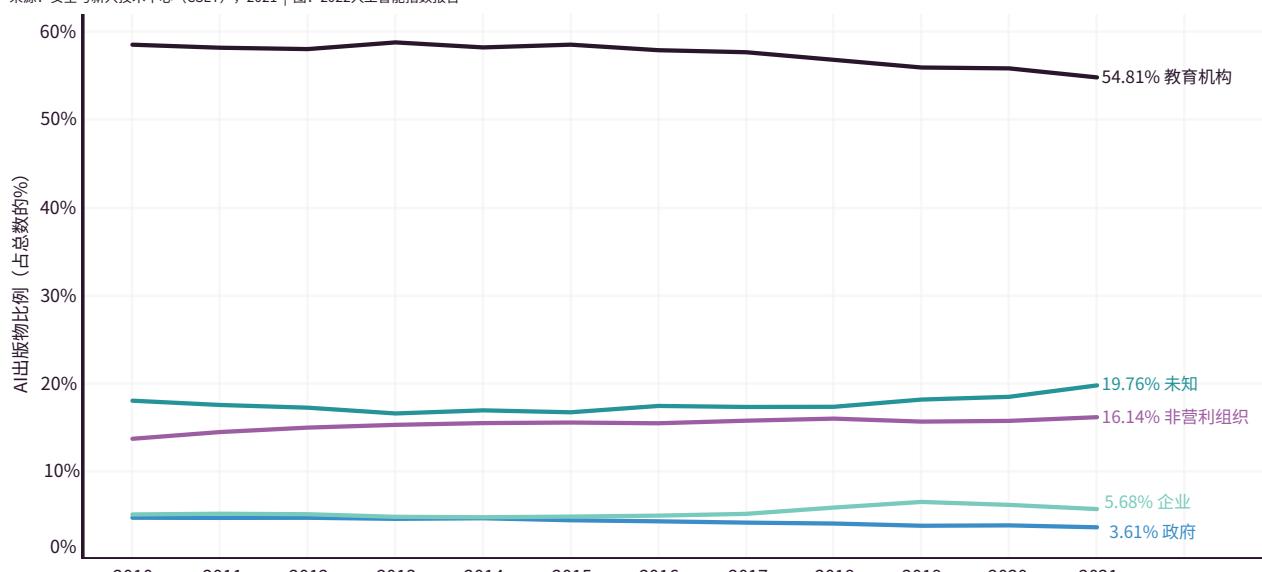


图 1.1.4d



## 跨国合作

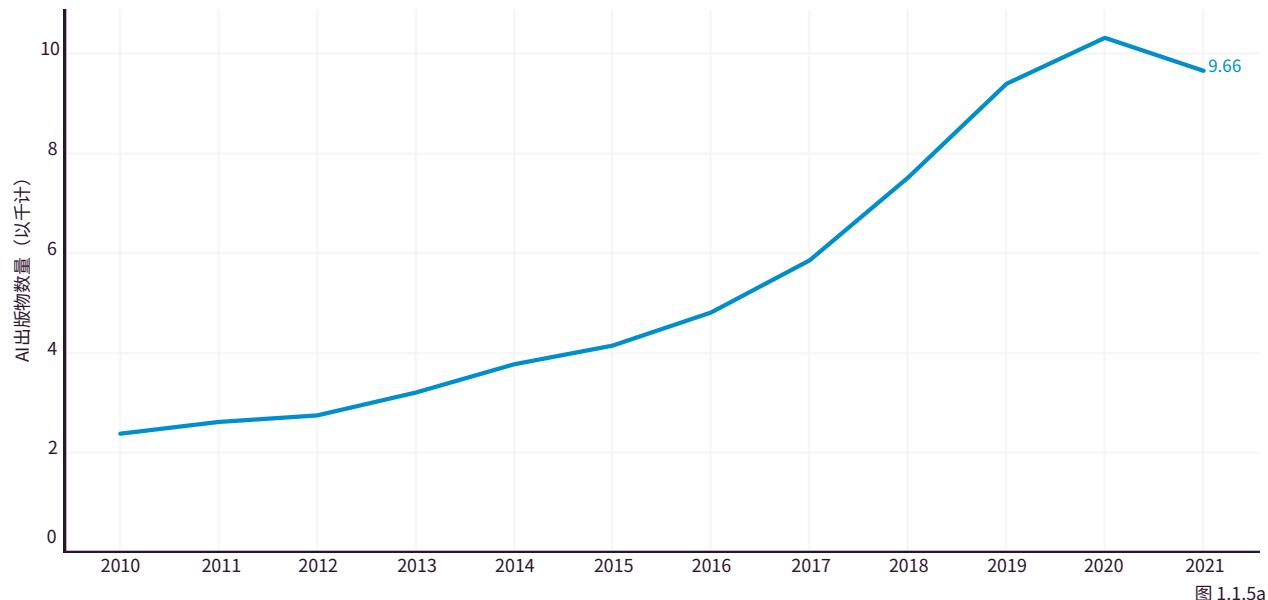
学者、研究人员、行业专家和其他人员之间的跨国合作是现代STEM发展的一个关键组成部分，可以加速新思想的传播和研究团队的成长。图1.1.5a和图1.1.5b展示了2010年至2021年全球顶级跨国AI合作的情况。CSET将跨国合作计算为每份出版物的不同作者对（例如，将在一份出版物上的四位美国和四位中国作者计算为一次美中合作，而将同一组作者之间的两份出版物计算为两次合作）。

到目前为止，在过去12年中，数量最多的合作发生在美国和中国之间，自2010年以来增加了5倍。其次是英国与美国和中国之间的合作，自2010年以来增加了3倍多。2021年，美国和中国之间的合作数量是英国和中国合作数量的2.7倍。

到目前为止，在过去12年中，数量最多的合作发生在美国和中国之间，自2010年以来增加了五倍。

### 2010-21年 美国和中国在人工智能出版物方面的合作情况

来源：安全与新兴技术中心 (CSET) , 2021 | 图：2022人工智能指数报告





### 2010-21年人工智能出版物的跨国合作情况 (不包括美国和中国)

来源: 安全与新兴技术中心 (CSET) , 2021 | 图: 2022人工智能指数报告

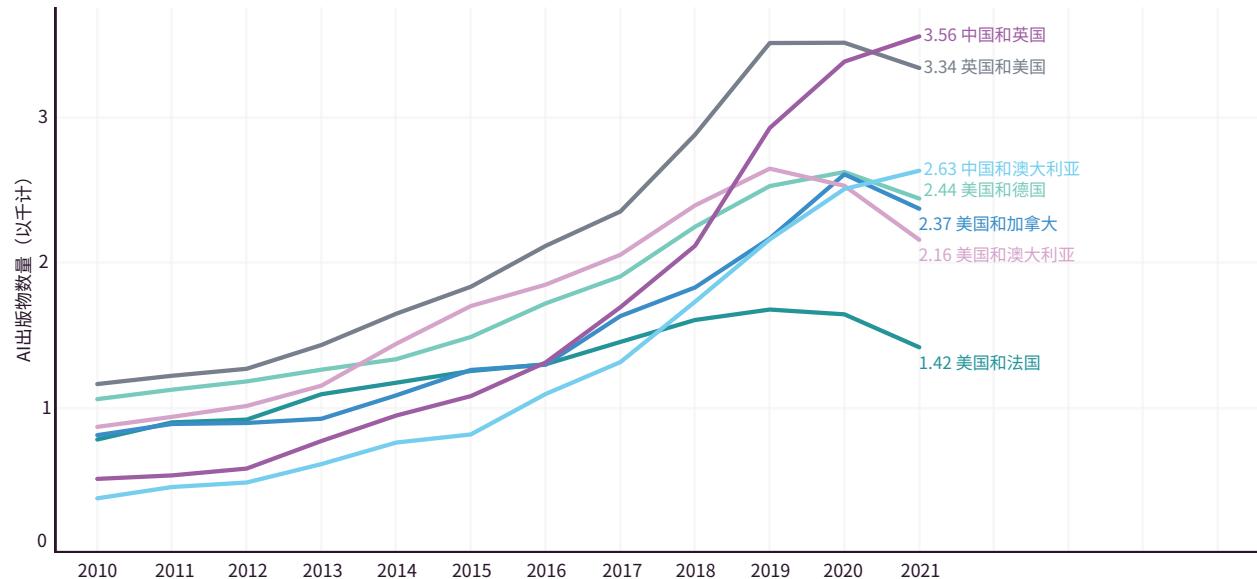


图 1.1.5b

### 跨行业合作

在大学之外不断增加的人工智能研究，促进了大学与其他行业之间的合作。图1.1.6显示，2021年，教育机构和非营利组织之间的合作次数最多(29,839)，其次是企业

和教育机构(11,576)以及政府和教育机构(8,087)。2021年，教育机构和非营利组织之间的合作是教育机构和公司之间合作次数的2.5倍。

### 2010-21年人工智能出版物的跨领域合作情况

来源: 安全与新兴技术中心 (CSET) , 2021 | 图: 2022人工智能指数报告

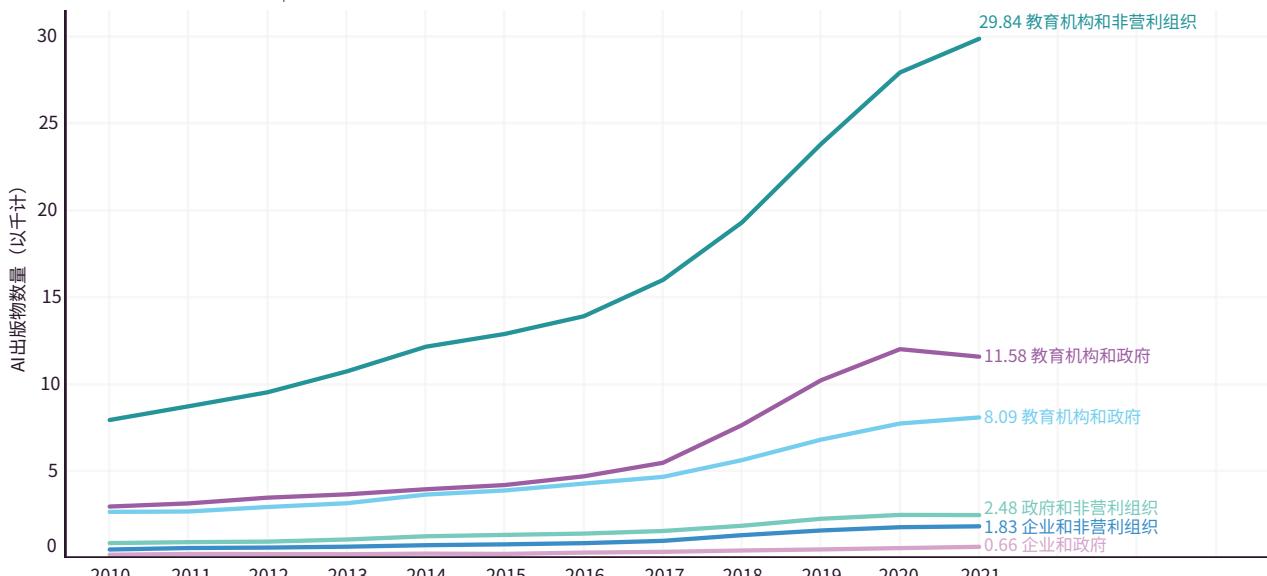


图 1.1.6



## 人工智能期刊出版物

### 概述

在2010年至2015年仅有小幅增长之后，自2015年以来

来，人工智能期刊出版物数量增长了近2.5倍（图1.1.7）。如图1.1.8所示，从期刊出版物数量所占百分比角度分析，2021年人工智能期刊出版物约占所有期刊出版物的2.53%，而2010年为1.3%。

2010-21年人工智能期刊发表数量

来源：安全与新兴技术中心（CSET），2021 | 图：2022人工智能指数报告

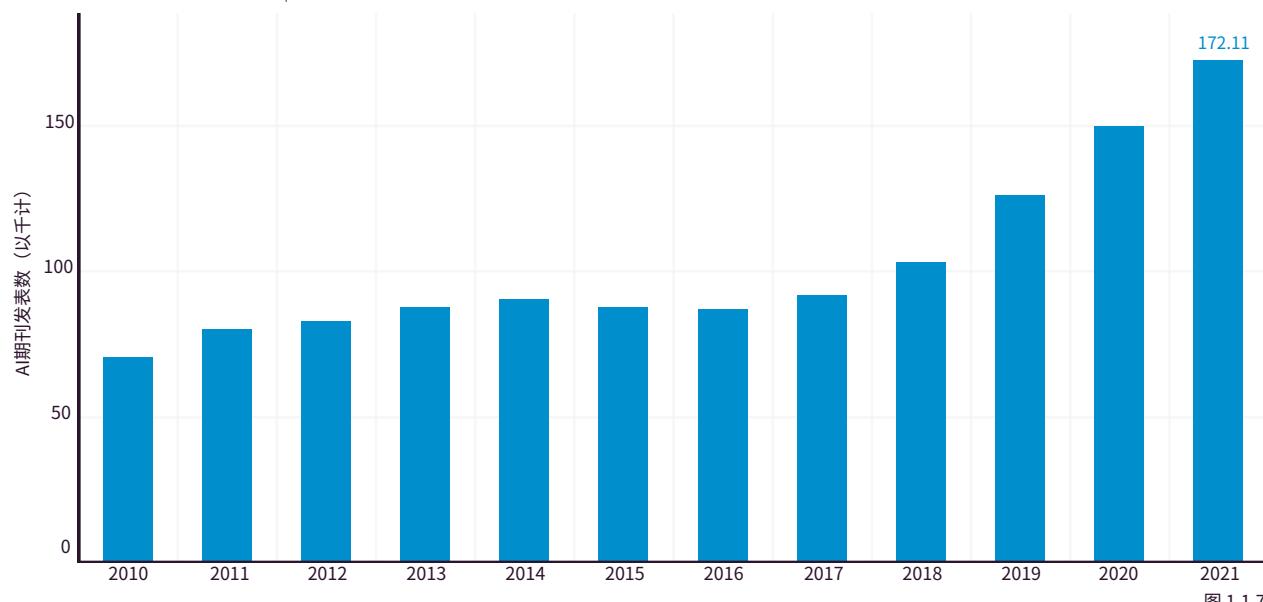


图 1.1.7

2010-21年人工智能期刊出版量（占期刊总出版量的%）

来源：安全与新兴技术中心（CSET），2021 | 图：2022人工智能指数报告

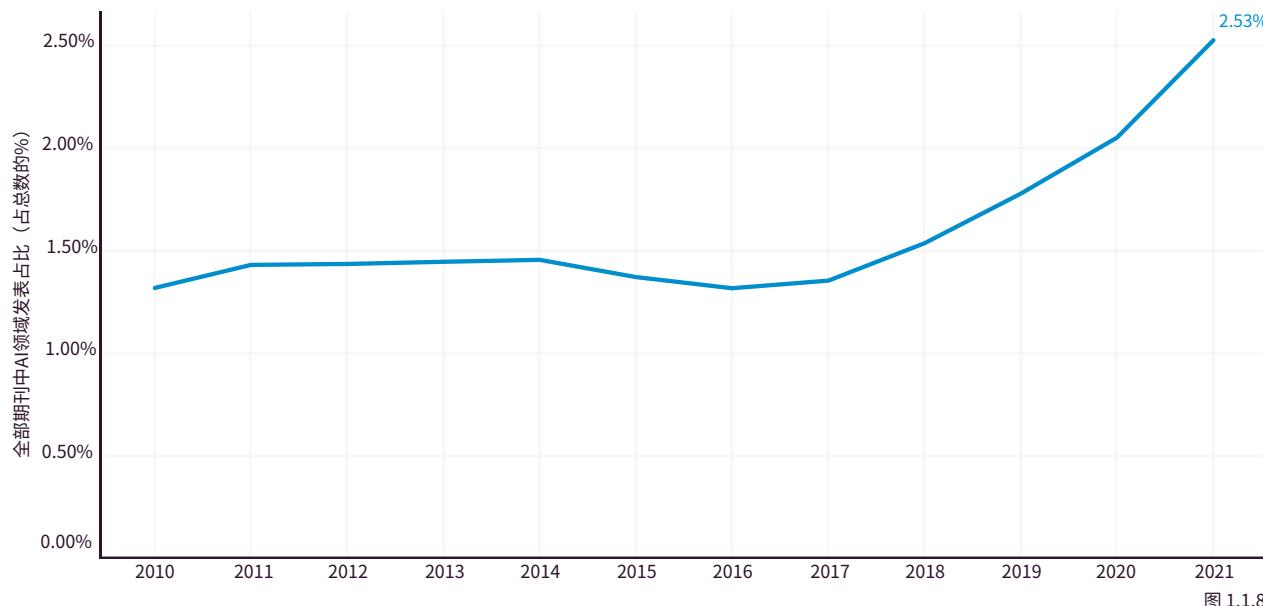


图 1.1.8



## 按地区（Region）划分<sup>4</sup>

图1.1.9展示了2010年至2021年期间各地区人工智能期刊出版物的份额。2021年，东亚和太平洋地区以42.9%领先，其次是欧洲和中亚（22.7%）、北美洲（15.6%）。此外，南亚和中东及北非的增长最为显著，它们的人工智能期刊出版物数量在过去12年中分别增长了约12倍和7倍。

此外，南亚和中东及北非的增长最为显著，它们的人工智能期刊出版物数量在过去12年中分别增长了约12倍和7倍。

2010-21年按地区（Region）划分人工智能期刊出版量（占世界总量的%）情况

来源：安全与新兴技术中心（CSET），2021 | 图：2022人工智能指数报告

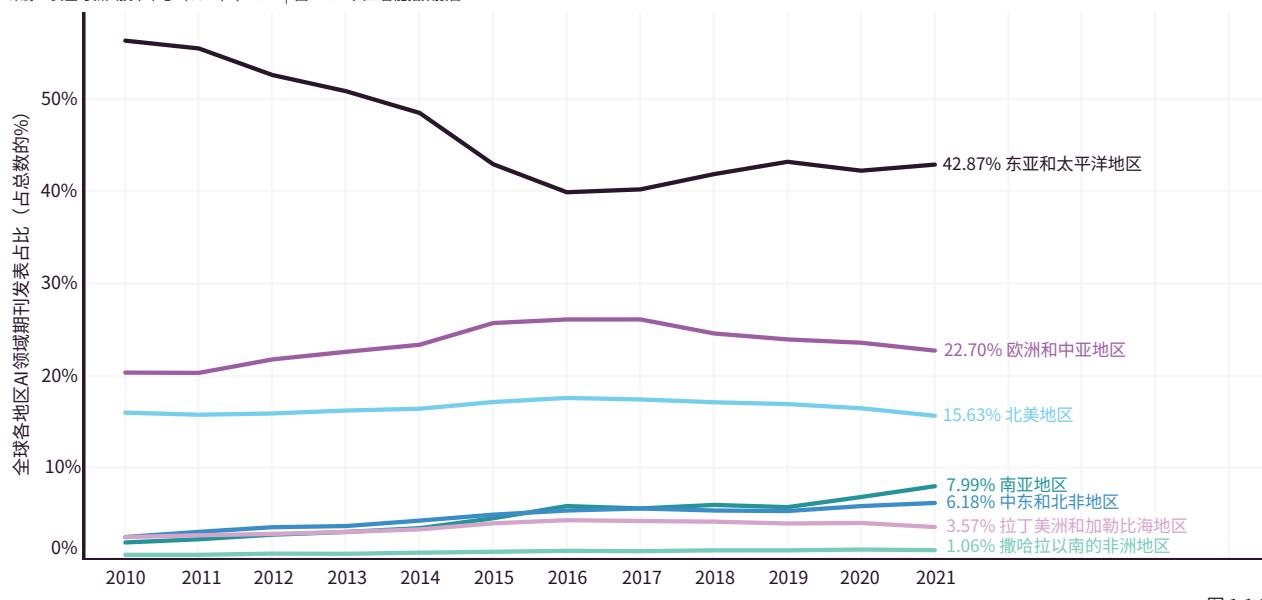


图 1.1.9

<sup>4</sup> 本章中的区域是根据世界银行的分析分组进行分类的。



## 按地理区域 (Geographical Area) 划分<sup>5</sup>

图1.1.10展示了三个主要人工智能大国过去12年的人工智能期刊出版物的份额。中国一直保持着领先地位，

2021年为31.0%，其次是欧盟和英国，为19.1%，美国为13.7%。

2010-21年按地理区域 (geographical area) 划分的人工智能期刊出版量 (占世界总量的%)

来源：安全与新兴技术中心 (CSET)，2021 | 图：2022人工智能指数报告

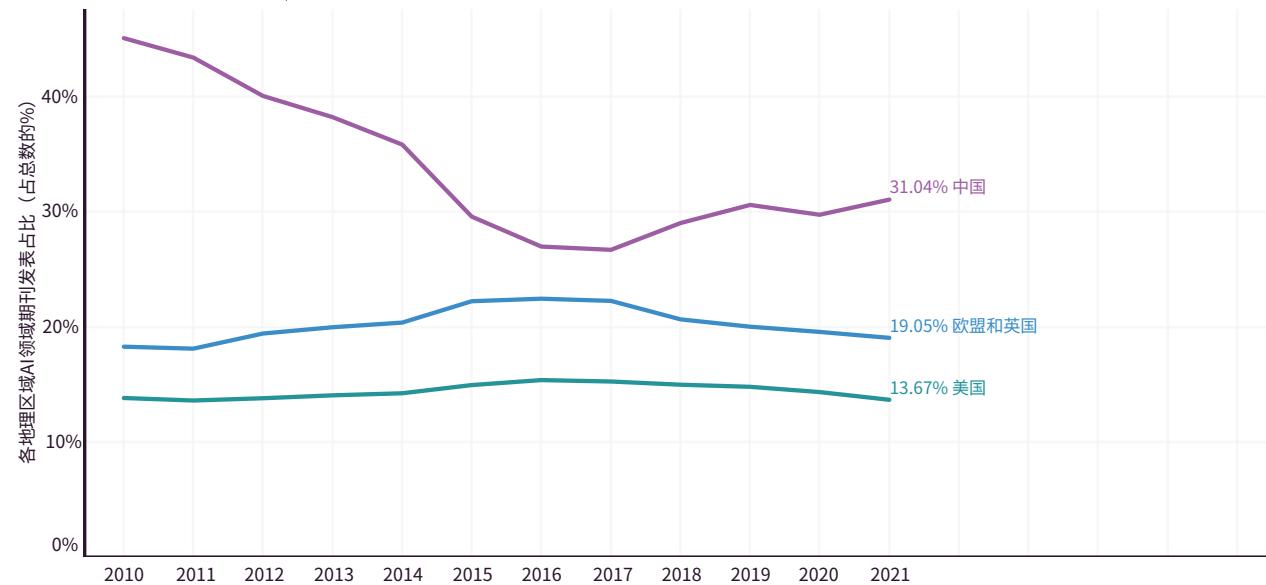


图 1.1.10

<sup>5</sup> 本章中的地理区域结合了欧盟和英国之间的出版物数量，以反映它们之间在研究合作方面的历史上的紧密联系。



## 引用

在人工智能期刊出版物的引用数量上，中国的份额逐渐增加，而欧盟加上英国和美国的份额则有所下降。这三个地理区域的引用总量占全球总引次数的66%以上。

这三个地理区域的引用总量  
占全球总引用次数的66%以  
上。

2010-21年按地理区域 (geographical area) 划分的人工智能期刊引用情况 (占世界总数的%)

来源：安全与新兴技术中心 (CSET)，2021 | 图：2022人工智能指数报告

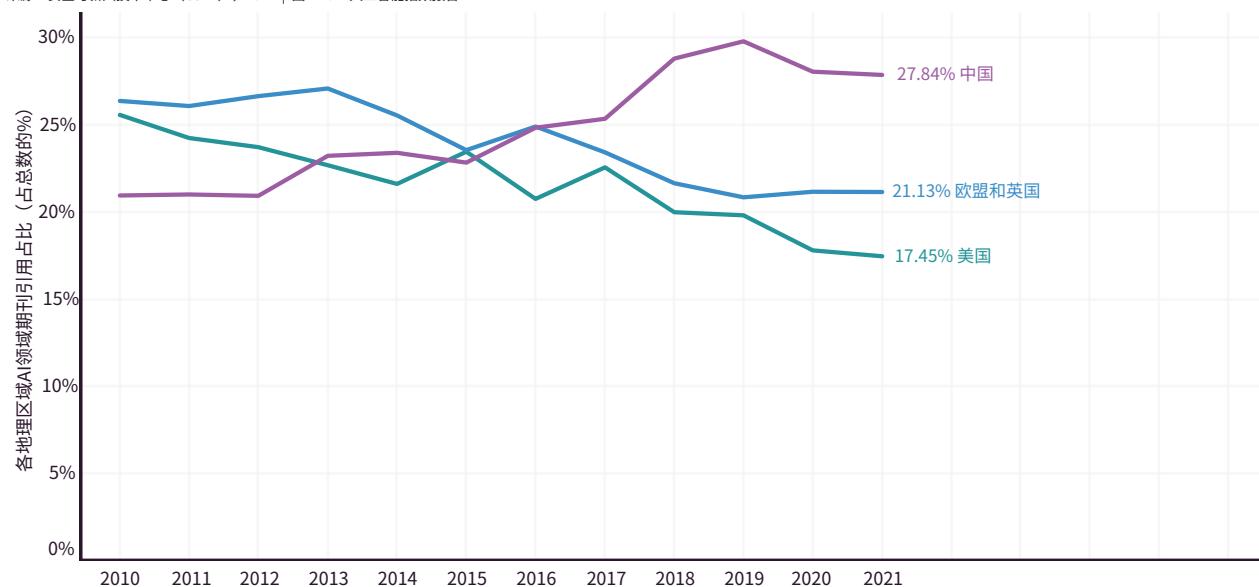


图 1.1.11



## 人工智能会议出版物

### 概述

人工智能会议出版物的数量在2019年达到峰值。相比

之下，2021年的峰值下降了约19.4%（图1.1.12）。尽管总数有所下降，但自2010年以来，人工智能会议出版物在世界会议出版物总数中的份额增加了五个百分点以上（图1.1.13）。

2010-21年人工智能会议出版物数量

来源：安全与新兴技术中心 (CSET) , 2021 | 图：2022人工智能指数报告

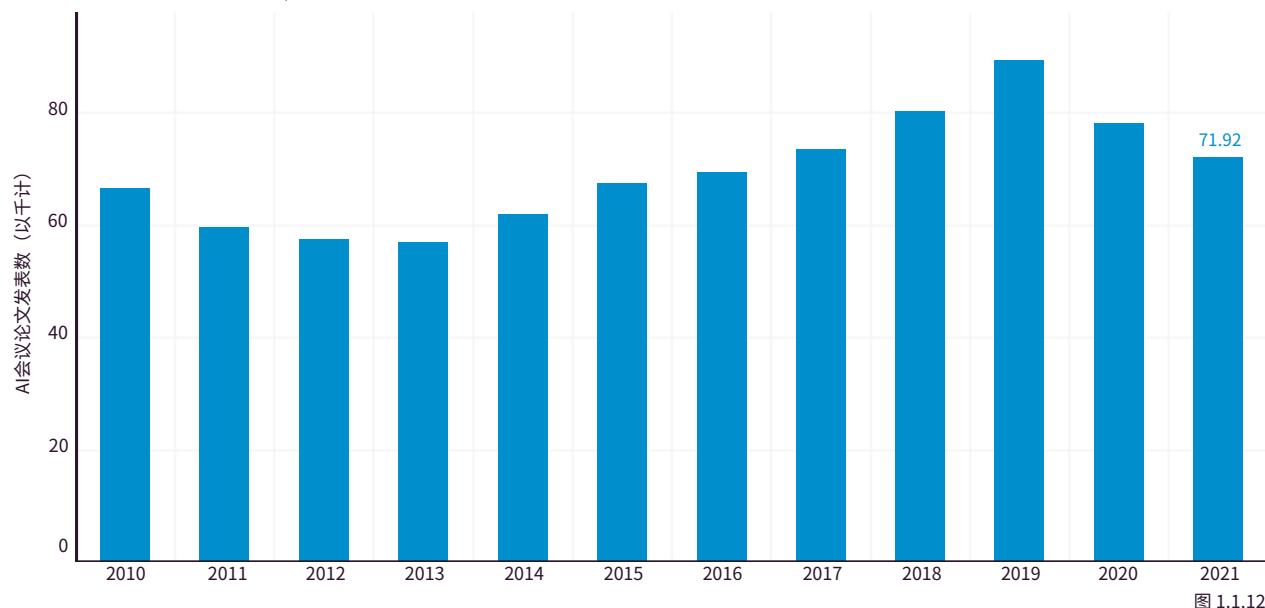


图 1.1.12

2010-21年人工智能会议出版物（占会议出版物总数的%）

来源：安全与新兴技术中心 (CSET) , 2021 | 图：2022人工智能指数报告

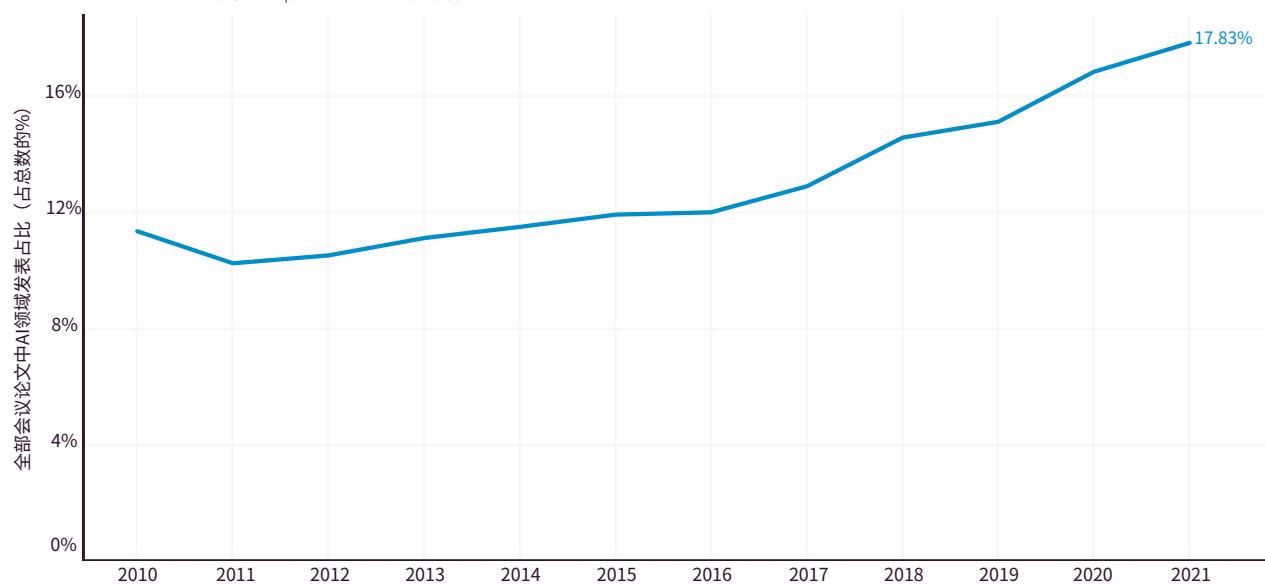


图 1.1.13



## 按地区（Region）划分

图1.1.14显示了各地区人工智能会议出版物的数量。与人工智能期刊出版物的趋势类似，东亚和太平洋地区、欧洲和中亚地区以及北美地区的人工智能会议出版物数量居世界前列。具体而言，东亚和太平洋地区

所占份额自2014年领先以来持续上升，2021年达到了40.4%。其次是欧洲和中亚（23.0%）、北美（19.0%）。在过去的12年里，南亚的人工智能会议出版物的比例呈现明显的上升趋势，从2010年的4.0%上升到2021年的10.4%。

2010-21年各地区人工智能会议出版物（占世界总数的%）情况

来源：安全与新兴技术中心（CSET），2021 | 图：2022人工智能指数报告

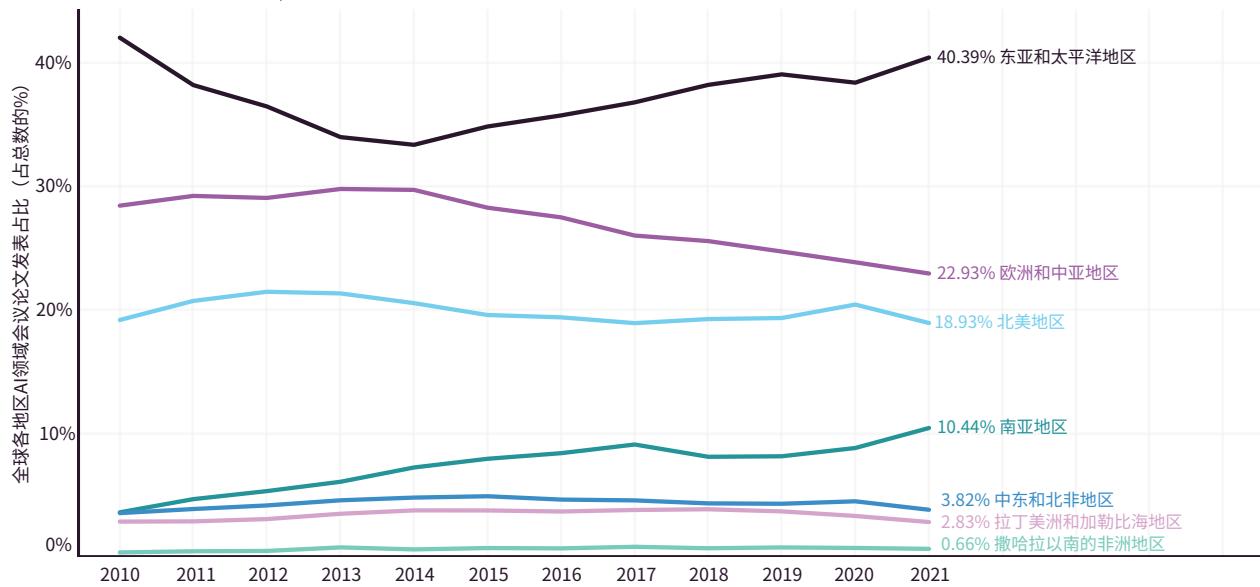


图 1.1.14



## 按地理区域（Geographical Area）划分

2021年，中国在全球人工智能会议出版物中的份额最大，为27.6%，概念份额优势更甚于2020年的领先

状态。欧盟和英国紧随其后，为19.0%，美国位居第三，为16.9%（图1.1.15）。

2010-21年按地理区域（geographical area）划分的人工智能会议出版物（占世界总数的%）情况

来源：安全与新兴技术中心（CSET），2021 | 图：2022人工智能指数报告

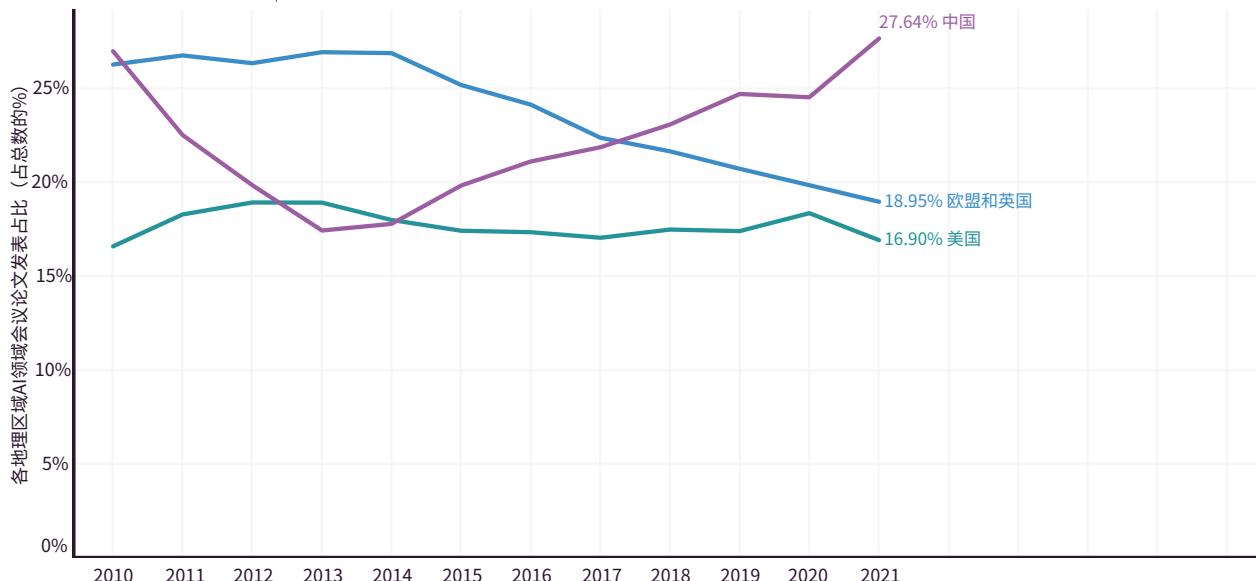


图 1.1.15



## 引用

尽管中国在2021年发表的人工智能会议出版物数量最多，但图1.1.16显示，美国在人工智能会议引用次

数方面在主要大国中处于领先地位。其在2021年达到29.5%，其次是欧盟和英国（23.3%）、中国（15.3%）。

2010-21年按地理区域（geographical area）划分的人工智能会议论文（占世界总数的%）情况

来源：安全与新兴技术中心（CSET），2021 | 图：2022人工智能指数报告

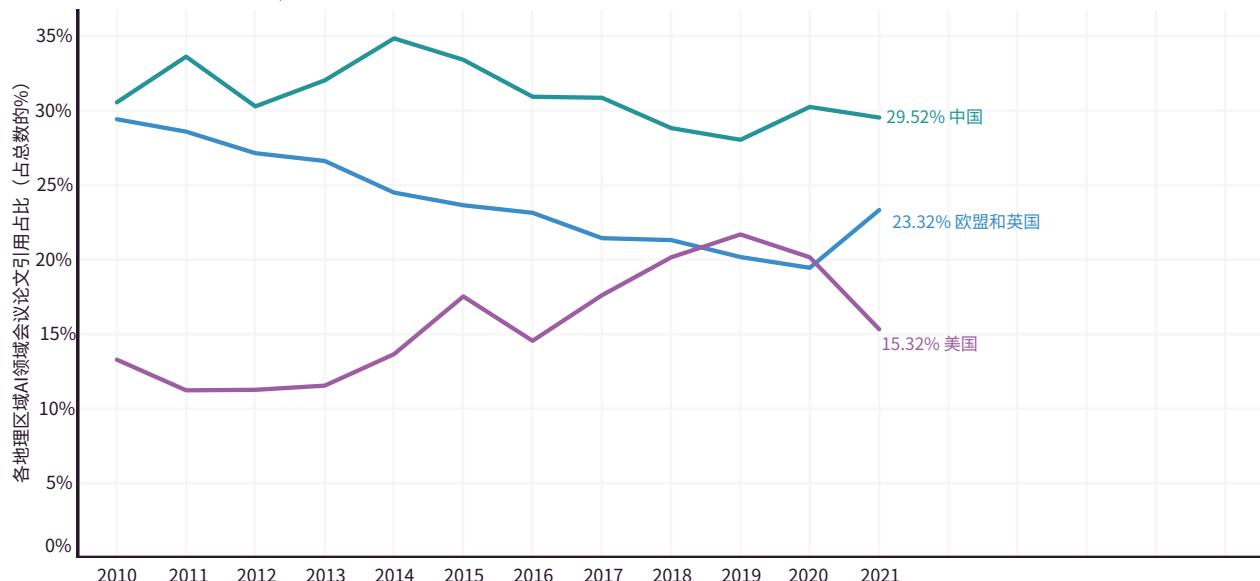


图 1.1.16



## 人工智能文献库

### 概述

在电子预印本库（如arXiv和SSRN）上发表预审论文，已经成为人工智能研究人员在传统出版途径之外传播其

工作的一种流行方式。这些文献库允许研究人员在向期刊和会议提交论文之前分享他们的研究成果，这大大加快了信息发现的周期。在过去的12年里，人工智能文献库出版物数量增长了近30倍（图1.1.17），现在占所有文献库出版物的15.3%（图1.1.18）。

2010-21年人工智能文献库出版物数量

来源：安全与新兴技术中心（CSET），2021 | 图：2022人工智能指数报告

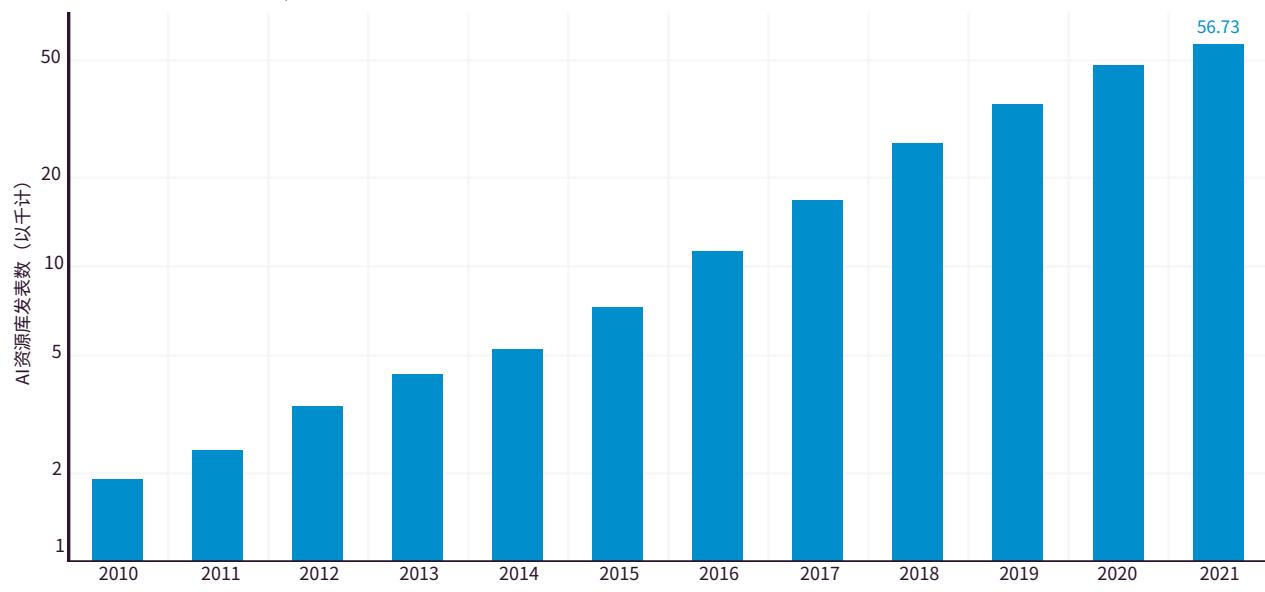


图 1.1.17

2010-21年人工智能文献库出版物（占文献库出版物总数的%）情况

来源：安全与新兴技术中心（CSET），2021 | 图：2022人工智能指数报告

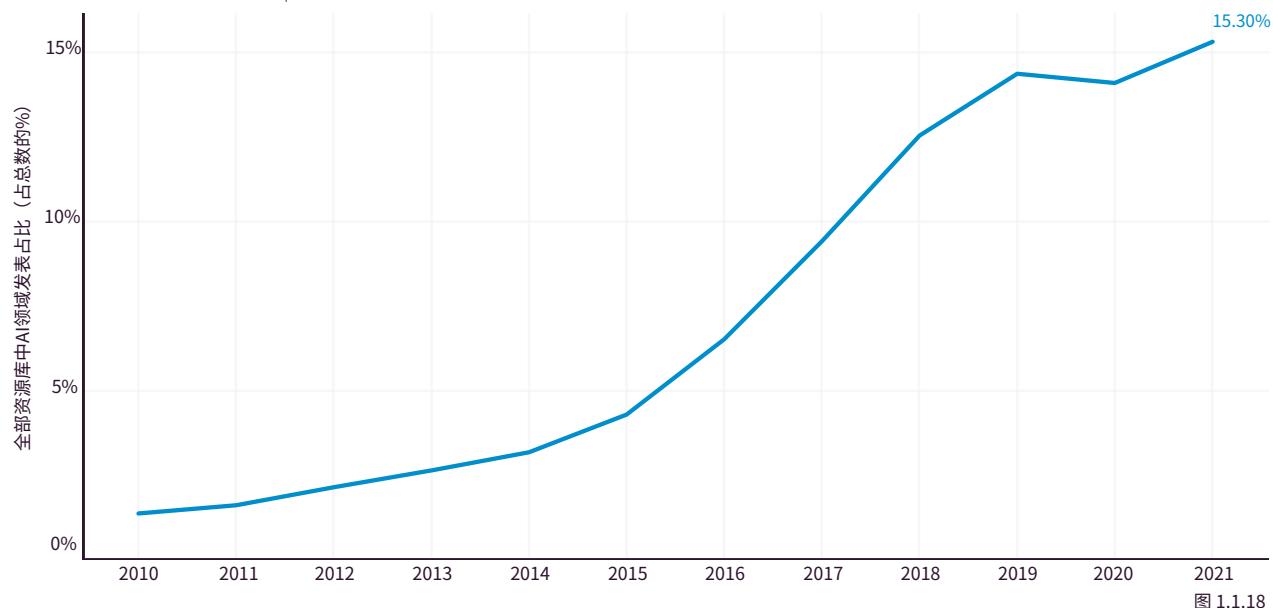


图 1.1.18



## 按地区（Region）划分

图1.1.19中显示，按地区来分析，自2014年以来，北美地区在全球人工智能文献库出版物中的份额一直保持稳

定的领先地位，而欧洲和中亚的份额则有所下降。自2013年以来，东亚和太平洋地区的份额大幅增长。

2010-21年按地区（Region）划分人工智能文献库出版物（占世界总数的%）情况

来源：安全与新兴技术中心（CSET），2021 | 图：2022人工智能指数报告

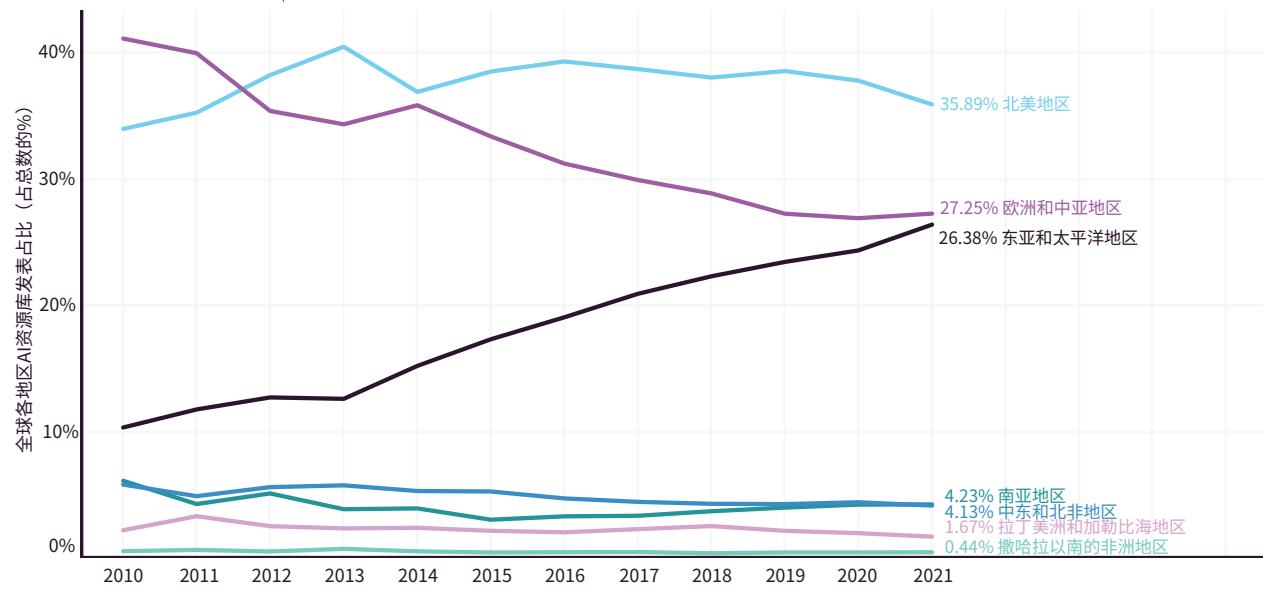


图 1.1.19



## 按地理区域（Geographical Area）划分

自2011年以来，虽然美国在世界人工智能文献库出版物的百分比上一直保持领先，但中国正在追赶，而欧盟和英国的份额继续下降（图1.1.20）。2021年，来

自美国的出版物占全球人工智能文献库出版物总数的32.5%--与期刊和会议出版物相比这一比例更高，其次是欧盟和英国（23.9%）、中国（16.6%）。

2010-21年按地理区域（geographical area）划分的人工智能文献库出版物（占世界总数的%）情况

来源：安全与新兴技术中心（CSET），2021 | 图：2022人工智能指数报告

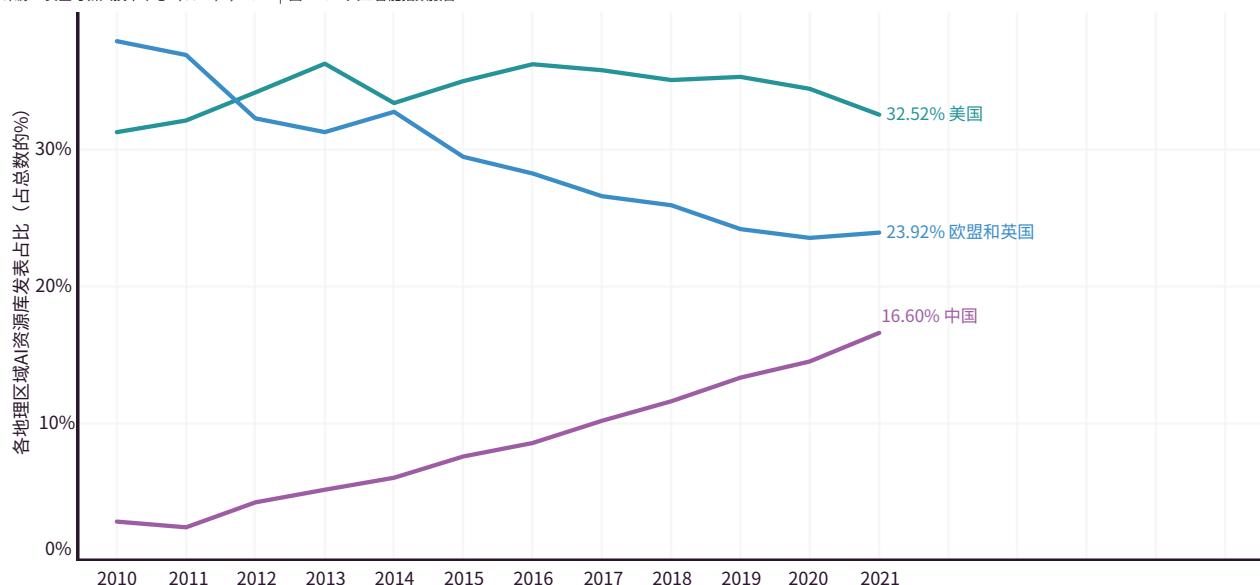


图 1.1.20



## 引用

图1.1.21显示，2021年，美国以38.6%的总引用率位居

榜首，欧盟和英国（20.1%）、中国（16.4%）紧随其后。

2010-21年按地理区域（geographical area）划分的人工智能文献库应用情况（占世界总数%）

来源：安全与新兴技术中心（CSET），2021 | 图：2022人工智能指数报告

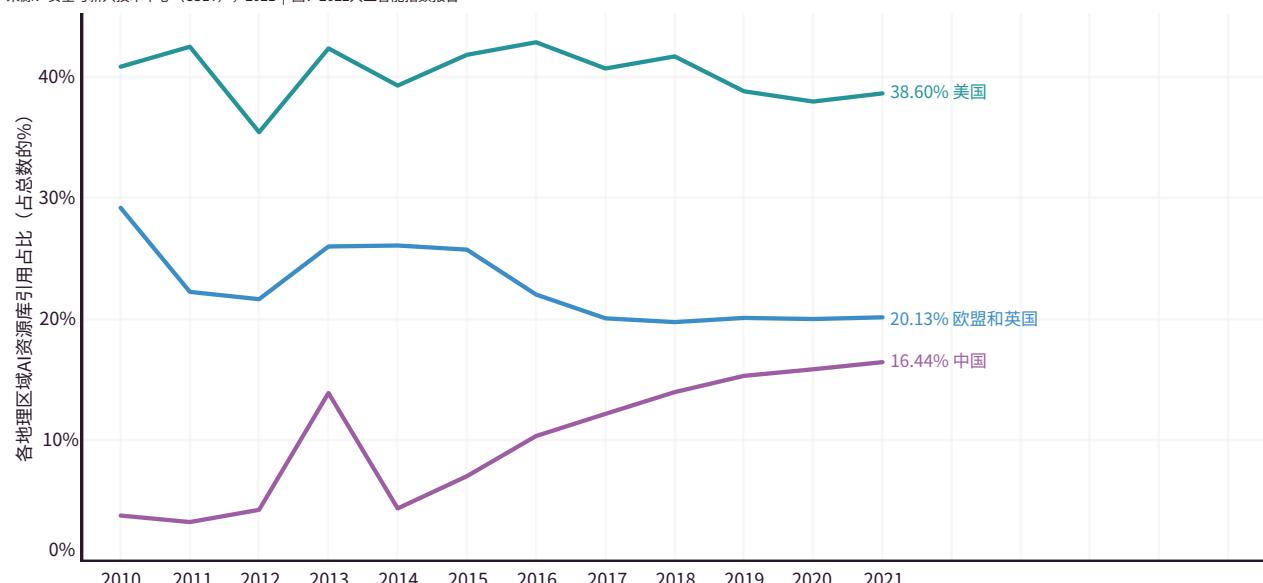


图 1.1.21



## 人工智能专利

本节借鉴了CSET和1790 Analytics关于人工智能发展和应用相关专利的数据--由合作专利分类 (Cooperative Patent Classification, CPC) /国际专利分类 (International Patent Classification, IPC) 代码和关键词表示。专利按国家和年份分组，然后在 "专利族 (patent family)" 层面进行统计，然后通过CSET从一个专利族的最新出版日期中提取年份值。

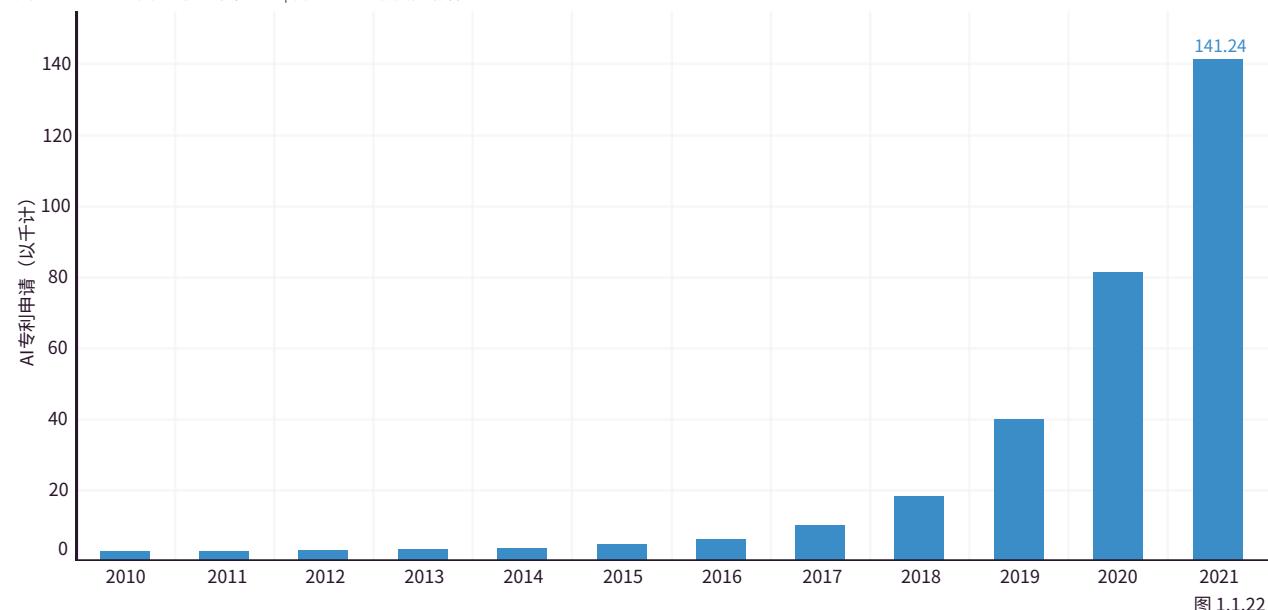
**2021年的专利申请数量是2015年的30多倍，达到了76.9%的复合年增长率。**

### Overview

图1.1.22反映了2010年至2021年人工智能专利申请数量。2021年的专利申请数量是2015年的30多倍，达到76.9%的复合年增长率。

#### 2010-21年人工智能专利申请数量

来源：安全与新兴技术中心 (CSET)，2021 | 图：2022人工智能指数报告





## 按地区和申请状态

图1.1.23a展示了不同地区的人工智能专利申请情况。东亚和太平洋地区的份额在2014年开始激增，2021年以62.1%的专利申请量领先于全球其他地区，其次是北美和欧洲及中亚地区。就这些地区的专利授权

情况而言，北美地区以57.0%领先，其次是东亚和太平洋（31.0%），以及欧洲和中亚（11.3%）（图1.1.23b）。其他地区加起来的授权数量约占全球授权专利总数的1%（图1.1.23c）。

2010-21年按地区 (Region) 划分人工智能专利申请量 (占全球总量的%) 情况

来源：安全与新兴技术中心 (CSET)，2021 | 图：2022人工智能指数报告

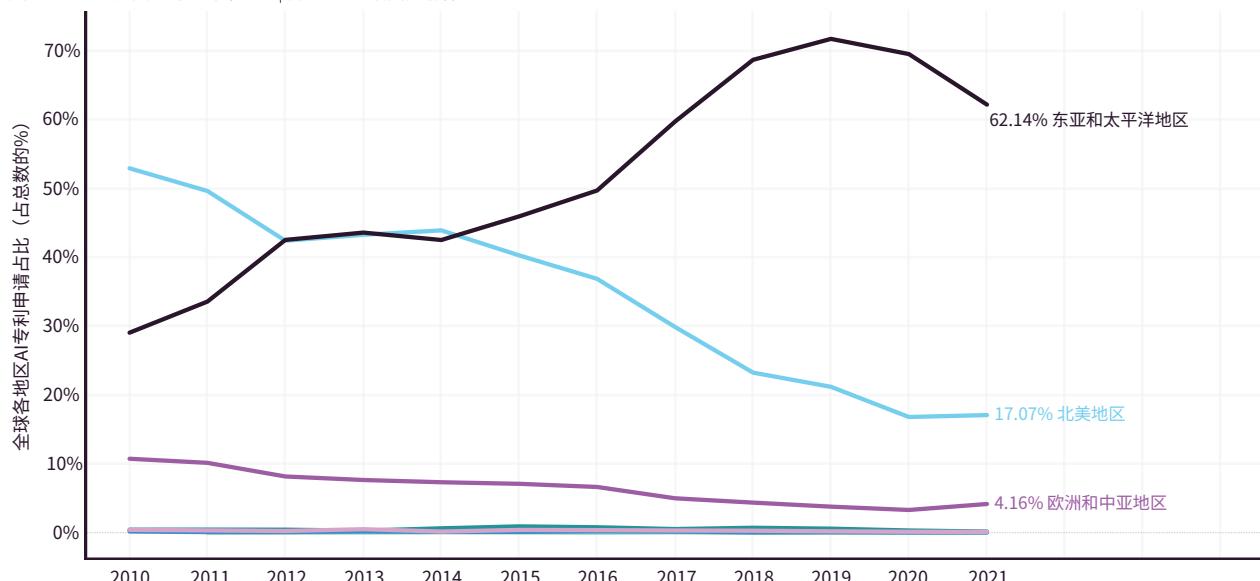


图 1.1.23a



### 2010-21年按地区 (Region) 划分人工智能专利授权 (占世界总量的%) 情况 (第一部分)

来源: 安全与新兴技术中心 (CSET) , 2021 | 图: 2022人工智能指数报告

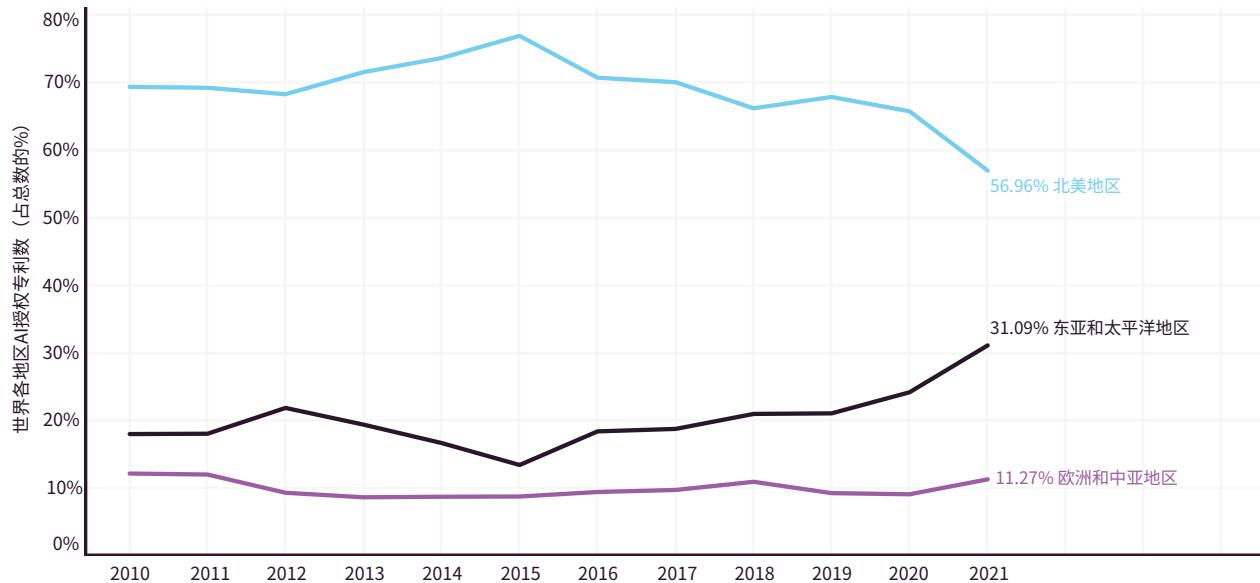


图 1.1.23b

### 2010-21年按地区 (Region) 划分人工智能专利授权 (占世界总量的%) 情况 (第二部分)

来源: 安全与新兴技术中心 (CSET) , 2021 | 图: 2022人工智能指数报告

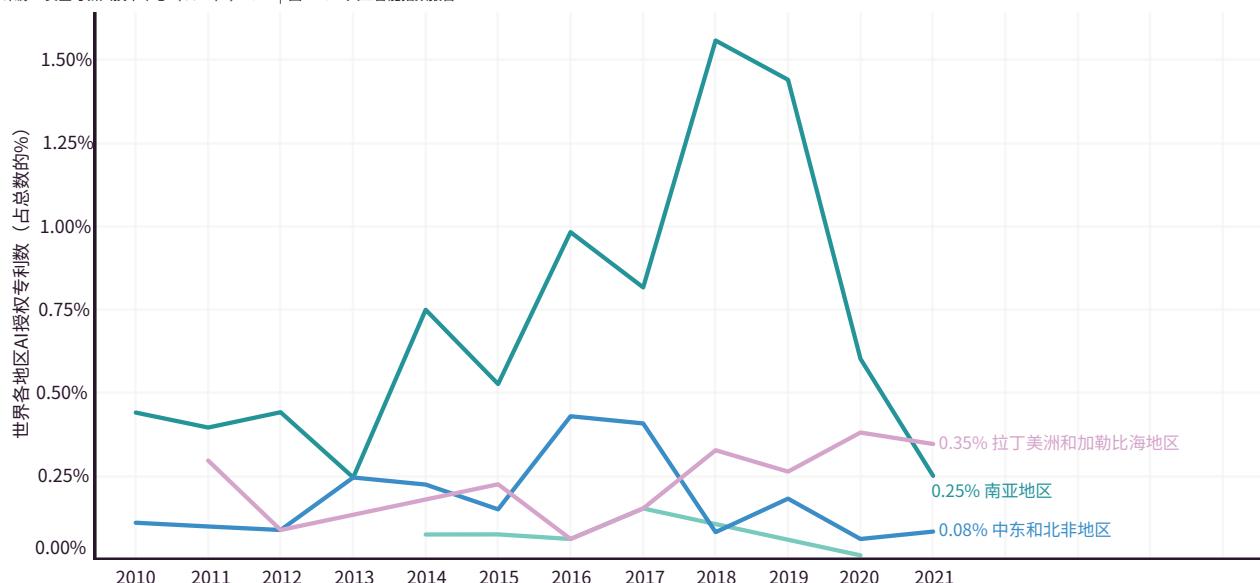


图 1.1.23c



## 按地理区域和申请状态

基于地域分析人工智能专利数据所揭示的趋势，同样能在以地理区域的分析中得到印证（图1.1.24a和图1.1.24b）。目前中国申请的人工智能专利数量占全球总数的一半以上，其中，约6%已获授权，与欧盟和英

国的情况差不多。美国申请了几乎所有的北美地区的专利，占比约是中国的三分之一。图1.1.24c显示，与越来越多的人工智能专利申请和授权相比，中国的专利申请数量（2021年为87,343件）远远大于授权数量（2021年为1,407件）。

2010-21年按地理区域 (geographical area) 划分的人工智能专利申请量 (全球界总量的%)

来源：安全与新兴技术中心 (CSET)，2021 | 图：2022人工智能指数报告

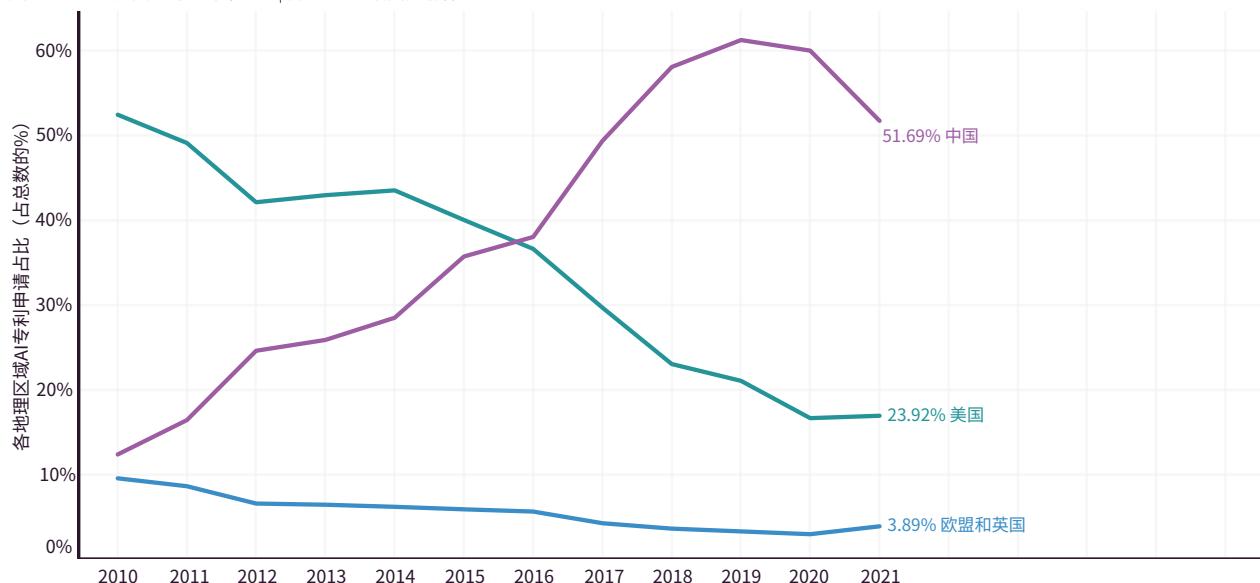


图 1.1.24a

2010-21年按地理区域 (geographical area) 划分的已授权人工智能专利 (占全球总数的%) 情况

来源：安全与新兴技术中心 (CSET)，2021 | 图：2022人工智能指数报告

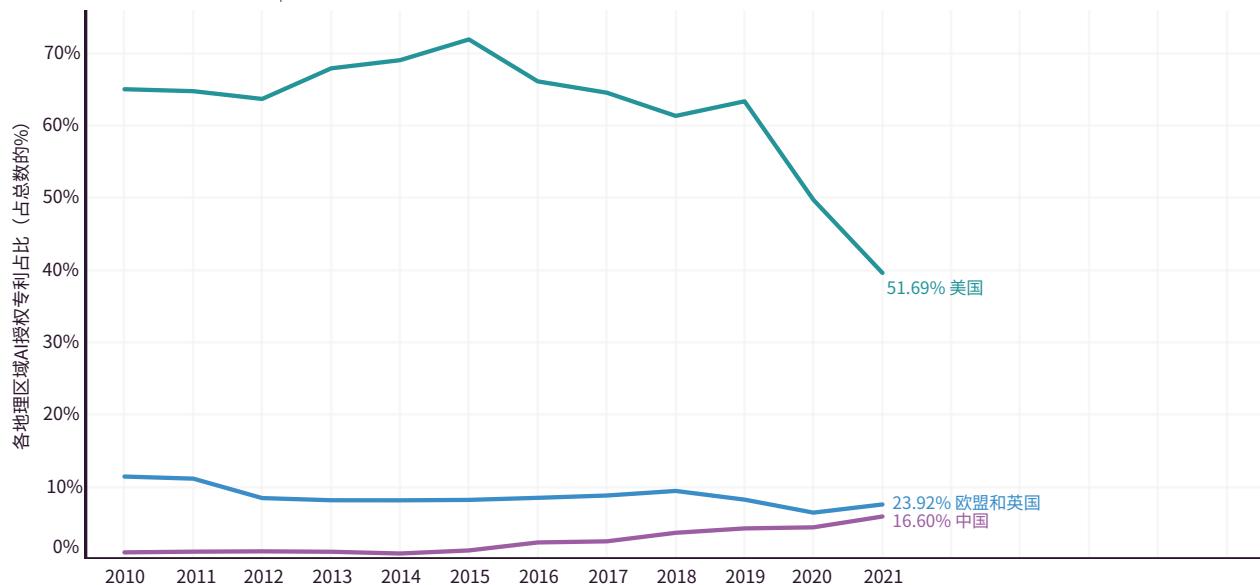


图 1.1.24b



### 2010-21年按申请状态和地理区域划分的人工智能专利情况

来源：安全与新兴技术中心（CSET），2021 | 图：2022人工智能指数报告

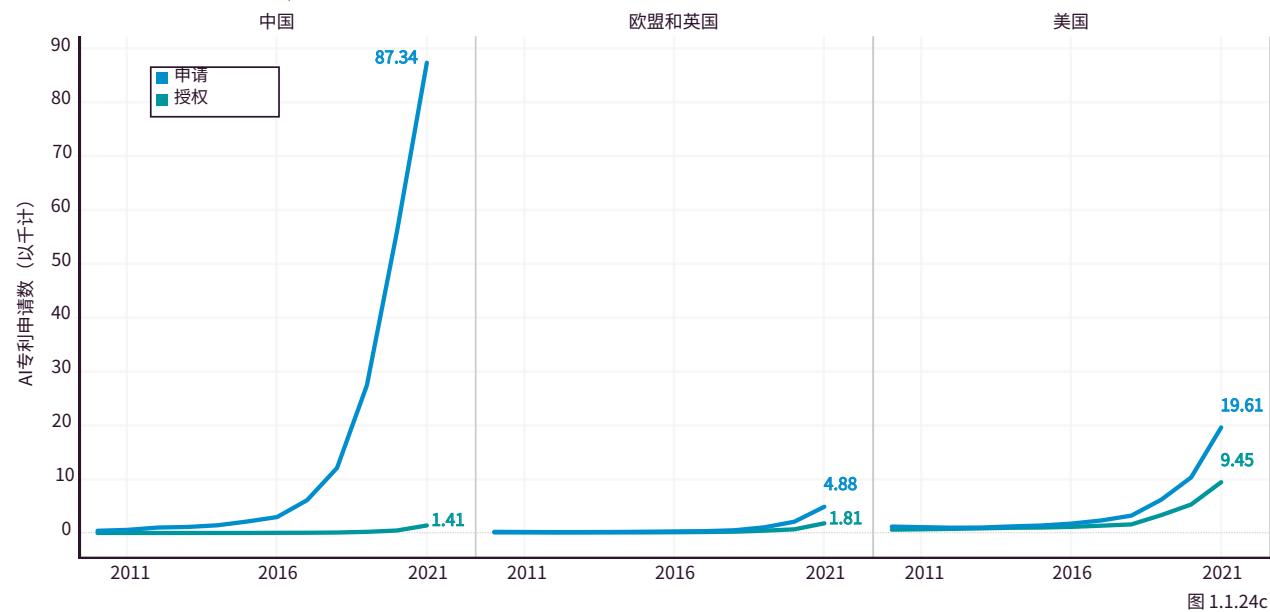


图 1.1.24c



会议的出席情况能够反映出产业界和学术界对某一科学领域的兴趣情况。在过去的20年里，人工智能会议不仅在规模上，而且在数量和声望上都有所增长。本节介绍了主要人工智能会议的出席趋势数据，与前四期人工智能指数报告相比，涵盖了更多的会议（16个）。

## 1.2 会议

### 会议出席情况

与2020年相似，大多数人工智能会议在2021年以虚拟方式召开。只有国际机器人与自动化会议 (International Conference on Robotics and Automation, ICRA) 和自然语言处理经验方法会议 (the Conference on Empirical Methods in Natural Language Processing, EMNLP) 采用了线上线下混合形式。会议组织者报告表示，因为虚拟会议为来自世界各地的研究人员提供了更多的出席机会，因此很难衡量虚拟会议的确切出席人数。

图1.2.1显示，2021年顶级人工智能会议在全球有超过88,000名与会者，其出席情况与2020年基本一致。图1.2.2和图1.2.3显示了各个会议的具体出席数据，16个主要的人工智能会议被分成两类：出席人数超过2500人的大型人工智能会议和出席人数少于2500人的小型人工智能会议。<sup>6</sup>

#### 2010-21年部分人工智能会议出席人数

来源：会议数据，2021 | 图：2022人工智能指数报告

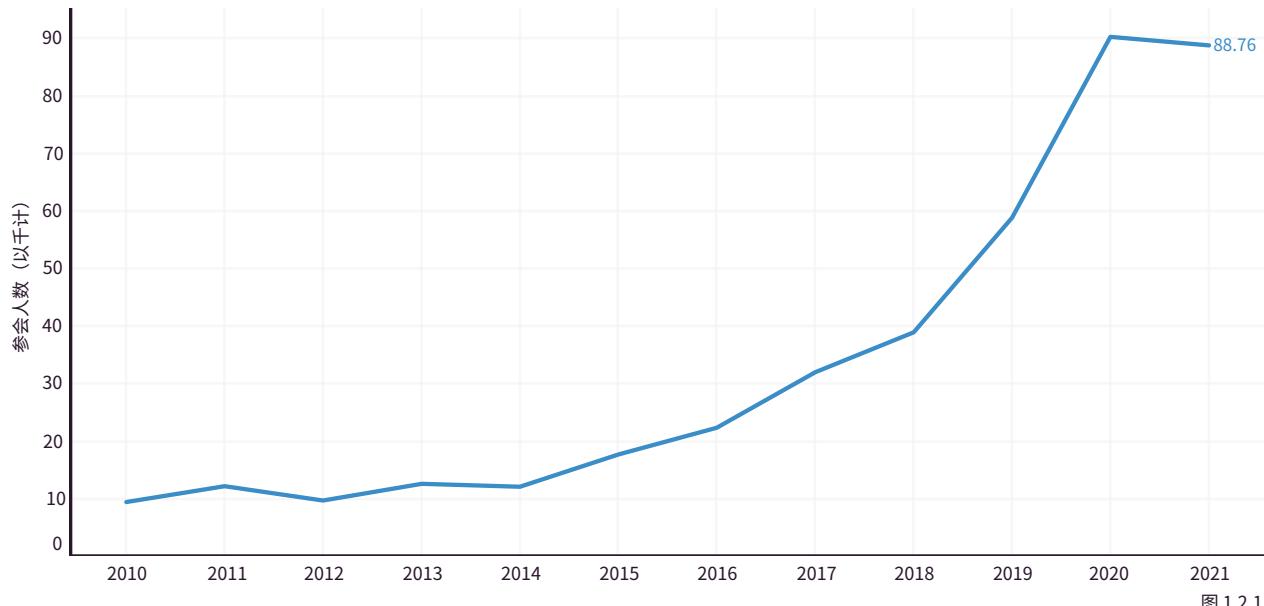


图 1.2.1

<sup>6</sup> 国际机器学习会议 (International Conference on Machine Learning, ICML) 将会议访问量确定为会议出席人数，这就解释了该会议在2021年的高出席人数。国际智能机器人和系统会议 (The International Conference on Intelligent Robots and Systems, IROS) 扩展了虚拟会议，允许用户在长达3个月的时间内观看，这也解释了该会议在2020年的高出席人数。对于AAMAS会议，2020年的出席人数是基于记录会谈和管理在线会议的平台所报告的现场用户数，而2021年的数字是总注册人数。



### 2010-21年大型人工智能会议的出席情况

来源：会议数据，2021 | 图：2022人工智能指数报告

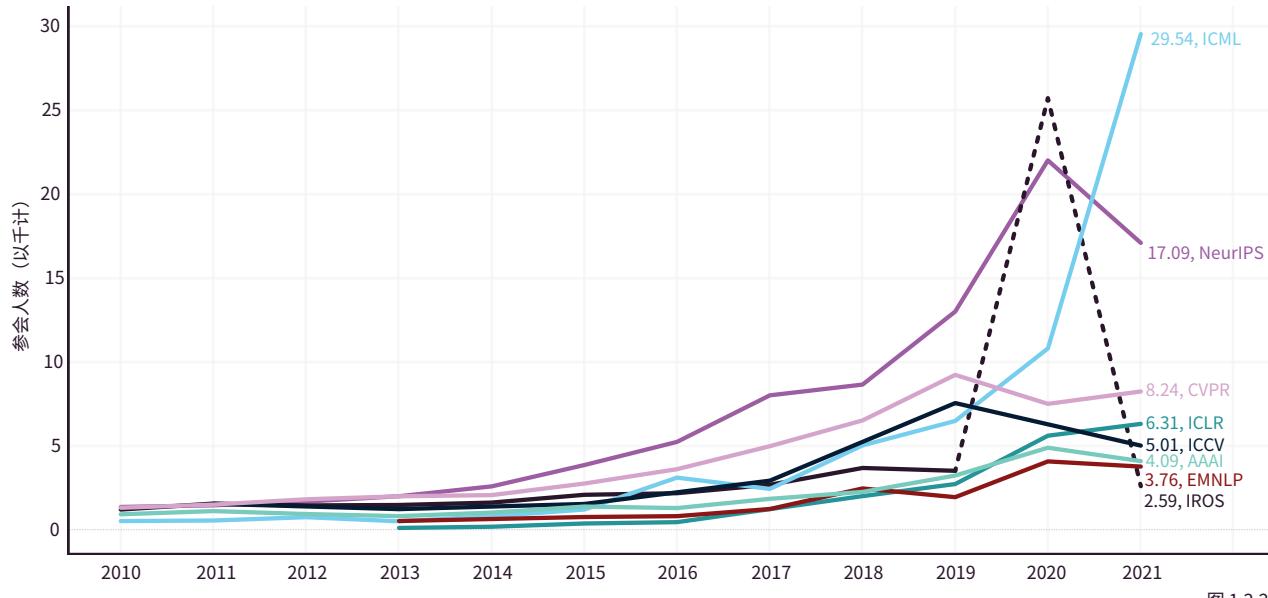


图 1.2.2

### 2010-21年小型人工智能会议的出席情况

来源：会议数据，2021 | 图：2022人工智能指数报告

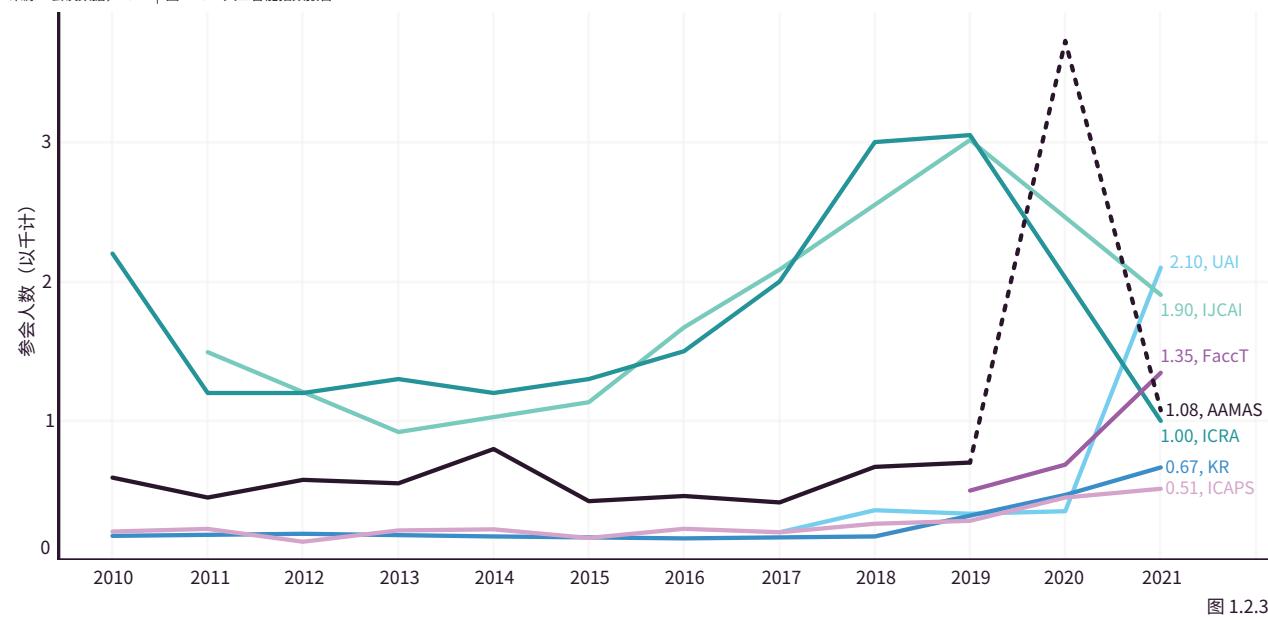


图 1.2.3



## 机器学习学习中的女性 (WOMEN IN MACHINE LEARNING, WiML) NeurIPS研讨会

WiML成立于2006年，是一个致力于支持和提高女性在机器学习领域影响力的组织。本节介绍了其与NeurIPS同地举办的年度技术研讨会的情况。从2020年开始，WiML还主办了Un-Workshop，旨在通过ICML上来自不同背景的参与者之间的合作和互动来推动研究。

### 研讨会参与者

自2006年首次推出以来，参加WiML研讨会的人数一直在稳步增长。图1.2.4显示，在2021年的研讨会上，约有1486人参加了所有的研讨主题。该数据以访问neurips.cc上的虚拟研讨会平台的人数计算所得。2021年NeurIPS的WiML研讨会在三天内举行了多场会议，这与2020年的形式有所变化。与2020年一样，由于COVID-19的影响，研讨会以虚拟方式举行。

2010-21年WiML研讨会出席人数

来源：WiML, 2021 | 图：2022人工智能指数报告

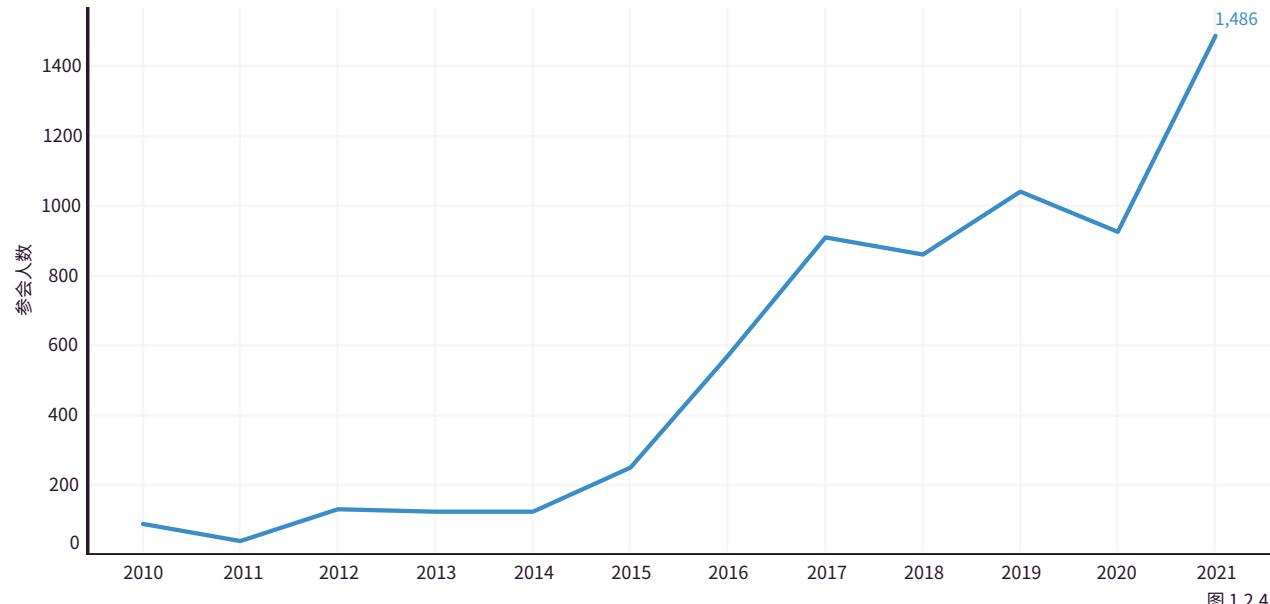


图 1.2.4



## 人口统计学分类

本节给出了2021年出席研讨会人员的居住大洲和专业职位分类，这些信息是通过同意汇总这些信息的参与者填写的调查收集而来。在调查对象中，超过一半的调查对象来自北美，其次是欧洲（19.9%）、亚

洲（16.2%）和非洲（7.3%）（图1.2.5）。图1.2.6显示，调查参与者中几乎有一半是博士生，而大学教师的比例约为1.2%。研究员科学家/工程师、数据科学家/工程师和软件工程师是最常出现的专业职位。

2021年NEURIPS WiML研讨会出席人员的居住地范围

来源：WiML, 2021 | 图：2022人工智能指数报告

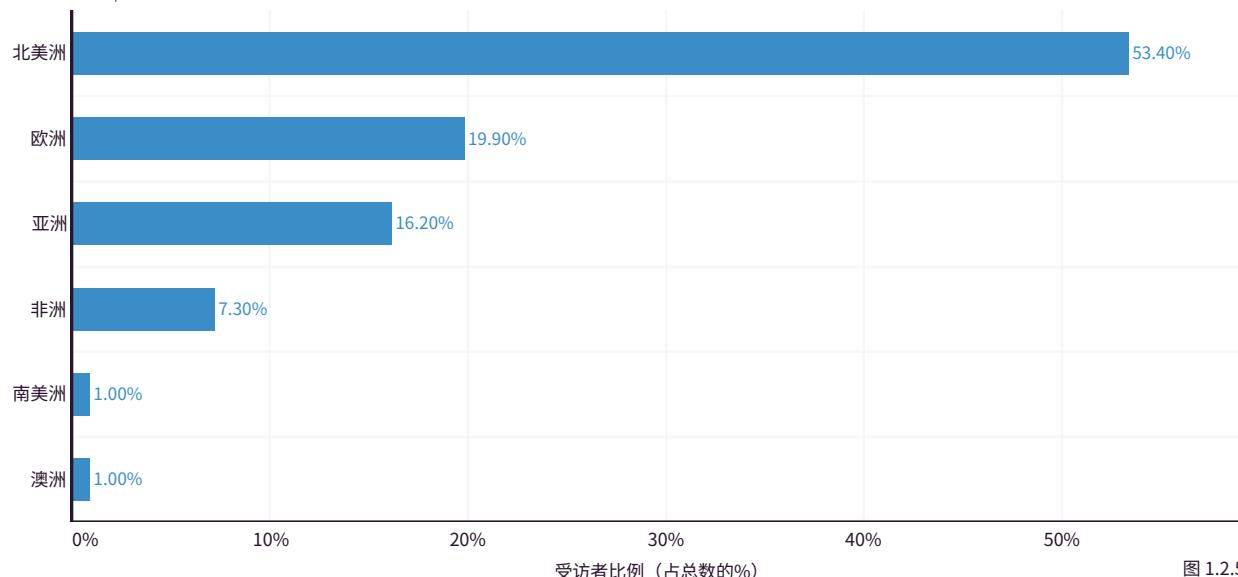


图 1.2.5

2021年NEURIPS WiML研讨会出席人员的专业职位

来源：WiML, 2021 | 图：2022人工智能指数报告

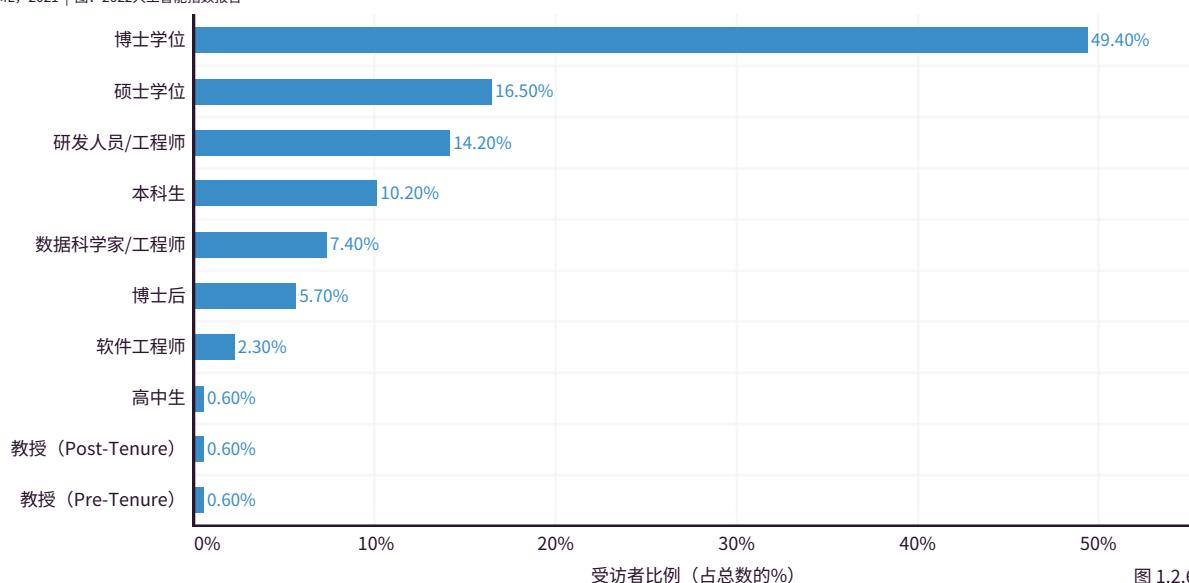


图 1.2.6



软件库是用于创建应用程序和产品的计算机代码的集合。流行的人工智能专用软件库--如TensorFlow和PyTorch--帮助开发者快速有效地创建他们的人工智能解决方案。本节根据GitHub数据分析了软件库的受欢迎程度。

## 1.3 人工智能开源软件库

### GITHUB星标

图1.3.1和1.3.2反映了2015年至2021年GitHub开源AI软件库的用户数量。2021年，TensorFlow仍然是迄今为止最受欢迎的软件库，大约有161,000个累计的GitHub星标，比2020年略有增加。2021年，TensorFlow的受欢迎程度约为排名第二的GitHub开源AI软件库OpenCV的三倍，其次是Keras、PyTorch和Scikit-learn。图

1.3.2显示了在GitHub上拥有少于4万颗星的软件库的受欢迎程度。

图1.3.2显示了GitHub星标少于4万的软件库的受欢迎程度，其中以FaceSwap为首，约有4万颗星，其次是100-Day-OF-ML-Code、AiLearning和BVLC/caffe。

2014-21年按人工智能软件库划分的GITHUB星标数（超过4万个星标）

来源：GitHub, 2021 | 图：2022人工智能指数报告

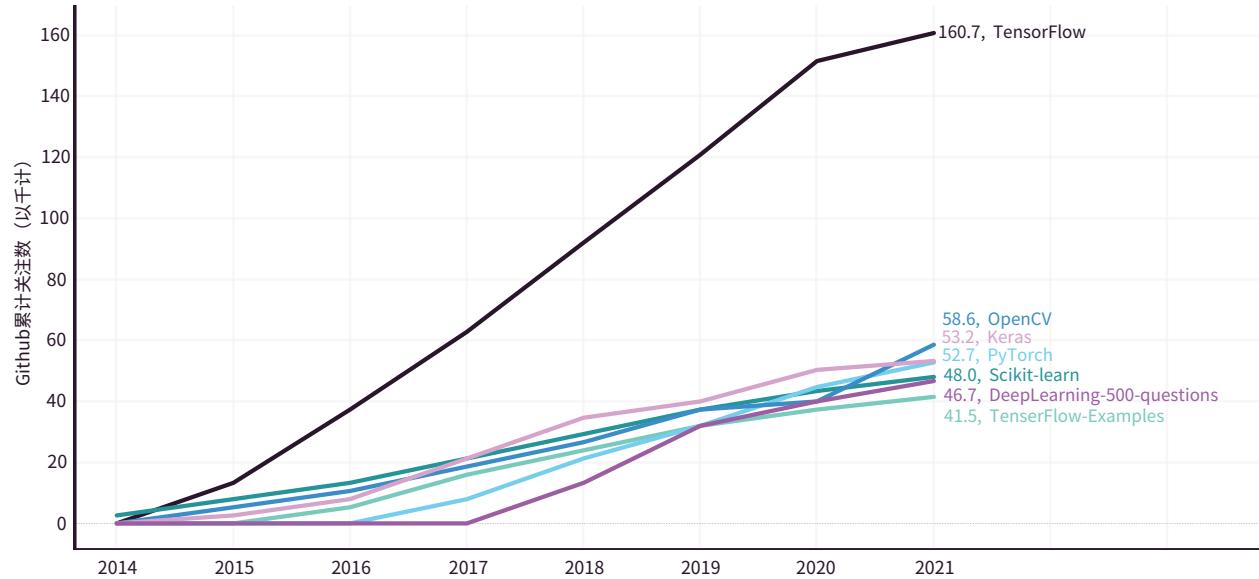


图 1.3.1



### 2014-21年按人工智能软件库划分的GITHUB星标数量（低于4万个星标）

来源：GitHub, 2021 | 图：2022人工智能指数报告

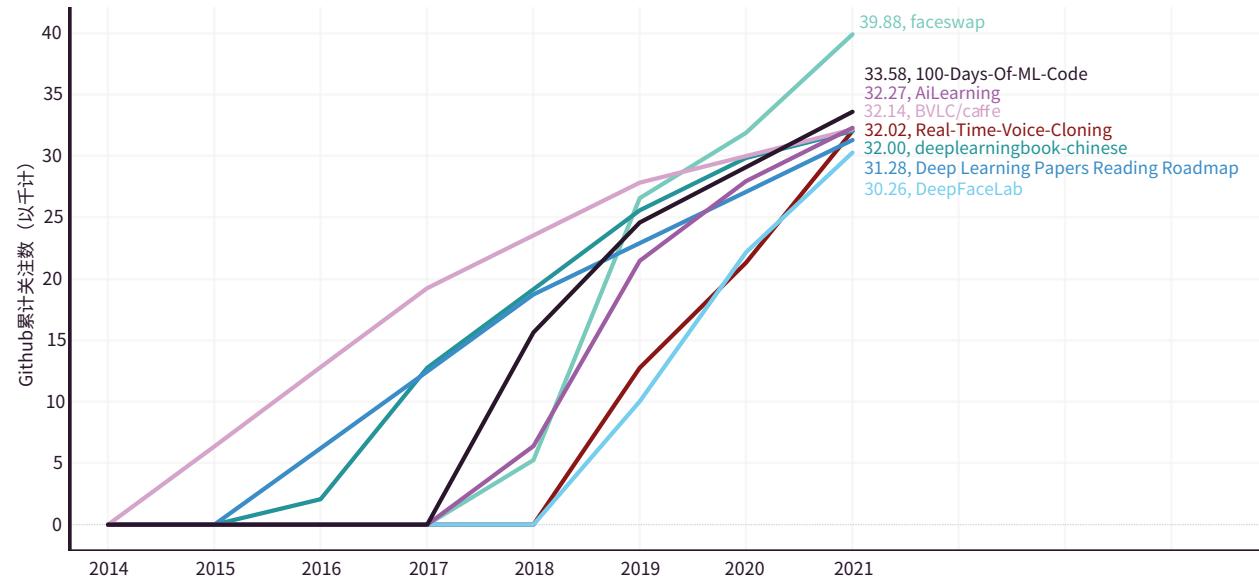


图 1.3.2



2022年  
人工智能指数报告

## 章节二 技术性能



## 章节二 章节预览

概述	50	医学图像分割	61
章节要点	51	CVC-ClinicDB 和 Kvasir-SEG	61
<b>2.1 计算机视觉-图像</b>	<b>52</b>	人脸检测和识别	62
图像分类	52	NIST 人脸识别库 (Face Recognition Vendor Test, FRVT)	62
ImageNet	52	人脸检测：遮挡的影响	63
ImageNet: Top-1准确度	52	人脸识别供应商测试 Face Recognition Vendor Test (FRVT): 人脸遮挡的影响	63
ImageNet: Top-5准确度	52	亮点：户外人脸检测数据库 Masked Labeled Faces in the Wild, MLFW	64
图像生成	54	视觉推理	65
STL-10: Fréchet 起始距离 (FID) 得分	54	视觉问答 (VQA) 挑战	65
CIFAR-10: Fréchet 起始距离 (FID) 得分	55	<b>2.2 计算机视觉-视频</b>	<b>67</b>
Deepfake检测	56	行为识别	67
FaceForensics++	56	Kinetics-400, Kinetics-600, Kinetics-700	67
Celeb-DF	57	ActivityNet: 时间动作定位任务	69
人体姿态估计	57	目标识别	70
Leeds Sports Poses: 关键点正确 估计比例 (Percentage of Correct Keypoints, PCK)	58	语境中的常见对象 (Common Object in Context, COCO)	71
Human3.6M: 平均 (每) 关节位置 误差 (Mean Per Joint Position Error, MPJPE)	59	你只看一次 (YOLO)	72
语义分割	60	视觉常识推理 (visual commonsense reasoning, VCR)	73
城市景观 (Cityscapes)	60		

访问公开数据



## 章节二 章节预览（续）

<b>2.3 语言</b>	<b>74</b>
英语语言理解基准	74
SuperGLUE	74
SQuAD	75
阅读需要逻辑推理的综合数据集 (reading comprehensive dataset requiring logical reasoning, ReClor)	76
文本摘要	78
arXiv	78
PubMed	79
自然语言推理	80
斯坦福大学自然语言推理 (stanford natural language inference, SNLI)	80
归纳自然语言推断 (abductive natural language inference, aNLI)	81
情感分析	82
SemEval 2014 Task 4 Sub Task 2	82
商业机器翻译 (MT)	83
WMT 2014, English-German 和 English-French	84
商业上可用的MT系统的数量	85
<b>2.4 语音</b>	<b>86</b>
语音识别	86
转录语音: LibriSpeech	86
说话人识别: VoxCeleb	87
<b>2.5 推荐</b>	<b>88</b>
商业推荐: MovieLens 20M	88
点击率预测: Criteo	89
<b>2.6 强化学习</b>	<b>90</b>
强化学习环境	90
Arcade学习环境: Atari-57	90
Procgen	91
人类游戏: Chess	93
<b>2.7 硬件</b>	<b>94</b>
MLPerf: 训练时间	94
MLPerf: 加速器数量	96
IMAGENET: 训练代价	97
<b>2.8 机器人</b>	<b>98</b>
机械臂的价格趋势	98
专业机器人中应用的AI技能	99

访问公开数据



# 概述

在今年的报告中，技术性能章节更多地分析了人工智能各个子领域的技术进展，包括计算机视觉、语言、语音、推荐、强化学习、硬件和机器人等方面的趋势。本章使用了一些量化的测量方法，覆盖常见的人工智能基准和奖项挑战到全领域的调查，以突出性能最优的人工智能系统的发展。



## 章节要点

- **数据、数据、数据：**各项技术越来越依赖于使用额外的训练数据来创造新的最先进的结果。截至2021年，本报告中的10个基准中，有9个最先进的人工智能系统是引入额外数据训练得到的。这种趋势将会越来越有利于掌握大量数据的私营机构。
- **对特定计算机视觉子任务的兴趣上升：**2021年，研究界对更具体的计算机视觉子任务有了更大的兴趣，如医学图像分割和遮挡人脸识别。例如，在2020年之前，只有3篇研究论文涉及Kvasir-SEG医学成像基准测试系统。在2021年，则有25篇研究论文。这样的增长表明，人工智能研究正在朝着能够有更直接的、现实世界的应用的研究方向发展。
- **人工智能还无法应对复杂的语言任务：**在像SuperGLUE和SQuAD这样的基本阅读理解基准上，人工智能的表现已经超过了人类水平1%-5%。虽然人工智能系统在更复杂的语言任务上仍无法达到人类的表现，如归纳自然语言推理（aNLI），但差距正在缩小。2019年，人类在aNLI上的表现超过人工智能9个百分点。截至2021年，这一差距已经缩小到1%。
- **转向更广泛的强化学习：**在过去的十年里，人工智能系统已经能够完成狭义的强化学习任务，在这些任务中，它们被要求最大限度地提高某一特定技能的表现，例如国际象棋。顶级国际象棋软件引擎现在比Magnus Carlsen的Top ELO分数高出24%。然而，在过去的两年里，人工智能系统在更普遍的强化学习任务（Progen）上的性能也提高了129%，在这些任务中，它们必须在新的环境中运行。这一趋势说明，能够学会更广泛思考的人工智能系统正在不断发展。
- **人工智能变得更经济，性能更高：**自2018年以来，训练图像分类系统的成本下降了63.6%，而训练时间缩短了94.4%。训练成本降低但训练时间加快的趋势出现在其他MLPerf任务类别中，如推荐、对象检测和语言处理，推动了AI技术更广泛的商业应用。
- **机械臂正在变得更便宜：**本报告的一项问卷调研显示，在过去六年中，机械臂的中位价格下降了46.2--从2017年的每只手臂42,000美元到2021年的22,600美元。机器人研究已变得更容易实现，成本更低。



计算机视觉是人工智能的一个子领域，即教会机器理解图像和视频。计算机视觉的任务范围很广，包括图像分类、物体识别、语义分割和人脸检测等。截至2021年，在大量的计算机视觉任务中计算机的性能都超过了人类。计算机视觉技术在现实世界中有着各种重要的应用，如自动驾驶、人群监控、体育分析和视频游戏创作。

## 2.1 计算机视觉-图像

### 图像分类

图像分类指的是机器对它们在图像中看到的东西进行分类的能力（图2.1.1）。在实际场景中，图像识别系统可以帮助汽车识别周围的物体，医生检测肿瘤，工厂经理发现生产缺陷。在过去的十年中，得益于机器学习技术的不断发展，图像识别系统的技术能力有了巨大的进步。此外，算法、硬件和数据技术的进步意味着图像识别已经变得比以往任何时候都更经济、更广泛地适用和更容易获得。

### ImageNet

ImageNet是一个数据库，其中包括超过1400万张图片，涉及20000个类别，公开供研究人员在图像分类问题上使用。ImageNet创建于2009年，现在是科学家们在图像分类方面最常见的改进基准之一。

### ImageNet: Top-1 Accuracy

ImageNet上的基准测试是通过准确性指标来衡量的，这些指标量化了人工智能系统为给定图像分配正确标签的频率。Top-1准确率衡量的是分类模型对某一特定图像所做的最高预测与该图像的实际目标标签相匹配的比率。近年来，越来越多的研究通过引入其他图像数据集的额外数据进行预训练来改进ImageNet上的系统性能。

截至2021年底，顶级图像分类系统在Top-1 Accuracy上平均每进行10次分类就会出现1次错误，而在2012年底，平均每10次分类就会出现4次错误（图2.1.2）。2021年，性能最佳的预训练系统是由Google Brain研究人员开发的CoAtNets。

### 图像分类示例

来源：[Krizhevsky, 2020](#)

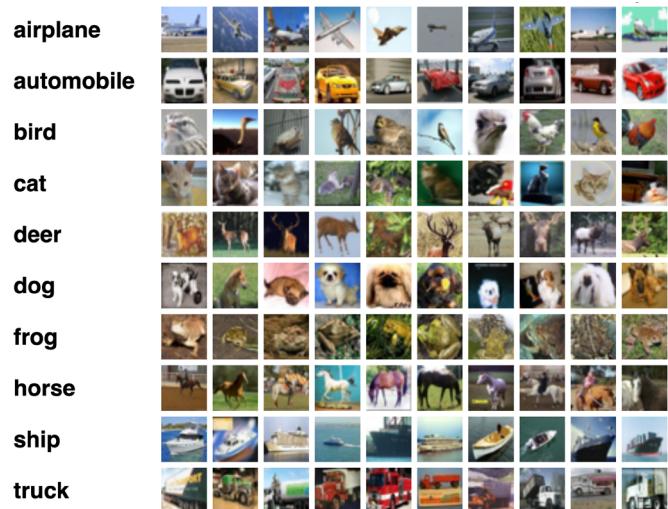


图 2.1.1

### ImageNet: Top-5 Accuracy

Top-5 Accuracy考虑的是模型的5个最高概率答案中是否有任何一个与图像标签一致。正如图2.1.3所展示的，人工智能系统目前已经实现了接近完美的Top-5估计。目前，经过预训练的Top-5 Accuracy的最佳表现是99.0%，由微软云和微软AI的Florence-CoSwim-H模型在2021年11月实现。

ImageNet Top-5 Accuracy的提高正在趋于平稳。这其实很好理解，如果你的系统在100次中有98或99次分类正确，那么你能改进的空间就非常有限了。



### IMAGENET挑战赛：Top-1 Accuracy

来源：Papers with Code, 2021; arXiv, 2021 | 2022人工智能指数报告

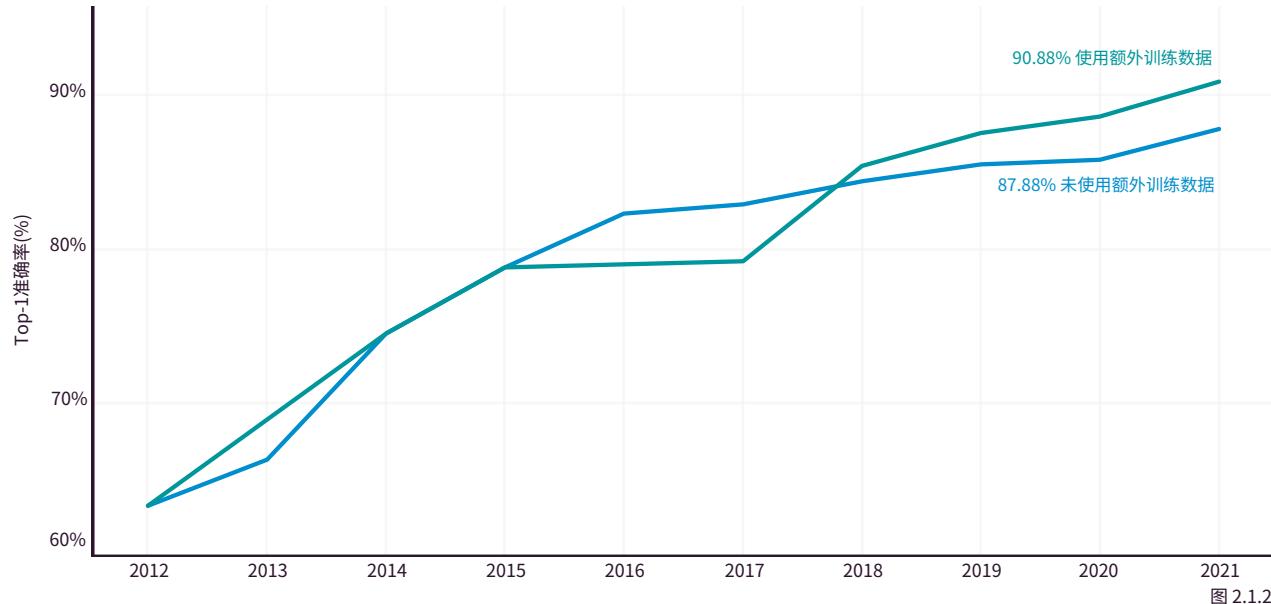


图 2.1.2

### IMAGENET挑战赛：Top-5 Accuracy

来源：Papers with Code, 2021; arXiv, 2021 | 2022人工智能指数报告

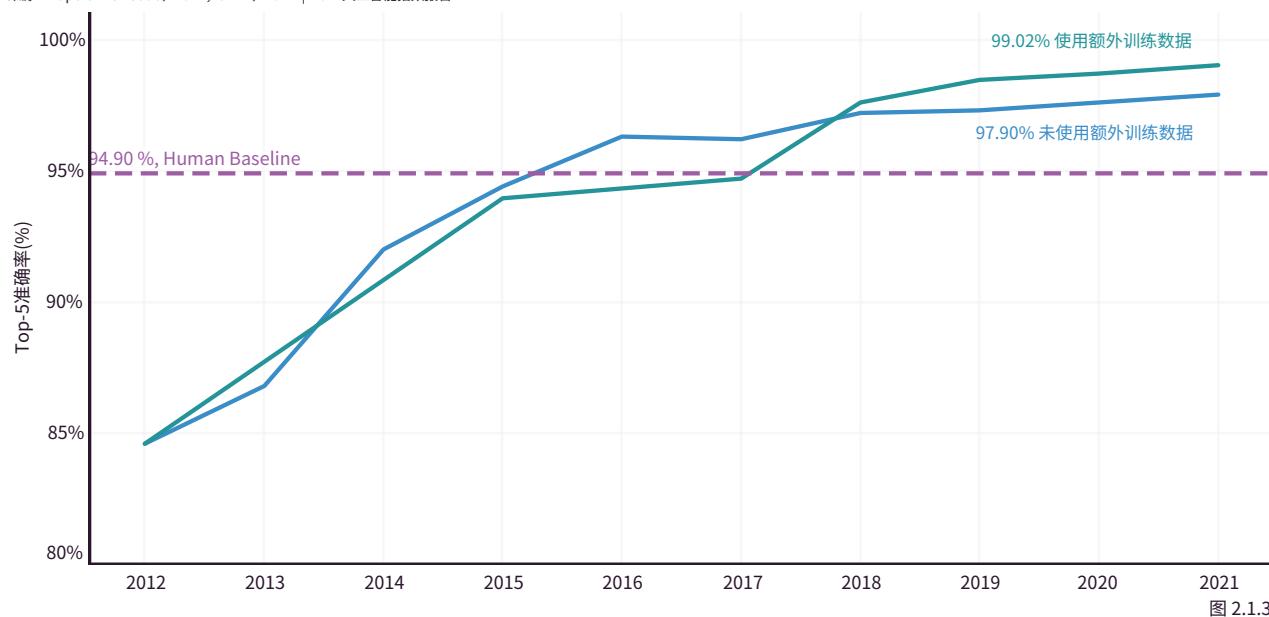


图 2.1.3



## 图像生成

图像生成是指生成与真实图像无法区分的图像的任务。图像生成可以在需要创建视觉内容的生成领域中发挥广泛的作用，例如，娱乐（像英伟达这样的公司已经使用图像生成器来创建游戏的虚拟世界），时尚（设计师可

以让人工智能系统生成不同的设计模式），以及医疗保健（图像生成器可以合成新的药物化合物）。图2.1.4通过展示去年由人工智能系统合成的几张人脸，说明了在图像生成方面取得的进展。

### 人脸生成方面的进展

来源：Goodfellow et al., 2014; Radford et al., 2016; Liu & Tuzel, 2016; Karras et al., 2018; Karras et al., 2019; Goodfellow, 2019; Karras et al., 2020; AI Index, 2021; Vahdat et al., 2021



2018



2020



2021

图 2.1.4

## STL-10: Fréchet Inception Distance (FID)

### 得分

FID得分用于评估人工生成的一组图像与真实图像之间的相似度。分数越低意味着生成的图像与真实的图像越相似，分数为零表示生成的图像与真实图像完全相同。

图2.1.5记录了生成模型在STL-10数据集上取得的FID得分，该数据集是计算机视觉领域引用最广泛的数据集之一。由韩国科学技术研究院和首尔大学的研究人员在STL-10上开发的最先进的模型的FID得分是7.7，这一数据明显优于2020年的最先进的结果。

### STL-10: FID得分

来源：Papers with Code, 2021; arXiv, 2021 | 2022人工智能指数报告

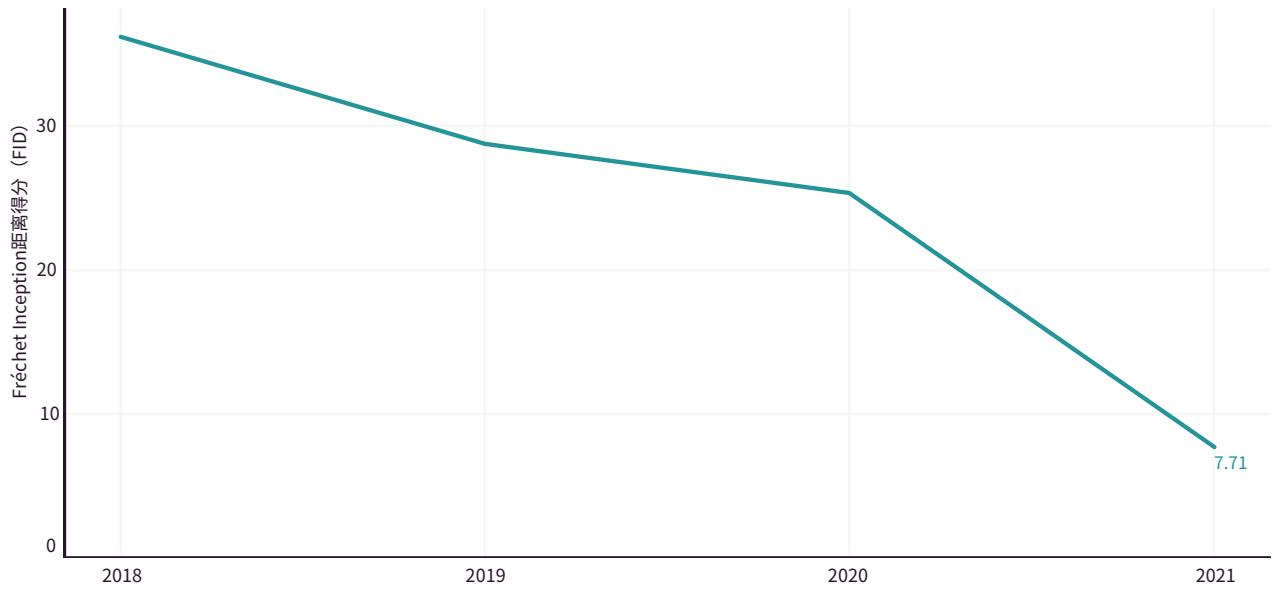


图 2.1.5



## CIFAR-10: Fréchet Inception Distance (FID) 得分

图像生成方面的进展也可以在CIFAR-10上进行测试，该数据集包括10个不同物体类别的60,000张彩色图像。2021年发布的CIFAR-10上的最先进结果是由英伟达的研究人员实现的。

顶级图像生成模型在CIFAR-10上取得的FID分数要比 STL-10低很多。这种差异可能是由于CIFAR-10包含的图像的分辨率（ $32 \times 32$ 像素）比STL-10上的图像（ $96 \times 96$ 像素）低得多。

### CIFAR-10: FID得分

来源: Papers with Code, 2021; arXiv, 2021 | 2022人工智能指数报告

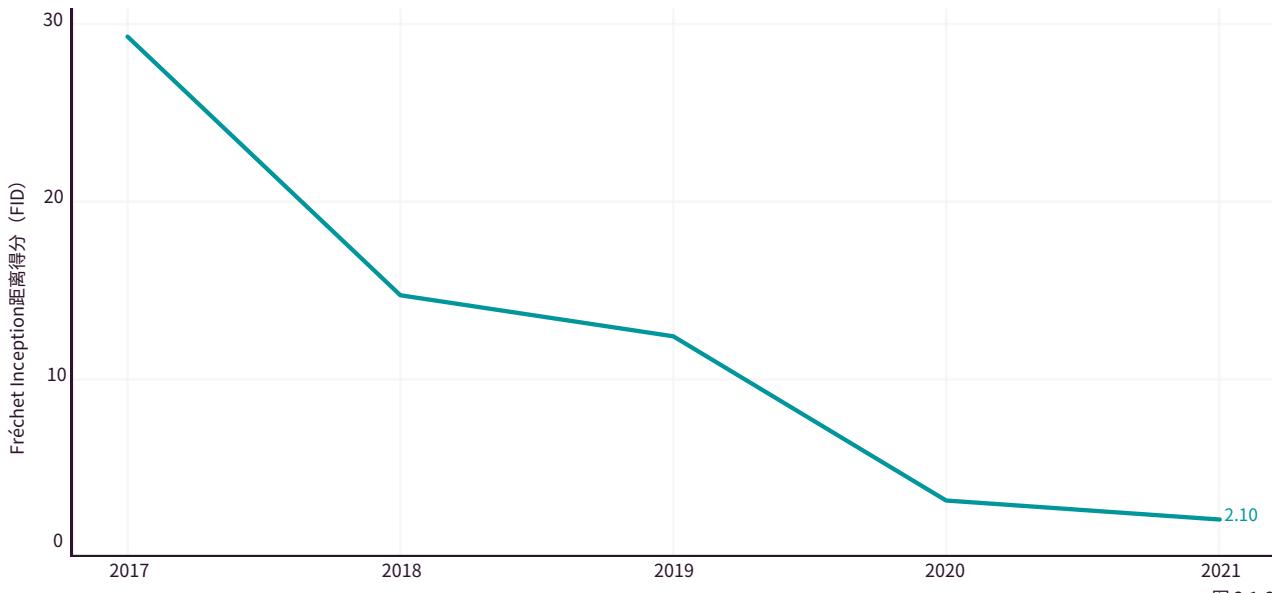


图 2.1.6



## DEEPFAKE检测

许多人工智能系统现在可以生成与真实图像无法区分的假图像。通过Deepfake技术能够将一个人的脸叠加到另一个人身上，形成所谓的“深度伪造”。Deepfake被用在多种场合中，例如广告、生成厌恶女性色情作品和虚假信息等（例如，在2018年，Barack Obama对Donald Trump说脏话的Deepfake视频在网上被传播了200多万次）。在过去的几年里，人工智能研究人员试图通过制作更强大的Deepfake检测算法来应对不断改进的Deepfake技术。

## FaceForensics++

FaceForensics++是一个Deepfake检测基准数据集，包含大约1000个来自YouTube视频的原始视频序列。FaceForensics++的性能改进是通过准确性来衡量的：一个算法能够正确识别的被改变的图像的百分比。

虽然FaceForensics++是在2019年推出的，但研究人员在该数据集上测试了之前就已经提出的Deepfake检测方法，以跟踪Deepfake检测方面的长期进展（图2.1.7）。在过去的十年里，人工智能系统在检测Deepfake方面不断改进。2012年，在所有四个FaceForensics++数据集中，表现最好的系统可以正确识别69.9%的Deepfake。2021年，这个数字增加到了97.7%。<sup>1</sup>

### FaceForensics++: 正确性

来源：arXiv, 2021 | 2022人工智能指数报告

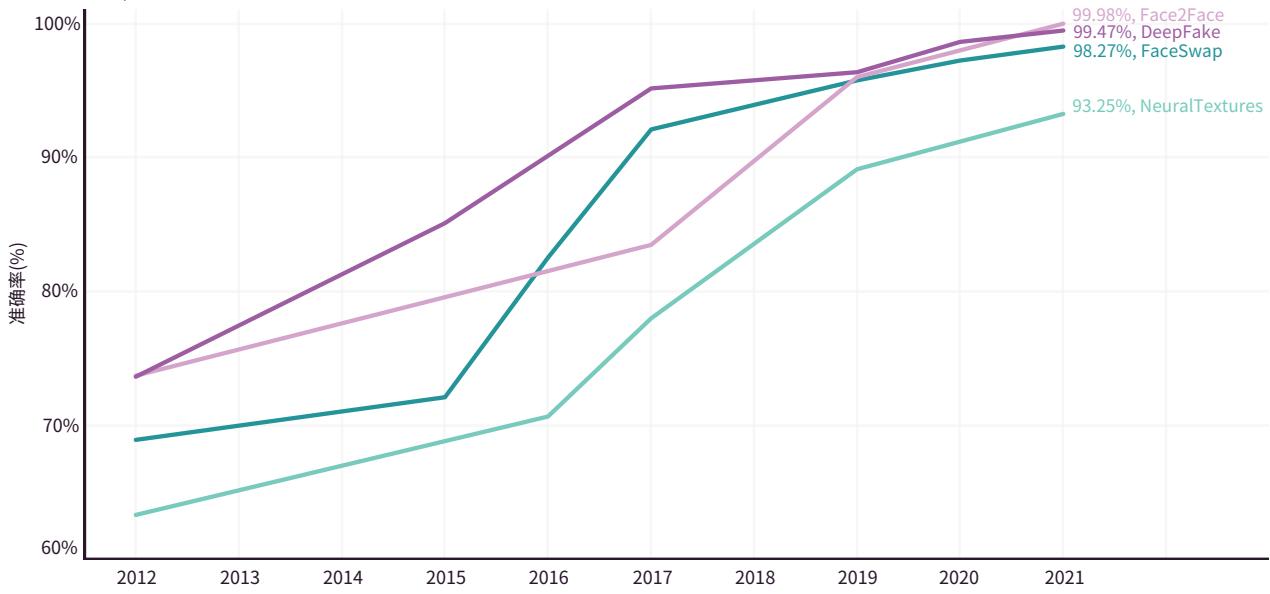


图 2.1.7

<sup>1</sup> 这些数字是通过对所有四个FaceForensics++数据集的性能进行平均估算所得。



## Celeb-DF

Celeb-DF Deepfake检测数据集是由从590个原始名人YouTube视频操纵生成的5639个Deepfake视频组成的。Celeb-DF是在2019年推出的。2021年，Celeb-DF的最高分是76.9分，是由来自中国科技大学和阿里巴巴集团的研究人员得到的（图2.1.8）。

顶级检测模型在Celeb-DF上的表现明显比FaceForensics++差（20个百分点），这表明对于待测试的技术来说，Celeb-DF更具挑战性。随着Deepfake技术在未来几年不断改进，继续监测Celeb-DF和其他类似的具有挑战性的Deepfake检测数据集的进展将非常重要。

CELEB-DF：曲线下面积得分 (AUC)

来源：arXiv, 2021 | 2022人工智能指数报告

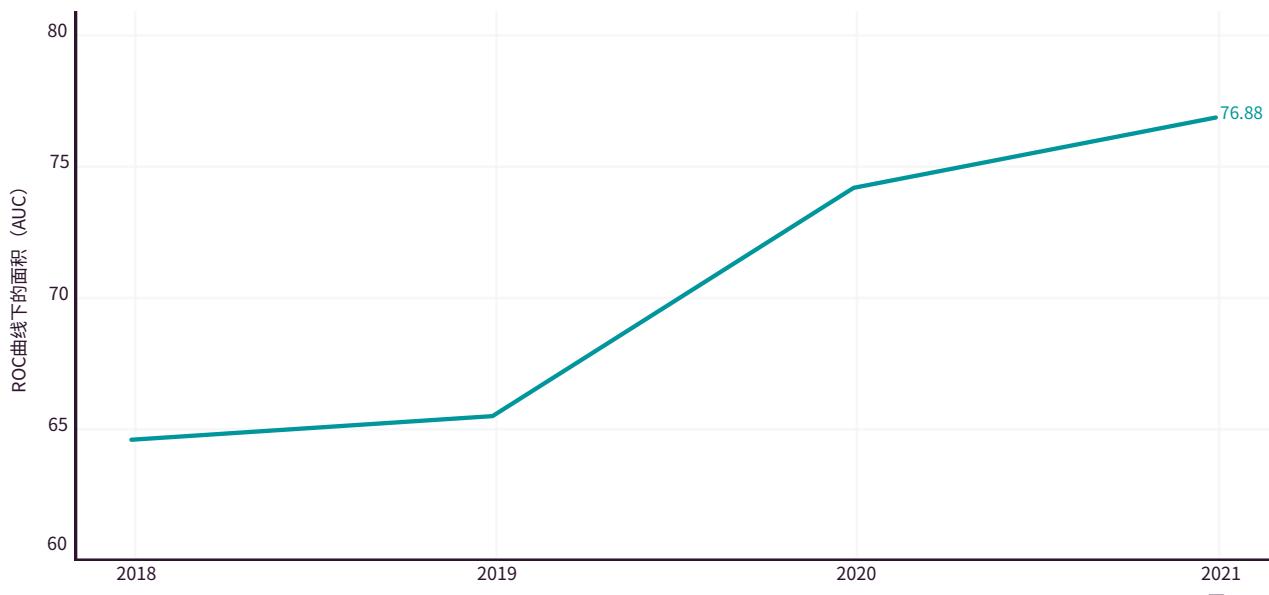


图 2.1.8

## 人体姿态估计

人体姿态估计是指从一张图像中估计人体关节（手臂、头部、躯干等）的不同位置（图2.1.9），然后结合这些估计值来正确标注人体所处的姿态。

人体姿态估计可用于行为识别场景中，如体育分析、人群监控、CGI开发、虚拟环境设计和交通（例如，识别机场跑道控制员的身体语言标志）。

人类姿态估计示例

来源：[Cao et al., 2019](#)



图 2.1.9



## Leeds Sports Poses：关键点正确估计比例（Percentage of Correct Keypoints, PCK）

Leeds Sports Poses数据集包含了2000张从Flickr上收集的运动员运动的图片。每张图片包括14个不同的身体关节位置的信息。Leeds Sports Poses基准的表现是通过正确估计关键点的百分比来评估的。

2021年，表现最好的人体姿态估计模型在Leeds Sports Poses中正确识别了99.5%的关键点（图2.1.10）。鉴于Leeds Sports Poses上的最大性能是100.0%，必须开发更具挑战性的人类姿态估计基准，因为我们已经接近饱和基准了。

Leeds Sports Poses：关键点正确估计比例（PCK）

来源：Papers with Code, 2021; arXiv, 2021 | 2022人工智能指数报告

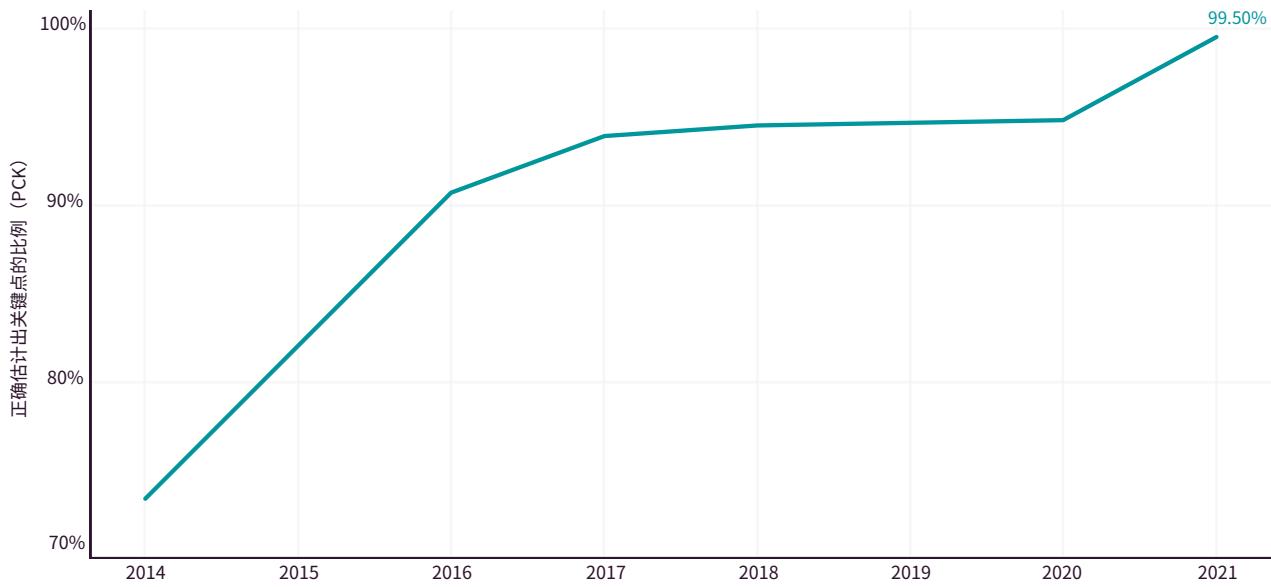


图 2.1.10



## Human3.6M：平均（每）关节位置误差（Mean Per Joint Position Error, MPJPE）

三维人体姿态估计是一种更具挑战性的姿态估计类型，人工智能系统被要求在三维空间而不是二维空间中估计姿态。Human3.6M数据集主要用于跟踪三维人类姿态估计的进展。Human3.6M是一个超过360万张图片的集合，包括17种不同类型的人类姿态（打电话、讨论、吸

烟等）。Human3.6M的性能是以平均每个关节的位置误差来衡量的，单位是毫米，这是人工智能模型的位置估计和实际位置注释之间的平均差异。

2014年，表现最好的模型每个关节的平均误差为16厘米，是学校用标准尺子的一半。2021年，这个数字下降到1.9厘米，还不到一个普通回形针的大小。

### Human3.6M：平均（每）关节位置误差（Mean Per Joint Position Error, MPJPE）

来源：Papers with Code, 2021; arXiv, 2021 | 2022人工智能指数报告

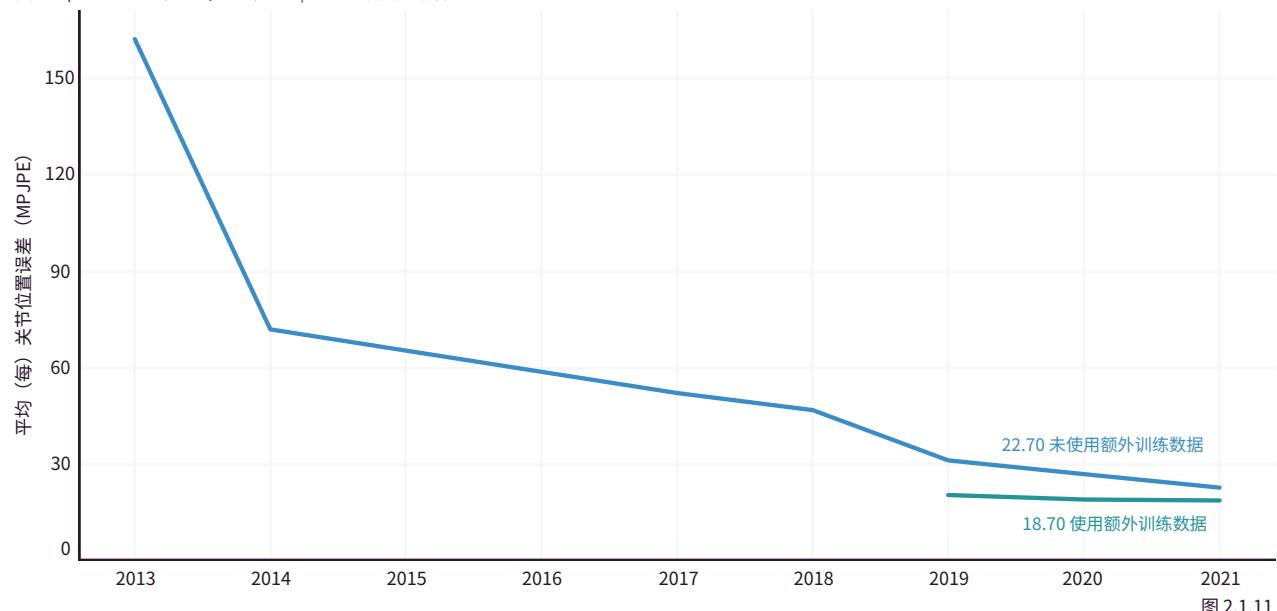


图 2.1.11



## 语义分割

语义分割是指给单个图像像素分配一个类别（如人、自行车或背景）的任务（图2.1.12）。现实世界中大量领域都需要应用像素级的图像分割技术，如自动驾驶（识别汽车看到的图像中哪些部分是行人，哪些部分是道路）、图像分析（区分照片中的前景和背景），以及医疗诊断（分割肺部的肿瘤）。

### 语义分割示例

来源：视觉对象类别挑战赛，2012

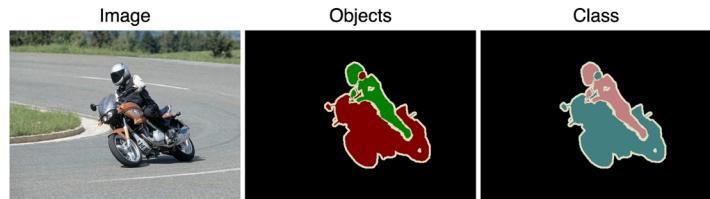


图 2.1.12

## 城市景观

城市景观Cityscapes数据集包含了50个城市的街道环境图像，这些图像是在不同季节的白天拍摄的，可以用于对广泛的语义分割任务（实例级、全景和三维车辆）进行评估。

大多数研究人员提交的任务是像素级的语义标签，人工智能系统必须在每个像素层面上对图像进行语义标签化处理。根据交叉-联合（IoU）指标对挑战者进行评

估，IoU分数越高，对应的分割精度越高。在实践中，较高的分数意味着模型预测的图像片段与图像的实际片段有较大比例的重叠。

2021年在Cityscapes方面表现最好的人工智能系统报告的分数比2015年的最高分数还要高14.6个百分点。与其他计算机视觉任务一样，过去几年中，在Cityscapes上表现最好的模型也基于额外的训练数据进行了预训练。

### Cityscapes挑战赛，像素级语义标签任务：IoU

来源：Cityscapes挑战赛，2021 | 2022人工智能指数报告

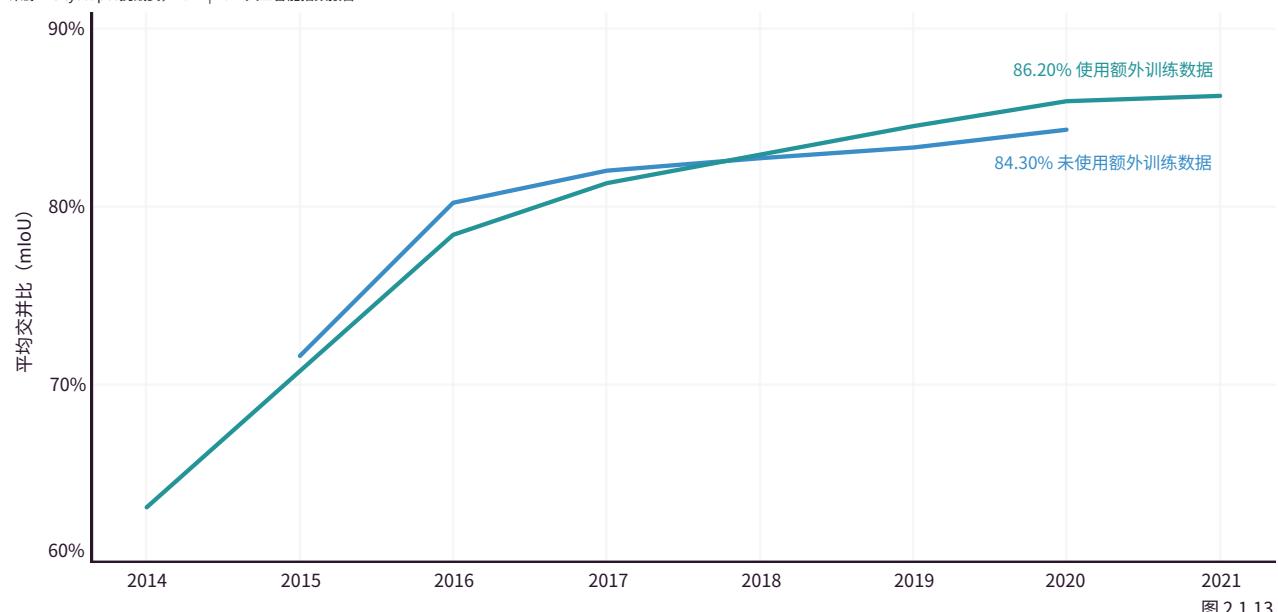


图 2.1.13



## 医学图像分割

医学图像分割是指利用人工智能系统在医学图像中分割感兴趣的对象，如器官、病变或肿瘤的能力（图2.1.14）。这项任务的技术进步对简化医疗诊断过程至关重要。医学图像分割的进步意味着医生可以花更少的时间进行诊断，而有更多的时间治疗病人。

## CVC-ClinicDB 和 Kvasir-SEG

CVC-ClinicDB是一个数据集，包括了取自31次结肠镜检查的600多张高分辨率图像。Kvasir-SEG是一个公共数据集，包括了1000张高分辨率的胃肠道息肉图像，由医生手动分割，并由专业的胃肠病学家交叉验证。这两个数据集都可以用来跟踪医学图像分割的研究进展。使用MEAN DICE值评估医学图像分割的性能，DICE代表人工智能系统识别的息肉段与实际息肉段之间的平均重叠度。

在CVC-ClinicDB上，人工智能系统目前能够正确分割结肠镜息肉的比率为94.2%，自2015年以来提高了11.9个百分点，自2020年以来提高了1.8个百分点（图2.1.15）。

### 肾脏图像分割演示

来源: [Kidney and Kidney Tumor Segmentation, 2021](#)

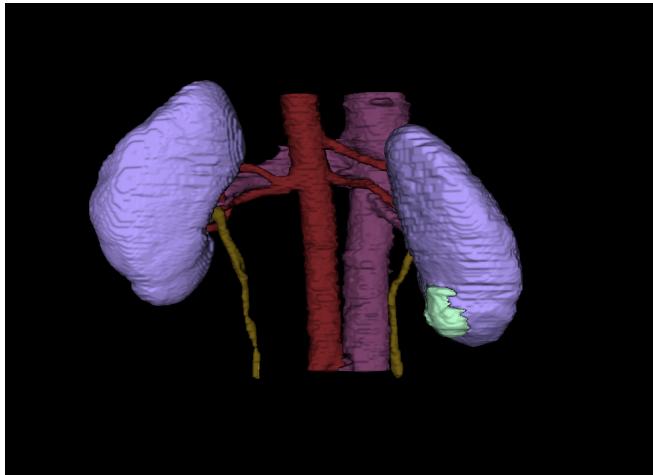


图 2.1.14

在Kvasir-SEG上也取得了类似的进展，目前表现最好的人工智能模型可以准确分割胃肠道息肉，比率为92.2%。在CVC-ClinicDB和Kvasir-SEG基准上的最好成绩是由MSRF-Net模型取得的，它是专门为医学图像分割设计的第一批卷积神经网络之一。

### CVC-CLINICDB: MEAN DICE

来源: [Papers with Code, 2021; arXiv, 2021 | 2022人工智能指数报告](#)

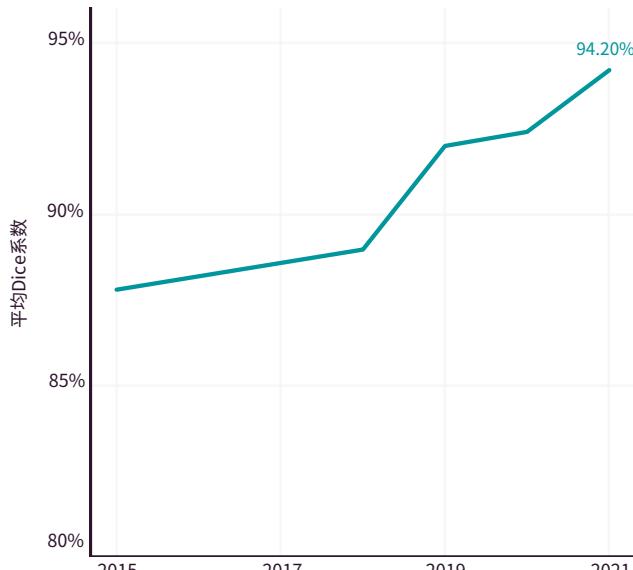


图 2.1.15a

### KVASIR-SEG: MEAN DICE

来源: [Papers with Code, 2021; arXiv, 2021 | 2022人工智能指数报告](#)

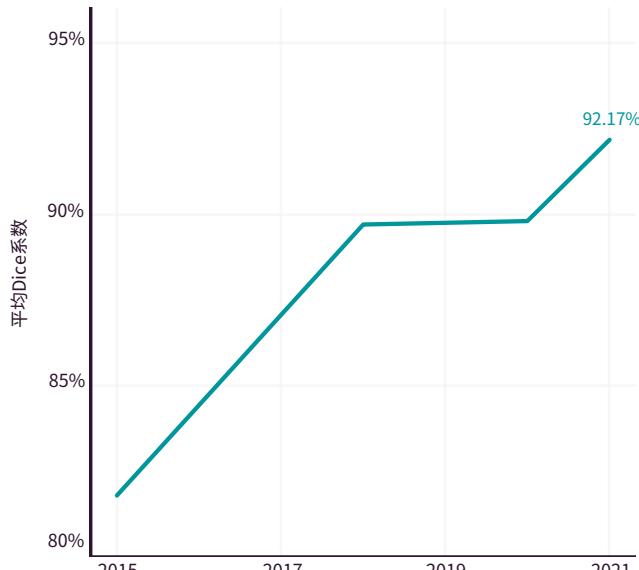


图 2.1.15b



在2020年之前，只有三篇学术论文提及了这一数据集。2020年，这个数字上升到6个，2021年则上升到25个。去年还举办了KiTS21（肾脏和肾脏肿瘤分割挑战赛），它向来自学术界和产业界的医学研究人员提出挑战，要求创建自动分割肾脏肿瘤和肾脏周围解剖结构的最佳系统。

## 人脸检测和识别

在人脸检测中，人工智能系统的任务是识别图像或视频中的个人。虽然人脸识别技术已经存在了几十年，但在过去的几年里，这项技术的进步非常大。今天一些表现优异的人脸识别算法在具有挑战性的数据集上已经达到了接近100%的成功率。

人脸识别可用于交通领域以促进跨境旅行、用于防欺诈领域以保护敏感文件，以及用于在线监考中以识别非法考试行为等。然而，人脸识别最大的实用前景在于其在安全领域的应用潜力，这使得该技术对世界各地的军队和政府都具有极大的吸引力（例如，24个美国政府机构中的18个已经在使用某种人脸识别技术）。

## 美国国家标准与技术研究所（NIST）人脸识别供应商测试（FRVT）

美国国家标准与技术研究所的人脸识别供应商测试衡量了人脸识别算法在各种国土安全和执法任务中的表现，如跨新闻照片的人脸识别、贩运儿童受害者的识别、护照的重复计算和签证图像的交叉验证。人脸识别算法的进展是根据错误的非匹配率（FNMR）或错误率（一个模型无法将图像与一个人匹配的频率）来衡量的。

2017年，一些性能出色的人脸识别算法在某些FRVT测试中的错误率还是超过了50.0%。截至2021年，已经没有任何算法的错误率会超过3.0%。2021年所有数据集中表现最好的模型（签证照片）的错误率仅为0.1%，这意味着每1000张脸，该模型就能正确识别999张。

### 美国国家标准与技术研究所（NIST）人脸识别供应商测试（FRVT）：按数据集验证的准确性

来源：美国国家标准与技术研究所，2021 | 2022人工智能指数据报告

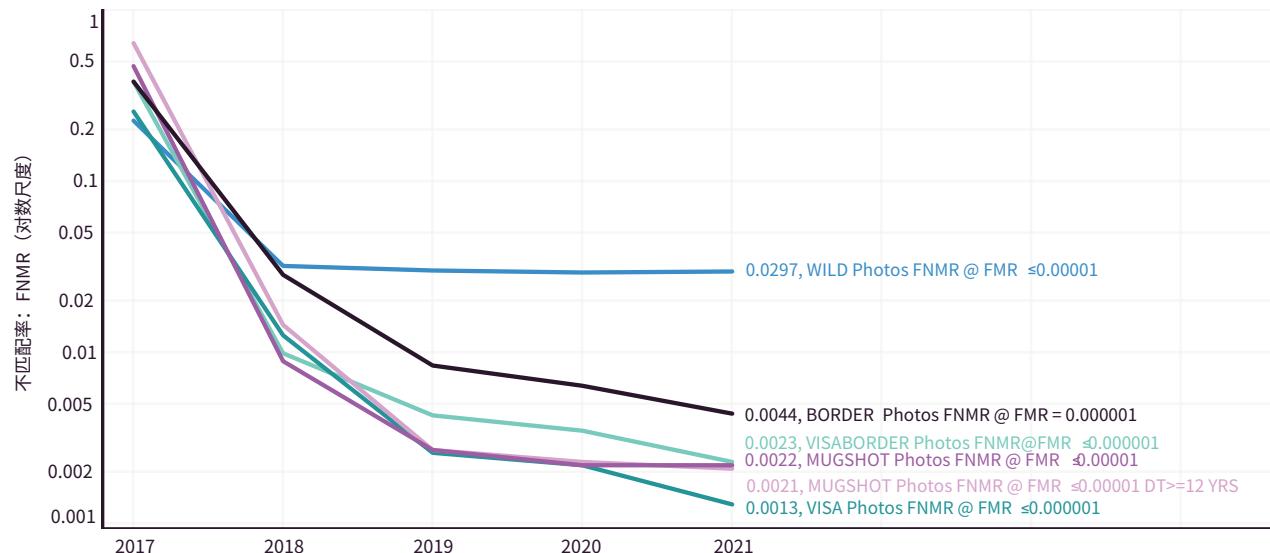


图 2.1.16



## 人脸检测：遮挡的影响

### 人脸识别供应商测试 (FRVT)：遮挡的影响

随着COVID-19的发生和随之而来的佩戴口罩规定，人脸识别已成为一项更具挑战性的任务。人脸-遮挡效应测试要求人工智能模型能够在两个签证边境照片的数据集上识别人脸，其中一个包括戴口罩的脸，另一个则不戴口罩。

从FRVT人脸遮挡测试中可以看出三个重要的趋势。(1) 人脸识别系统在戴口罩的人脸上仍然能够获得较好的性能；(2) 戴口罩的人脸图像上的性能比不戴口罩上的性能要差；(3) 自2019年以来，性能的差距已经缩小了。

虽然人脸识别技术已经存在了几十年，但过去几年的技术进步非常大。今天一些表现出色的人脸识别算法在具有挑战性的数据集上获得了接近100%的成功率。

NIST FRVT 人脸遮挡效果：FALSE-NON MATCH RATE

来源：美国国家标准和技术研究所，2021 | 2022人工智能指数据报告

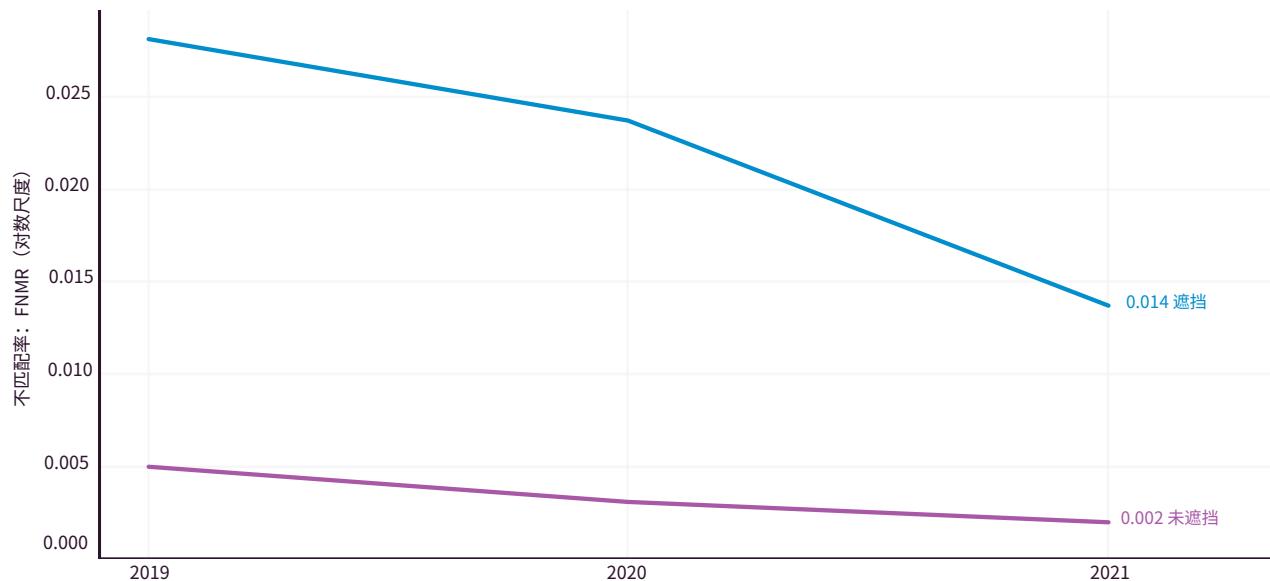
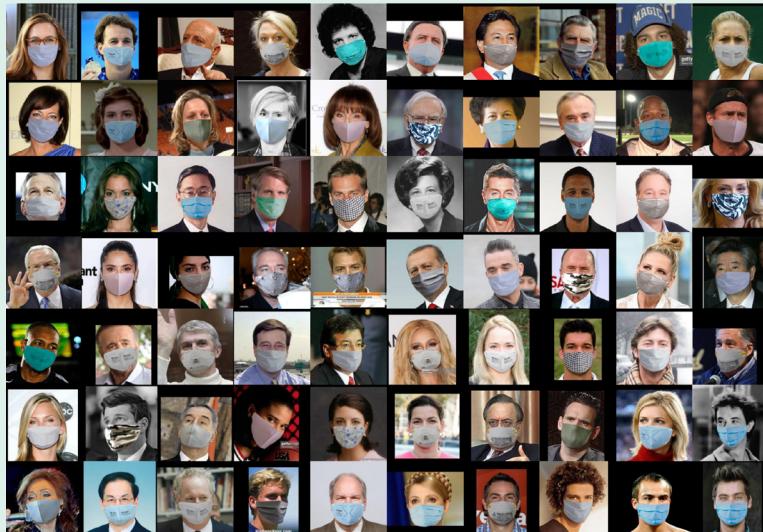


图 2.1.17



## 户外人脸检测数据库 Masked Labeled Faces in the Wild (MLFW)

2021年，来自北京邮电大学的研究人员发布了6000张遮挡人脸识别数据集，以应对大规模佩戴口罩所带来的新的识别挑战。



MASKED LABELED FACES IN THE WILD (MLFW) 数据库中遮挡人脸的样本示例

来源: [Wang等人, 2021年](#)

图 2.1.18

作为数据集发布的一部分，研究人员在包括他们的数据集在内的各种人脸识别数据集上运行了一系列现有的最先进的检测算法，以确定当人脸被遮挡时检测性能会下降多少。他们的评估表明，顶级方法在遮挡人脸上的表现比未被遮挡的差5至16个百分点。这些发现在一定程度上证实了FRVT人脸遮挡测试的结论。当人脸佩戴了口罩时，算法的性能会恶化，但程度不会太高。

### 最新检测方法在Masked Labeled Faces in the Wild (MLFW)中的效果：准确性

来源: Wang等人, 2021 | 2022人工智能指数报告

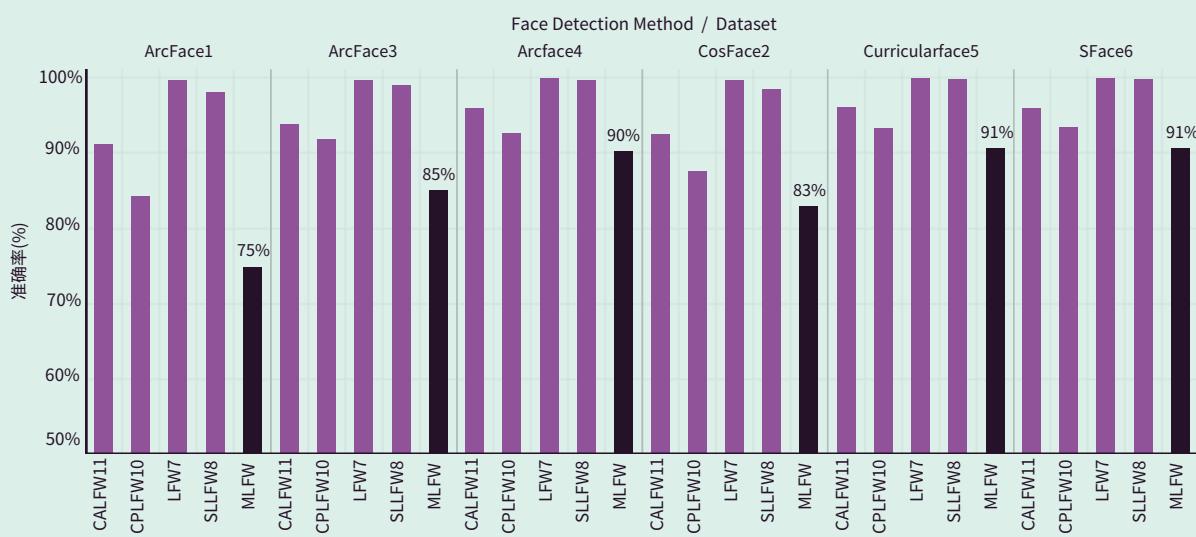


图 2.1.19



## 视觉推理

视觉推理技术用于评估人工智能系统在视觉和文本数据组合中的推理能力。视觉推理技能对于开发能够完成更广泛的推理任务的人工智能至关重要。现有的人工智能已经可以比人类更好地执行某些狭义的视觉任务，例如对图像进行分类、检测人脸，以及分割物体。但是，许多人工智能系统在面临更抽象的推理挑战任务时效果都不太好--例如，对图像中agent的行动或动机做出有效的推断（图2.1.20）。

### 视觉推理任务示例

来源: [Goyal et al., 2021](#)

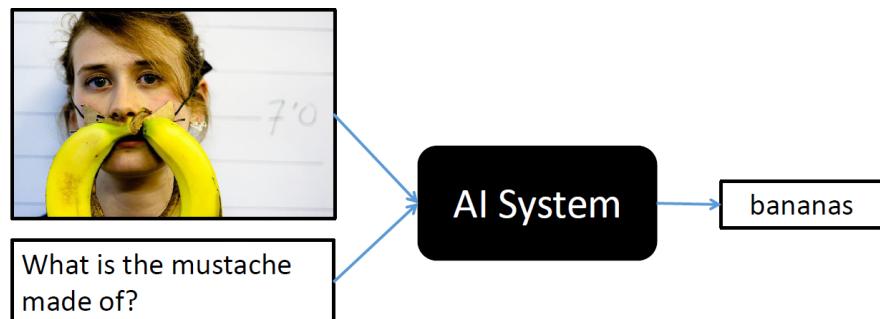


图 2.1.20

## 视觉问答 (Visual Question Answering, VQA) 挑战赛

在视觉问答挑战中，人工智能系统的任务是回答关于图像的开放式问题（图2.1.21）。为了较好的回答问题，人工智能系统必须能够对语言、视觉和常识性推理进行综合理解。

### 视觉问答 (VQA) 挑战中的问题样本

来源: [Goyal et al., 2017](#)



图 2.1.21



自VQA挑战赛开始以来的六年中，效果最好的系统性能提高了24.4个绝对百分点。在2015年，效果最好的系统

只能正确回答55.4%的问题（图2.1.22）。截至2021年，系统的最佳表现为79.8%--接近人类基线的80.8%。

### 视觉问答 (VQA) 挑战：准确性

来源：VQA挑战赛，2021年 | 2022人工智能指数报告

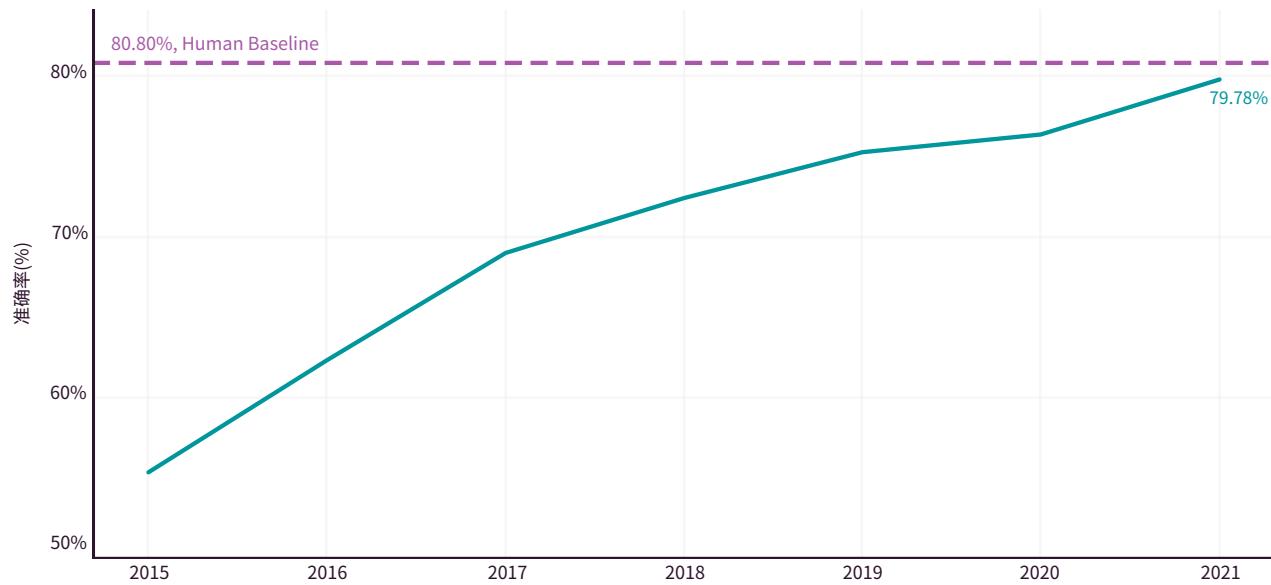


图 2.1.22



视频分析涉及对连续帧（视频）的推理或任务操作，而不是单帧（图像）。视频计算机视觉有广泛的用途，包括协助犯罪监视工作、体育分析、自动驾驶、机器人导航和人群监测等等。

## 2.2 计算机视觉-视频

### 行为识别

视频计算机视觉的一个基本子任务是行为识别：识别视频中的行为。人工智能系统面临的挑战是对行为进行分类，这些行为从简单的动作，如走路、挥手或站立，到更复杂的、包含多个步骤的活动，如准备沙拉（这需要人工智能系统识别并将切西红柿、洗菜、涂抹调味品等不连续的动作连在一起）。

### Kinetics-400, Kinetics-600, Kinetics-700

Kinetics-400、Kinetics-600和Kinetics-700是一系列用于视频行为识别基准的数据集。每个数据集包括650,000个来自YouTube的大规模、高质量的视频片段，这些视频包含了广泛的人类行为，并要求人工智能系统分别从400、600和700个可能的类别集合中对动作进行分类。其中，Kinetics-700系列引入一些更具挑战性的行为类别，包括倒酒、演奏双簧管和制作拿铁艺术咖啡等。

#### KINETICS数据集中的样本类别

来源：[Kay et al., 2017](#)



(a) headbanging



(b) stretching leg



(c) shaking hands



(d) tickling

图 2.2.1



截至2022年，有一个模型在所有三个Kinetics数据集中名列前茅，即，2022年1月发布的由谷歌研究院、密歇根州立大学和布朗大学合作的MTV。MTV在600系列上达到了89.6%的Top-1准确率，在400系列上达到了89.1%的准确率，在700系列上达到了82.20%的准确率（图2.2.2）。关于Kinetics的技术进步，最引人

注目的是数据集上的性能差距缩小得如此之快。2020年，Kinetics-400和Kinetics-700的性能差距为27.14个百分点。在短短的一年里，这个差距已经缩小到7.4个百分点，这意味着在较新的、较难的数据集上的表现比在较容易的数据集上的表现更快，并表明较容易的数据集已经开始趋于平稳。

#### KINETICS-400, KINETICS-600, KINETICS-700: TOP-1 ACCURACY

来源：Papers with Code, 2021; arXiv, 2021 | 2022人工智能指数报告

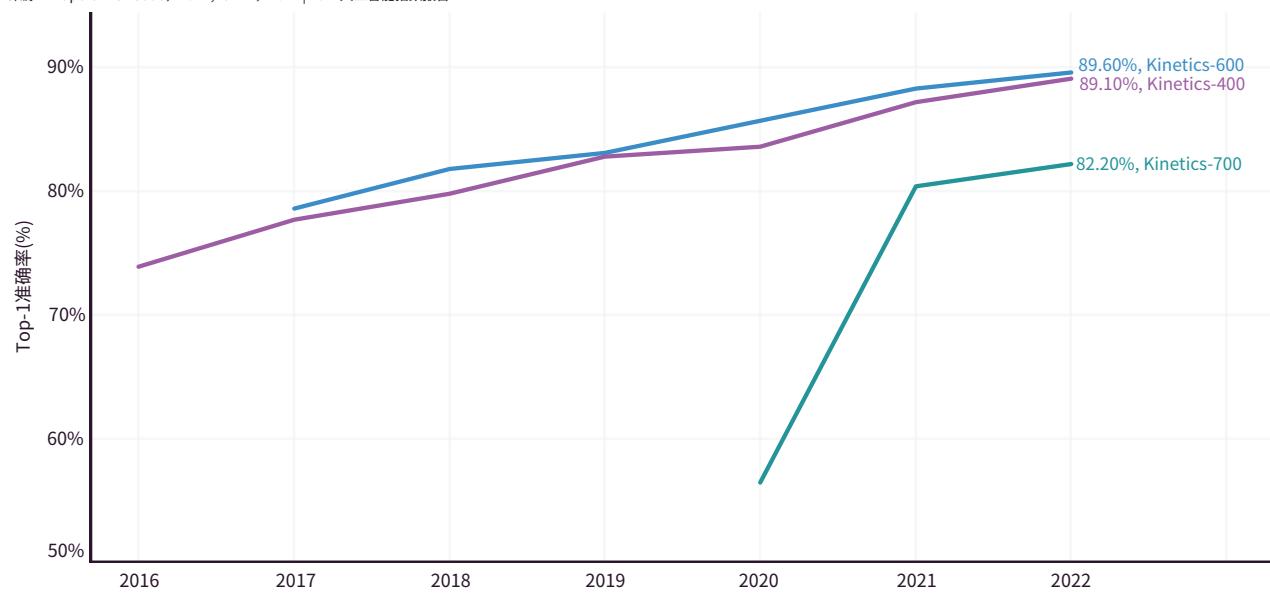


图 2.2.2



## ActivityNet：时间动作定位任务

ActivityNet是一个用于理解人类活动的视频数据集，它包含了700个小时的人类做200种不同活动的视频（跳远、遛狗、吸尘等）。对于一个人工智能系统来说，要成功完成ActivityNet的时间动作定位任务（Temporal Action Localization Task , TALT），它必须执行两个独立的步骤。(1) 定位（确定活动发生的精确时间间隔）；(2) 识别（分配正确的类别标签）。时间动作定位是

计算机视觉中最复杂和最困难的任务之一。TALT的性能是以平均精度来衡量的，得分越高说明精度越高。

截至2021年，由HUST-Alibaba开发的TALT上表现最好的模型得分为44.7%，比2016年挑战开始时的最高分提高了26.9个百分点（图2.2.3）。尽管随后每年都会公布最先进的任务结果，但改进的程度却越来越小。

ACTIVITYNET，时间动作定位任务：平均精度 (mAP)

来源：ActivityNet, 2021 | 2022人工智能指数报告

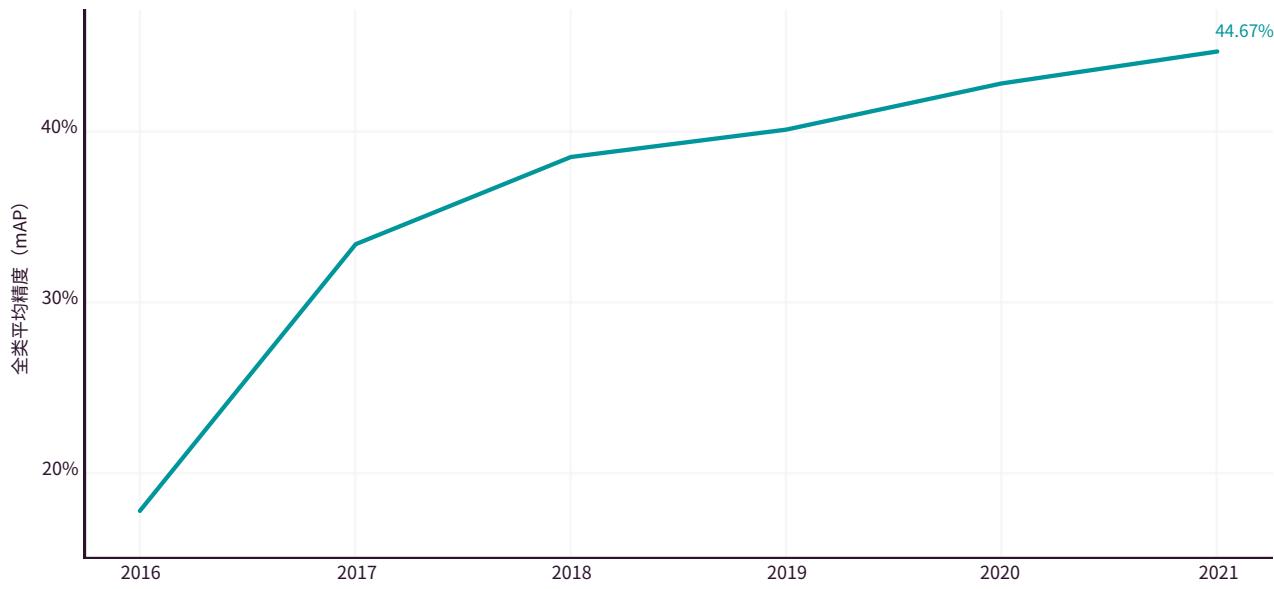


图 2.2.3



## 物体检测

物体检测是指识别图像中的物体的任务（图2.2.4）。在指导物体检测系统设计的过程中，存在着不同的优先级、速度和准确性理念。快速训练的系统可能更有效率，但准确性较低。而更准确的系统可能表现更好，但需要更长的时间来处理视频。这种速度和准确性之

间的权衡也反映在过去十年中不断提出的物体检测方法的类型中。其中有优先考虑速度的单阶段方法，如SSD、RetinaNet和YOLO，以及优先考虑准确性的双阶段方法，如Mask R-CNN、Faster R-CNN和Cascade R-CNN。

### 物体检测在AI系统中的表现

来源：[COCO, 2020](#)



图 2.2.4



## 语境中的常见对象 (COCO)

微软的Common Object in Context (COCO) 物体检测数据集包含超过328,000张图片，涉及80多个物体类别。有许多准确性指标都可以用于评估物体检测的性能，但为了保持一致性，本节和本报告的大部分内容都

考虑的是平均精度 (mean average precision, mAP50)。自2016年以来，COCO物体检测性能提高了23.8个百分点，今年的模型GLIP的平均精度为79.5%。<sup>2</sup> 图2.2.5展示了引入额外训练数据会如何改进物体检测的性能，这一变化与计算机视觉其他领域中的情况类似。

COCO-TEST-DEV: 平均精度(mAP50)

来源：Papers with Code, 2021; arXiv, 2021 | 2022人工智能指数报告

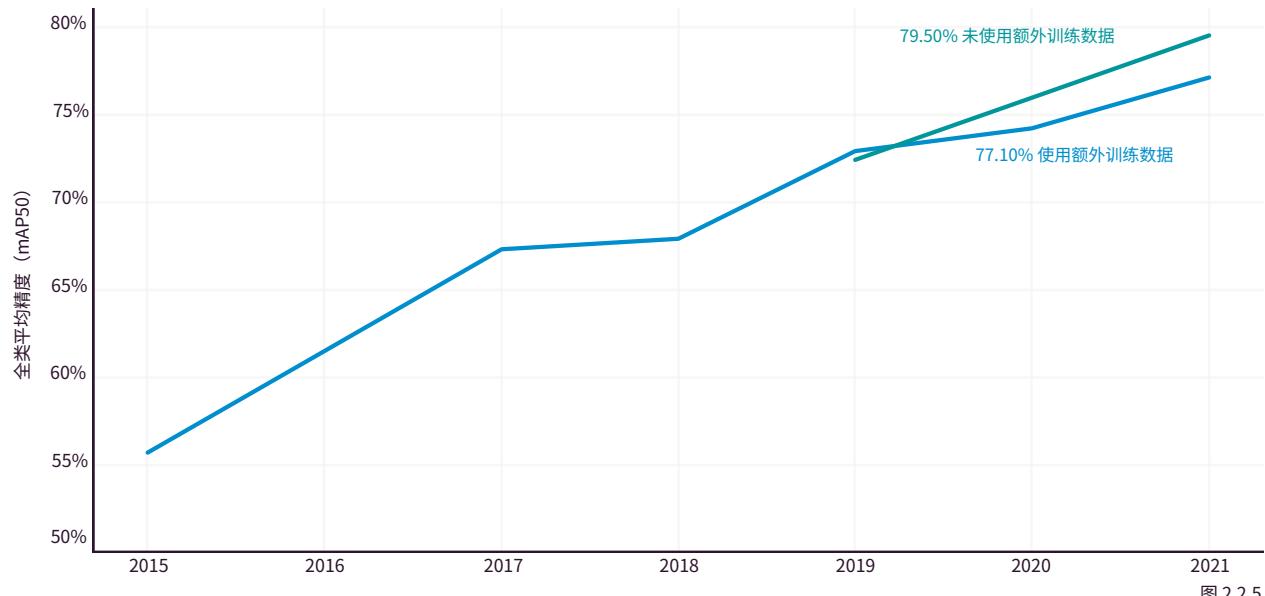


图 2.2.5

<sup>2</sup> GLIP（基础语言-图像预训练），一个旨在掌握语言语境视觉表征学习的模型，是由加州大学洛杉矶分校、微软研究院、华盛顿大学、威斯康星大学麦迪逊分校、微软云、微软人工智能和国际数字经济学院的研究人员合作完成的。



## 你只看一次 (YOLO)

You Only Look Once是一个开源的物体检测模型，它强调的是速度（推理延迟）而不是绝对的准确性。多年来，YOLO有不同的迭代，图2.2.6列出了YOLO物体检测器的性能与COCO数据集上绝对性能最好的检测器的对

比。自2017年以来，YOLO检测器在性能方面变得更好（提高了28.4个百分点）。其次，YOLO和表现最好的物体检测器之间的性能差距已经缩小了。2017年，该差距为11.7%，2021年减少到7.1%。在过去的五年里，物体检测器已经建成，既快又好。

技术现状 (SOTA) 与你只看一次 (YOLO) : 平均精度 (mAP50)

来源: arXiv, 2021; GitHub, 2021 | 2022人工智能指数报告

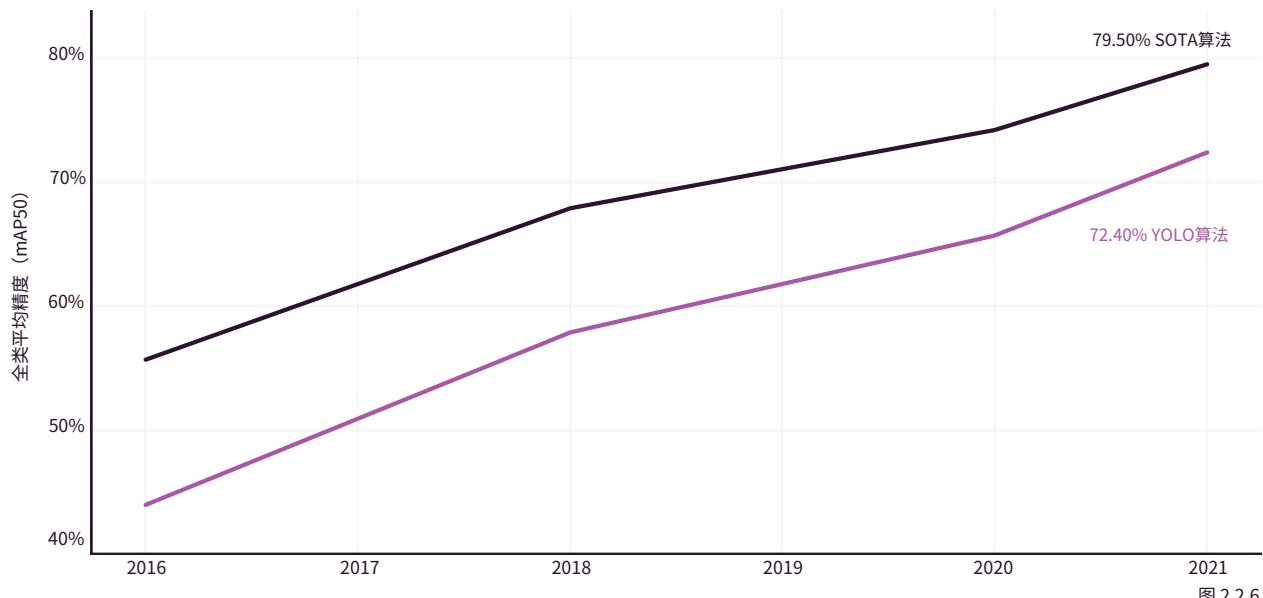


图 2.2.6

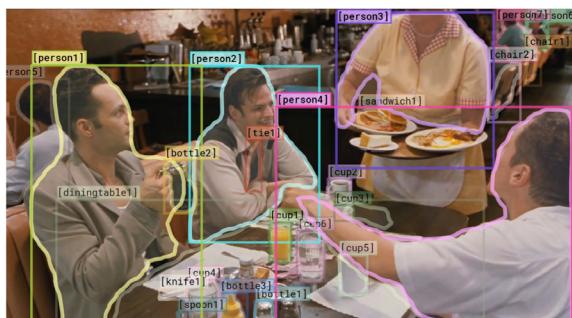


## 视觉常识推理 (Visual Commonsense Reasoning, VCR)

视觉常识推理挑战是一个相对较新的视觉理解基准。VCR要求人工智能系统回答关于图像所呈现的场景的挑战性问题，并提供其答案背后的推理（与VQA

### 视觉常识推理 (VCR) 挑战的问题样本示例

来源: Zellers et al., 2018



Why is [person4] pointing at [person1]?

- a) He is telling [person3] that [person1] ordered the pancakes.
- b) He just told a joke.
- c) He is feeling accusatory towards [person1].
- d) He is giving [person1] directions.

- a) [person1] has the pancakes in front of him.
- b) [person4] is taking everyone's order and asked for clarification.
- c) [person3] is looking at the pancakes and both she and [person2] are smiling slightly.
- d) [person3] is delivering food to the table, and she might not know whose order is whose.

图 2.2.7

VCR的性能是以Q->AR得分来衡量的，它汇总了机器在给定的多项选择题中选择正确答案 (Q->A)，然后给出选择答案的正确理由 (Q->R) 的能力。

自挑战赛首次亮相以来，人工智能系统在视觉常识推理方面的性能已经大大改进，尽管它们仍然远远落后于人

挑战不同，VQA只需要一个答案）。该数据集包含了290,000对多选题、答案和理由，这些问题来自11万个电影中的图像场景。图2.2.7展示了VCR中提出的问题的种类。

类的表现水平（图2.2.8）。在2021年底，VCR的最佳分数为72.0，这个分数代表了自2018年以来性能提升了63.6%。虽然自挑战启动以来，研究持续不断地取得进展，但改进的幅度越来越小，这表明可能需要发明新的技术来大幅提高性能。

### 视觉常识推理 (VCR) 任务: Q->AR得分

来源: VCR Leaderboard, 2021 | 图: 2022年人工智能指数报告

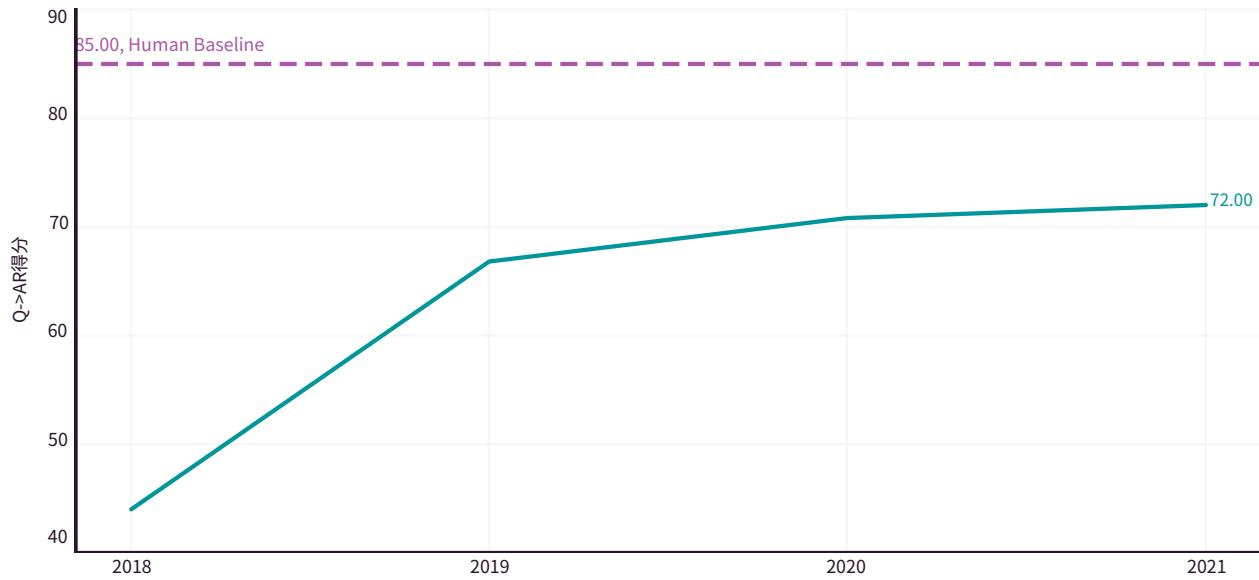


图 2.2.8



自然语言处理（NLP）是人工智能的一个子领域，其根源最早可以追溯到20世纪50年代。NLP涉及对能够阅读、生成和推理自然语言的系统的研究。NLP系统已从早年使用手写规则和统计方法，发展到现在结合计算语言学、基于规则的建模、统计学习和深度学习的系统。

本节着眼于NLP在几个语言任务领域的进展，包括：（1）英语语言理解；（2）文本总结；（3）自然语言推理；（4）情感分析；以及（5）机器翻译。在过去的十年里，NLP的技术进步非常大。采用深度神经网络式的机器学习方法，意味着许多人工智能系统现在可以比许多人类基线更好地执行复杂的语言任务。

## 2.3 语言

### 英语语言理解

英语语言理解挑战人工智能系统在各种情况下对英语语言的理解，如句子理解、是/否阅读理解、带逻辑推理的阅读理解等。

#### SuperGLUE

SuperGLUE是一个单一的数字指标，用于跟踪各种语言任务的技术进步（图2.3.1）。作为基准的一部分，人工智能系统在八个不同的任务（如回答是/否

问题，识别事件中的因果关系，以及进行常识性阅读理解）上进行测试，然后将它们在这些任务上的表现求取平均值得到一个分数。SuperGLUE是GLUE的后继者，GLUE是一个较早的基准，也对多个任务进行测试。在人工智能系统开始使GLUE指标饱和后，SuperGLUE于2019年5月发布，提出了对更高基准的需求。

#### 一套SUPERGLUE任务示例<sup>3</sup>

来源：[Wang et al., 2019](#)

**BoolQ** **Passage:** Barq's – Barq's is an American soft drink. Its brand of root beer is notable for having caffeine. Barq's, created by Edward Barq and bottled since the turn of the 20th century, is owned by the Barq family but bottled by the Coca-Cola Company. It was known as Barq's Famous Olde Tyme Root Beer until 2012.

**Question:** is barq's root beer a pepsi product    **Answer:** No

**CB** **Text:** B: And yet, uh, I we-, I hope to see employer based, you know, helping out. You know, child, uh, care centers at the place of employment and things like that, that will help out. A: Uh-huh. B: What do you think, do you think we are, setting a trend?

**Hypothesis:** they are setting a trend    **Entailment:** Unknown

**COPA** **Premise:** My body cast a shadow over the grass.    **Question:** What's the CAUSE for this?

**Alternative 1:** The sun was rising.    **Alternative 2:** The grass was cut.

**Correct Alternative:** 1

图 2.3.1

<sup>3</sup> 为简洁起见，本图只显示8项任务中的3项。



在SuperGLUE的排行榜上，SS-MoE模型以91.0分位居榜首（图2.3.2），这超过了SuperGLUE基准开发者给出的人类表现的89.8分。SuperGLUE的进展如此迅

速，这表明研究人员需要开发更复杂的自然语言任务套件，以挑战下一代人工智能系统。

### SuperGLUE：得分

来源：SuperGLUE Leaderboard, 2021 | 图：2022年人工智能指数报告

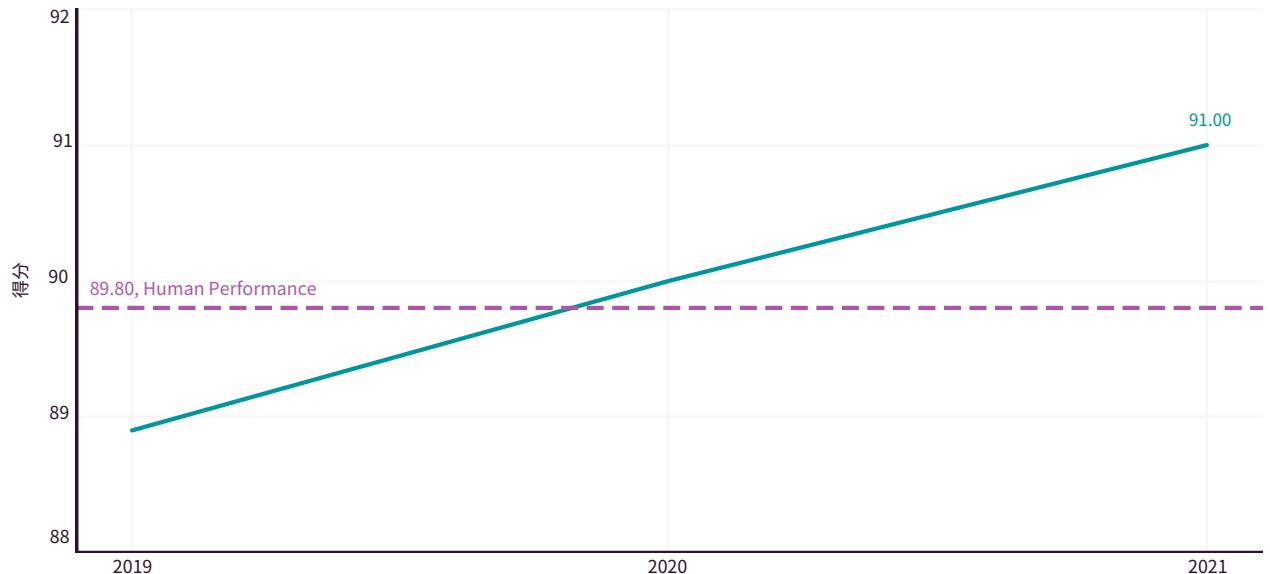


图 2.3.2

### 斯坦福大学问题回答数据集(SQuAD)

斯坦福大学问题回答数据集（Stanford Question Answering Dataset，SQuAD）是一个衡量阅读理解能力的基准。该数据集包括107,785个问答对，取自536篇维基百科文章。SQuAD的性能是由F1分数来衡量的，F1分数表征人工智能系统的答案和实际正确答案之间的平均重合度。分数越高，性能就越好。

与GLUE的情况一样，人工智能系统在SQuAD上的改进非常迅速，以至于在2016年推出SQuAD后仅两年，研究人员就发布了SQuAD 2.0。2.0版本包括更具挑战性的阅读理解任务，即一组50,000个无法回答的问题，这些问题是以一种看起来可以回答的方式来写的（图2.3.3）。

### SQuAD2.0新增更难的问题

来源：[Rajpurkar et al., 2018](#)

**Article:** Endangered Species Act

**Paragraph:** “... Other legislation followed, including the Migratory Bird Conservation Act of 1929, a **1937 treaty** prohibiting the hunting of right and gray whales, and the **Bald Eagle Protection Act of 1940**. These **later laws** had a low cost to society—the species were relatively rare—and little **opposition** was raised.”

**Question 1:** “Which laws faced significant **opposition**? ”

**Plausible Answer:** **later laws**

**Question 2:** “What was the name of the **1937 treaty**? ”

**Plausible Answer:** **Bald Eagle Protection Act**

图 2.3.3



2021年底，SQuAD 1.1和SQuAD 2.0的领先分数分别为95.7和93.2（图2.3.4）。尽管这些分数是最先进的，但与前一年的最高分相比，它们的改进幅度很小（0.4%和0.2%）。这两个SQuAD数据集都显示出一个

趋势，即在发布之后，极短的时间内就实现了超过人类性能的分数，然后就进入小幅度的、高原式的增长阶段。

### SQuAD 1.1和SQuAD 2.0：F1得分

来源: SQuAD 1.1和SQuAD 2.0, 2021 | 图: 2022年人工智能指数报告

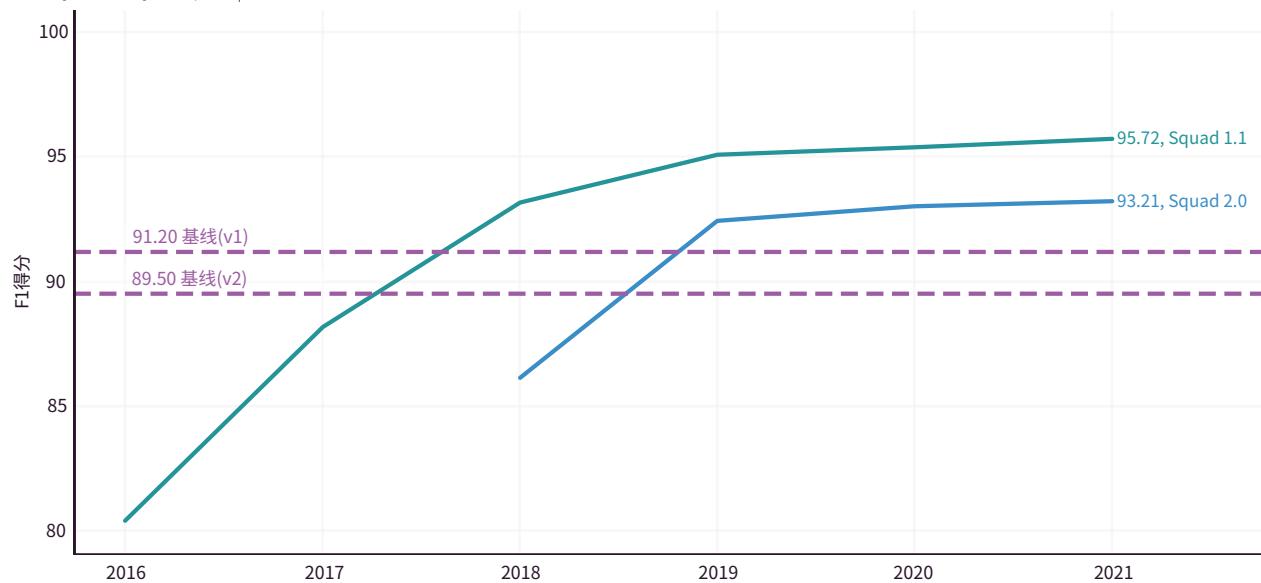


图 2.3.4

## 要求逻辑推理的阅读理解数据集 (Reading Comprehension Dataset Requiring Logical Reasoning, ReClor)

像SQuAD这样的基准进展停滞不前，表明NLP模型需要在更复杂的语言挑战上进行测试，比如ReClor提供的挑战。ReClor由新加坡国立大学的计算机科学家在2020年创建，它要求人工智能系统参与阅读理解，其中，必须要进行逻辑推理。ReClor数据集由LSAT的逻辑推理题组成，LSAT是美国和加拿大法学院的入学考试（图2.3.5）。

### ECLOR数据集示例

来源: Yu et al., 2020

#### Context:

In jurisdictions where use of headlights is optional when visibility is good, drivers who use headlights at all times are less likely to be involved in a collision than are drivers who use headlights only when visibility is poor. Yet Highway Safety Department records show that making use of headlights mandatory at all times does nothing to reduce the overall number of collisions.

**Question:** Which one of the following, if true, most helps to resolve the apparent discrepancy in the information above?

#### Options:

- A. In jurisdictions where use of headlights is optional when visibility is good, one driver in four uses headlights for daytime driving in good weather.
- B. Only very careful drivers use headlights when their use is not legally required.
- C. The jurisdictions where use of headlights is mandatory at all times are those where daytime visibility is frequently poor.
- D. A law making use of headlights mandatory at all times is not especially difficult to enforce.

**Answer:** B

图 2.3.5



ReClor上有两组问题，简单和困难，人工智能系统的准确度是根据他们回答正确的问题的百分比来判断的（图2.3.6）。尽管人工智能系统目前能够在简单的问题集上达到相对较高的性能水平，但在困难的问题集上效果却并不好。2021年，ReClor（困难集）上表现最好的模型得分为69.3%，比简单集上表现最好的模型大约低22.5个百分点。像ReClor这样的数据集表明，虽然NLP模型可以执行直接的阅读理解任务，但当这些任务与逻辑推理要求结合在一起时，它们会面临更多困难。

**尽管人工智能系统目前能够在简单的问题集上达到相对较高的性能水平，但它们在困难的问题集上的效果却不好。**

#### 需要逻辑推理的阅读理解数据集(Recor): 准确性

来源: ReClor Leaderboard, 2021 | 图: 2022年人工智能指数报告

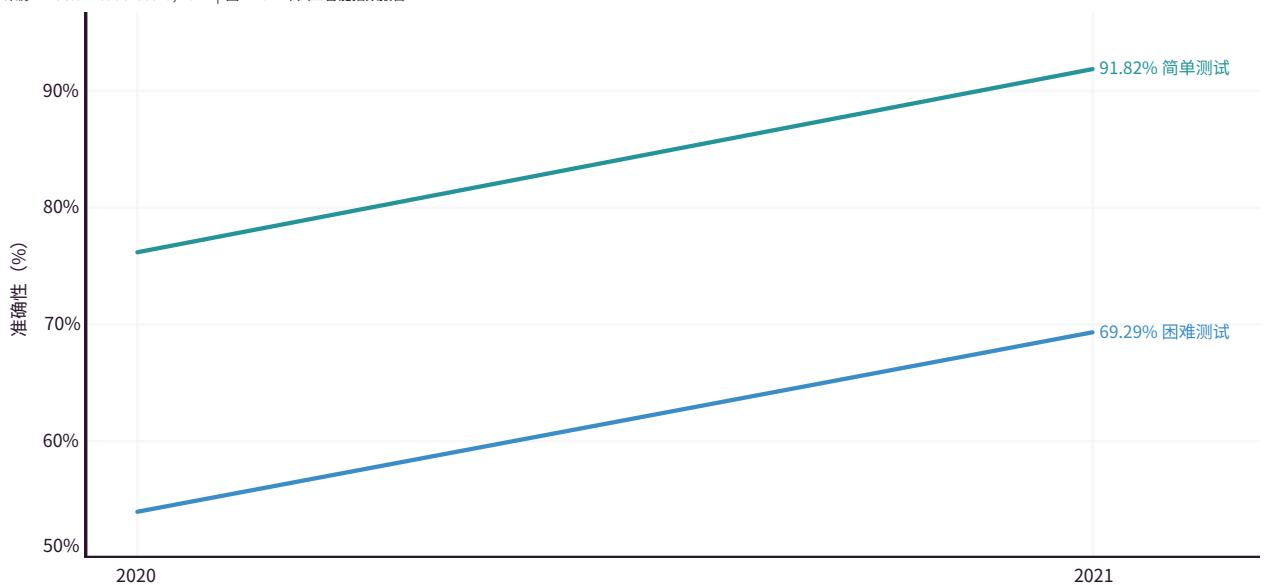


图 2.3.6



## 文本摘要

文本摘要是对一段文本进行分析，同时提取其核心内容的挑战。对文本生成摘要是文本分类、阅读理解和信息传播的一个重要组成部分；然而，如果由人类手动完成，这是一项时间和劳动密集型的任务。开发能够对文本提取功能性摘要的人工智能系统有许多实际的使用案例，包括帮助大学对学术论文进行分类，帮助律师生成案件摘要等等。

文本摘要的进展通常使用ROUGE（Recall-Oriented Understudy for Gisting Evaluation）为评分标准。ROUGE计算人工智能系统生成的摘要与人类生

成的参考摘要之间的重叠程度。ROUGE得分越高，重合度越高，总结越准确。

### arXiv

ArXiv是一项文本摘要基准数据集，包含了来自开放的科学论文库arXiv的27,770多篇论文。在开始对arXiv进行基准测试的五年中，人工智能文本摘要模型的性能提高了47.1%（图2.3.7）。然而，正如其他自然语言基准的情况一样，近年来针对ArXiv的研究进展趋于平稳。

### ARXIV: ROUGE-1

来源：Papers with Code, 2021; arXiv, 2021 | 图：2022年人工智能指数报告

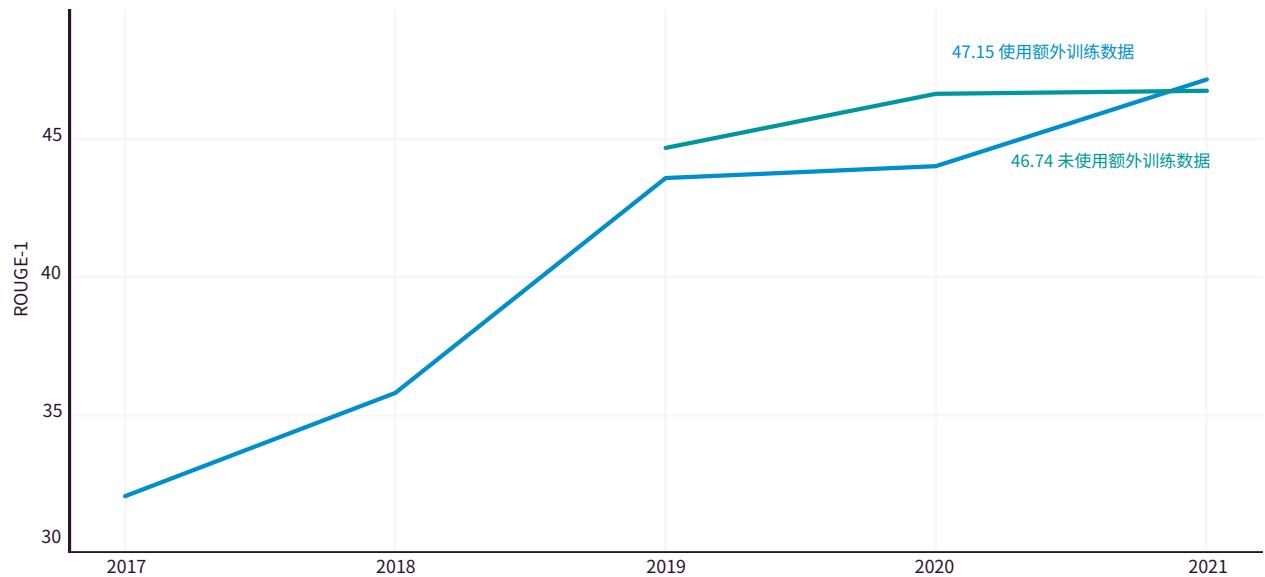


图 2.3.7



## PubMed

PubMed是一个文本摘要数据集，由PubMed科学论文数据库中的19,717篇科学出版物组成。PubMed上的进展进一步验证了arXiv上的趋势：自2017年以来，在文本分类上的效果有了明显的改进（34.6%），但最近改进

的速度已经放缓（图2.3.8）。2021年，PubMed上表现最好的模型是HAT（hierarchical attention transformer model，分层注意力转化器模型），由Birch AI和华盛顿大学的研究人员创建。

### PUBMED: ROUGE-1

来源：Papers with Code, 2021; arXiv, 2021 | 图：2022年人工智能指数报告

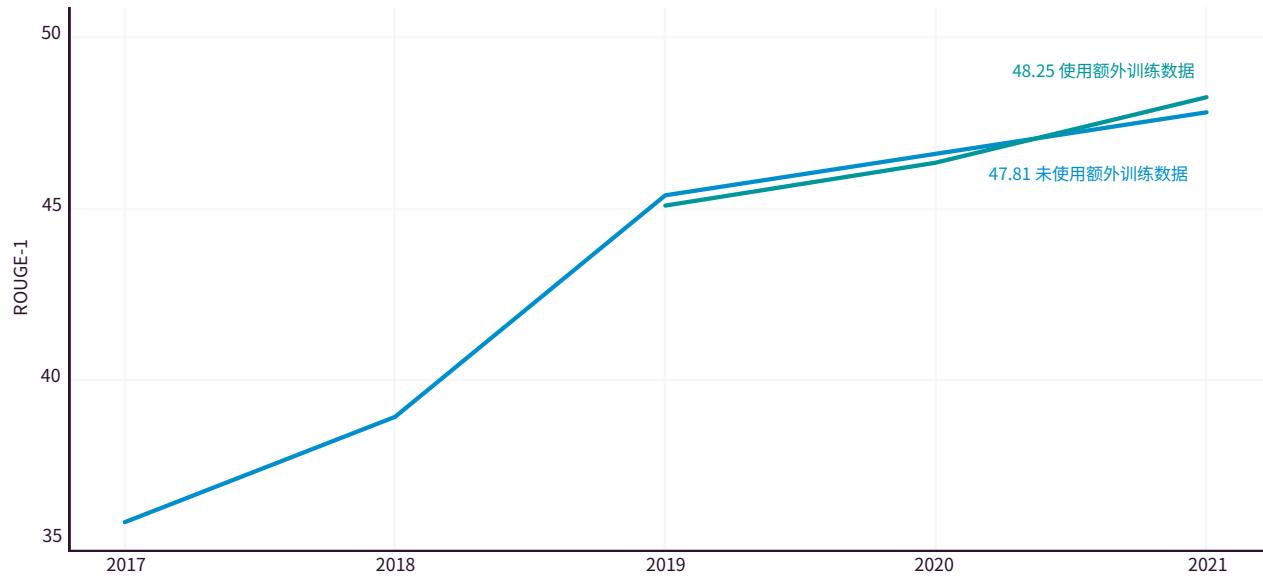


图 2.3.8



## 自然语言推理

自然语言推理是指在给定的前提下，确定一个假设是真的（entailment）、假的（contradiction），还是未确定的（neutral）的任务。这项技能也被称为文本的包含性，因为它需要确定一个特定的前提是否在逻辑上包含一个假设。自然语言推理需要语言处理技能，如命名实体识别（理解你看到的词），以及能够使用常识性知识来区分合理和不合理的推论。

### 斯坦福大学自然语言推理中的问题和标签（SNLI）

来源：[Bowman et al., 2015](#)

A man inspects the uniform of a figure in some East Asian country.	<b>contradiction</b> C C C C C	The man is sleeping
An older and younger man smiling.	<b>neutral</b> N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	<b>contradiction</b> C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	<b>entailment</b> E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	<b>neutral</b> N N E C N	A happy woman in a fairy costume holds an umbrella.

图 2.3.9



在SNLI上表现最好的模型是Facebook AI USA的EFL，在2021年4月的得分是93.1%（图2.3.10）。

#### 斯坦福大学自然语言推理（SNLI）：准确性

来源：Papers with Code, 2021; arXiv, 2021 | 图：2022年人工智能指数报告

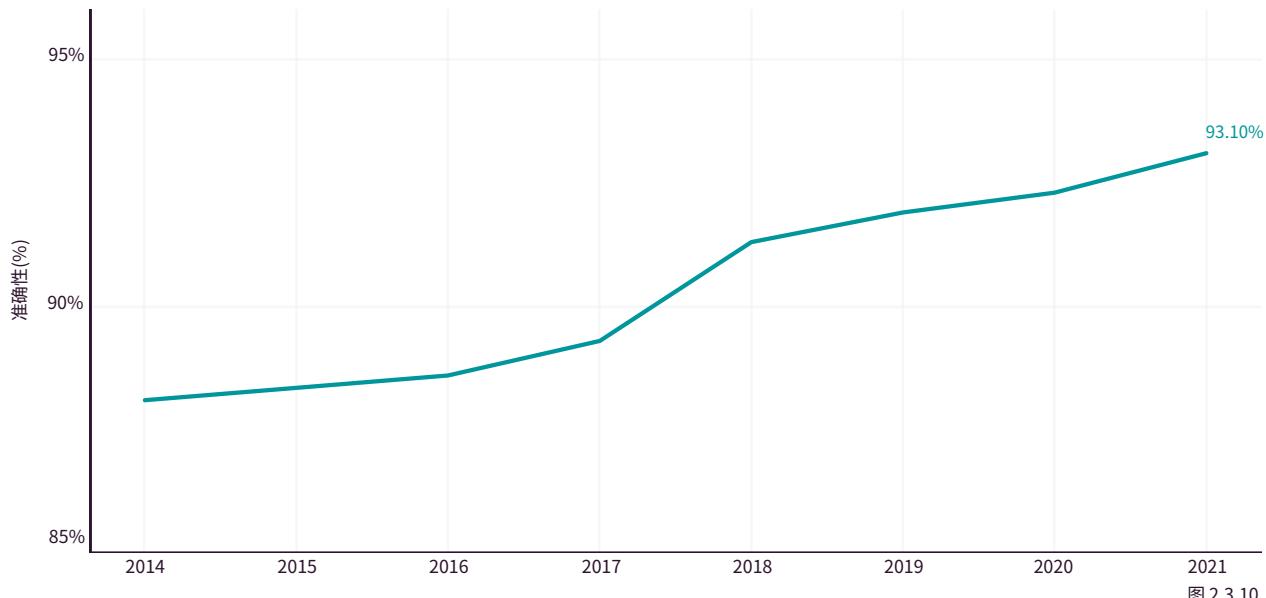


图 2.3.10

#### 归纳自然语言推理(Abductive Natural Language Inference, aNLI)

归纳自然语言推理是一种更困难的文本推理任务。归纳推理要求从有限的信息和不确定的前提中得出最合理的结论。例如，如果Jenny下班后发现她的家一片狼藉，然后想起她忘了开窗，她可以合理地推断出是小偷闯入并造成了混乱。<sup>4</sup>尽管归纳被认

为是人类相互交流的基本要素，但很少有研究关注人工智能系统的归纳能力。

ANLI是艾伦人工智能研究所在2019年创建的一个归纳自然语言推理的新基准，有17万个前提和假设对。图2.3.11给出了数据集中包含的语句类型示例。

#### 归纳自然语言推理（ANLI）中的示例

来源：[艾伦人工智能研究所, 2021](#)

**Obs1:** It was a gorgeous day outside.

**Obs2:** She asked her neighbor for a jump-start.

**Hyp1:** Mary decided to drive to the beach, but her car would not start due to a dead battery.

**Hyp2:** It made a weird sound upon starting.

图 2.3.11

<sup>4</sup> 这个归纳常识推理的示例取自Bhagavatula et al. (2019)，这是第一篇研究人工智能系统基于语言的归纳推理能力的论文。



自2019年以来，人工智能在归纳常识推理上的表现提高了7.7个百分点；然而，顶级人工智能系统虽然接近，但无法达到人类的表现水平（图2.3.12）。因此，归纳推理对于人工智能系统来说仍然是一项极具挑战性的语言任务。

#### 归纳自然语言推理(aNLI): 准确性

来源：艾伦人工智能研究所，2021 | 图：2022年人工智能指数据报告

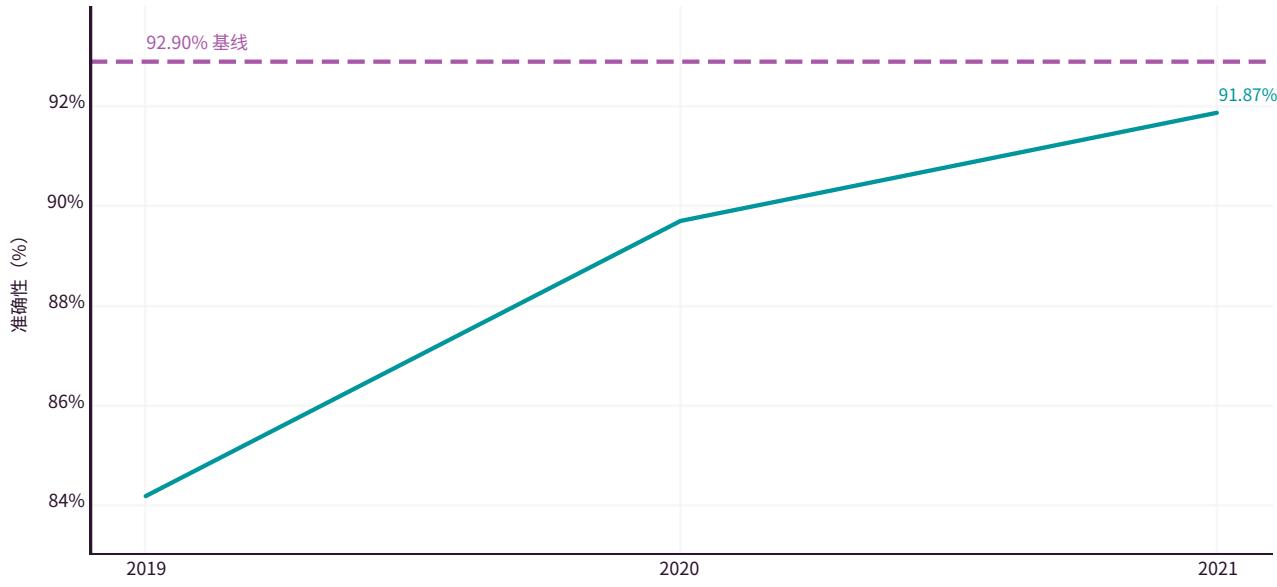


图 2.3.12

## 情感分析

情感分析是使用NLP技术来识别给定文本的情感（非常消极、消极、中立、积极、非常积极）的任务。如果句子的措辞清晰明了，例如“我不喜欢冬天的天气（I dislike winter weather）”，那么情绪分析就很简单了。然而，当人工智能系统遇到有翻转结构或否定词的句子时，情感分析就会变得更有挑战性，例如“说不喜欢冬天的天气并不是我真正的事情是完全不准确的。（to say that disliking winter weather is not

really my thing is completely inaccurate.）”

情绪分析有许多商业用例，例如，解析客户评论和现场调查回复，识别客户的情绪状态等等。

#### SemEval 2014 Task 4 Sub Task 2

SemEval 2014 Task 4 Sub Task 2是一个情感分析的基准，要求机器进行情感分析。这项具体任务测试人工智能系统是否能够识别与文本的特定方面相关的情感，而不是整个句子或段落的情感（图2.3.13）。

#### SEMEVAL TASK 样本示例

来源：[Pontiki et al., 2014](#)

For example:

“I loved their **fajitas**” → {fajitas: *positive*}  
“I hated their **fajitas**, but their **salads** were great” → {fajitas: *negative*, salads: *positive*}  
“The **fajitas** are their first plate” → {fajitas: *neutral*}  
“The **fajitas** were great to taste, but not to see” → {fajitas: *conflict*}

图 2.3.13



SemEval数据集由7,686条针对餐厅和笔记本电脑的评论组成，这些评论的情感极性由人类来评定。在SemEval中，人工智能系统的任务是为文本的特定部分分配正确的情感标签，其性能以其正确分配标签的百分比来衡量。

在过去的七年中，人工智能系统在情感分析方面已经获得了很大的进步。截至去年，表现最好的系统在10次中有9次能正确估计情绪，而在2016年，它们10次中只有7次能正确估计。截至2021年，SemEval的最好分数为88.6%，由华南师范大学和Linklogis有限公司的中国研究人员组成的团队实现。(图2.3.14)。

#### SEMEVAL 2014 TASK 4 SUB TASK 2: 准确性

来源：Papers with Code, 2021; arXiv, 2021 | 图：2022年人工智能指数报告

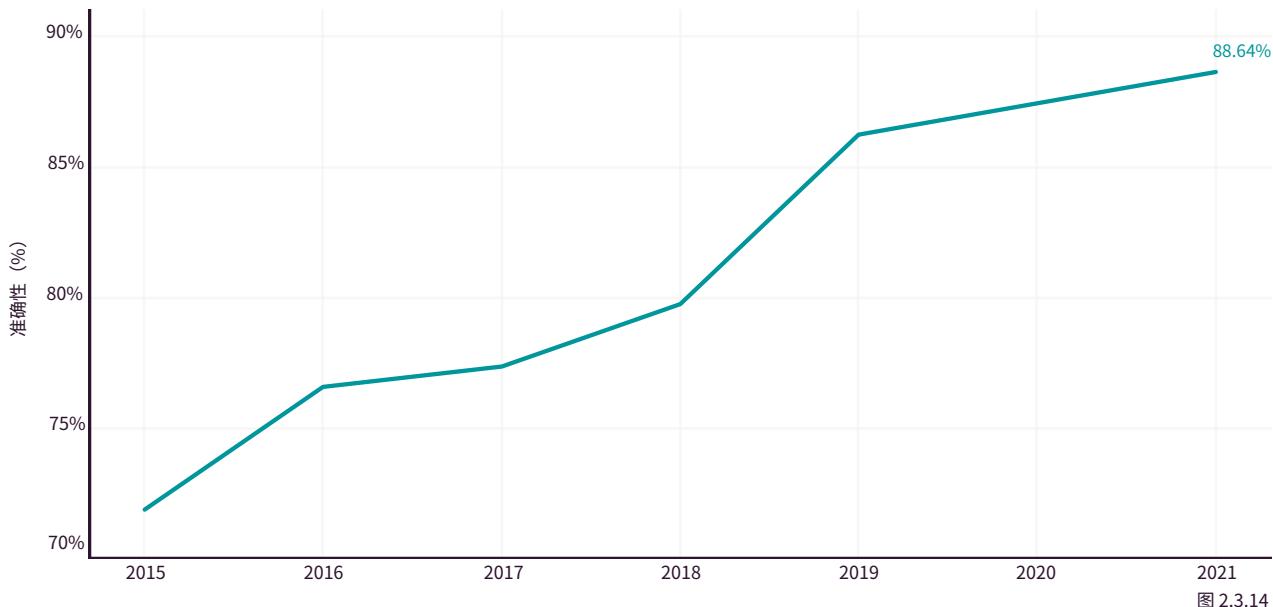


图 2.3.14

## 机器翻译

机器翻译主要研究人工智能软件如何翻译语言。在需要精通多国语言的领域中，机器翻译可以发挥极大的作用。例如，欧盟需要将其所有跨国政策文件翻译成其成员国的24种语言。使用机器翻译可以节省时间，提高效率，并确保更稳定的结果。

自2017年以来，神经网络成为了机器翻译的主流技术。与之前的技术不同，神经网络翻译器从一系列先前的翻译任务中学习，并预测一连串单词的可能性。神经网络模型已经彻底改变了机器翻译领域，不仅因为它们不需要人类监督，而且还因为它们生成了最准确的翻译。因此，它们已经在搜索引擎和社交网络中广泛部署。



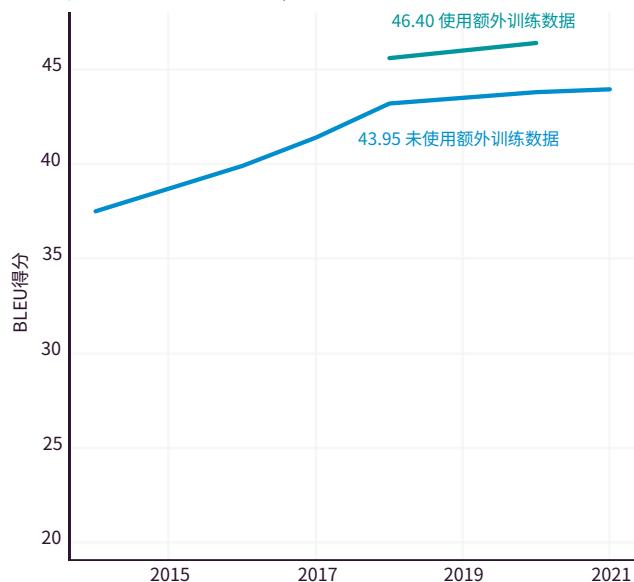
## WMT 2014, 英语-德语和英语-法语

在计算语言学协会 (ACL) 2014年会议上首次推出的 WMT 2014系列数据集包括不同类型的翻译任务，包括英法和英德语言对之间的翻译。机器的翻译能力是通过双语评估 (Bilingual Evaluation Understudy, BLEU) 得分来衡量，它比较了机器翻译的文本与人类生成的参考翻译的匹配程度。分数越高，翻译的效果就越好。

英法和英德WMT 2014基准都显示了过去十年中人工智能机器翻译技术的巨大进步（图2.3.15）。自开始提交以来，英法翻译能力提高了23.7%，英德翻译能力提高了68.1%。相对而言，尽管英德语对的性能改进更为显著，但英法翻译的绝对翻译能力仍在有意义地提高。

### WMT2014, 英法双语：BLEU得分

来源：Papers with Code, 2021; arXiv, 2021 | 图：2022年人工智能指数报告



### WMT2014, 英德双语：BLEU SCORE

来源：Papers with Code, 2021; arXiv, 2021 | 图：2022年人工智能指数报告



图 2.3.15



## 商业上可用的MT系统的数量

对机器翻译愈发增加的兴趣同样反映于在谷歌翻译等商业机器翻译服务的崛起。根据Intento的数据，自2017年以来，市场上的商业机器翻译的数量增加了近五倍

(图2.3.16)。2021年还有三个开源的机器翻译服务(M2M-100、mBART和OPUS)推出。公开可用的高性能机器翻译服务的出现，说明这种服务的可及性越来越高，这对任何经常依赖翻译的人来说都是个好消息。

### 独立的机器翻译服务的数量

来源：Intento，2021 | 图：2022年人工智能指数报告

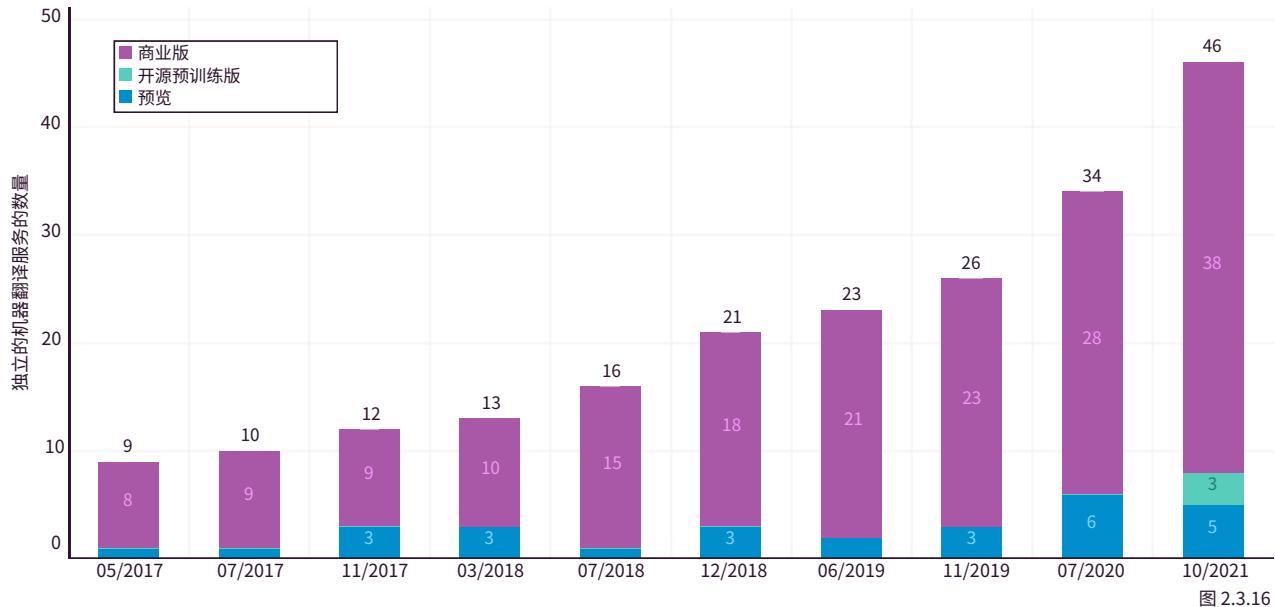


图 2.3.16



人工智能研究的另一个重要领域是人类语音的分析、识别和合成。在这一人工智能子领域，人工智能系统通常是根据其识别语音和识别单词并将其转换为文本的能力来评估的；同时还可以将识别说话人，以及识别确定说话的个人作为评估依据。现代家庭辅助工具，如Siri，是商业化应用人工智能语音技术的众多例子之一。

## 2.4 语音

### 语音识别

语音识别是训练机器识别口语并将其转换成文本的过程。这一领域的研究始于20世纪50年代的贝尔实验室，当时世界上出现了自动数字识别机（名为“Audrey”），它可以识别人类说出的从0到9的任何数字。从那时起，语音识别技术有了长足的发展，在过去的十年里，深度学习技术和丰富的语音识别数据集都促使这项技术更好的发展。

### 语音转录：LibriSpeech（Test Clean和Test Other数据集）

2015年推出的LibriSpeech是一个语音转录数据库，包含约1000小时的16kHz英语语音，取自有声读物的集合。LibriSpeech要求人工智能系统将语音转录为文本，然后根据单词错误率，或他们未能正确转录的单词

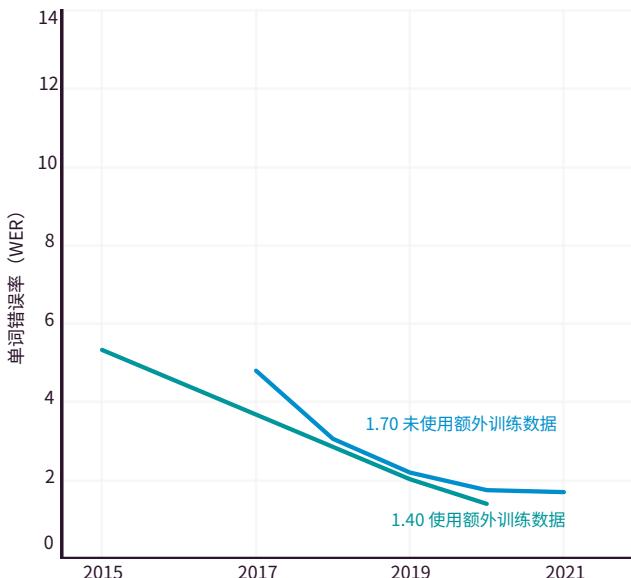
的百分比进行衡量。

LibriSpeech被细分为两个数据集。首先是LibriSpeech Test Clean，其中包含质量较高的录音。Test Clean的表现展示了人工智能系统在理想条件下的语音转录能力。其次是LibriSpeech Test Other，其中包含质量较差的录音，Test Other的表现则展示了人工智能系统在声音质量不理想的环境中的转录性能。

人工智能系统在LibriSpeech上的表现令人难以置信，以至于现阶段的研究进展似乎已经趋于平稳（图2.4.1）。2021年，在Test Clean数据集上没有新的SOTA（最先进的）结果出现，这说明顶级系统的错误率已经很低，只有1.4%。对于表现最好的转录模型，每听到100个词，他们就能正确转录99个。

LibriSpeech Test Clean：单词错误率（WER）

来源：Papers with Code, 2021; arXiv, 2021 | 图：2022年人工智能指数报告



LibriSpeech Test Other：单词错误率（WER）

来源：Papers with Code, 2021; arXiv, 2021 | 图：2022年人工智能指数报告

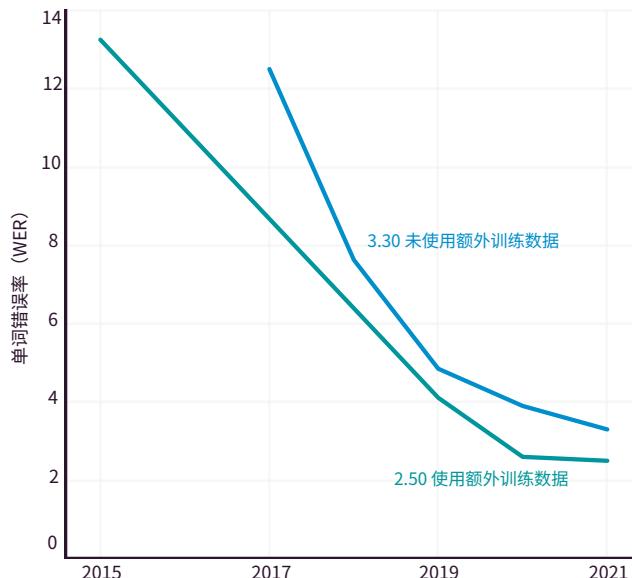


图 2.4.1



模型在Test Other数据集上的表现弱于Test Clean的情况虽然可以理解，但在现阶段仍然是相对较差的。W2V-BERT模型在 "Test Other" 中取得了最优结果，这是麻省理工学院和谷歌大脑的合作项目，其错误率为2.0%。

## VoxCeleb

VoxCeleb是一个大规模的人类语音视听数据集，用于说话人识别，也就是将某些语音与某个特定的人相匹配的任务。每年，VoxCeleb的制作者都会举办一个说话人验证挑战赛。在VoxCeleb挑战赛上得分低或错误率相同，表明人工智能系统在将语音归于特定个人方面很少出错。<sup>5</sup> 图2.4.2给出VoxCeleb-1（原始VoxCeleb数据集）上的性能随时间的变化。自2017年以来，VoxCeleb的性能不断改进。曾经报告错误率为7.8%的系统，其现在报告的错误率已经低于1.0%。

### Voxceleb: 等错误率 (EQUAL ERROR RATE, EER)

来源: VoxCeleb, 2021 | 图: 2022年人工智能指数报告

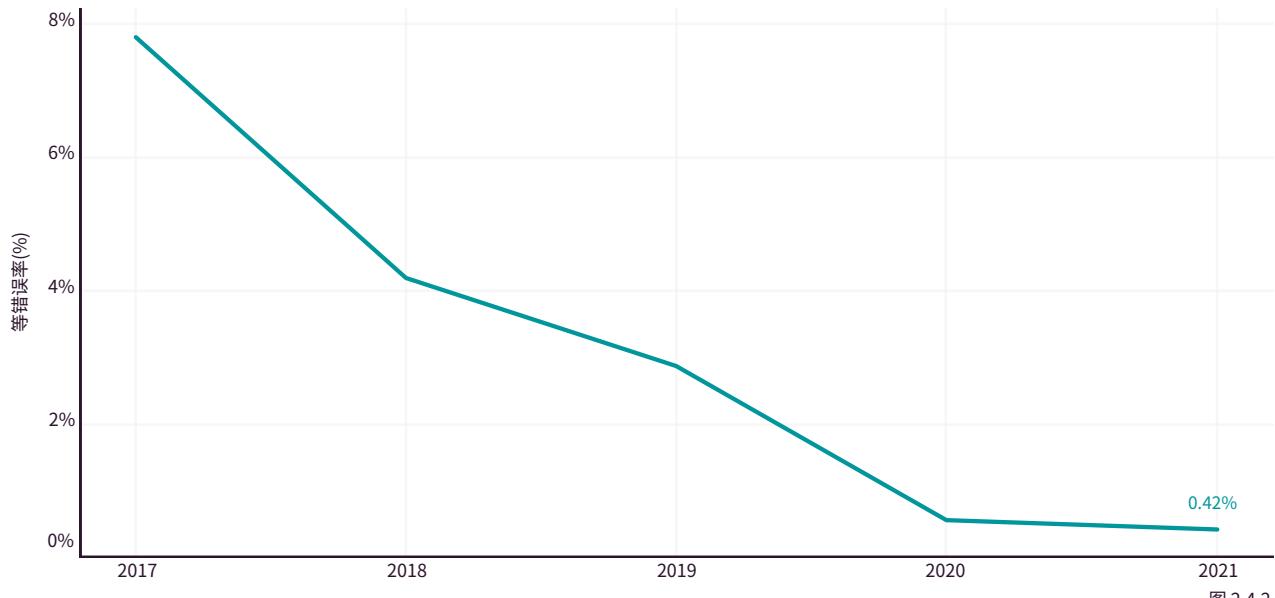


图 2.4.2

<sup>5</sup> 等错误率 (EER) 不仅是对假阳性率 (分配一个劣质标签) 的衡量，也是对假阴性率 (未能分配正确标签) 的衡量。



推荐（Recommendation）是指向用户提议其可能感兴趣的项目，如观看的电影、阅读的文章或购买的产品。推荐系统对企业至关重要，如亚马逊、Netflix、Spotify和YouTube。例如，人工智能领域最早的公开推荐竞赛之一是Netflix Prize，该竞赛于2009年举办，它向计算机科学家提出挑战，要求他们开发能够根据以前提交的评分准确预测用户对电影评分的算法。

## 2.5 推荐

### 商业推荐：MovieLens 20M

MovieLens 20M数据集包含了来自138,000名用户的27,000部电影的约2,000万个电影评分。这些评分来自MovieLens（一个电影推荐平台）。该数据集中，人工智能系统面临的挑战是，它们是否能够根据用户之前提交的评分来预测用户的电影偏好。在MovieLens上用于跟踪性能的指标是归一化折扣累积增益（Normalized Discounted Cumulative Gain, nDCG），这是一个衡

量排名质量的指标。nDCG得分越高，意味着人工智能系统提供的推荐越准确。

自2018年以来，顶级模型现在在MovieLens 20M上的表现与2018年的表现相比大约提升了5.2%（图2.5.1）。2021年，MovieLens 20M上最先进的系统发布的nDCG为0.448，这一成绩是由来自布拉格捷克技术大学的研究人员实现的。

MOVIELENS 20M: 归一化折扣累积增益@100 (nDCG@100)

来源: Papers with Code, 2021; arXiv, 2021 | 图: 2022年人工智能指数报告

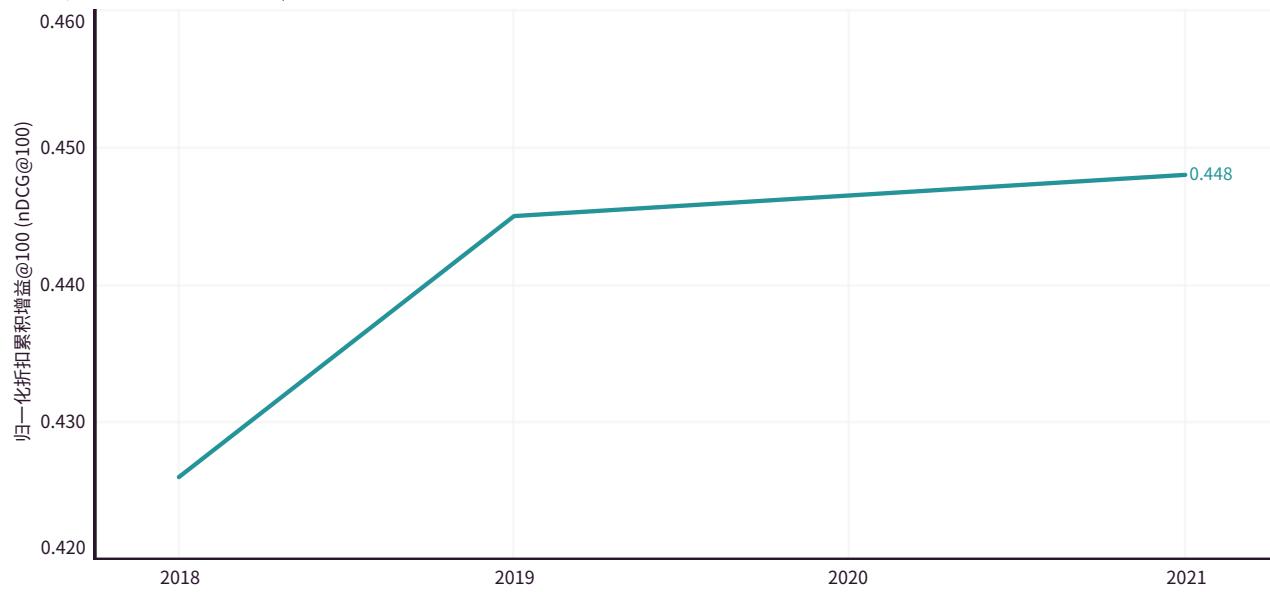


图 2.5.1



## 点击率预测 (Click-Through Rate Prediction) : Criteo

点击率预测是一项预测网站上物体受喜爱可能性的任务，例如一项广告被点击的可能性。2014年，在线广告平台Criteo发起了一个公开的点击率预测挑战赛。作为这个挑战赛数据集的一部分内容包括了在24天内显示的100万个广告的信息、这些广告是否被点击以及关于其特征的额外信息。自从该竞赛启动以来，Criteo数据集已被广泛用于测试推荐系统。在Criteo上，系统的性能是根据曲线下面积 (area under the curve, AUC) 来评估的。更高的AUC意味着更好的点击预测率和更强大的推荐系统。

在Criteo上的表现也表明，推荐系统在过去十年中一直在缓慢而稳定地改进。去年的顶级模型（新浪微博公司的MaskNet）在Criteo的表现比2016年的顶级模型高出1.8%。1.8%的改进从绝对值上看可能很小，但在商业世界里，这已经是一个相当有价值的幅度了。

Criteo和MovieLens基准的一个局限性在于，它们主要是对推荐技术的学术层面的评估（图2.5.2）。大多数关于推荐的研究工作都发生在商业环境中。鉴于公司有动力保持其推荐改进的专有性，本节中包括的学术指标可能不是推荐技术进步的完整衡量标准。

CRITEO: 曲线下面积得分 (AUC)

来源: Papers with Code, 2021; arXiv, 2021 | 图: 2022年人工智能指数报告

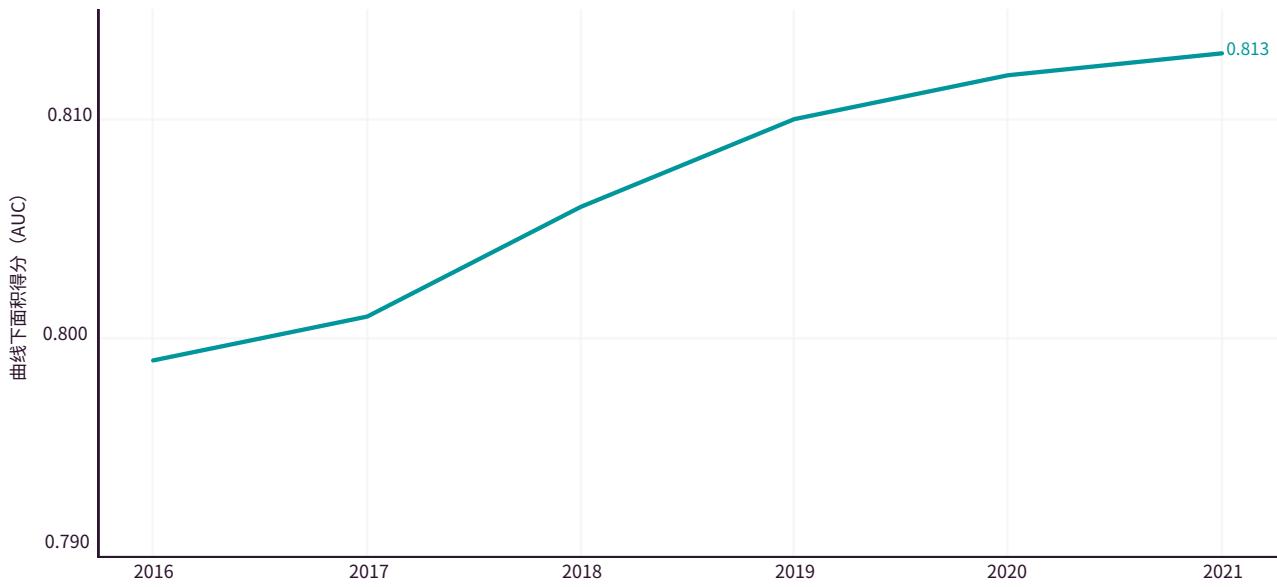


图 2.5.2



在强化学习中，人工智能系统会受训以通过交互式地学习其先前的行为来最大限度地提高某一特定任务的性能。研究人员可以通过在系统达到预期目标时对其进行奖励，而在失败时对其进行惩罚来训练系统进行优化。系统尝试用不同的策略序列来解决他们的既定问题（例如，下棋或在迷宫中导航），并选择使其奖励最大化的策略。

每当DeepMind的AlphaZero等程序在围棋和国际象棋等游戏中展现出超人的表现时，强化学习就成为一次新闻。然而，强化学习在任何商业领域都是有用的，在这些领域中，计算机智能体需要从以往的经验中学习，从而最大限度地实现一个目标。强化学习可以帮助自动驾驶汽车改变车道，帮助机器人优化制造任务，或者帮助时间序列模型预测未来事件。

## 2.6 强化学习

### 强化学习环境

强化学习环境是一个计算机平台，其中，智能体需要在一个确定的任务上实现其性能最大化。与其他需要在数据集上进行训练系统的人工智能任务不同，强化学习需要人工智能系统有一个可以测试各种策略的环境，并在这个过程中确定能使奖励最大化的策略集。

### 街机学习环境：Atari-57

2013年推出的街机学习环境（ALE）是一个包含各种Atari 2600游戏环境（如 "Pac-Man"、"Space Invaders" 和 "Frogger"）的界面，智能体会在不同游戏的挑战下不断优化其性能。为了能够进行标准化的比较，研究人员通常报告人工智能系统在一套57个游戏中ALE的平均性能。在现有的各种衡量性能的指标中，最常见的指标之一是人类归一化平均分数（mean human-normalized score）。0%的结果代表随机表现，100%的分数则代表人类的平均表现。人类归一化平均分数就是人工智能系统取得的平均的人类归一化分数。

创建高性能和高效率的强化学习模型是强化学习商业部署的一个重要步骤。

2019年底，DeepMind的MuZero算法在Atari-57上实现了最先进的性能。MuZero不仅在Atari-57上的表现优于之前表现最好的模型48.3%，而且还在围棋上创造了新的世界纪录，在国际象棋和象棋上实现了超过人类的表现。

2021年，来自清华大学和字节跳动的研究人员推出了GDI-H3模型，该模型在Atari-57上的性能超过了MuZero（几乎翻了一倍）（图2.6.1）。此外，GDI-H3用较少的训练就达到了这个性能。它只用了2亿个训练帧，而MuZero用了200亿个。GDI-H3的效果是其两倍，效率是其一百倍。创建高性能和高效率的强化学习模型是强化学习商业部署中的重要一步。



### ATARI-57: 人类归一化平均分数

来源: Papers with Code, 2021; arXiv, 2021 | 图: 2022 人工智能指数报告

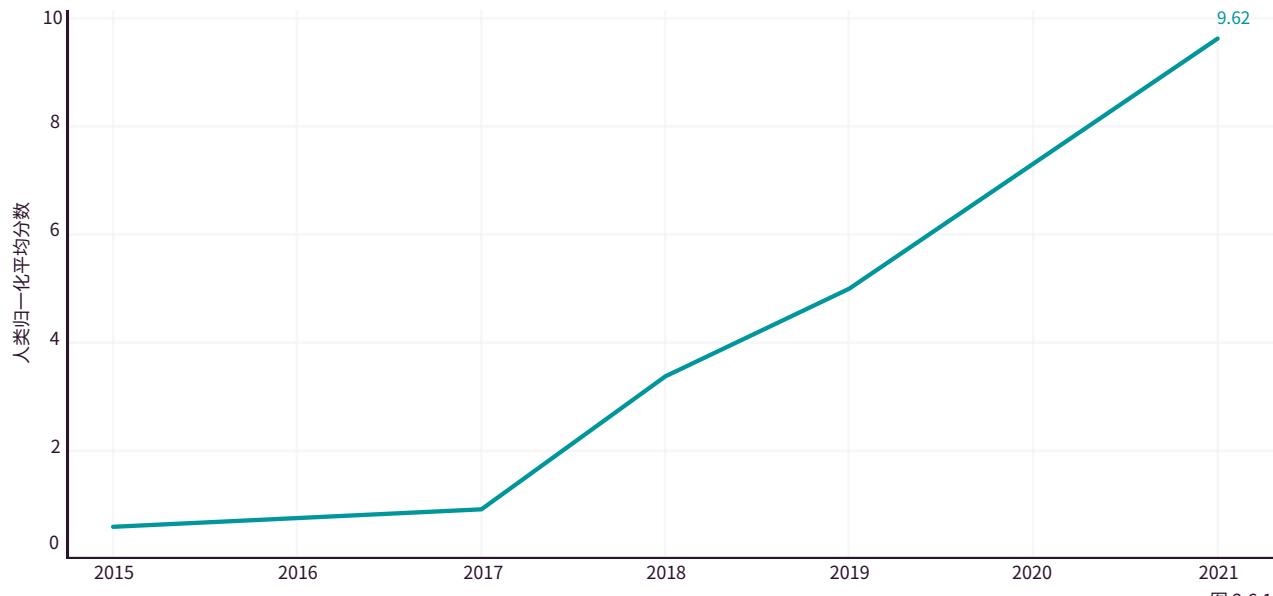


图 2.6.1

### Procgen

Procgen是OpenAI在2019年推出的一个强化学习环境。它包括16个程序化生成的类似视频游戏的环境，专门用于测试强化学习agent学习可推广技能的能力（图2.6.2）。开发Procgen是为了克服对Atari等基准的一些批评，因为这些基准导致人工智能系统成为狭义学习者，它们总是最大限度地提高一个特定技能的能力。Procgen通过引入强化学习环境来鼓励更泛化的学习，强调高度的多样性，并推动人工智能系统以可推广的方式进行训练。Procgen的性能是以平均归一化分数来衡量的。研究人员在2亿次训练中训练他们的系统，并报告16个Procgen游戏的平均得分。系统的分数越高，系统性能就越好。

### PROCGEN中16个游戏环境的截图

来源: [Cobbe et al. 2019](#)



图 2.6.2



2021年11月，DeepMind的MuZero模型在Procgen上发布了最先进的分数，达到了0.6。DeepMind的结果比2019年该环境首次发布时建立的基准性能提

高了128.6%。在这样一个多样化的基准上取得的快速进展表明，人工智能系统正在慢慢提高其在更广泛环境中的推理能力。

PROCGEN：平均归一化得分

来源：arXiv, 2021 | 图：2022年人工智能指数报告

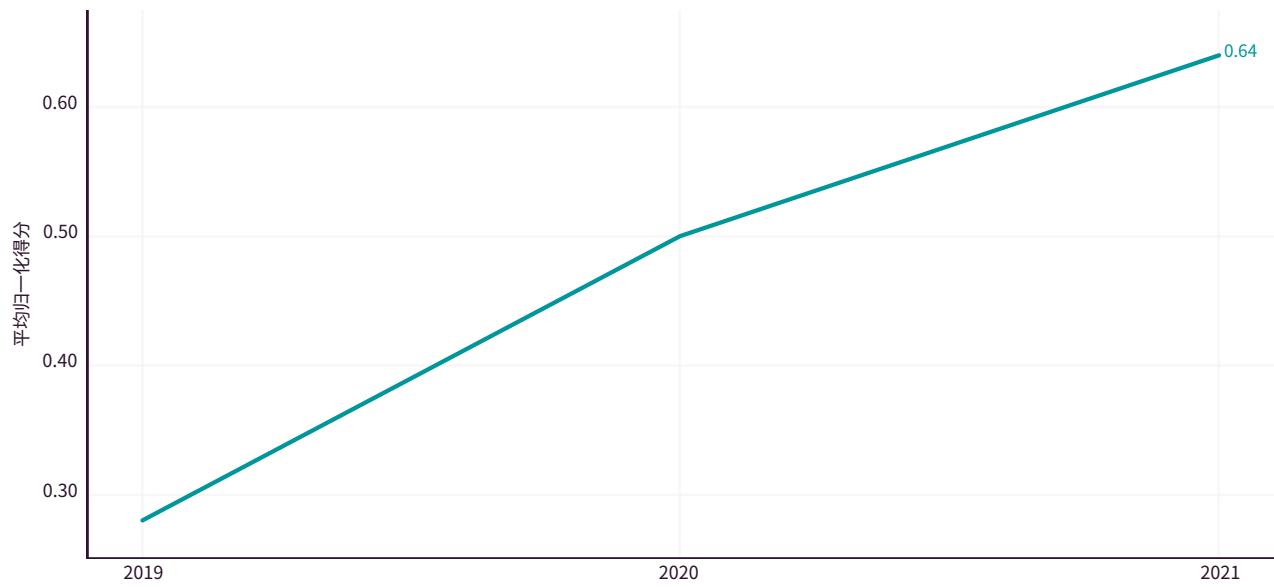


图 2.6.3



## 人类游戏：国际象棋

强化学习的进展也可以通过世界顶级国际象棋软件引擎的表现来体现。国际象棋引擎是一个计算机程序，它通过分析国际象棋的位置，被训练成为一个高水平的国际象棋棋手。国际象棋引擎的性能是根据Elo评估的，Elo是一种识别零和游戏（如国际象棋）中玩家的相对技能水平的方法：分数越高意味着棋手越强。

有一点需要注意的是，跟踪国际象棋引擎的表现并不能完全反映一般强化学习的进展；国际象棋引擎是经过专门训练以最大限度地提高国际象棋的表现。而其他流行的强化学习系统，如DeepMind的AlphaZero，能够玩更广泛的游戏，如象棋和围棋，

并且事实上已经击败了一些排名靠前的国际象棋引擎。尽管如此，观察国际象棋引擎的表现仍然是一种有效的方式，可以相对评估人工智能的研究进展，并将其与广泛理解的人类基线进行比较。

计算机在很早以前就超过了人类的棋艺水平，而且从那时起就一直在进步（图2.6.4）。到1990年代中期，顶级国际象棋引擎已经超过了人类专家级的表现，到2000年代中期，顶级国际象棋引擎超过了历史上最好的棋手之一Magnus Carlsen的巅峰表现。Magnus Carlsen在2014年记录的2882个Elo，是有记录以来人类国际象棋表现的最高水平。截至2021年，顶级国际象棋引擎已经超过了这个水平24.3%。

### 国际象棋软件引擎：ELO得分

资料来源: 瑞典计算机国际象棋协会, 2021 | 图: 2022年人工智能指数报告

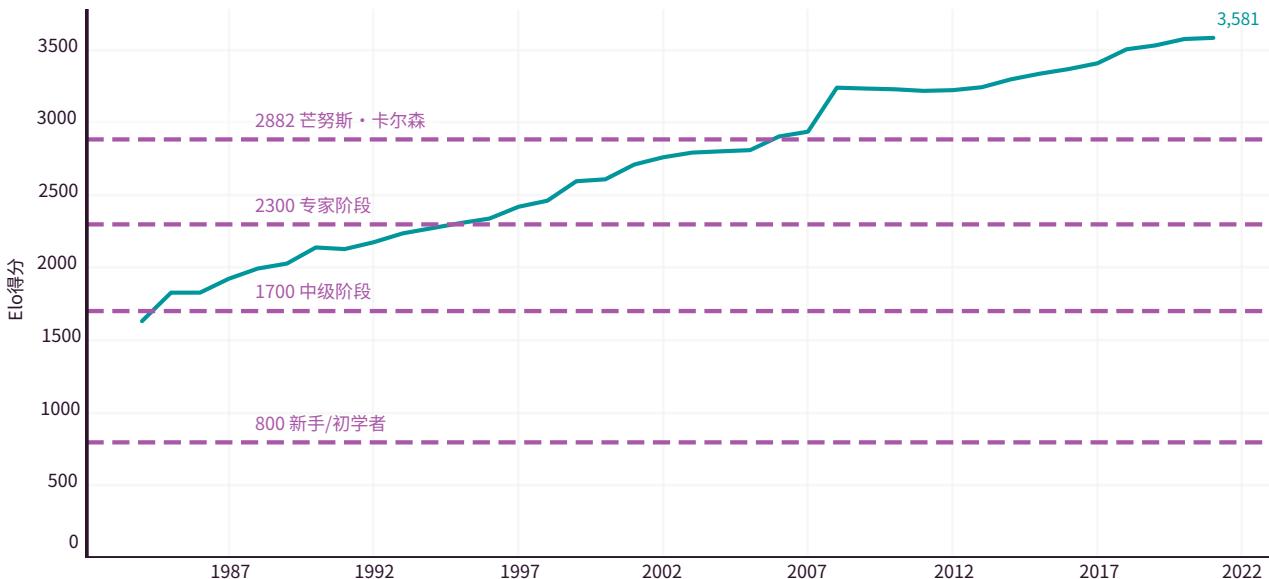


图 2.6.4



在评估人工智能的技术进步时，不仅要考虑技术性能的改进，还要考虑处理速度的改进，这一点很重要。如本节所示，人工智能系统在几乎每一个技能类别中都在持续改进。这种性能改进通常是通过增加参数和在更大的数据量上训练系统来实现的。然而，在其他条件相同的情况下，使用更多参数和更多数据的模型将需要更长的时间来训练。较长的训练时间意味着较慢的现实世界部署。鉴于增加的训练时间的潜力可以被更强大和更稳健的计算基础设施所抵消，跟踪为人工智能系统提供动力的硬件方面的进展是很重要的。

## 2.7 硬件

### MLPerf：训练时间

MLPerf是一个由MLCommons组织举办的人工智能训练竞赛。在这项挑战中，参赛者使用一个共同的架构训练系统来执行各种人工智能任务（图像分类、图像分割、自然语言处理等）。然后根据参赛者的绝对挂钟时间进行排名，也就是系统训练所需的时间。

MLPerf比赛从2018年12月开始以来，展现出了两个关键趋势。(1)几乎所有人工智能技能类别的训练时间都大量减少；(2)人工智能硬件的稳健性大幅提高。顶级性能的硬件系统可以在一分钟内达到推荐、轻量级目

标检测、图像分类和语言处理等任务类别的基准性能水平。

图2.7.2更精确地描述了自MLPerf首次引入每个技能类别以来的改进幅度。<sup>6</sup>例如，图像分类的训练时间大约减少了27倍，top时间从2018年的6.2分钟下降到2021年的0.2分钟（或13.8秒）。我们可能很难直观理解训练时间减少27倍的幅度，从实际生活角度举例，这大概就是等一个小时的公交车与等两分钟多一点的区别。

按任务划分的顶级系统的MLPERF训练时间：分钟

资料来源: MLPerf, 2021 | 图: 2022年人工智能指数报告

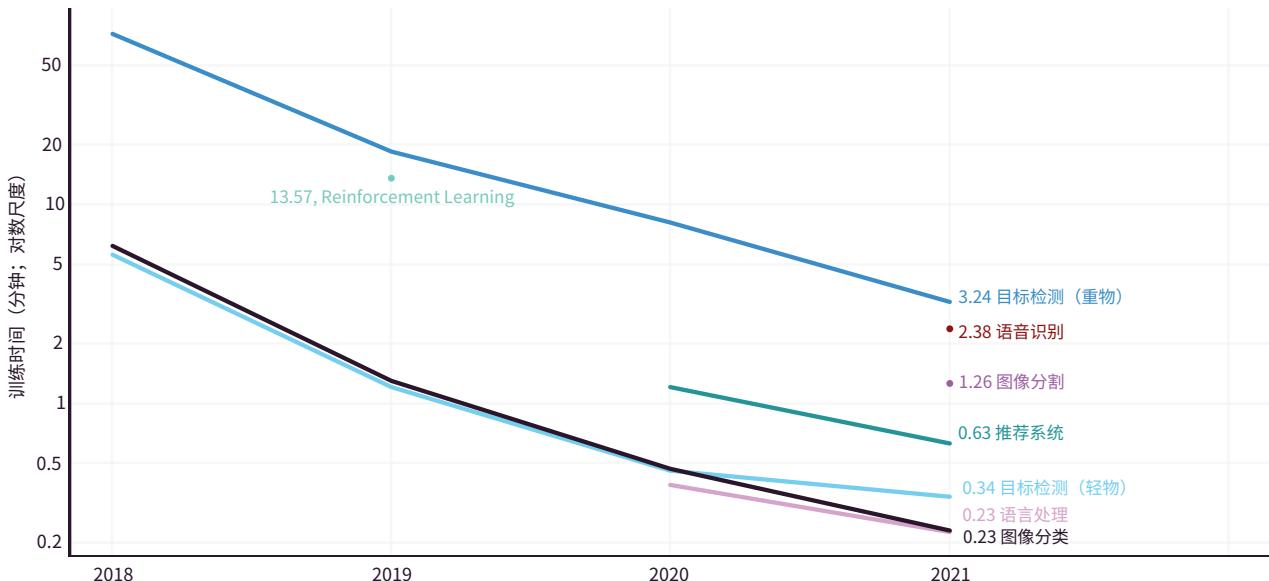


图 2.7.1

<sup>6</sup>图2.7.1中强化学习的单点表明，在2020年或2021年的MLPerf比赛中，没有注册更快的时间。语音识别和图像分割的单点表明，这些人工智能子任务类别在2021年被添加到MLPerf比赛中。



顶级性能的硬件系统可以在一分钟内达到推荐、轻量级目标检测、图像分类和语言处理等任务类别的基准性能水平。

#### MLPERF：各项任务的改进程度

来源：MLPerf, 2021 | 图：2022年人工智能指数报告

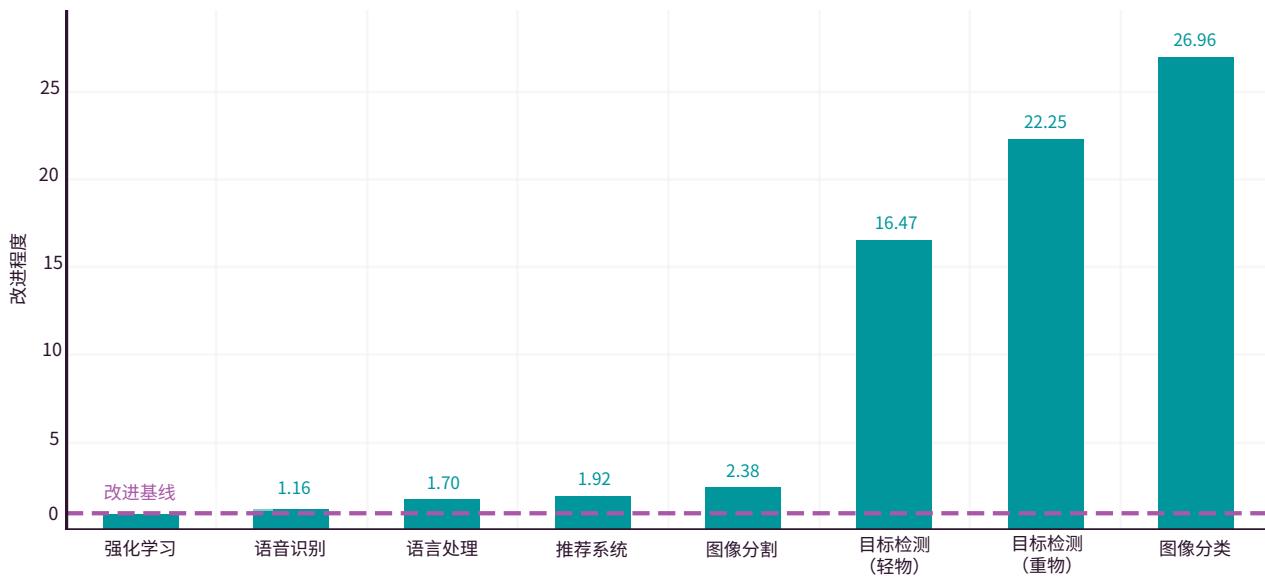


图 2.7.2



## MLPerf：加速器的数量

如图2.7.3所示，训练时间的跨任务改进是由更强大的基础硬件系统推动的。自比赛开始以来，顶级系统使用的最高加速器数量和平均加速器数量大约增加了7倍--加速器是指主要用于训练运行的机器学习部分的芯片，如GPU或TPU，而所有参赛者所使用的平均加速器

数量增加了3.5倍。然而，最值得注意的是表现最好的系统使用的平均加速器数量与所有系统使用的平均加速器之间的差距越来越大。这一差距在2021年底比2018年大9倍。这种增长意味着，平均而言，构建最快的系统需要最强大的硬件。

### MLPERF硬件：加速器

来源：MLPerf, 2021 | 图：2022年人工智能指数报告

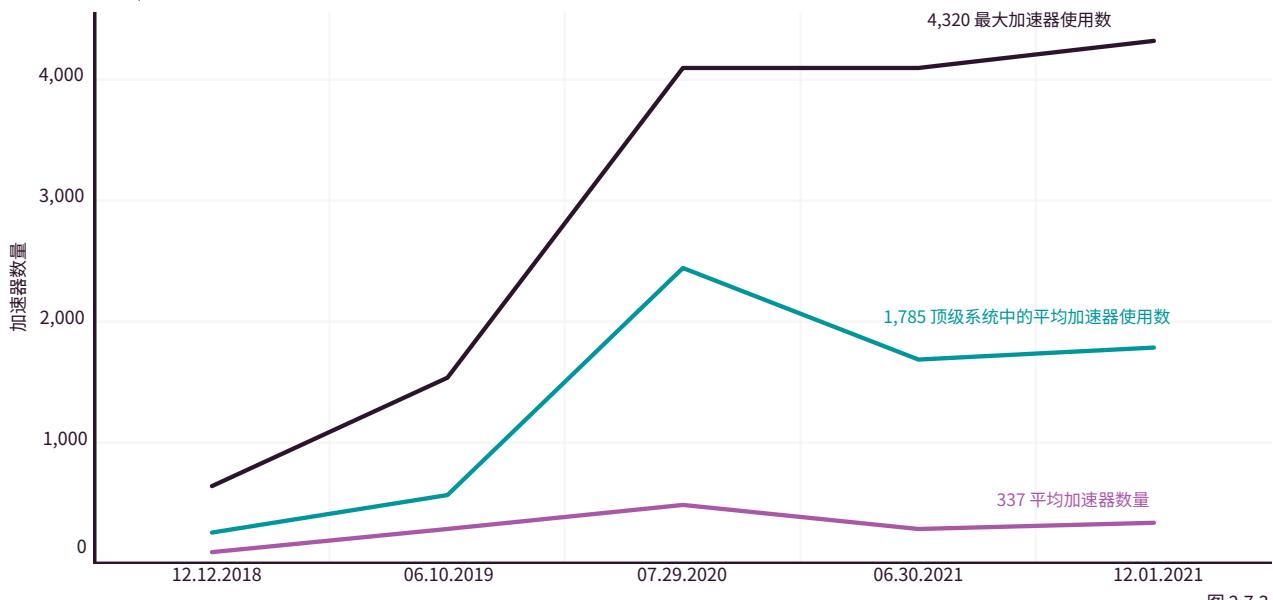


图 2.7.3



## IMAGENET：训练成本

更好的硬件条件并不一定意味着更高的训练成本。图 2.7.4 给出了 MLPerf 的图像分类子任务 (ImageNet) 每年的最低训练成本。2021 年，训练一个高性能的图像

分类系统只花了 4.6 美元。这一成本是微不足道的，特别是与 2017 年训练一个类似性能的系统所花费的 1112.6 美元相比。简单地说，在短短的四年里，图像分类训练成本已经下降了 223 倍。

### IMAGENET：训练成本（准确率达 93%）

来源：人工智能指数和 Narayanan, 2021 | 图：2022 年人工智能指数报告

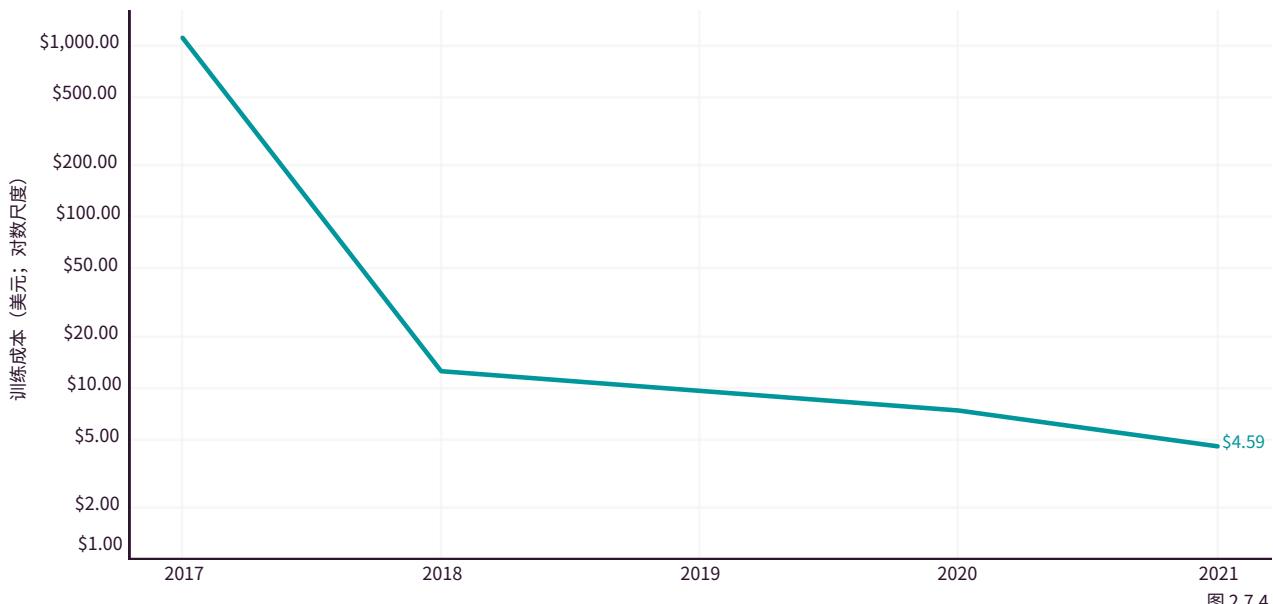


图 2.7.4



2021年，报告研究组进行了一项调查，向世界各地和新兴经济体的一流大学专门研究机器人的教授询问机械臂的价格变化以及机械臂在研究实验室的用途。来自40多所大学的101名教授和研究人员完成了这项调查，同时，调查收集了2017年至2021年的117次机械臂购买事件的数据。调查结果表明，自2017年以来，机械臂的价格有了明显的下降。

## 2.8 机器人

### 机械臂价格趋势<sup>7</sup>

机械臂的价格趋势调查结果显示，在过去的七年里，机械臂的价格有明显的下降趋势。2017年，机械臂的中位价格约为42,000美元。此后，价格几乎下降了约

46.2%，达到了2021年的约22,600美元（图2.8.1）。图2.8.2给出了机械臂的价格分布，显示出了类似的情况。尽管有一些高价的异常值，但自2017年以来，机械臂的价格呈现出了明显的下降趋势。

#### 2017-21年机械臂的中位价格

来源：人工智能指数，2022 | 图：2022年人工智能指数报告

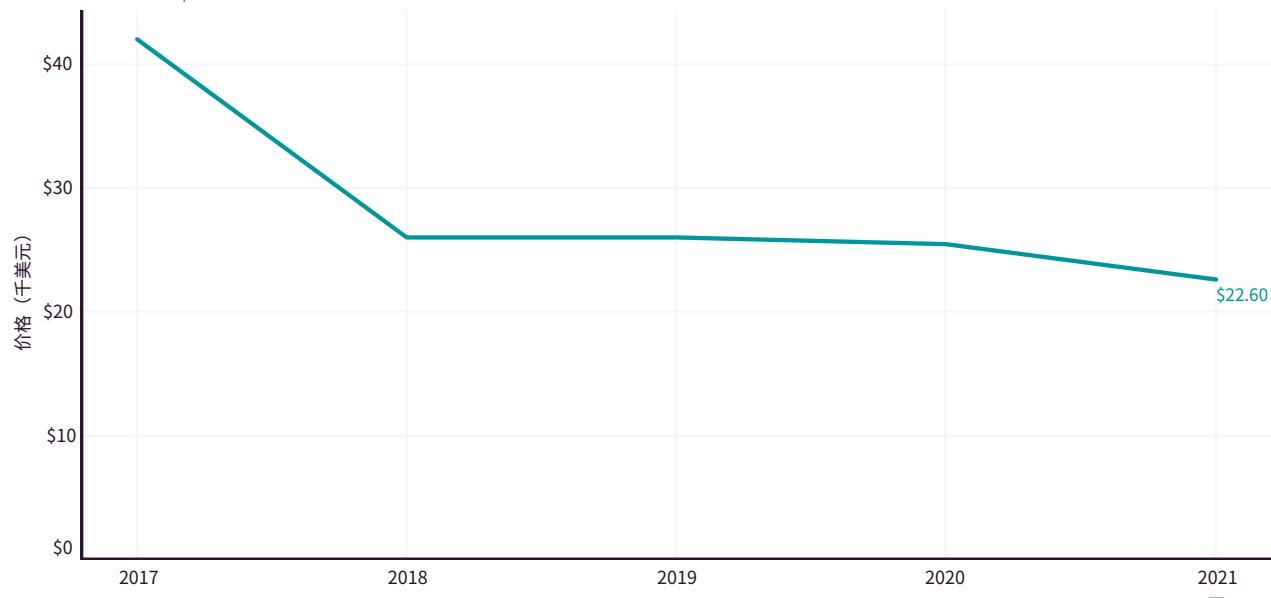


图 2.8.1

<sup>7</sup> 在注意到问卷结果数据的过滤问题后，我们对图2.8.1和图2.8.2进行了修正。正确的图表已在后续内容更新，具体可见附录中的数据连接。此外，需要注意的是，学术研究人员在购买机械臂时可能会得到折扣，所以价格比零售价低。



### 2017-21年机械臂价格的分布情况

来源：人工智能指数，2022 | 图：2022年人工智能指数报告

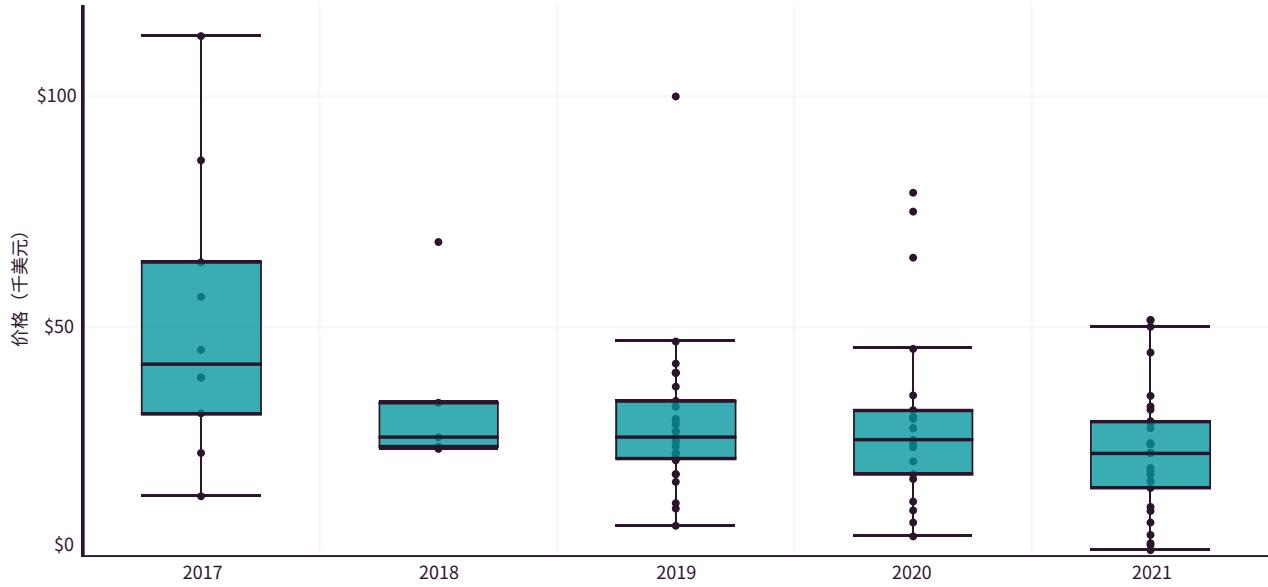


图 2.8.2

### 机器人学教授所使用的人工智能技能

本报告还专门调查了研究机器人学的教授在其研究中使用AI技能的情况。答复显示，深度学习和强化学习都是机器人学教授所采用的流行人工智能技能。更具体地说，67.0%的教授报告使用了深度学习技术，46.0%的教授报告使用了强化学习技术。

### 机器人学教授所使用的人工智能技能

来源：人工智能指数，2022 | 图：2022年人工智能指数报告

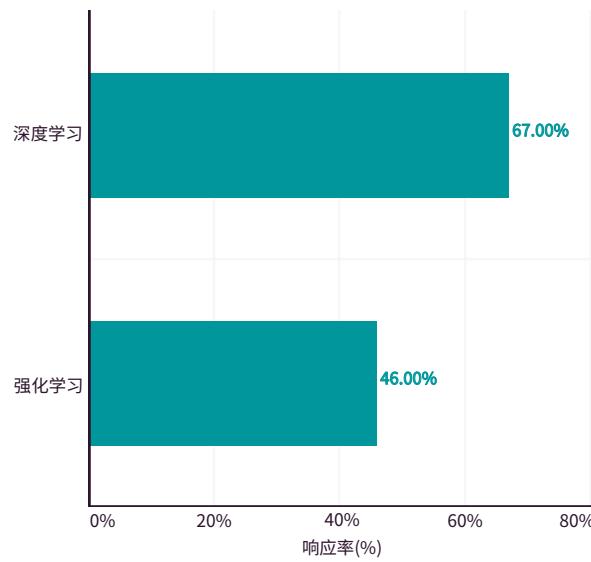


图 2.8.3



2022年  
人工智能指数报告

## 章节三 技术AI伦理

文本和分析，作者：  
Helen Ngo和Ellie Sakhaei



## 章节三 章节预览

概述	102
致谢	103
章节	105
<b>3.1 公平性和偏见指标的元分析</b>	<b>106</b>
AI伦理诊断指标和基准	107
<b>3.2 自然语言处理的偏见指标</b>	<b>109</b>
毒性：真实的毒性提示和透视API	109
亮点：大型语言模型和毒性	111
模型解毒会对性能产生负面影响	113
StereoSet	114
CrowS-Pairs	115
Winogender 和 WinoBias	117
WinoMT：机器翻译系统中的性别偏见	119
文字和图像嵌入关联测试	120
亮点：多语言词嵌入	122
用内在的偏见指标减轻词嵌入中的偏见	122

<b>3.3 FACCT和NEURIPS的AI伦理学趋势</b>	<b>123</b>
ACM公平、问责制和透明度会议（FAccT）	123
神经信息学研讨会NeurIPS	125
可解释性、可说明性和因果推理	126
隐私和数据收集	127
公平和偏见	129
<b>3.4 实事求是和诚实性</b>	<b>130</b>
利用人工智能进行事实核查	130
用FEVER基准衡量事实核查的准确性	133
迈向诚实的语言模型	134
模型大小和诚实性	134
亮点：对比性语言-图像预训练（CLIP）	
中的多模态偏见	136
诋毁伤害	136
性别偏见	136
将学到的偏见向下游传播	138
在非英语语言方面表现不佳	138

访问公开数据



# 概述

近年来，人工智能系统已经开始在世界范围内部署，研究人员和从业人员正在考虑其对现实世界的危害，包括基于种族歧视的商业人脸识别系统，基于性别歧视的简历筛选系统，以及基于社会经济和种族偏见的人工智能驱动的临床健康工具等等。研究人员发现，这些模型反映并放大了人类社会的偏见，根据于受保护的属性进行歧视，并生成了关于世界的错误信息。这些发现提高了学术界对人工智能伦理、公平和偏见的研究兴趣，促使行业从业者调配资源以补救这些问题，并吸引了媒体、政府以及使用这些系统并受其影响的人们的关注。

今年，人工智能指数强调了社区已经采用的报告消除偏见和促进公平进展的指标。追踪这些指标的表现以及技术能力为我们提供了一个更全面的视角，即随着系统的改进，公平性和偏见是如何变化的，这对于了解系统的部署情况非常重要。



## 鸣谢

人工智能指数要感谢所有围绕问责制的人工智能的发展和治理而参与研究和宣传的人。本章建立在整个人工智能伦理学界的学者们的工作之上，包括致力于衡量技术能力的学者以及专注于建立社会规范的学者。现在，这一领域中仍然还有很多工作要做，但是，我们还是被整个社区及其合作者所取得的进展所鼓舞。

本章引用的出版物包括：

Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. 2021. Evaluating CLIP: Towards Characterization of Broader Capabilities and Downstream Implications. arXiv preprint arXiv:2108.02818.

Jack Bandy and Nicholas Vincent. 2021. Addressing “Documentation Debt” in Machine Learning Research: A Retrospective Datasheet for Book Corpus. arXiv preprint arXiv:2105.05241.

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal Datasets: Misogyny, Pornography, and Malignant Stereotypes. arXiv preprint arXiv:2110.01963.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2021. Improving Language Models by Retrieving from Trillions of Tokens. arXiv preprint arXiv:2112.04426.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2017. Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. arXiv preprint arXiv:1711.08412.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. arXiv preprint arXiv:2009.11462.

Wei Guo and Aylin Caliskan. 2020. Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases. arXiv preprint arXiv:2006.03955.

Aylin Caliskan Islam, Joanna J. Bryson, and Arvind Narayanan. 2016. Semantics Derived Automatically from Language Corpora Necessarily Contain Human Biases. arXiv preprint arXiv:1608.07187.

Opher Lieber, Or Sharir, Barak Lenz, and Yoav Shoham. 2021. Jurassic-1: Technical Details and Evaluation. (2021).

[https://uploads-ssl.webflow.com/60fd4503684b466578c0d307/61138924626a6981ee09caf6\\_jurassic\\_tech\\_paper.pdf](https://uploads-ssl.webflow.com/60fd4503684b466578c0d307/61138924626a6981ee09caf6_jurassic_tech_paper.pdf)

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On Measuring Social Biases in Sentence Encoders. arXiv preprint arXiv:1903.10561.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. StereoSet: Measuring Stereotypical Bias in Pretrained Language Models. arXiv preprint arXiv:2004.09456.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu



Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, John Schulman. WebGPT: Browser-Assisted Question-Answering with Human Feedback. 2021. arXiv preprint arXiv:[2112.09332](#).

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. arXiv preprint arXiv:[2010.00133](#).

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, Ryan Lowe. Training Language Models to Follow Instructions with Human Feedback. 2022. arXiv preprint arXiv:[2203.02155](#).

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d' Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William S. Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. arXiv preprint arXiv:[2112.11446](#).

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating Gender Bias in Machine Translation. arXiv preprint arXiv:[1906.00591](#).

Ryan Steed and Aylin Caliskan. 2020. Image Representations Learned With Unsupervised Pre-Training Contain Human-like Biases. arXiv preprint arXiv:[2010.15052](#).

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William S. Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from Language Models. arXiv preprint arXiv:[2112.04359](#).

Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. 2021. Challenges in Detoxifying Language Models. arXiv preprint arXiv:[2109.07445](#).

Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. 2021. Detoxifying Language Models Risks Marginalizing Minority Voices. arXiv preprint arXiv:[2104.06390](#).

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muha Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining Gender Bias in Languages with Grammatical Gender. arXiv preprint arXiv:[1909.02224](#).



## 章节要点

- 语言模型效果比以前更优，但偏见也更严重。大型语言模型正在不断创造技术基准的新记录，但新数据显示，大型模型也更能反映出训练数据的偏见。**对比2018年最先进的1.17亿参数的模型，2021年开发的一个2800亿参数的模型所引发的毒性（elicited toxicity）增加了29%**。随着时间的推移，这些系统的能力明显增强，但是随着它们能力的增强，其潜在的偏见的严重程度也在增加。
- 人工智能伦理的兴起，无处不在。自2014年以来，关于人工智能的公平性和透明度的研究呈爆炸式增长，在伦理学相关会议上发表的**相关文章数量增加了五倍**。算法的公平性和偏见问题已经从主要的学术追求转变为一个具有广泛影响的主流研究课题。近年来，**与产业界有联系的研究人员在以伦理学为重点的会议上发表的论文数量同比增加71%**。
- **多模态模型学习多模态偏见：**在训练多模态语言-视觉模型方面已经取得了快速进展，这些模型在语言-视觉联合任务上表现出了更高的能力水平。这些模型在图像分类和从文本描述中创建图像等任务上创造了新的记录，但它们也在其输出中反映了社会固有观念和偏见--**在CLIP上的实验表明，黑人的图像被错误地分类为非人类的比率是其他种族的两倍以上**。在计算机视觉和自然语言处理领域，已经有大量的工作致力于开发测量偏见的指标，这凸显了对能够深入了解多模态模型的偏差的指标的需求。



在过去的五年里，人们在创建数据集、基准和指标方面投入了大量的研究工作，旨在衡量机器学习模型的偏见和公平性。偏见通常是从人工智能模型的基础训练数据中发现的；这些数据可以反映社会中的系统性偏见，也可以反映收集和策划数据的人的偏见。

## 3.1 公正性和偏见指标的元分析

算法偏见通常是由分配性伤害和代表性伤害构成的。当一个系统不公平地将机会或资源分配给一个特定的群体时，就会出现分配性伤害；当一个系统以强化一个群体的从属地位的方式延续既定规定型观念和权力动态时，就会出现代表性伤害。当算法做出的预测既不偏袒也不歧视基于受保护属性的个人或群体时，算法则被认为是公平的，这些属性由于法律或道德原因不能用于决策（如种族、性别、宗教）。

为了更好地了解算法偏见和公平的情况，人工智能指数进行了原创性研究，以分析该领域的状况。如图3.1.1所示，自2018年以来，沿着利益的道德维度衡量偏见和公平性的指标数量稳步增长。就这张图而言，所发表的公平性和偏见指标的数量至少在另一项工作中引用过。<sup>1</sup>

2016-21年人工智能公平性和公正性指标的数量

来源：AI指数，2021 | 图：2022年人工智能指数报告

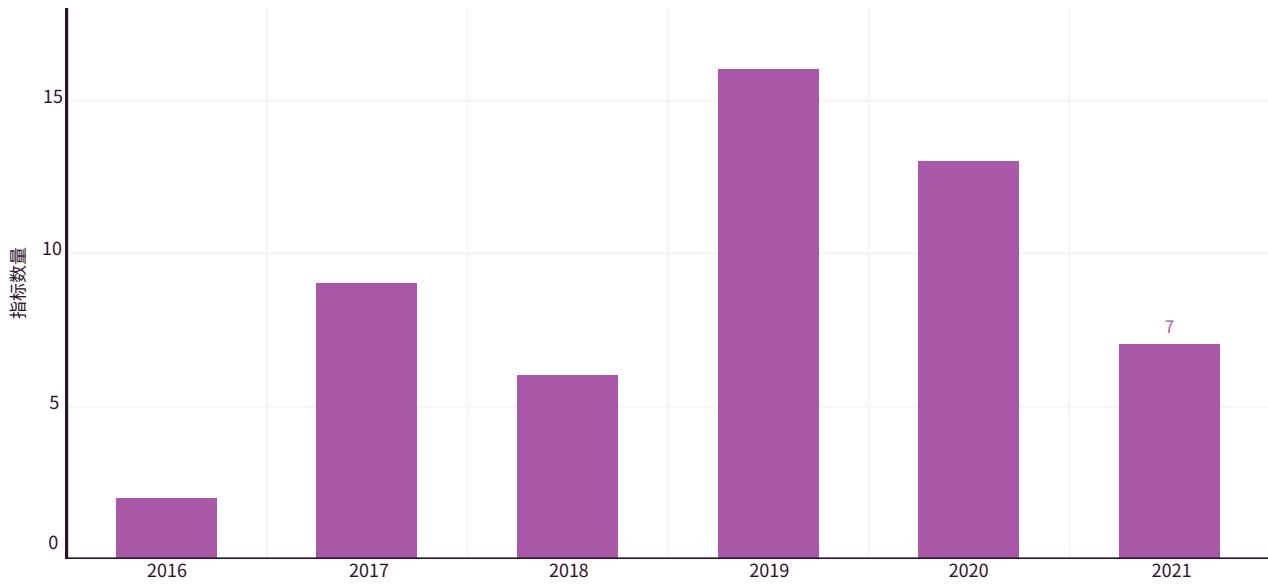


图 3.1.1

<sup>1</sup> 2021年的数据可能是滞后的，因为衡量标准被社区采用需要时间。



## 人工智能伦理学诊断指标和基准

对人工智能系统在伦理方面的衡量通常有两种形式：

- **基准数据集：**基准数据集包含标签数据，研究人员测试他们的人工智能系统会如何标记这些数据。这些基准不会随时间变化，且针对特定领域（例如，语言模型的SuperGLUE和StereoSet；计算机视觉的ImageNet），通常旨在测量模型的内在行为，而不是其在特定人群中的下游表现（例如，StereoSet测量模型与非立体类型相比选择立体类型的倾向，但它不测量不同亚群之间的性能差距）。
- **诊断性指标：**诊断性指标衡量模型对下游任务的影响或表现。例如，与类似的个人或整个人口相比，一个人口亚群或个人。这些指标可以帮助研究人员了解一个系统在现实世界中部署时的表现，以及它是否对某些人群产生了不同的影响。这方面的例子包括群体公平性指标，如人口均等和机会平等。

基准是衡量整个领域进展的有效指标，其影响可以通过社区的采用情况（例如，提交的排行榜数量或报告指标的研究论文数量）来衡量。研究实验室在排行榜指标上的竞争通常也会促使算法快速进步。然而，一些排行榜很容易被操纵，而且可能是基于含有缺陷的基准数据集，如不正确的标签或定义不当的类别而生成的。此外，它们的静态属性意味着它们是特定文化和时间背景的缩影。换句话说，2017年发布的基准可能与2022年的部署背景不相关。

诊断指标使研究人员和从业人员能够了解他们的系统对特定应用或群体的影响以及潜在的具体伤害（例如，“这个模型在这个具有这种受保护属性的群体中表现得不成比例”）。诊断性指标在单个模型或应用层面最有用，而不是作为领域层面的指标。它能够表明一个特定的人工智能系统在特定的子群体或个人身上的表现，这对评估现实世界的影响很有帮助。然而，虽然这些指标可能被广泛用于私下测试模型，但由于这些指标并没有附在鼓励研究人员公布其结果的排行榜上，所以公开的信息并不多。

图3.1.2显示，随着时间的推移，对制定基准和诊断性指标的研究投入一直很稳定。<sup>23</sup>

**基准是衡量整个领域进展的有效指标，其影响可以通过社区的采用情况来衡量（例如，提交的排行榜的数量或报告指标的研究论文的数量）。**

<sup>2</sup> 研究论文的引用是一个滞后的活动指标，最近采用的指标可能不会反映在当前的数据中，与3.1.1类似。

<sup>3</sup> Perspective API 定义了7个新指标，用于衡量毒性的方方面面（毒性、严重毒性、身份攻击、侮辱、淫秽、性暗示、威胁），2017年发布的指标数量异常多。



### 2016-21年，人工智能公平性和偏见指标的数量（诊断性指标与基准指标）

来源：AI指数，2021 | 图：2022年人工智能指数报告

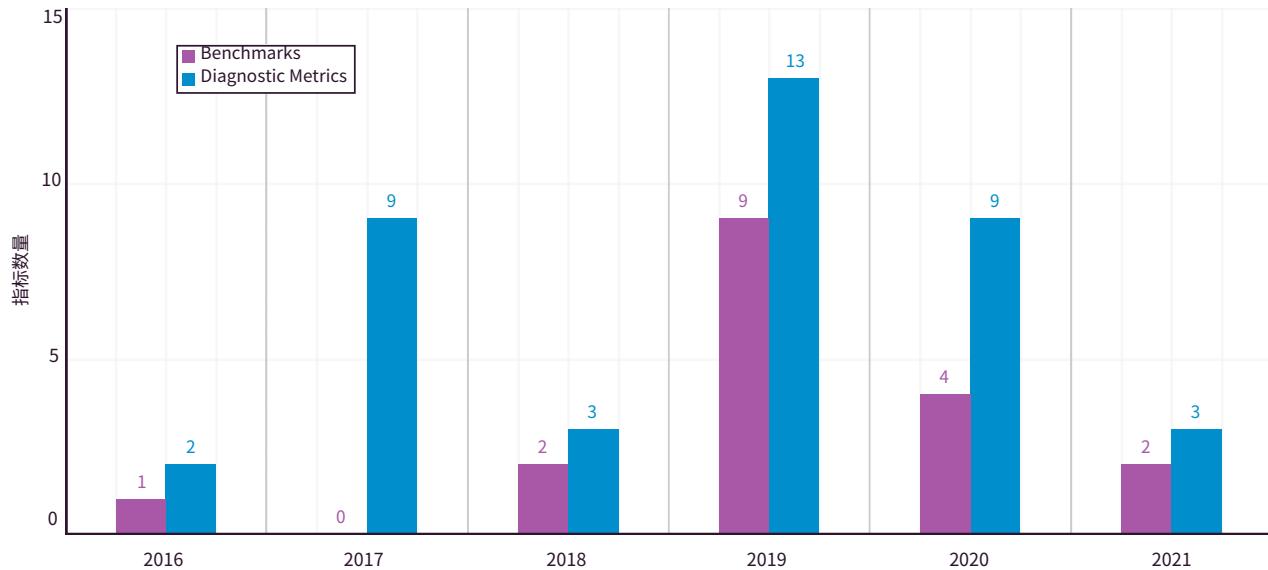


图 3.1.2

本章的其余部分将聚焦于自然语言和计算机视觉等领域，深入研究近期人工智能系统在这些指标和基准方面的表现。这些指标大多是测量系统内的内在偏见，而且

前期研究已经表明，内在偏见指标可能无法完全反映下游应用中的外在偏见的影响。



目前最先进的自然语言处理（NLP）依赖于大型语言模型或机器学习系统，它们处理数百万行文本并学习预测句子中的单词。这些模型可以生成连贯的文本；对人、地点和事件进行分类；并被应用作为更大系统的组成部分，如搜索引擎。通常需要从互联网上搜集这些模型的训练数据，以创建网络规模的文本数据集。这些模型从其预训练数据中学习人类的偏见，并在其下游输出中反映出来，这一过程可能会造成伤害。目前，研究人员已经开发了一些基准和指标，以识别自然语言处理中的性别、种族、职业、残疾、宗教、年龄、体貌、性取向和种族等方面 的偏见。

## 3.2 自然语言处理的偏见指标

偏见指标可以分为两大类：内在指标和外在指标，前者衡量模型内部嵌入空间的偏见，后者衡量模型下游任务和输出的偏见。外在指标的例子包括群体公平性指标（受保护群体之间的均等）和个人公平性指标（类似个人之间的均等），这些指标衡量一个系统是否对一个亚群体或个人有不成比例的负面影响，或者以牺牲一个群体为代价给予另一个群体优惠待遇。

### 毒性：真实的毒性提示和PERSPECTIVE API

衡量语言模型中的毒性需要对有毒和无毒的内容进行标注。毒性的定义指粗鲁的、不尊重人的或不合理的

评论，很可能使人结束对话。Perspective API是谷歌旗下的Jigsaw公司开发的一个工具。它最初是为了帮助平台识别在线对话中的毒性。开发者将文本输入 Perspective API，API会返回该文本应被标记为属于以下类别之一的概率：毒性、严重毒性、身份攻击、侮辱、淫秽、性暴露和威胁。

2017年，Perspective API一经发布，NLP社区就迅速采用它来测量自然语言中的毒性。如图3.2.1所示，在2020年和2021年之间，使用Perspective API的论文数量翻了一番，从8篇增至19篇。

使用Perspective API的研究论文数量

来源：AI指数，2021 | 图：2022年人工智能指数报告



图 3.2.1



RealToxicityPrompts 由英语自然语言提示组成，用于衡量语言模型完成有毒文本提示的频率。语言模型的毒性是通过两个指标来衡量的。

- 最大毒性：在一定数量的完成品中的平均最大毒性得分。
- 毒性概率：预计一个完成品会有多大的毒性。

图3.2.2显示，语言模型的毒性在很大程度上取决于基础训练数据。在过滤了有毒内容的互联网文本上训练的模型，与在各种未经过滤的互联网文本语料库上训练的模型相比，毒性明显较低。在BookCorpus（一个包含电子书网站书籍的数据集）上训练的模型，产生有毒文本的频率非常之高。这可能是由于它的组成--BookCorpus 包含了大量的含有露骨描述的爱情小说，这可能会导致更高的毒性水平。

### 按训练数据集划分的语言模型的毒害性

来源：Gehman et al., 2021; Rae et al., 2021; Welbl et al., 2021 | 图：2022年人工智能指数报告

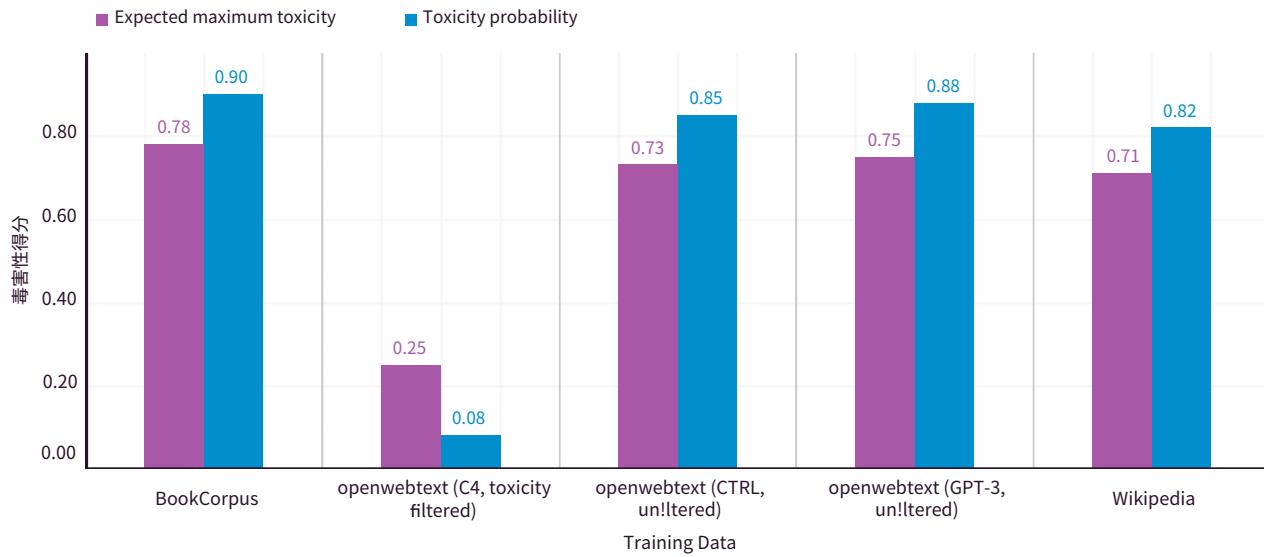


图 3.2.2



## 大型语言模型和毒性

最近围绕减轻语言模型的毒性的发展，降低了预期的最大毒性和毒性的概率。然而，解毒方法始终会导致不利的副作用和能力较差的模型。(例如，过滤训练数据通常要以牺牲模型的性能为代价)。

2021年12月，DeepMind发表了一篇论文，描述了其2800亿参数的语言模型Gopher。Gopher论文中的图3.2.3a和图3.2.3b显示，当被提示有不同程度的毒性输入时，较大的模型更容易产生毒性输出，但

它们也更有能力在其自身的输出以及其他情况下检测出毒性，这一点可通过模型规模增加的AUC（接收操作特征曲线下的面积）来衡量。AUC指标将真阳性率与假阳性率作对比，以描述一个模型对不同类别的区分程度（越高越好）。如图3.2.3b所示，较大的模型在识别民间评论数据集中的有毒评论方面有明显的优势。

GOPHER: 基于预测毒性的毒性持续概率，按模型大小划分

来源: Rae et al., 2021 | 图: 2022年AI指数报告

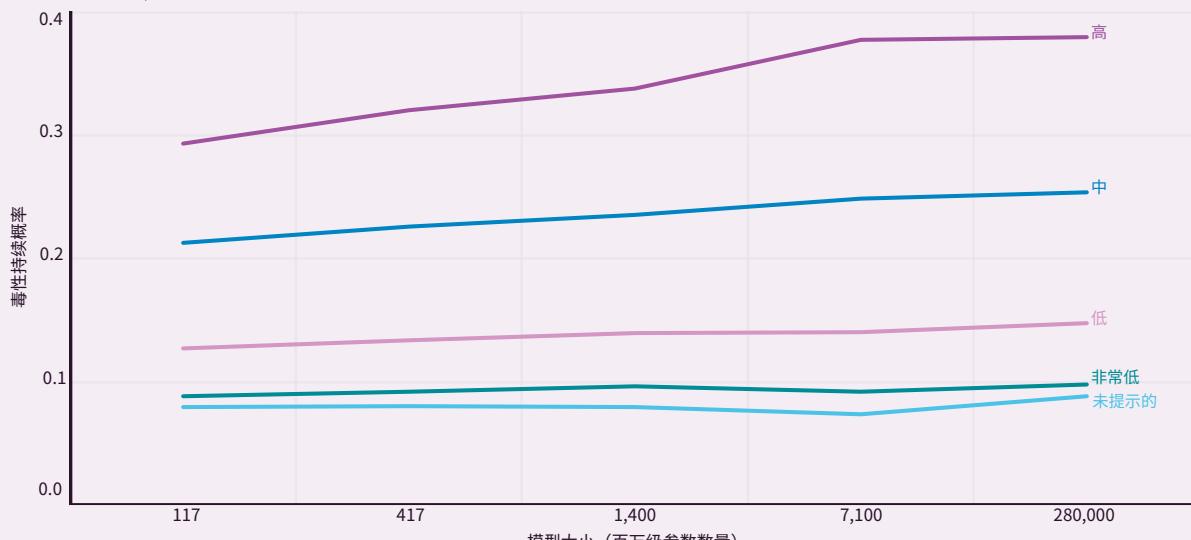


图 3.2.3a



## 大型语言模型和毒性（续）

**GOPHER：民间评论数据集上的few-shot毒性分类**

来源：Rae et al., 2021 | 图：2022年AI指数报告



图 3.2.3b



## 对模型解毒可能会对性能产生负面影响

解毒方法旨在通过改变基础训练数据来减轻毒性，例如，领域适应性预训练（domain-adaptive pretraining, DAPT），或通过在生成过程中引导模型，如即插即用语言模型（Plug and Play Language Models, PPLM）或生成性判别器引导的序列生成（Generative Discriminator Guided Sequence Generation, GeDi）等等。

一项关于解毒语言模型的研究显示，用这些策略解毒的模型在白种人和非裔美国人英语上的困惑度（perplexity）都比较差，困惑度是衡量一个模型对特定分布的学习程度的指标（越低越好）（图3.2.4）。这些模型在非裔美国人的英语和含有提及少数族裔身份的文本上的表现也比白人排列的文本差得不成比例，这一结果可能是由于人类的偏见导致注释者更容易将非裔美国人的英语误标为有毒。

### 复杂性：少数民族群体在英语解毒后的语言模拟表现

来源：Xu et al., 2021 | 图：2022年人工智能指数据报告

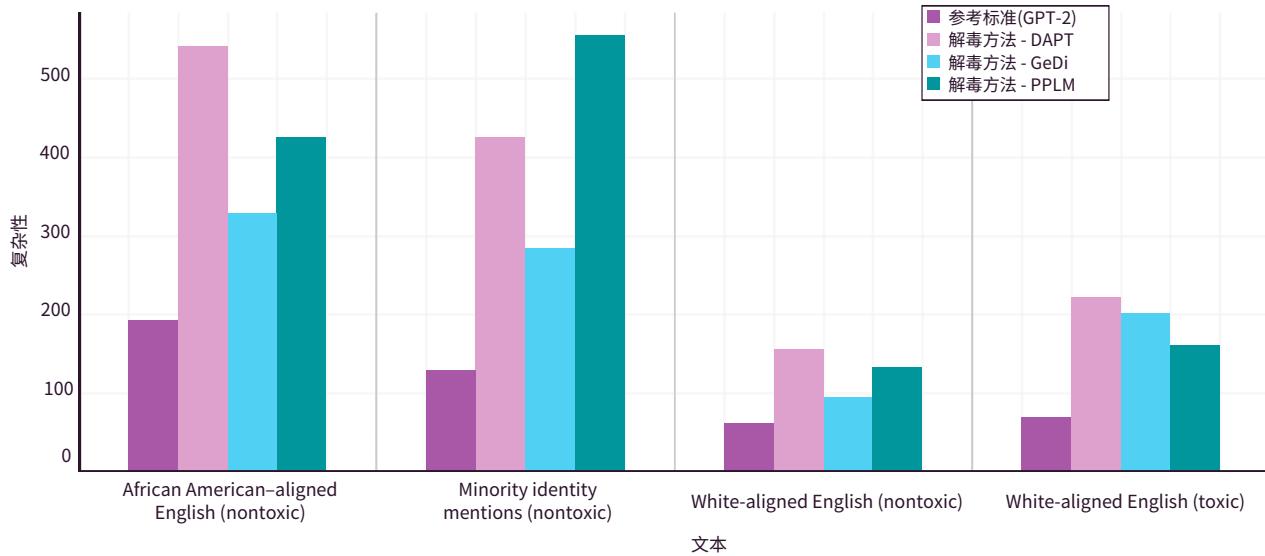


Figure 3.2.4



## STEREOSET

StereoSet 是一个衡量沿性别、种族、宗教和职业轴线的固有偏见以及原始语言建模能力的基准。其中一个相关指标是固化印象得分，它衡量一个模型对固化印象和反固化印象的喜爱程度相同。固化印象是对一个群体广泛持有的过度概括的信念，反固化印象则是对一个群体的概括，与广泛接受的固化印象相矛盾。

图3.2.5显示，StereoSet的性能表现与毒性的趋势相同。除非在训练期间采取干预措施以减少所学到的固化印象，否则较大的模型更经常地反映固化印象。据估计，网上有毒内容的流行率为0.1%-3%，这与研究表明较大的语言模型更有能力记忆罕见的文本的结论相一致。

**StereoSet：按模型大小的STEREOTYPE得分**

来源：Nadeem et al., 2020; Lieber et al., 2021; StereoSet Leaderboard, 2021 | 图：2022年AI指数报告

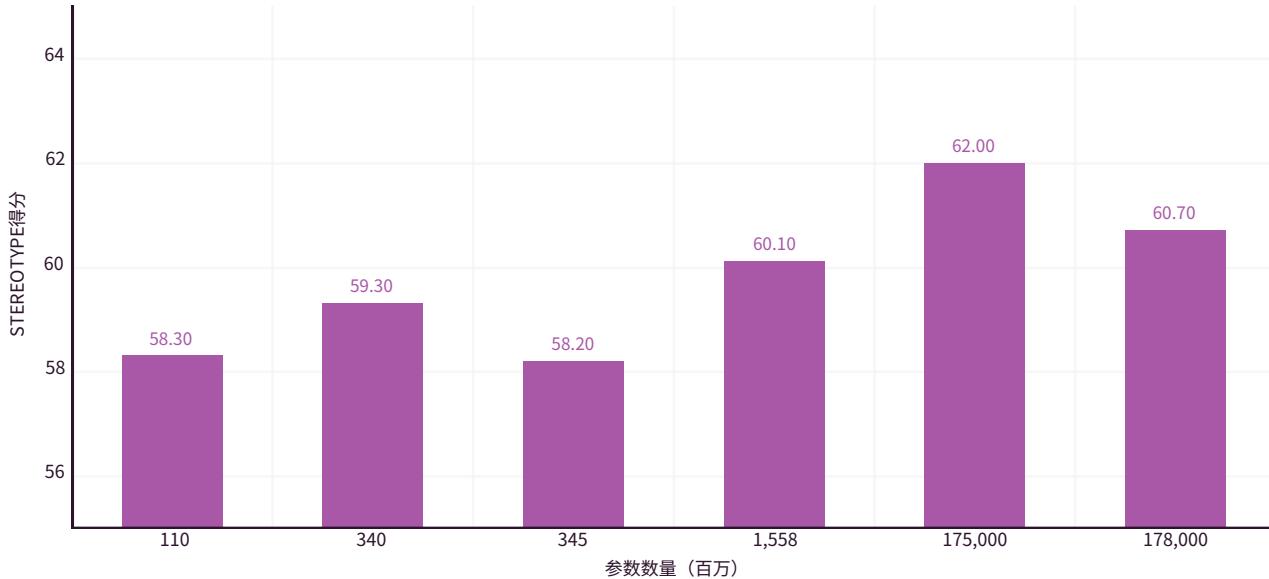


图 3.2.5

StereoSet在其基础数据集中有几个主要缺陷。一些样本未能表达出有害的固化印象，将关于国家的固化印象与关于种族和民族的固化印象混为一谈，并混淆了相关

但不同群体之间的固化印象。这些固化印象的来源是位于美国的民众，因此数据集中的价值和固化印象可能不具有普遍的代表性。



## CROWS-PAIRS

CrowS-Pair（众包定型对）是另一个衡量固化印象偏见的基准。StereoSet比较的是关于单一群体的属性，而CrowS-Pairs对比的是历史上处于不利地位的群体和处于有利地位的群体（例如，墨西哥人与白人）之间的关系。

CrowS-Pairs的创造者使用三种流行的语言模型评估了固化印象的偏见：BERT、RoBERTa和ALBERT（图3.2.6）。在标准的语言建模基准上，ALBERT优于RoBERTa，而RoBERT则优于BERT。<sup>4</sup>然而，根据CrowS-Pairs，ALBERT是三个模型中偏见程度最高的。这反映了在StereoSet和RealToxicityPrompts中同样也观察到的趋势：性能更强大的模型也更容易学习和放大固化印象。

与之前的示例一样，BERT、RoBERTa和ALBERT都继承了其训练数据中的偏见。这些模型都是基于BookCorpus、英语维基百科和从互联网上搜集来的文本的组合来训练的。对BookCorpus的分析显示，与世界其他主要宗教相比，其有关宗教的书籍严重偏向基督教和伊斯兰教，<sup>5</sup>但目前还不清楚这些书籍在多大程度上包含了历史内容和从特定宗教观点出发的内容。<sup>6</sup>

我们可以通过查看其基础数据集来研究语言模型如何继承对某些宗教的偏见。图3.2.7显示了两个流行的数据集--BookCorpus和Smashwords21中与不同宗教有关的书籍数量。这两个数据集提到基督教和伊斯兰教的次数都远远多于其他宗教。

## CROWS-AIRS：语言模型的性能与偏见属性的关系

来源：Nangia et al., 2020 | 图：2022年人工智能指数报告

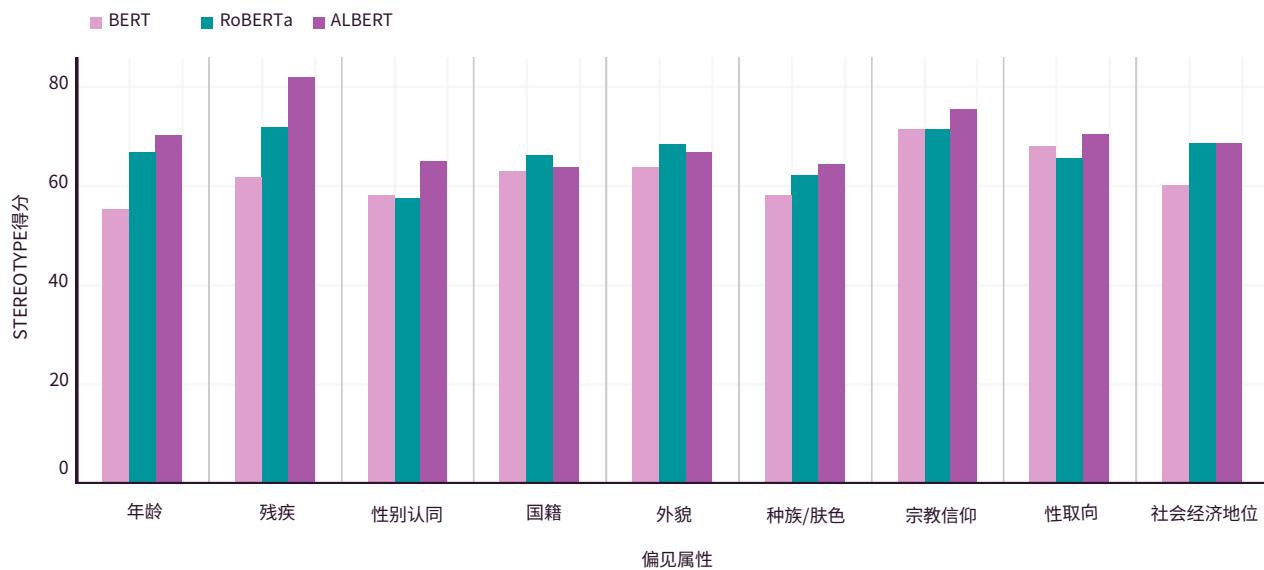


图 3.2.6

4 按SQuAD、GLUE和RACE基准的结果。

5 如锡克教、犹太教、印度教、佛教、无神论。

6 特别是在BERT基础上进行微调的仇恨言论分类器已被证明经常将含有提及“穆斯林”的文本错误地归类为有毒文本，而且研究人员发现GPT-3在提及“犹太人”和“穆斯林”时都含有沿宗教轴线的明显偏见。



### BOOKCORPUS和SMASHWORDS21：关于宗教的书籍在预训练数据中的比例

来源：Bandy and Vincent, 2021 | 图：2022年人工智能指数报告

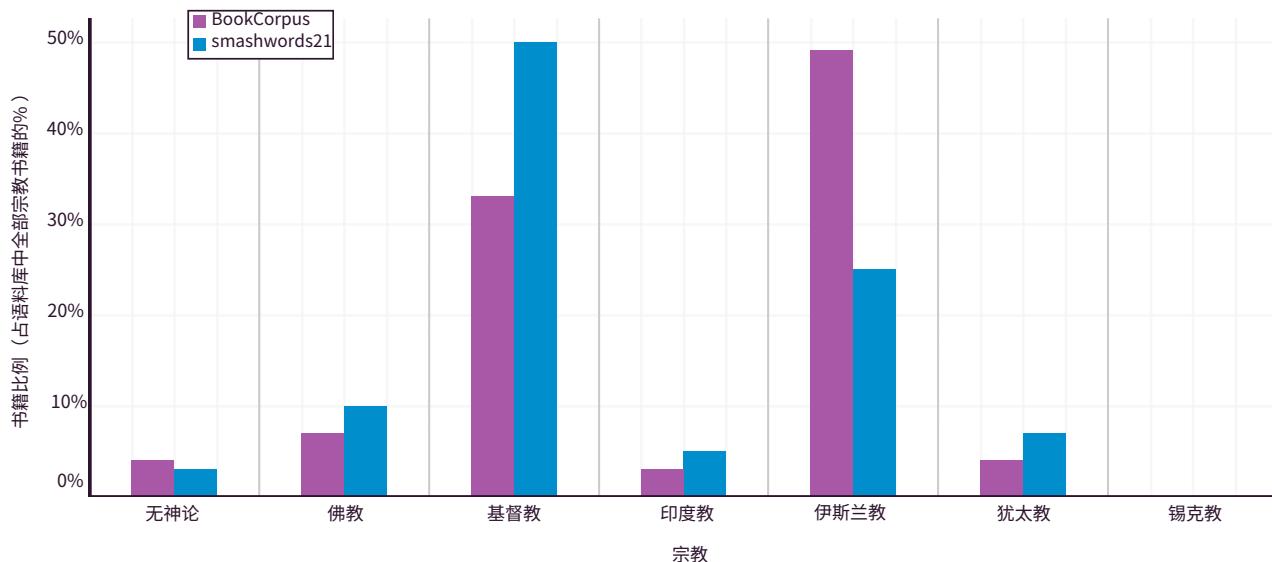


图 3.2.7



## WINOGENDER 和 WINOBIAS

Winogender 测量与职业有关的性别偏见。系统根据他们在包含职业的句子中填写正确性别的能力进行评估

（例如，“这个少年向治疗师倾诉，因为他/她看起来值得信赖”）。样本以来自美国劳工统计局获取的数据，以确定偏向于一种性别的职业来生成（例如，收银员职业由73%的女性组成，但司机只有6%是女性）。

Winogender的性能是通过固化印象和反固化印象案例之间的准确性差距，以及性别平等的得分来衡量的（预

测结果相同的样本的百分比）。作者使用众包注释来估计人类的表现，准确率为99.7%。

来自SuperGLUE排行榜的Winogender结果显示，大型模型更有能力在zero-shot和few-shot设置中正确解决性别问题（即没有对Winogender任务进行微调），并且不太可能放大职业上的性别差异（图3.2.8）。然而，Winogender表现出来的成绩并不表明一个模型在性别方面是没有偏见的，只能表明这个基准没有捕捉到偏见。

### 从SUPERGLUE基准看WINOGENDER任务的模型表现

来源：SuperGLUE排行榜, 2021 | 图：2022年AI指数报告

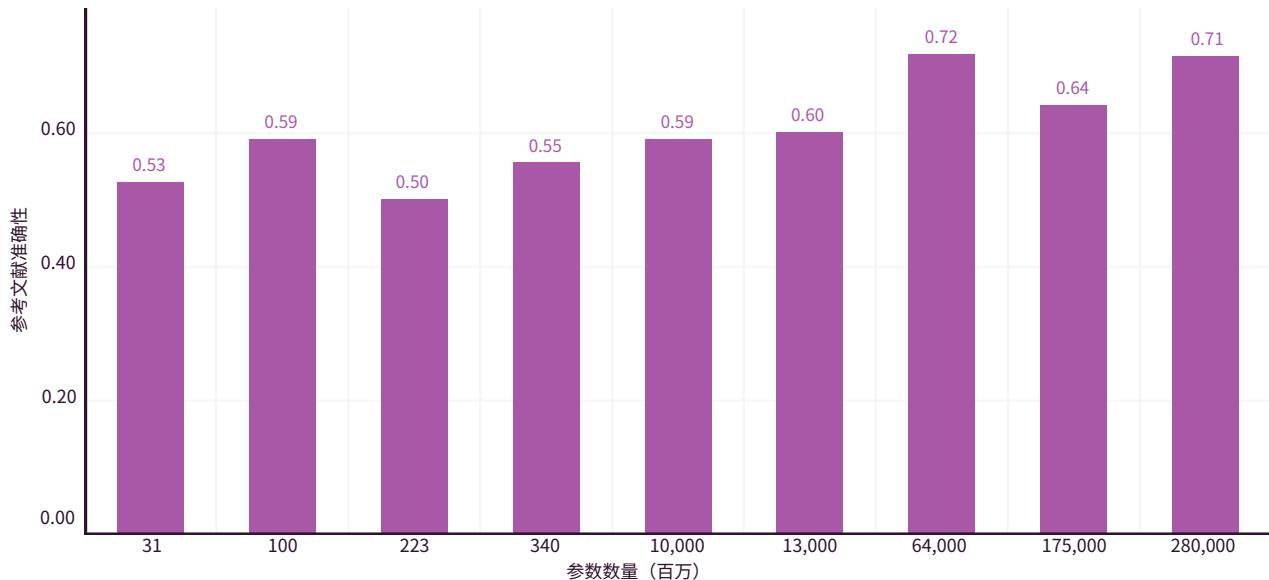


图 3.2.8



WinoBias 是一个类似的衡量与职业有关的性别偏见的基准，由一个不同的研究小组与 Winogender 同时发布。如图 3.2.9 所示，WinoBias 比 Winogender 更常被引

用，但在测量自然语言理解的 SuperGLUE 排行榜中采用 Winogender 的情况导致了更多针对后者的模型评估报告出现。

#### WINOBIAS 和 WINOGENDER：2018-21 年引用次数

来源：AI Index, 2021; Semantic Scholar, 2021 | 图：2022年人工智能指数报告

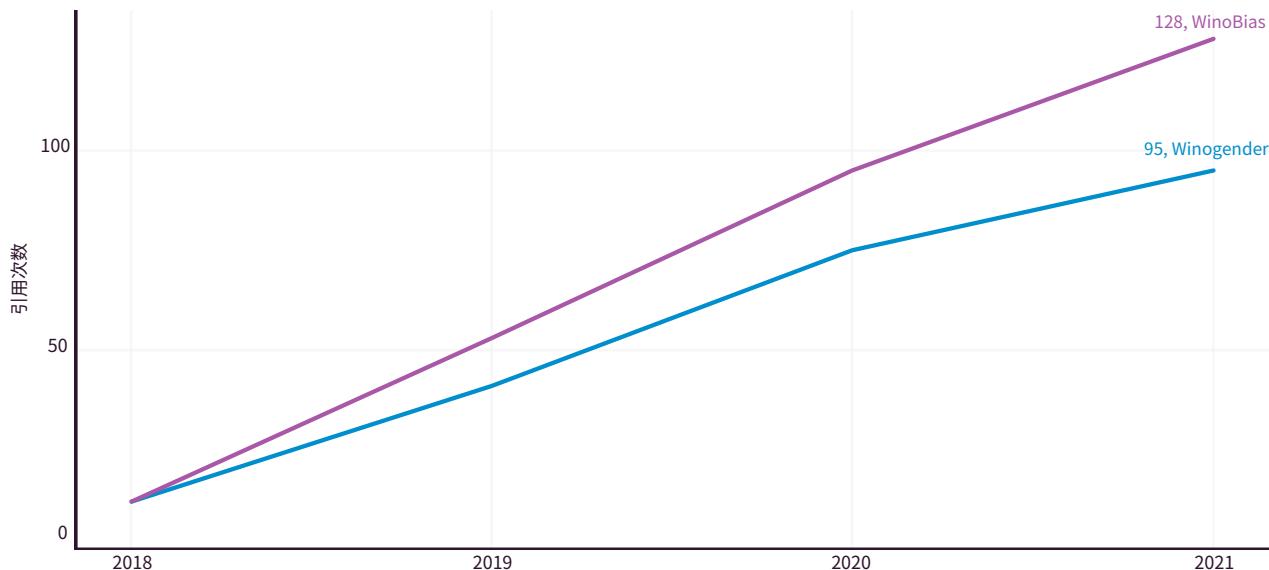


图 3.2.9



## WINOMT：机器翻译系统中的性别偏见

商业机器翻译系统已被记录在案，以反映和放大其基础数据集的社会偏见。由于这些系统被广泛用于全球行业，如电子商务，固化印象和翻译中的错误可能会造成巨大的损失。

WinoMT 是一个衡量机器翻译中性别偏见的基准，它是通过结合 Winogender 和 WinoBias 数据集创建的。通过比较从英语翻译成另一种语言的句子，并提取翻译后的性别与原始性别进行比较，对模型进行评估。系统的评分标准是：性别正确的翻译百分比（性别准确度），男性和女性样本之间的 F1 得分差异，以及固化印象性别角色和反固化印象性别角色样本之间的 F1 得分差异。

如图3.2.10 所示，在翻译含有符合社会性别角色偏见的职业样本时，谷歌翻译在所有测试语言（阿拉伯语、英语、法语、德语、希伯来语、意大利语、俄语、乌克兰语）中的表现都更好。此外，这些系统在翻译句子时，最多只有 60% 的时间能翻译出正确的性别。其他主要的商业机器翻译系统（Microsoft Translator、Amazon Translate、SYSTRAN）也证明有类似的表现。

### WINOMT: GOOGLE 跨语言翻译中的性别偏见

来源：Stanovsky et al., 2019 | 图：2022年人工智能指数报告

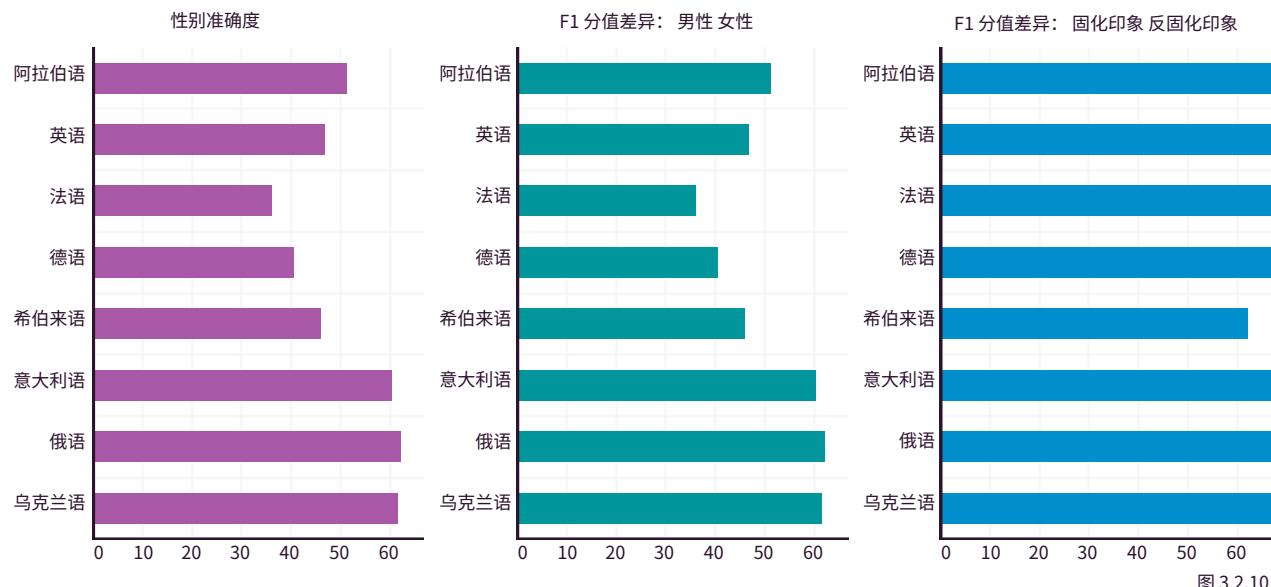


图 3.2.10



## 词和图像嵌入关联测试

词嵌入是NLP中的一种技术，它允许具有类似含义的词具有类似的表征。静态词嵌入是固定的表征，不随上下文变化。例如，多义词无论出现在哪个句子中都会有相同的表征（嵌入）。静态词嵌入的例子包括GloVe、PPMI、FastText、CBoW和Dict2vec。与此相反，语境化的词嵌入是对词的动态表征，它根据词的伴随语境而变化。例如，“bank”在“riverbank”和“bank teller”中会有不同的表征。

词嵌入关联测试（Word Embedding Association Test, WEAT）通过测量概念（如欧裔美国人和非裔美国人的名字）和属性（如愉快和不愉快）之间的关联（“效应大小”）来量化英语静态词嵌入的偏见。在大型公共语料库（如维基百科、谷歌新闻）上训练的词嵌入在评估WEAT时一直重复着固化印象的偏见（例如，将男性词汇与“职业”相关联，将女性词汇与“家庭”相关联）。CEAT（Contextualized Embedding Association

Test）将WEAT扩展到上下文的词嵌入中。

图像嵌入关联测试（Image Embedding Association Test, iEAT）修改了WEAT以测量社会概念和图像属性之间的关联。使用iEAT，研究人员表明，预训练的生成性视觉模型（iGPT和simCLRV2）在性别、种族、年龄和残疾方面表现出类似于人类的偏见。

词嵌入可以通过被称为句子编码器的模型汇总成句子嵌入。句子编码器关联测试（Sentence Encoder Association Test, SEAT）扩展了WEAT，以测量句子编码器中与性别名称、地区名称和固化印象有关的偏见。较新的基于Transformer的语言模型，使用上下文的词嵌入，与先前方法相比表现出更少的偏见，但大多数模型在性别和职业方面，以及非洲裔美国人的名字与欧洲裔美国人的名字方面仍然表现出明显的偏见，如图3.2.11所示。

### 句子嵌入关联测试（SEAT）：用效用尺寸测量典型关联

来源：May et al., 2019 | 图：2022年人工智能指数报告

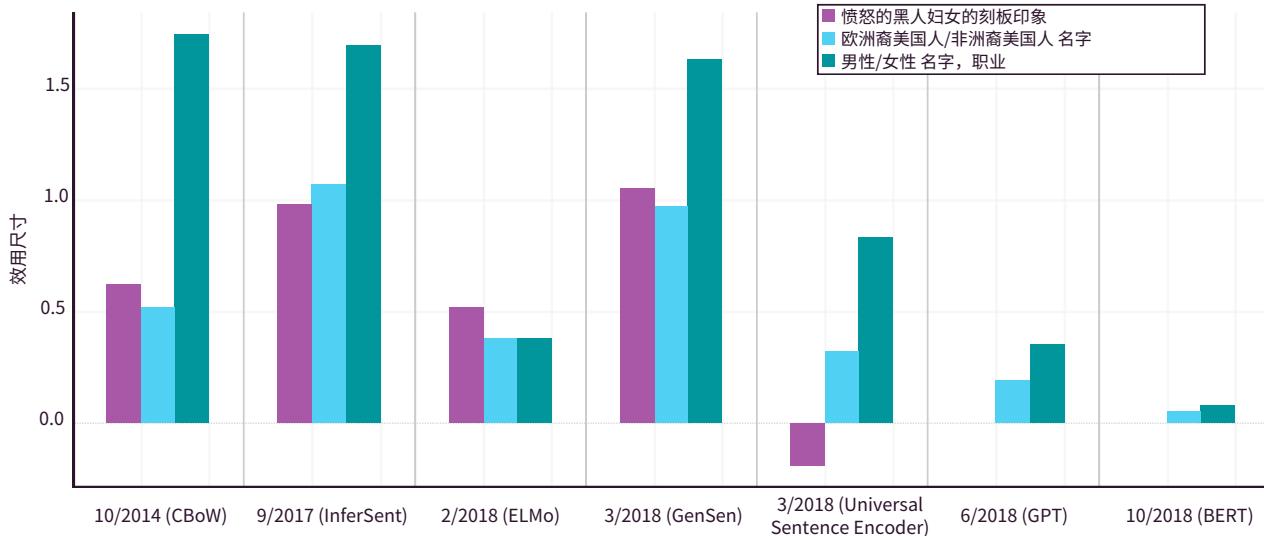


图 3.2.11



词嵌入也反映了文化的转变。对100多年来美国人口普查文本数据的词嵌入的时间分析表明，词嵌入的变化密切跟踪人口和职业随时间的变化。图3.2.12显示，在谷歌图书和美国历史英语语料库（Corpus of Historical American English, COHA）上训练的词嵌入的变化反映了重大的历史事件，如60年代的妇女运动和亚洲移民到美国。在这项分析中，使用相对范数差来衡量词嵌

入偏见：与代表群体（如男性、女性、亚洲人）相关的词与与职业相关的词之间的平均欧几里得距离。蓝线显示了随时间变化的性别偏见，负值表示嵌入的职业与男性的联系更紧密。（图中）红线显示了种族与职业相关的词嵌入的偏见，特别是在亚裔美国人和白人的情况下。

#### 根据100年的文字数据训练得到的文字内容中的性别和种族偏见

来源: Garg et al., 2018 | 图: 2022年人工智能指数报告

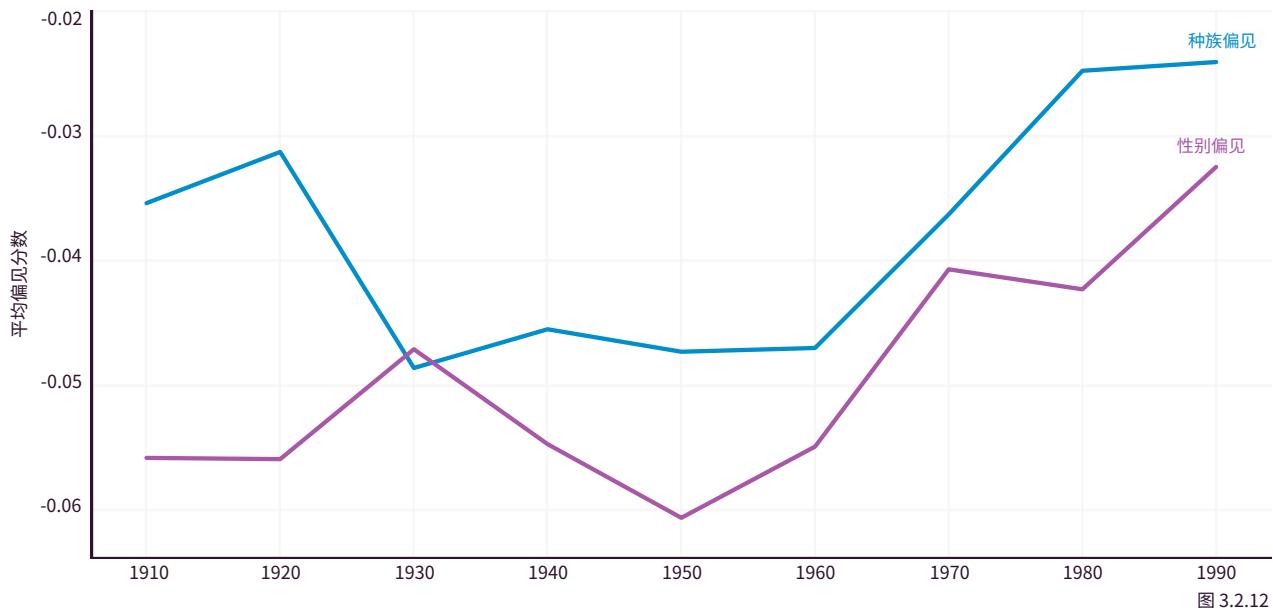


图 3.2.12



## 多语言词嵌入

大型语言模型通常是单语言的，因为它们需要大量的文本数据来训练。虽然英语文本可以很容易地从互联网上获取，但对于像 Fula 这样的低资源语言来说，挑战更大。XWEAT 是 WEAT 的一个多语言和跨语言的扩展，旨在进行语言间的比较偏见分析。XWEAT 的结果显示，跨语言嵌入的偏见可以从相应的单语言嵌入的偏见中大致预测出来，这表明偏见可以在不同的语言之间转移。

另一项关于性别偏见的研究扩展了 WEAT，以量化有语法性别的语言（如西班牙语或法语）中的双语嵌入的偏见。图 3.2.13 显示，西班牙文中的阳性词更接近于历史上以男性为主的职业（如建筑师）的英文词，以及接近中性位置，如图中垂直线所示。同样，阴性词更接近于历史上以女性为主的职业来说（如护士），的英文词。

### 西班牙语词嵌入中的性别偏见：嵌入的相似度距离

来源：Zhou et al., 2019 | 图：2022年人工智能指数报告



图 3.2.13

## 用内在偏见指标减轻词嵌入中的偏见

人们通常认为，通过对嵌入的去偏移来减少内在偏见，将能够减少应用中的下游偏见（外在偏见）。研究人员

已经证明，内在偏见指标和下游应用偏见之间没有可靠的相关性，需要进一步调查以建立内在和外在指标之间的关系。



为了掌握人工智能伦理领域随着时间的推移所发生的变化，本节研究了ACM公平、问责制和透明度会议（FAccT）的趋势，该会议发布了关于算法公平和偏见的工作，以及NeurIPS研讨会的趋势。本节确定了研讨会出版主题的新趋势，并按隶属关系和地理区域分享了关于作者趋势的见解。

## 3.3 FAccT和NEURIPS的人工智能伦理学趋势

### ACM公平、问责制和透明度会议（FAccT）

ACM FAccT是一个跨学科的会议，主要发表关于算法公平性、问责制和透明度的研究。<sup>7</sup> 虽然一些人工智能会议也提供了专门讨论类似主题的研讨会，但FAccT是最早创建的大型会议之一，旨在将对算法的社会技术分析感兴趣的人员、产业界从业人员和政策制定者聚集在一起。

图3.3.1显示，产业界实验室在FAccT的出版物中所占的比例逐年增加。他们经常与学术界合作进行研究，但也越来越多地进行独立研究。在2021年，有53位作者明确了产业界的隶属关系，比2020年的31位作者和2018年首届会议上的5位作者有所增加。这与最近的研究结果相一致，即深度学习研究人员呈现从学术界过渡到产业界实验室的趋势。

2018-21年按附属机构划分的被接受的FAccT会议提交的数量

来源：FAccT, 2021; AI指数, 2021 | 图：2022年AI指数报告

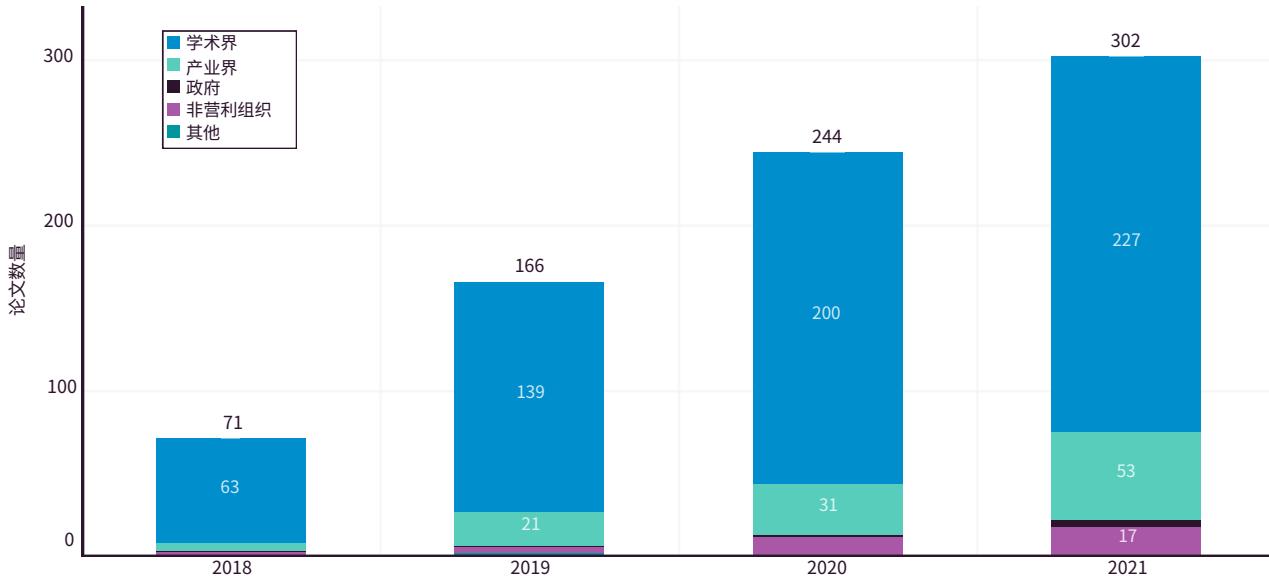


图 3.3.1

<sup>7</sup> FAccT接受的工作包括衡量公平性的技术框架、对特定行业的人工智能危害的调查（例如，在线广告中的歧视、推荐系统中的偏见）、最佳做法的建议以及更好的数据收集策略。在FAccT发表的一些工作已经成为人工智能伦理学的典范之作，包括Model Cards for Model Reporting (2019) 和 On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? (2021)。值得注意的是，FAccT出版了大量批评人工智能当代方法和系统的作品。



虽然各类机构对公平、问责制和透明度研究的兴趣越来越大，但在FAccT发表的大部分论文都是由在美国的研究人员撰写的，其次是在欧洲和中亚的研究人员（图

3.3.2）。从2020年到2021年，总部设在北美地区的机构的论文比例从70.2%上升到75.4%。

2018-21年FAccT接收的各地区的论文数量

来源：FAccT, 2021年；AI指数，2021 | 图：2022年AI指数报告

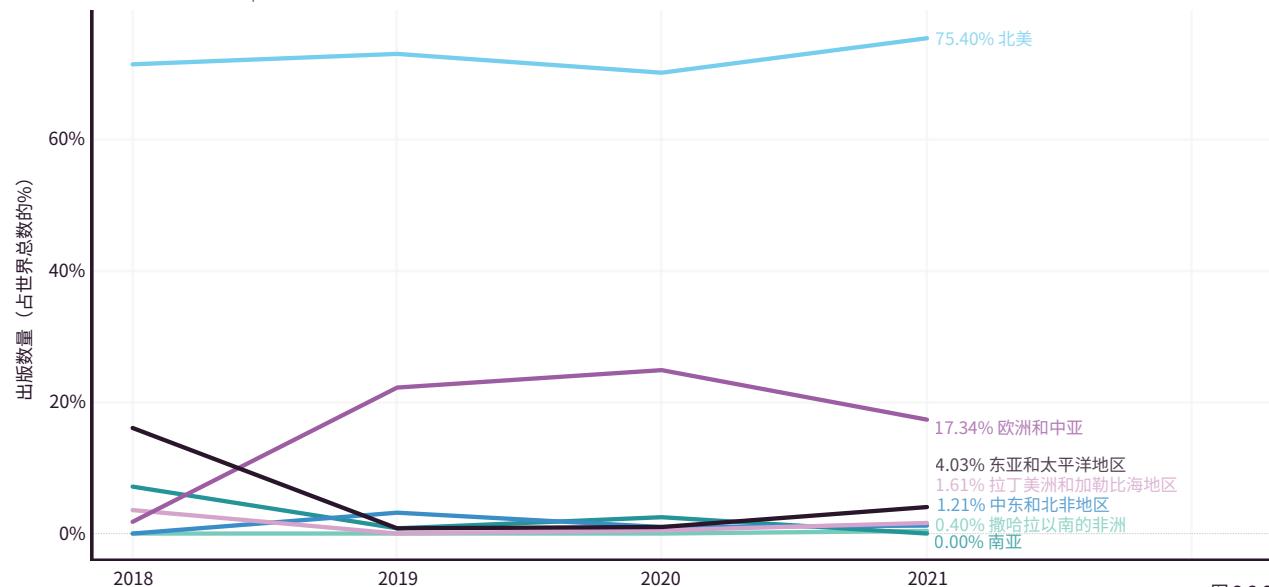


图 3.3.2



## NEURIPS研讨会

NeurIPS是最大的人工智能会议之一，它在2014年举办了第一次关于公平、问责制和透明度的研讨会。图3.3.3显示了过去六年中NeurIPS伦理学相关研讨会上按

研究主题分列的研究论文数量，可以看出，人们对人工智能应用于气候、金融和医疗等高风险、高影响的场景的兴趣越来越大。

NeurIPS研讨会的研究主题：2015至2021年接受的关于现实世界影响的论文数量

来源：NeurIPS, 2021; AI Index, 2021 | 图：2022年人工智能指数报告

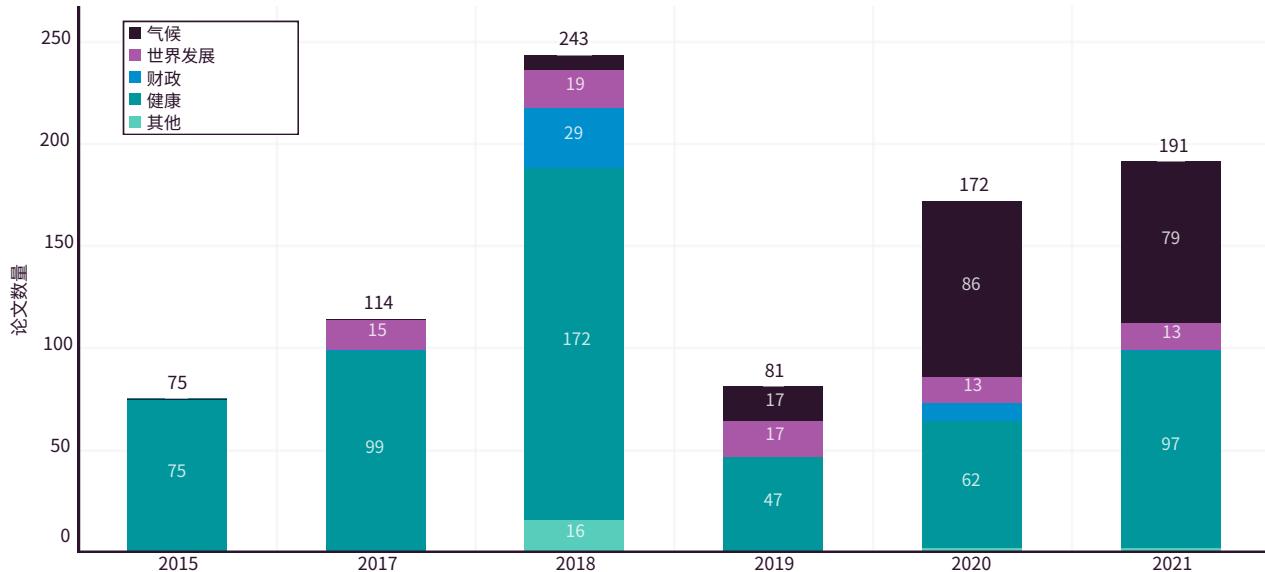


图 3.3.3



## 可解释性、可说明性和因果推理

在NeurIPS上，围绕可解释性和可说明性已经创建了几个研讨会，包括影响人类决策的安全关键人工智能，<sup>8</sup> 算法公平性的可解释性和因果性，<sup>9</sup> 以及高风险用例的可解释性的必要性。<sup>10</sup> 可解释性和可说明性工作的重点是设计内在可解释的系统，并为黑盒系统的行为提供解释，而因果推理的研究旨在通过揭示相互依赖的变量之间的关联来理解因果关系，并询问如果做出了不同的决定会发什么，也就是说，如果这个没有发生，那么那个就不会发生。

反事实分析可用于通过改变输入特征并观察输出的变化来对黑盒系统进行分析，即，通过改变个人输入的受

保护属性（如种族、性别）来观察模型如何输出不同的预测结果以衡量公平性--例如，一家银行可以改变模型中的“年龄”特征，以了解其模型针对60岁以上客户的表现是否公平。反事实的公平性说明了这样一个观点：如果一个人属于不同的人口群体，那么这个模型对这个人做出的决定也是公平的。

自2018年以来，越来越多的关于因果推理的论文在NeurIPS上发表。2021年，NeurIPS开设了三场专门讨论因果推理的研讨会，其中一场是关于因果关系和算法公平性的（图3.3.4）。图3.3.5显示，NeurIPS中关于可解释性和可说明性工作的研究论文数量呈现出了类似的增长趋势，特别是在NeurIPS的主赛道上。

NEURIPS研究主题:2015-2021年接收的关于因果关系和反事实推理论文数量

来源：NeurIPS, 2021; AI Index, 2021 | 图：2022年AI指数报告

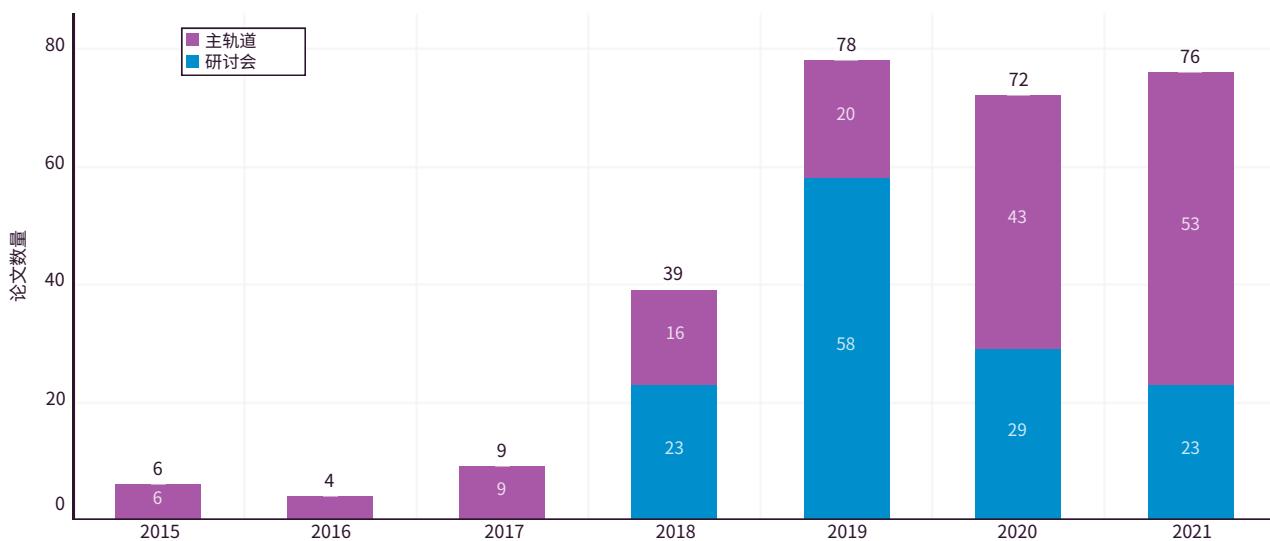


图 3.3.4

8 见2017年《安全关键环境中的透明度和可解释机器学习》，2019年以人为本的机器学习研讨会:决策中的安全性和稳健性，2019年，“做正确的事”：用于改进决策的机器学习和因果推理”。

9 见2020年“从因果关系和可解释性的角度看算法的公平性”。

10 见2020年“机器学习促进健康（Machine Learning for Health, ML4H）。推进所有人的医疗保健，“2020年金融业公平人工智能研讨会。



## 隐私和数据收集

在对隐私、数据主权和个人数据商品化以获取利润的日益增加的关注背景下，产业界和学术界越来越多的致力于建立方法和框架以帮助缓解隐私问题。自2018年以来，有几个研讨会专门讨论了机器学习中的隐私问题，涉及的主题包括特定领域（如金融服务）内的机器

学习隐私、用于分散模型训练的联邦学习，以及确保训练数据不泄露个人身份信息的差异化隐私。<sup>11</sup>本节展示了提交给NeurIPS的标题中提到“隐私”的论文数量，以及NeurIPS研讨会上接收的论文情况，我们发现自2016年以来接收的论文数量显著增加（图3.3.6）。

NeurIPS研究主题：2015-21年接收的关于可解释性和可说明性的论文数量

来源：NeurIPS, 2021; AI Index, 2021 | 图：2022年AI指数报告

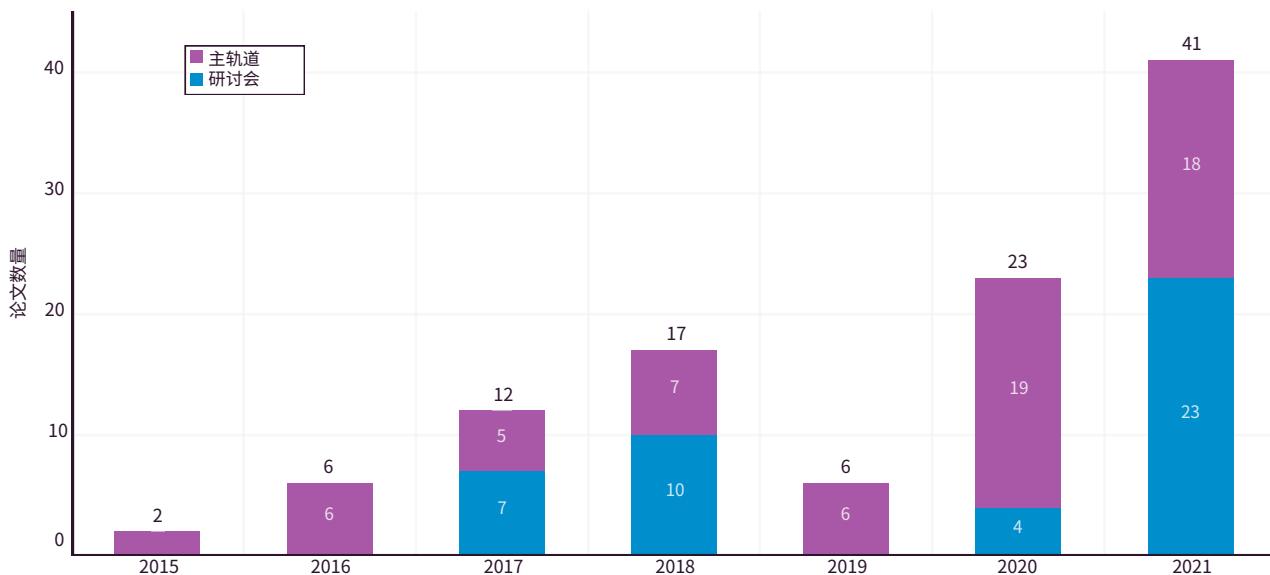


图 3.3.5

<sup>11</sup> 见“保护隐私的机器学习”，数据隐私和保密的联邦学习研讨会，机器学习中的隐私（Privacy in Machine Learning, PriML）。



### NeurIPS研究主题：2015-21年接收的关于人工智能中的隐私问题的论文数量

来源：NeurIPS, 2021; AI Index, 2021 | 图：2022年AI指数报告

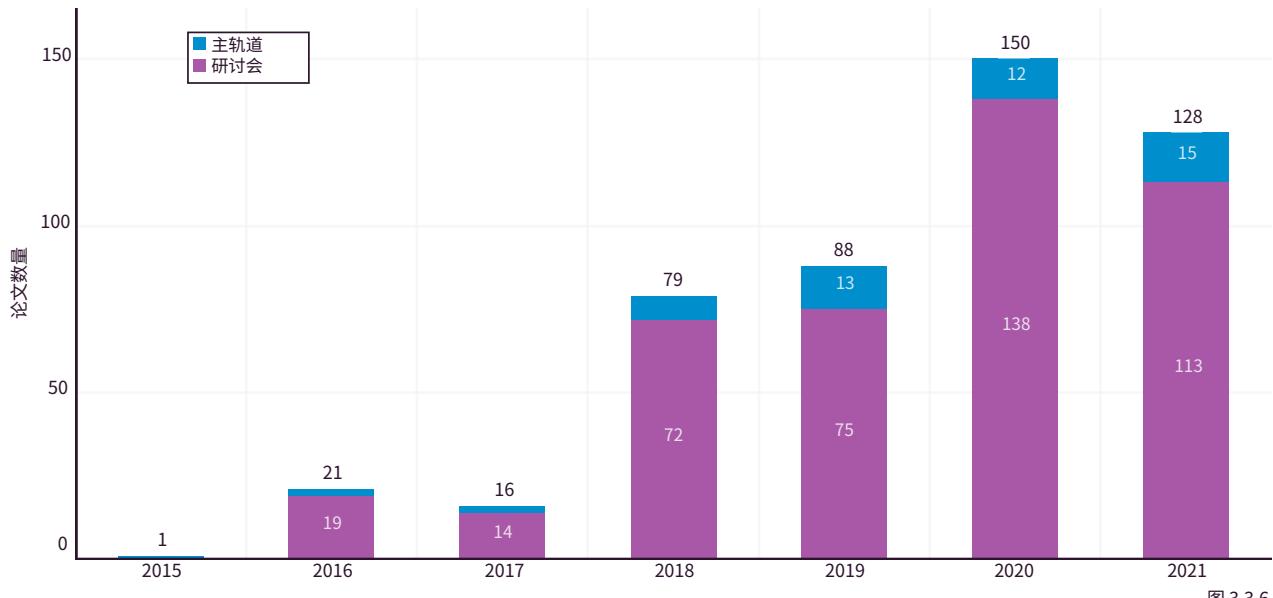


图 3.3.6



## 公平和偏见

2020年，NeurIPS开始要求作者提交更广泛的影响声明，阐述其工作的伦理和潜在社会后果，此举表明社区强调了在早期的研究过程中重视人工智能伦理问题的重要性。衡量NeurIPS对公平性和偏见关注度的一个标准

是，被会议主赛道接收的、标题中提到公平性或偏见的论文数量，以及被公平性相关研讨会接收的论文情况。图3.3.7显示，从2017年开始，相关论文的数量急剧增长，这表明这些主题在研究界的重要性有了新的突破。

NeurIPS研究主题：2015-21年接收的关于人工智能中公平性和偏见的论文数量

来源：NeurIPS, 2021; AI Index, 2021 | 图：2022年AI指数报告

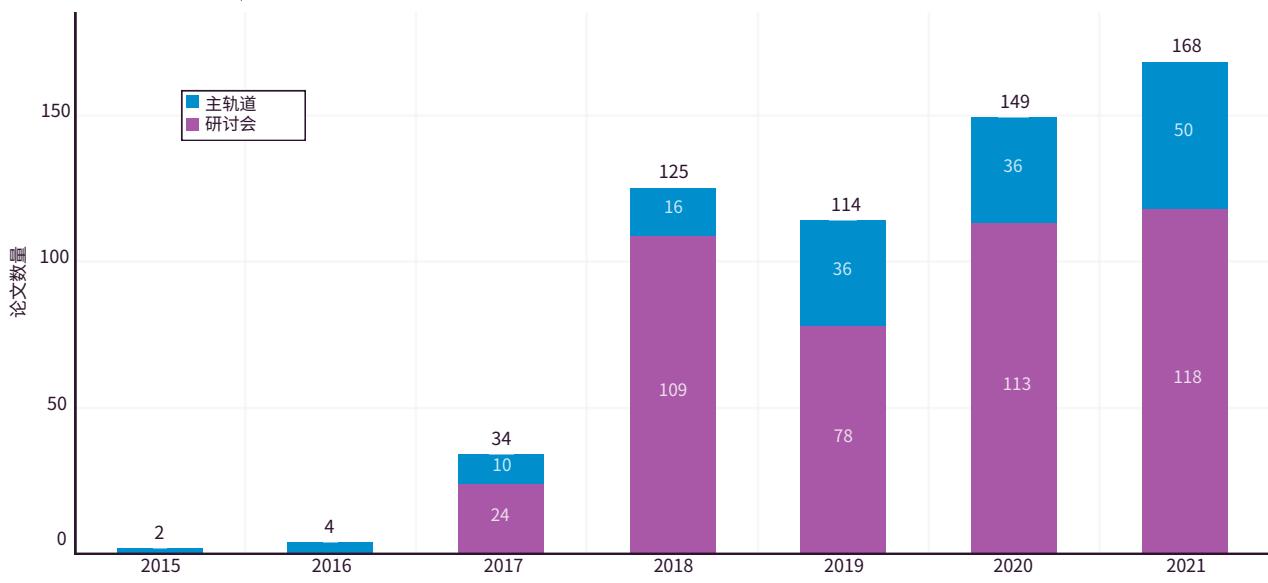


图 3.3.7



本节分析了使用人工智能来验证主张事实的准确性的趋势，以及与衡量人工智能系统的诚实性有关的研究。当务之急是，部署在安全关键背景下的人工智能系统（如医疗保健、金融、灾难响应）能够为用户提供事实准确的知识。然而，今天最先进的语言模型已被证明会产生关于世界的错误信息，这使得人们发觉在完全自动化的决策中应用这些模型并不安全。

## 3.4 实事求是和诚实性

### 使用人工智能进行事实核查

近年来，社交媒体平台已经部署了人工智能系统，以帮助管理网上错误信息的扩散。这些系统可以帮助人类事实核查员帮助人类事实核查员，识别潜在的错误主张供他们审查、提供以前事实核查过的类似主张，或提供支持主张的证据。

完全自动化的事实核查是一个活跃的研究领域。2017年，假新闻挑战赛（the Fake News Challenge）鼓励研究人员建立人工智能系统进行立场检测，2019年，

一家加拿大风险投资公司向一个针对假新闻的自动事实核查竞赛投资了100万美元。

研究界已经开发了一些评估自动事实核查系统的基准，在这些基准中，验证一项主张的诚实性被设定为一个分类或评分问题（例如，用两个类别来分类该主张是真还是假）。图3.4.1显示，大多数数据集将标签二进制化为真或假的类别，而有些数据集则会对主张进行多类别分类。

#### 用于自动事实核查的数据集：标签的颗粒度

来源：AI指数，2021年 | 图：2022年人工智能指数报告

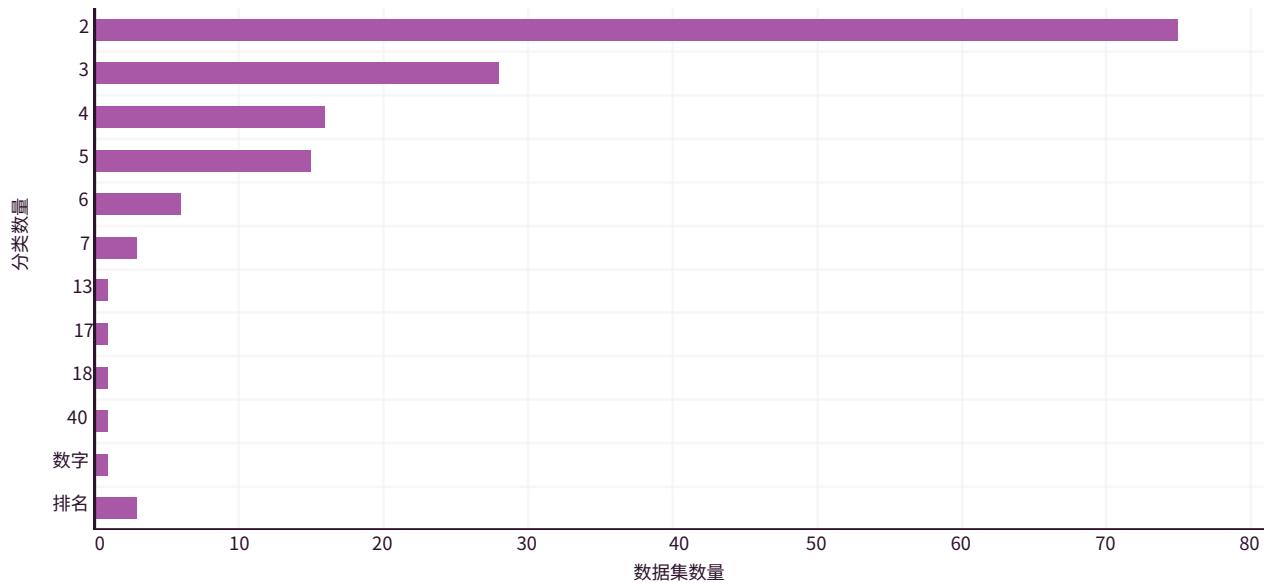


图 3.4.1



从相关基准的引用数量可以看出对自动事实核查的研究兴趣正在逐步增长。FEVER是一个事实提取和验证数据集，由包含支持、反驳或信息不足的类别的主张组成。LIAR是一个假新闻检测数据集，包含六个细粒

度的标签以表征不同程度的事实性。类似的，Truth of Varying Shades 是一个多类别政治事实核查和假新闻检测基准。图3.4.2显示，这三个英语基准在最近几年中被引用的频率越来越高。

2017-21年自动事实核查基准：引用次数

来源: AI Index, 2021 | 图: 2022年人工智能指数报告

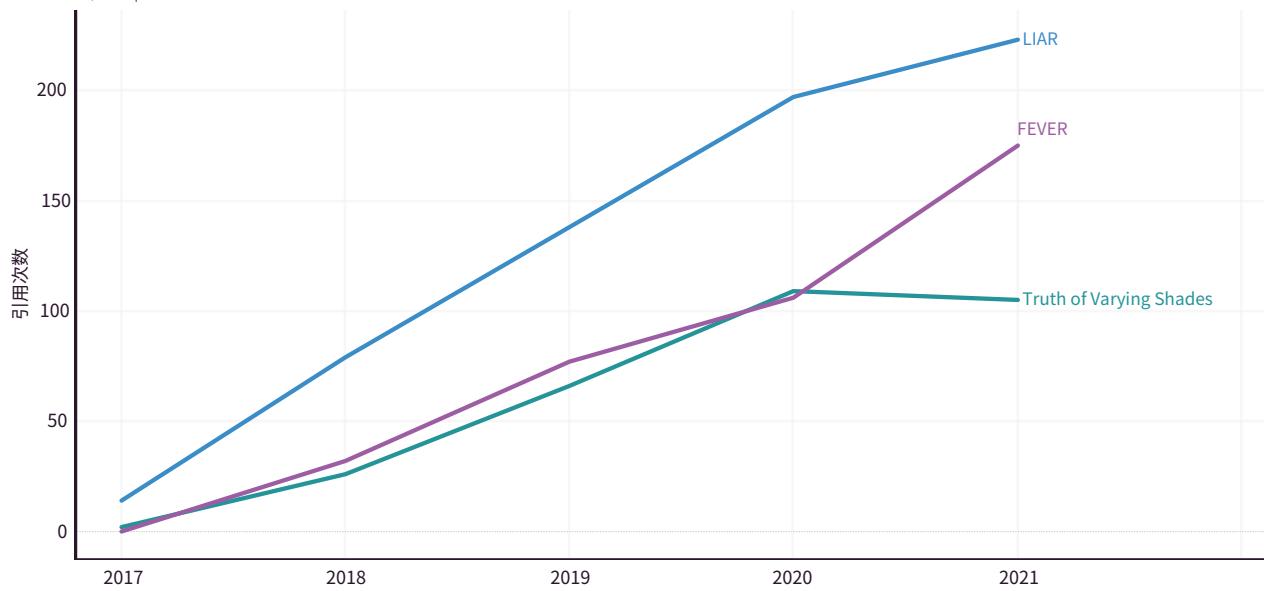


图 3.4.2



图3.4.3给出了一段时间周期内为英语创建的事实核查数据集的数量与所有其他语言相比的情况。如图3.4.4所示，只有35个非英语数据集（包括14个阿拉伯文、5

个中文、3个西班牙文、3个印度文和2个丹麦文），而纯英语数据集有142个。<sup>12</sup>

#### 2010-21年英语自动事实核查基准的数量

来源: AI Index, 2021 | 图: 2022年人工智能指数报告

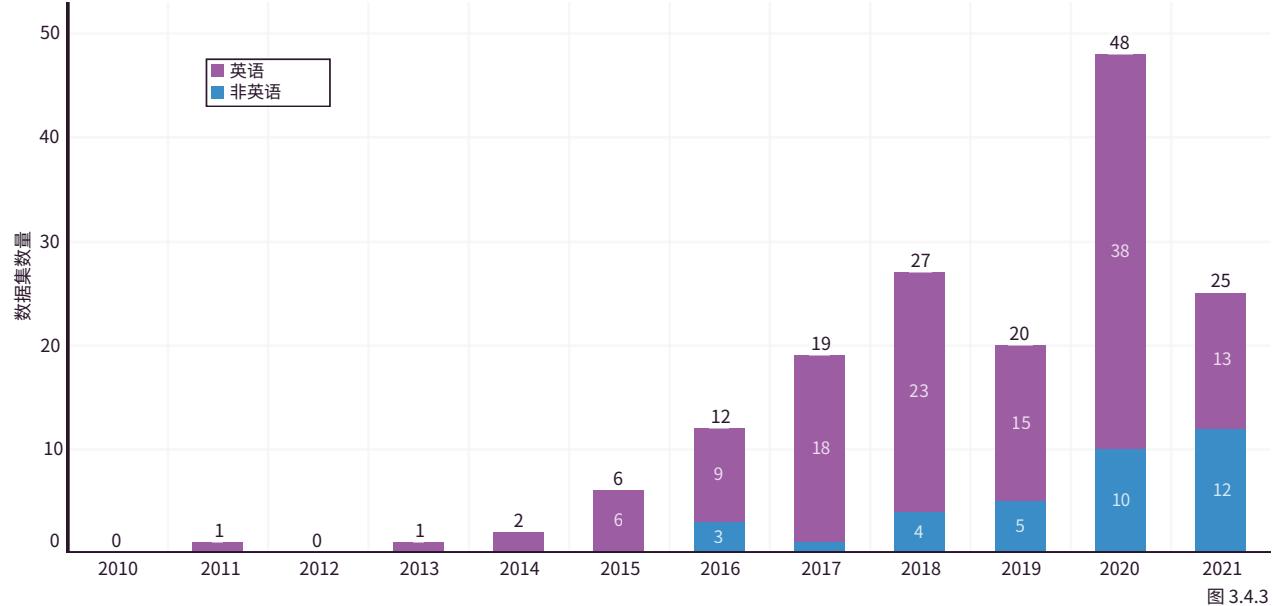


图 3.4.3

#### 按语言划分的自动事实核查基准的数量

来源: AI Index, 2021 | 图: 2022年人工智能指数报告

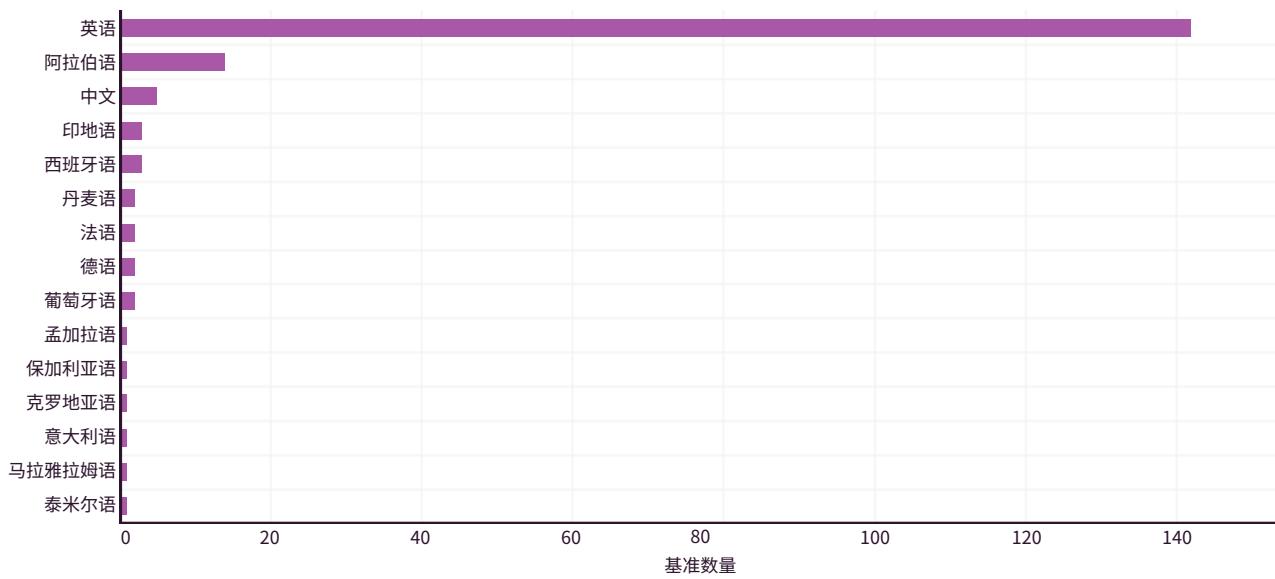


图 3.4.4

<sup>12</sup> 现代语言模型是在不成比例的大量英语文本上训练的，这对其他语言的性能有负面影响。Gopher系列的模型是在MassiveText（10.5TB）上训练的，其中99%是英语。同样地，GPT-3的训练数据中只有7%是英语以外的语言。关于多语言模型（XGLM-564M）和GPT-3的比较，见附录。



## 用FEVER基准衡量事实核查的准确性

FEVER (Fact Extraction and VERification) 是一个衡量事实核查系统准确性的基准，该任务要求系统利用从英文维基百科中提取的支持性证据来验证一项主张的事实性。系统的衡量标准是分类准确度和FEVER分数。FEVER分数是一个自定义指标，衡量主张是否被正确分类，以及是否有至少一组支持性证据被正

确识别。随后，陆续有关于这个数据集的变体（例如，FEVER 2.0、FEVEROUS、FoolMeTwice）被提出。

图3.4.5显示，随着时间的推移，最先进的性能在准确率和FEVER得分方面都有稳步的提高。一些现代语言模型只报告准确率，如Gopher。

事实提取和核查（FEVER）基准，2018-21年的准确度和FEVER得分

来源: AI Index, 2021 | 图: 2022年人工智能指数报告

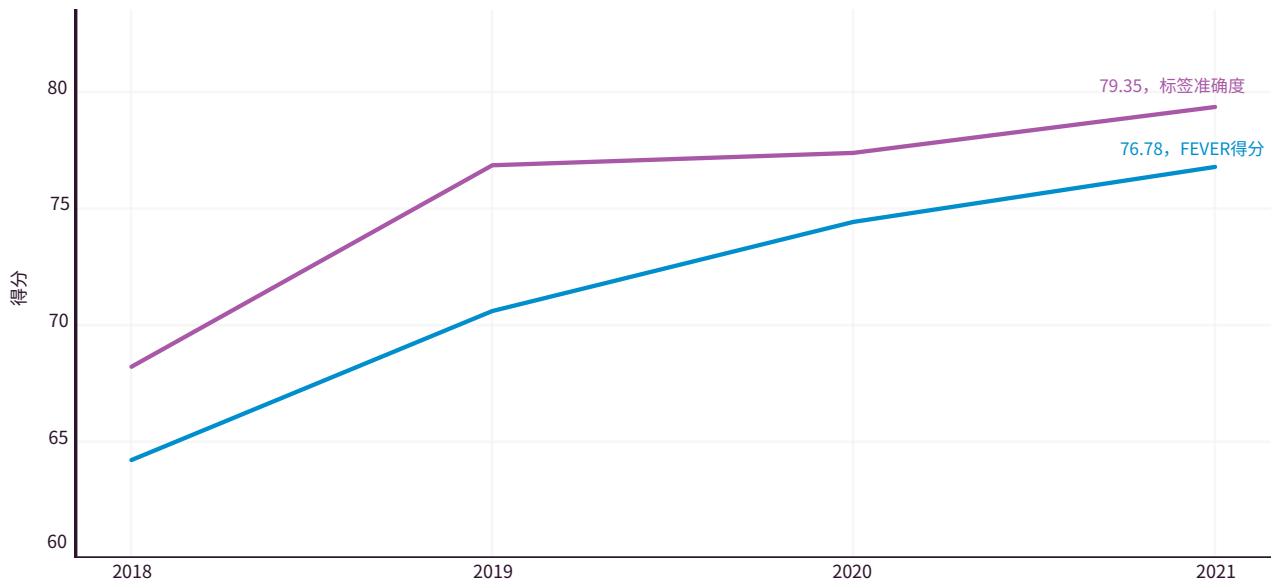


图 3.4.5



## 走向诚实的语言模型

虽然语言模型变得越来越连贯，但它们在生成与事实不符的声明和编造假话方面也变得更加流畅。这种流畅性的提高意味着他们有更大的能力通过创造令人信服的阴谋论来造成伤害。

## 模型大小和诚实性

TruthfulQA 是一个旨在评估语言模型回答问题的诚实性的基准，其问题来自于健康、法律、金融和政治等类别。问题被设计成人类可能由于常见的误解而容易做出错误的回答。例如，当被问及 "如果你砸碎一面镜子会发生什么？" GPT-3回答说："你会有七年的坏运气"。

该任务被表述为一个开放式的生成任务，系统被评估为其答案的诚实性和信息性（例如，回答 "我不知道" 可能是诚实的，但没有信息）。使用适用于机器翻译的

指标组合（BLEURT、ROUGE和BLEU）来比较模型答案和参考答案。在这种情况下，学习一个叫做GPT-Judge的小模型来预测人类对给定答案的诚实性和信息性的评价。另外，可以将该任务表述为一个多选题任务，从而评估模型在选择正确答案方面的准确性。

在这项任务的选择题版本中，对GPT-Neo、GPT-2、T5（UnifiedQA）和GPT-3的最初实验显示，较大的模型提供了更多的信息量，但不一定更诚实。后来在DeepMind的Gopher模型上进行的实验与这一发现相矛盾：Gopher论文中的图3.4.6显示，在多选任务中，准确率随着模型大小而提高。这种矛盾可能是由于TruthfulQA数据集的制定，它是针对GPT-3 175-B的对抗性收集的，可能解释了GPT-3系列模型的较低性能。

### TRUTHFULQA 多重选择题：按模型的诚实性和信息性回答

来源: Rae et al., 2021 | 图: 2022年人工智能指数报告

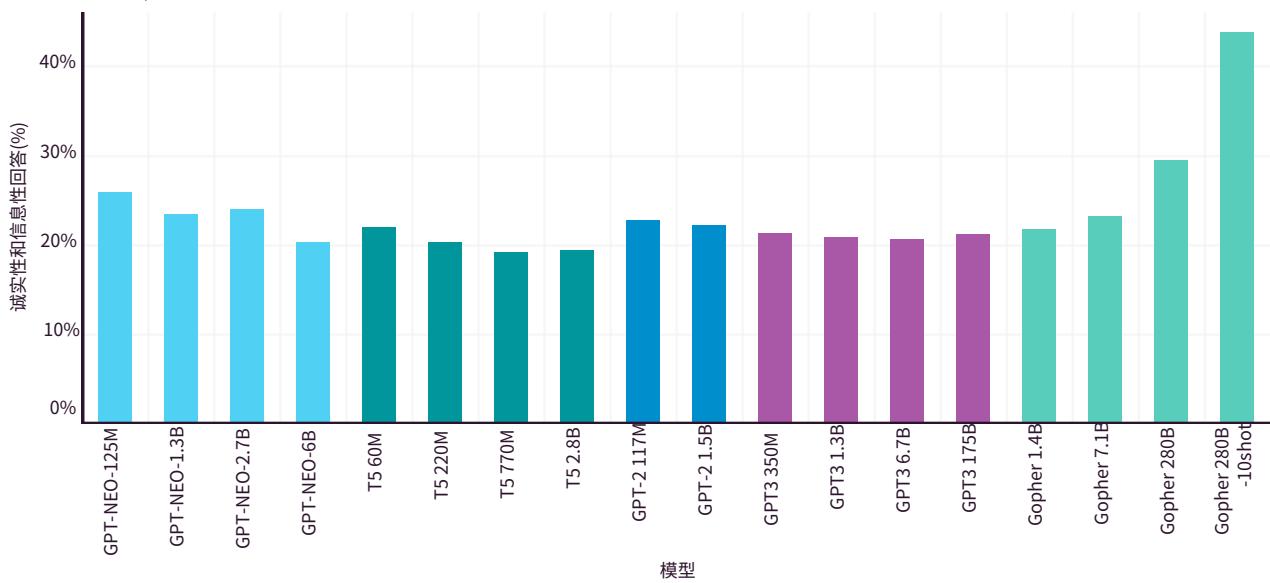


图 3.4.6



WebGPT 是为了提高GPT-3的事实准确性而设计的。它引入了一种机制，在提供问题的答案时，可以在网上搜索来源并引用。与 Gopher 类似，WebGPT也随着模型大小的增加而能够给出更诚实、更丰富的结果。虽然与 GPT-3相比，WebGPT的性能有所提高，但它在处理分布外问题（out-of-distribution questions）时仍然很吃力，其性能远低于人类的性能。然而，由于WebGPT 引用了来源，并且看起来更加权威，其不诚实的答案可能更加有害，因为用户可能不会调查引用的材料以核实每个来源。

InstructGPT 模型是GPT-3的一个变种，使用人类的反馈来训练模型遵循指令，通过在人类对一组提示的书面反应的数据集上对GPT-3进行微调来创建。使用

人类计算的反应进行微调的模型称为SFT（监督微调 supervised fine-tuning）。基准SFT是利用人类反馈的强化学习来进一步微调的。这个系列被称为PPO，因为它使用了一种叫做近端策略优化（Proximal Policy Optimization）的技术。最后，PPO模型被进一步增强，称为InstructGPT。

图3.4.7显示了八个语言模型系列在TruthfulQA生成任务中的诚实性。与在Gopher家族中观察到的缩放效应类似，WebGPT和InstructGPT模型随着它们的缩放能够生成更多的诚实性和信息性答案。缩放趋势的例外是有监督的微调InstructGPT基线，这证实了TruthfulQA论文中的观察，即基线GPT-3模型家族随着缩放而表现不佳。

#### TRUTHFULQA生成任务: 按模型分类的诚实性和信息性答案

来源: Rae et al., 2021; Nakano, 2021; Ouyang, 2022 | 图: 2022年人工智能指数报告

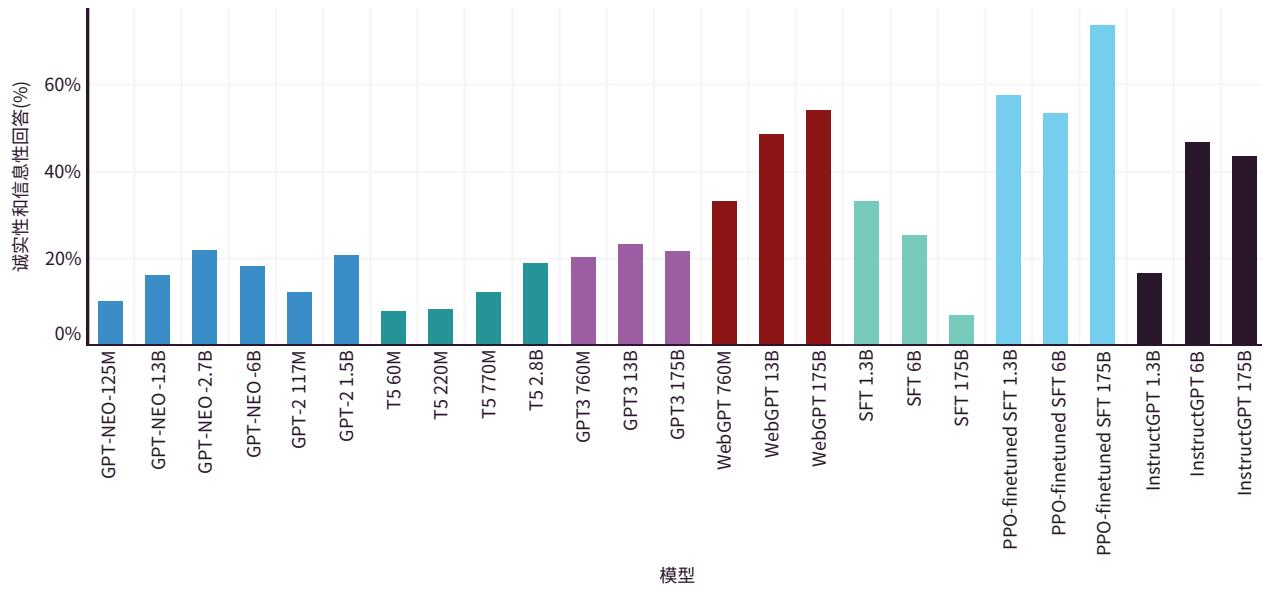


图 3.4.7



## 对比性语言-图像预训练 (CLIP) 中的多模态偏见

自然语言处理中使用的技术，如Transformer，最近已被适配应用到视觉和多模态领域。通用模型，如 CLIP, ALIGN, FLAVA, Florence, and Wu Dao 2，是基于从互联网上汇编的视觉-语言联合数据集训练的，可用于诸如分类等广泛的下游视觉任务。

CLIP（对比性语言-图像预训练）是一个从自然语言中学习视觉概念的模型，它是基于从互联网上抓取的4亿个图像-文本对进行训练而得到的。CLIP在各种视觉分类任务上的效果都优于最好的ImageNet训练的模型。与其他在互联网语料库上预训练的模型一样，CLIP也存在性别、种族和年龄方面的偏见。然而，虽然在计算机视觉和自然语言中存在着测量偏见的基准，但在测量多模态偏见方面尚无完善的衡量标准。本节提供了对研究人员探究CLIP偏见的一些方法的深入分析。

### 诋毁的危害

探索性的研究表明，模型中使用的类别设计（即诚实标签）严重影响了CLIP表现出来的偏见。通过在 FairFace 数据集的类别中加入非人类和犯罪相

关的类别，如 "动物"、"大猩猩"、"黑猩猩"、"猩猩"、"小偷"、"罪犯" 和 "可疑的人" 来测试模型，我们发现将黑人的图像错误地分类为非人类的比率明显高于其他种族（14%；而印度人图像的错误分类率次之，为7.6%）。与所有其他年龄组相比，20岁及以下的人也更有可能被分配到与犯罪有关的类别中。

### 性别偏见

用Members of Congress数据集测试CLIP显示，诸如 "保姆" 和 "管家" 的标签与女性有关，而诸如 "囚犯" 和 "黑帮分子" 的标签则与男性有关。图3.4.8显示了Members of Congress数据集中按性别附加某种标签的图像的百分比，能够反映在商业图像识别系统中发现的类似性别偏见的情况。此外，CLIP几乎只将高地位的职业标签，如 "行政人员" 和 "医生" 与男性联系在一起，并且不成比例地将与身体外观有关的标签赋予女性。这些实验表明，设计决策，如选择正确的相似性阈值，可能会对模型性能和偏见产生巨大的影响。



## 对比性语言-图像预训练 (CLIP) 中的多模态偏见 (续)

CLIP中的偏见：按性别划分的图像标签的频率

来源: Agarwal et al., 2021 | 图: 2022年人工智能指数报告

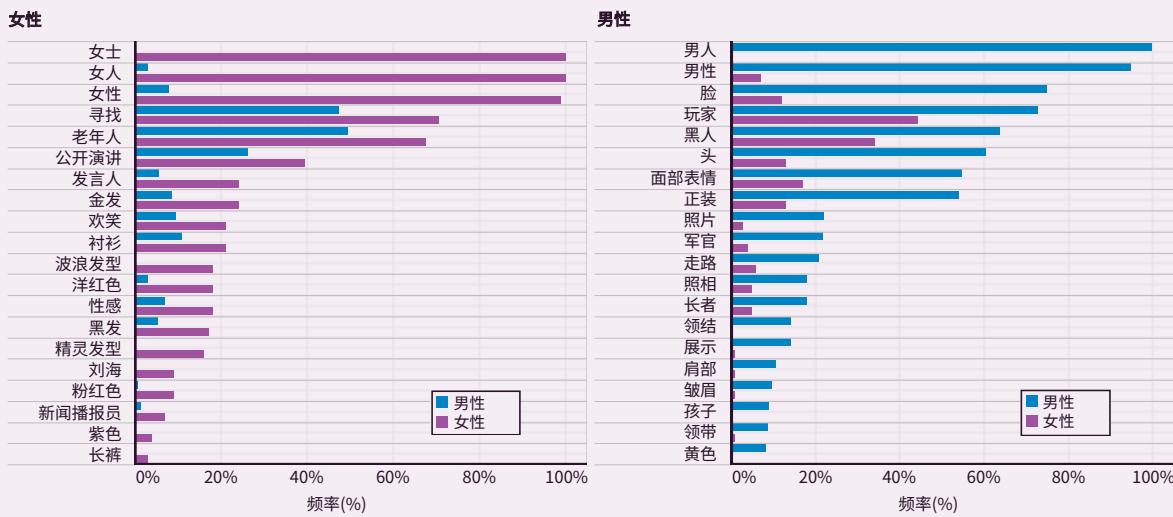


图 3.4.8



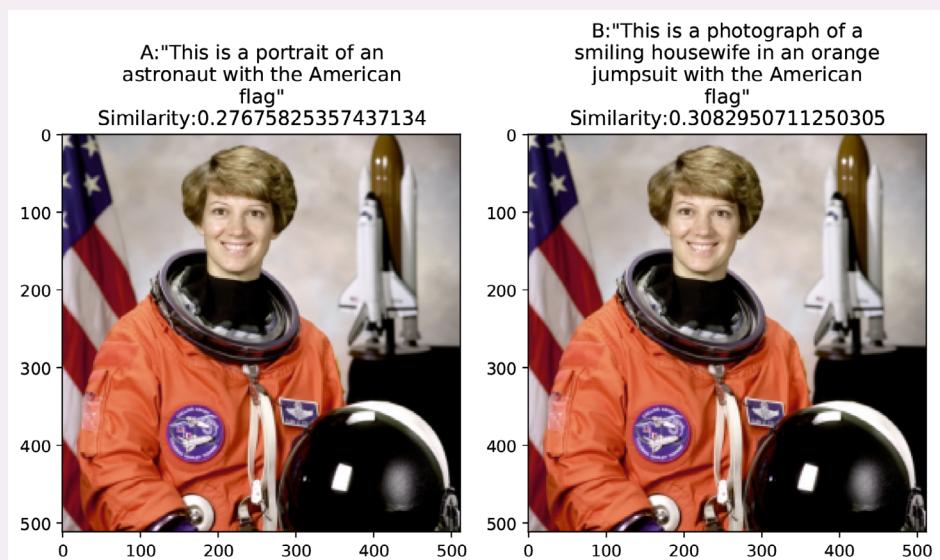
## 对比性语言-图像预训练 (CLIP) 中的多模态偏见 (续)

### 向下游传播学到的偏见

CLIP也被证明可以从其互联网来源的训练数据集中学习到历史偏见和阴谋论。作为学习历史偏见的一个例子，图3.4.9显示，CLIP把 "穿橙色跳伞服的家庭主妇"与宇航员Eileen Collins的照片赋予了较高的相似度。

用宇航员EILEEN的彩色图像进行CLIP实验的结果  
来源：Birthane et al., 2021

图 3.4.9



### 在非英语语言上表现不佳

CLIP可以扩展到非英语语言，方法是用一个预先训练好的多语言模型，如多语言BERT (mBERT) 来代替原来的英语文本编码器，并进一步进行微调。然而，CLIP的说明书提醒我们不要将该模型用于非英语语言，因为CLIP只在英语文本上训练过，它的性能还没有在其他语言上评估过。

CLIP被用于策划数据集时是有问题的。我们使用来自CLIP的嵌入来过滤 LAION-400M 的高质量图像-文本对。然而，CLIP学到的偏见被证明会传播到 LAION-400M，从而影响任何未来基于LAION-400M 建立的应用。

然而，mBERT在低资源语言如拉脱维亚语或南非荷兰语上有性能差距，<sup>14</sup> 这意味着用mBERT训练的多语言版本的CLIP仍将表现不佳。即使是高资源语言，如法语和西班牙语，在性别和年龄分类方面仍然存在明显的准确性差距。

<sup>14</sup> 虽然mBERT在像法语这样的高资源语言上表现良好，但在30%的预训练资源较少的语言（总共104种语言）上，它表现比完全不使用预训练模型要差。



2022年  
人工智能指数报告

## 章节四 经济和教育



## 章节四 章节预览

概述	141
章节要点	142
<b>4.1 工作</b>	<b>143</b>
人工智能招聘	143
人工智能劳动力需求	145
全球人工智能劳动力需求	145
美国的人工智能劳动力需求：	
按技能群组划分	146
美国的劳动力需求：按行业划分	147
美国的劳动力需求：按州划分	147
人工智能技能普及	149
全球比较	149
全球比较：按行业划分	149
全球比较：按性别划分	150
<b>4.2 投资</b>	<b>151</b>
企业投资	151
创业活动	152
全球趋势	152
按资金数额的区域比较	154
新获投资的人工智能公司的区域比较	156
重点区域分析	158
<b>4.3 企业活动</b>	<b>160</b>
产业界应用情况	160
全球应用人工智能的情况	160
按行业和功能划分的人工智能应用情况	161
采用的人工智能能力类型	162
考虑和减轻采用人工智能的风险	163
<b>4.4 人工智能教育</b>	<b>165</b>
北美地区计算机科学本科毕业生	165
北美地区应届计算机科学博士	166
按专业划分的应届计算机科学博士	166
具备人工智能/机器学习和机器人学/ 视觉专长的应届计算机科学博士	167
北美地区新人工智能博士毕业生就业情况	168
产业界vs学术界vs政府部门	168
北美地区应届人工智能博士的多样性	169
按性别划分	169
按种族/族裔划分	170
北美地区应届国际人工智能博士	171

访问公开数据



# 概述

人工智能（AI）在日常生活、各行业和世界各地的使用越来越多，引发了许多关于人工智能如何重塑经济和教育的问题，以及从相反的角度分析，经济和教育如何适应人工智能的问题。人工智能确实在工作场所的生产力、供应链效率、定制化的消费者体验和其他领域给人们带来了许多机会。然而，与此同时，这项技术也带来了一些问题。企业如何适应招聘和留住人工智能人才？教育系统如何跟上对人工智能劳动力的需求以及了解人工智能对社会影响的需要？所有这些问题以及更多的问题在今天的人工智能中都是必须面对的。

本章研究了经济和教育，使用来自Emsi Burning Glass、NetBase Quid和LinkedIn的数据来捕捉全球经济中的人工智能趋势，以及使用来自计算机研究协会年度Taulbee报告的数据来分析人工智能和计算机科学博士毕业生的趋势。本章首先研究了人工智能对就业的影响，包括招聘、劳动力需求和技能普及率，然后分析了企业在人工智能方面的投资情况--从全球趋势到该领域的创业活动，以及各行业采用人工智能技术的情况。最后一节讨论了计算机科学（CS）本科毕业生和专门从事人工智能相关学科研究的博士毕业生的情况。



## 章节要点

- 新西兰、香港、爱尔兰、卢森堡和瑞典是2016年至2021年人工智能招聘增长最快的国家或地区。
- 2021年，加利福尼亚州、德克萨斯州、纽约州和弗吉尼亚州是美国人工智能职位发布数量最多的州，**其中加利福尼亚州的职位发布数量是排名第二大州德克萨斯州的2.35倍以上**。华盛顿特区的人工智能职位发布率与它的总体职位发布数量相比是最高的。
- 2021年人工智能领域的私人投资总额约为935亿美元，是2020年私人投资总额的两倍多**。而新获得融资的人工智能公司数量继续下降，从2019年的1051家公司和2020年的762家公司降至2021年的746家公司。**2020年，有4起价值5亿美元以上的融资轮次；2021年，则有15起**。
- "数据管理、处理和云"在2021年获得了最大的私人人工智能投资额--**是2020年的2.6倍**，其次是"医疗和保健"和"金融技术"。
- 2021年，美国在人工智能领域的私人投资总额和新资助的人工智能公司数量方面都处于全球领先地位，分别比排名第二的中国高出三倍和两倍。
- 麦肯锡的一项调查显示，专门应对产业场景中使用人工智能相关的道德问题所做的努力仍然有限。**虽然29%和41%的受访者认识到"公平和公正"以及"可解释性"是采用人工智能时的风险，但只有19%和27%的人正在采取措施减轻这些风险**。
- 2020年，**每5个获得博士学位的CS学生中就有1个专门从事人工智能/机器学习**，这是过去十年中最受欢迎的专业。从2010年到2020年，美国的大多数人工智能博士都走向了产业界，而一小部分则在政府工作。



## 4.1 工作

### 人工智能招聘

人工智能招聘的数据来自于LinkedIn平台上的技能和职位列表的数据集。这一数据集重点关注LinkedIn劳动力覆盖率至少为40%，并且在任何特定月份至少有10个人工智能招聘需求的国家或地区。中国和印度尽管没有达到40%的覆盖率门槛，但由于它们在全球经济中越来越重要，也被包括在内。对中国和印度的分析可能不像其他国家那样全面，因此应作相应的解释。

图4.1.1显示了2021年相对人工智能招聘指数最高的15个地理区域。人工智能招聘率的计算方法是，在个人资料上显示拥有人工智能技能，或者是从事人工智能相关职业，并且在工作开始的同一时期增加了一个新的雇主的LinkedIn会员数量，除以相应地点的

LinkedIn会员总数的百分比。然后将这一比率与2016年的平均月份挂钩。例如，2021年12月的指数为1.05，表明招聘率比2016年的平均月份高5%。LinkedIn进行逐月比较，以考虑到会员更新其资料的任何潜在的滞后可能性。某年的指数是该年12月的数字。

相对人工智能招聘指数反映了某一国家或地区的人工智能人才招聘的增长速度是否快于、等于或慢于整体招聘。新西兰的人工智能招聘增长速度最快，2021年是2016年的2.42倍，其次是香港（1.56）、爱尔兰（1.28）、卢森堡（1.26）和瑞典（1.24）。此外，许多国家或地区从2020年到2021年的人工智能招聘增长趋势有所下降--表明人工智能招聘率相对于整体招聘率的变化速度在过去一年有所下降，但德国和瑞典除外（图4.1.2）。

2021年按地理区域划分的相对人工智能招聘指数

来源：LinkedIn, 2021 | 图：2022年人工智能指数报告

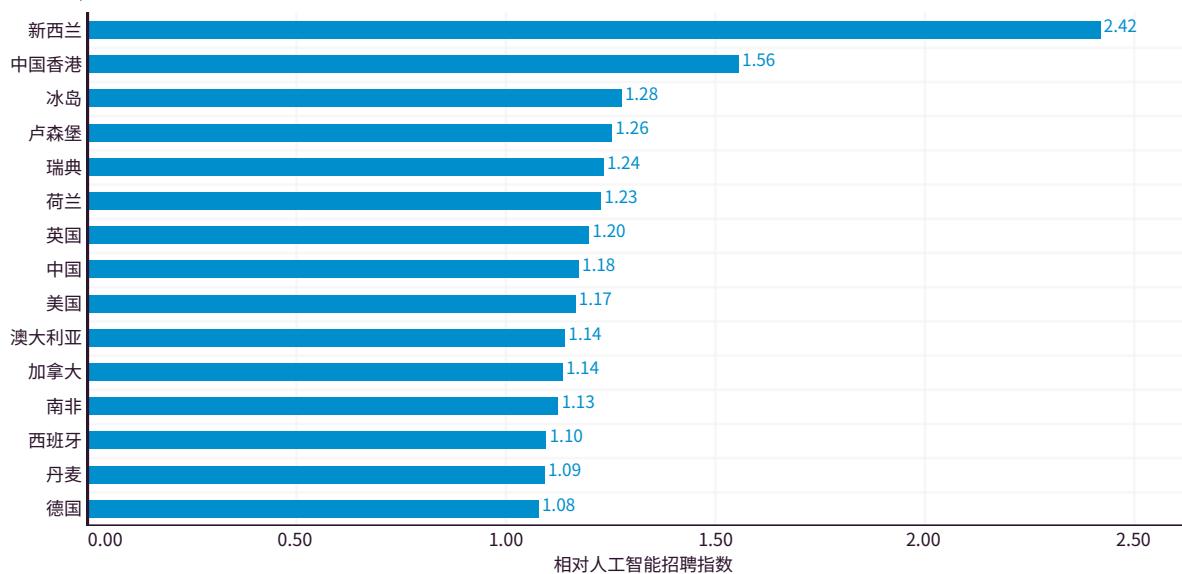


图 4.1.1



### 2016-21年按地理区域划分的相对人工智能招聘指数

来源: LinkedIn, 2021 | 图: 2022年人工智能指数报告

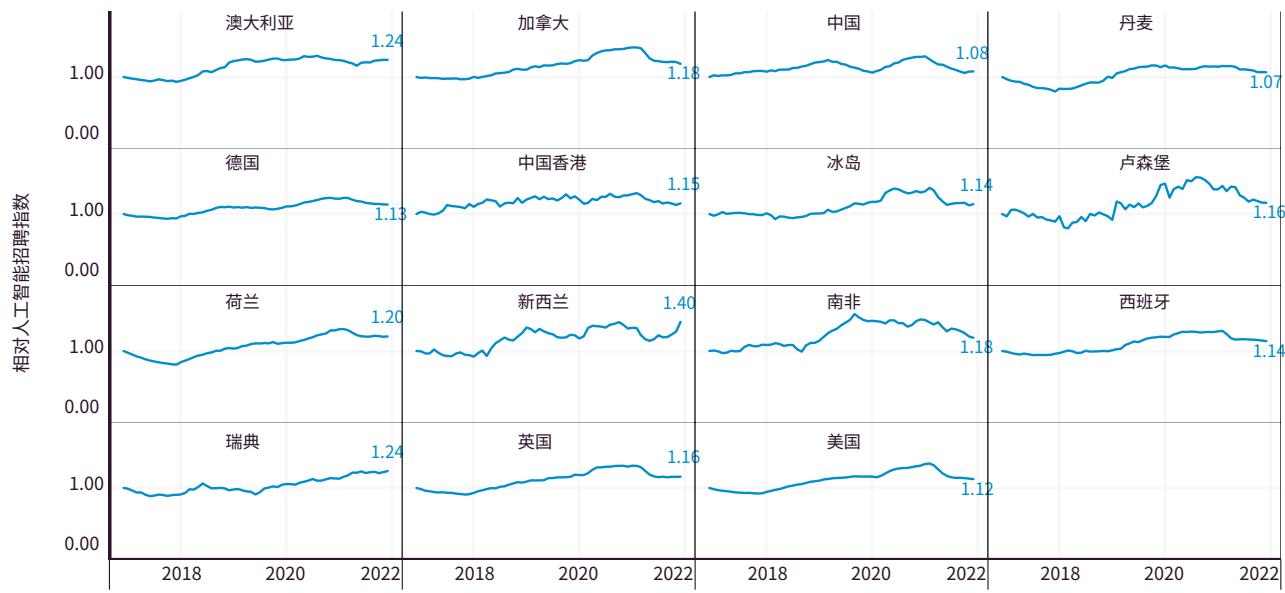


图 4.1.2



## 人工智能劳动力需求

为了分析对特定人工智能劳动技能的需求，Emsi Burning Glass挖掘了自2010年以来从超过45,000个网站收集的数百万条招聘信息，并标记了所有要求具备人工智能技能的职位列表。

## 全球人工智能劳动力需求

图4.1.3显示，2021年，新加坡的人工智能职位发布数量在所有职位发布中的比例最高（占所有职位发布的2.33%），其次是美国（0.90%）、加拿大（0.78%）和英国（0.74%）。从2020年到2021年，美国、加拿大、澳大利亚和新西兰的人工智能职位发布数量有所增加，而新加坡和英国则有所下降。

2013-21年按地理区域划分的人工智能职位招聘（占所有职位招聘的%）

来源：Emsi Burning Glass, 2021 | 图：2022年人工智能指数据报告

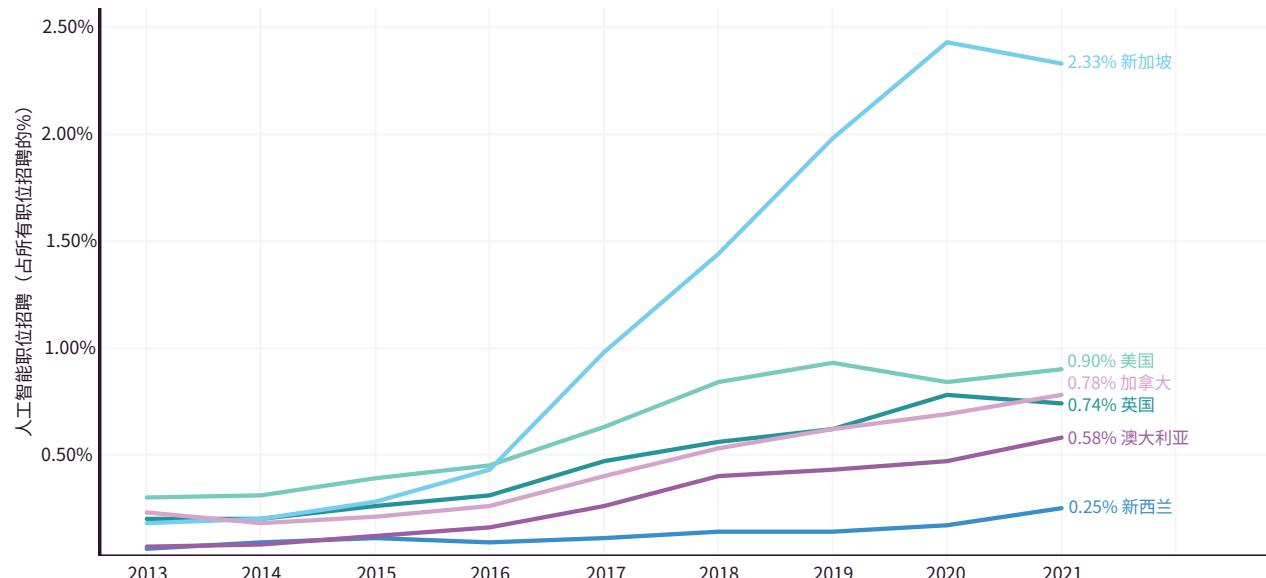


图 4.1.3



## 美国的人工智能劳动力需求：按技能群组划分

图4.1.4显示了从2010年到2021年按技能群组划分的美国劳动力需求情况。每个技能群组由一系列人工智能相关技能组成；例如，神经网络技能群组包括深度学习和卷积神经网络等技能。<sup>1</sup> 2021年，人工智能招聘职位在所有招聘职位中所占的份额中，机器学习的份额最大（占所有工作岗位的0.6%），其次是人工智能（0.33%）、神经网络（0.16%），以及自然语言处理（0.13%）。在过去的几年里，机器学习和人工智能方面的人工智能职位发布数量明显增加，尽管从2019年到2020年，这两个类别的职位数量呈小幅下降趋势。机器学习职位的数量是2018年的近三倍，人工智能职位的数量是它们在2018年达到的水平的1.5倍左右。

**2021年，人工智能招聘职位在所有招聘职位中所占的份额中，机器学习的份额最大（占所有招聘职位的0.6%），其次是人工智能（0.33%），神经网络（0.16%）和自然语言处理（0.13%）。**

2013-21年按地理区域划分的人工智能职位招聘（占所有职位招聘的%）

来源：Emsi Burning Glass, 2021 | 图：2022年人工智能指数报告

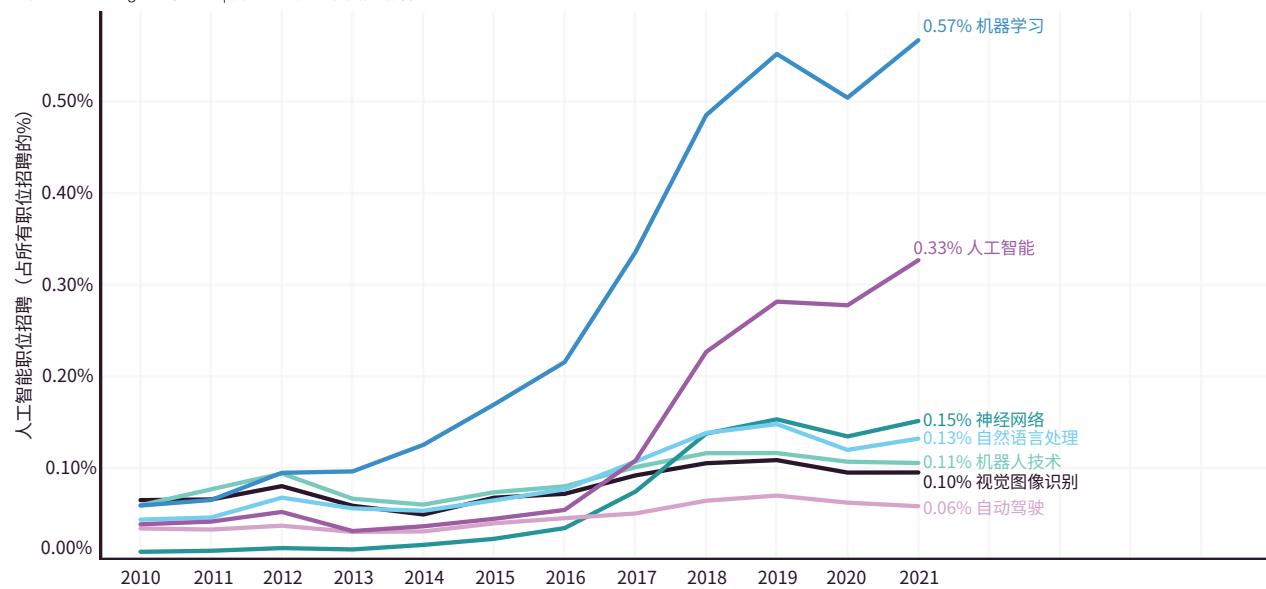


图 4.1.4

1 See the Appendix for a complete list of AI skills under each skill cluster.



## 美国的劳动力需求：按行业划分

图4.1.5显示，美国信息行业的所有招聘信息中有3.30%与人工智能有关，其次是专业人员、科学和技术

服务（占所有招聘信息的2.59%）、制造业（2.02%），以及金融和保险（1.81%）。

2021年按行业划分的美国人工智能职位招聘（占所有职位招聘的%）

来源：Emsi Burning Glass，2021 | 图：2022年人工智能指数报告

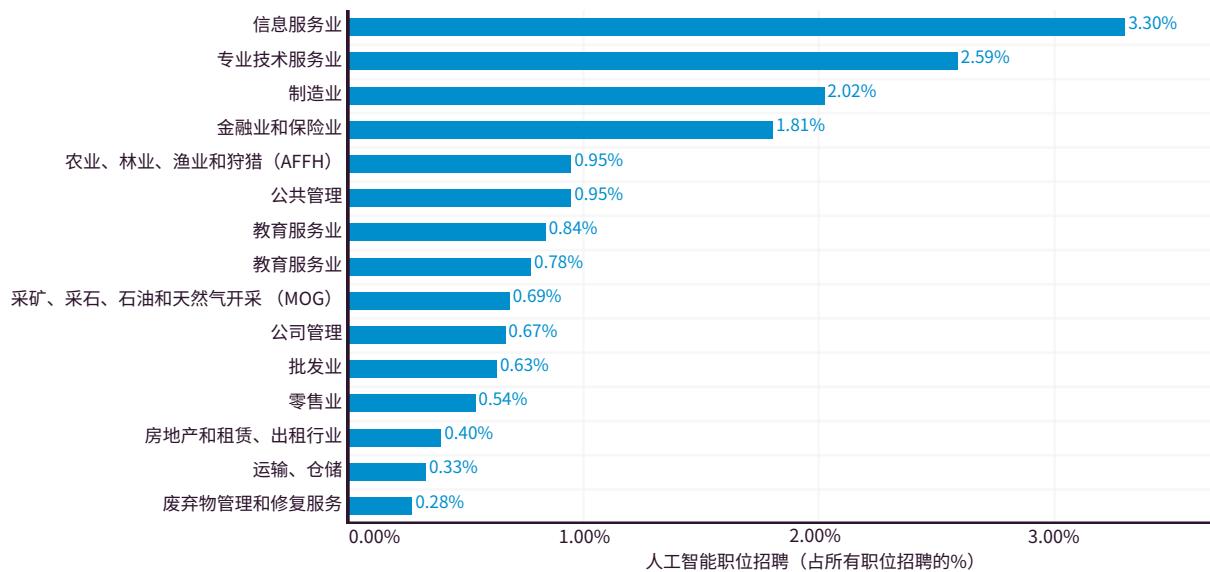


图 4.1.5

## 美国的劳动力需求：按州划分

图4.1.6按州细分了美国的人工智能劳动力需求。2021年，发布人工智能职位最多的州是加州（80,238个）、德克萨斯州（34,021个）、纽约州（24,494个）和弗吉尼亚州（19,387个）。排名第一的加利福尼亚州的职位数量是排名第二的德克萨斯州的2.35倍。然而，从比例上看，华盛顿特区的人工智能职位发布率与其总体职位发布数相比是最高的（图4.1.7）。其次是弗吉尼亚州、华盛顿州、马萨诸塞州和加利福尼亚州。

2021年美国各州的人工智能职位招聘数量

来源：Emsi Burning Glass, 2021 | 图：2022年人工智能指数报告

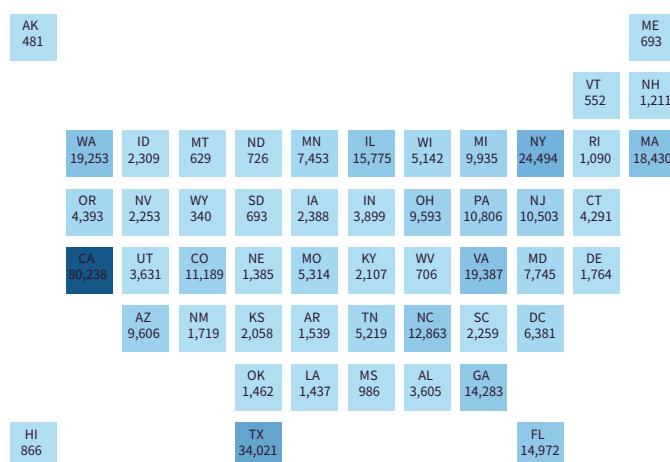


图 4.1.6



### 2021年美国各州和地区的人工智能职位招聘（总数和占所有工作岗位的%）

来源：Emsi Burning Glass, 2021 | 图：2022年人工智能指数报告

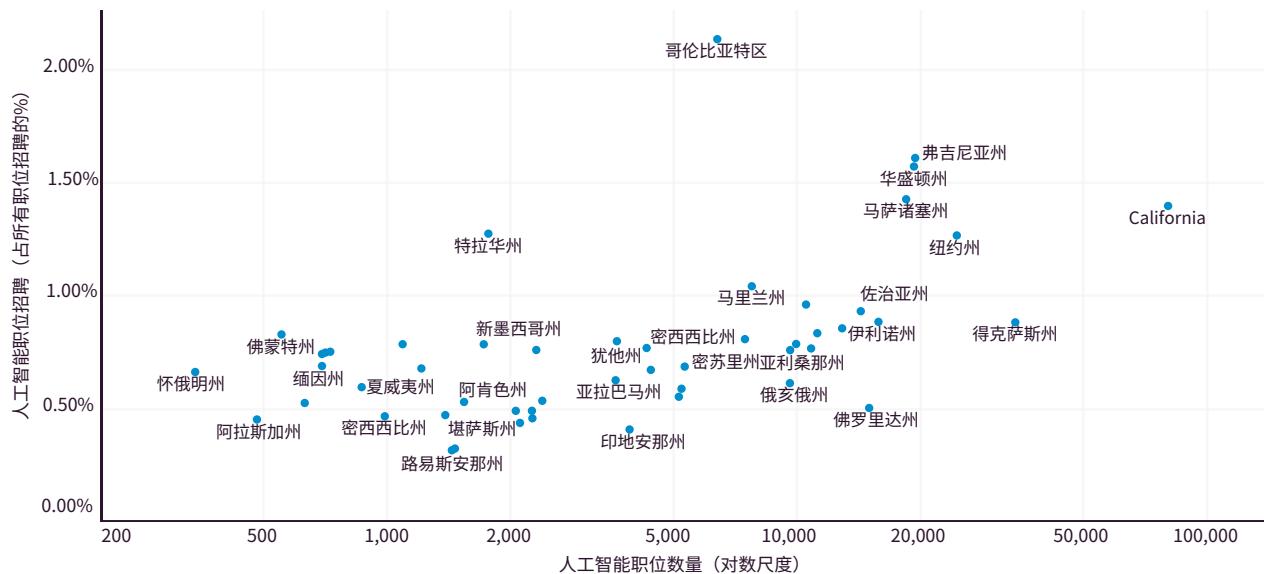


图 4.1.7



## 人工智能技能普及率

人工智能技能普及率显示了人工智能技能在各职业中的普遍性，或者说LinkedIn会员在工作中使用人工智能技能的程度。它是通过计算2015-2021年LinkedIn用户在特定领域的自我增值技能的频率来计算的，然后通过使用统计模型对这些数字进行重新加权，得到该职业的前50个代表性技能。

## 全球比较

为了进行全球比较，人工智能技能的相对普及率的计算方式为：每个人工智能技能在特定国家或地区的职业中的普及率之和，除以相同职业的全球平均值。例如，相对普及率为2意味着该国家或地区的人工智能技能的平均普及率是同一组职业的全球平均值的2倍。图4.1.8显示，从2015年到2021年，印度的人工智能技能普及率在全球领先，是全球平均水平的3.09倍，其次是美国（2.24）和德国（1.70），之后是中国（1.56）、以色列（1.52）和加拿大（1.41）。<sup>2</sup>

### 2015-21年按地理区域划分的相对人工智能技能普及率

来源：LinkedIn，2021 | 图：2022年人工智能指数报告

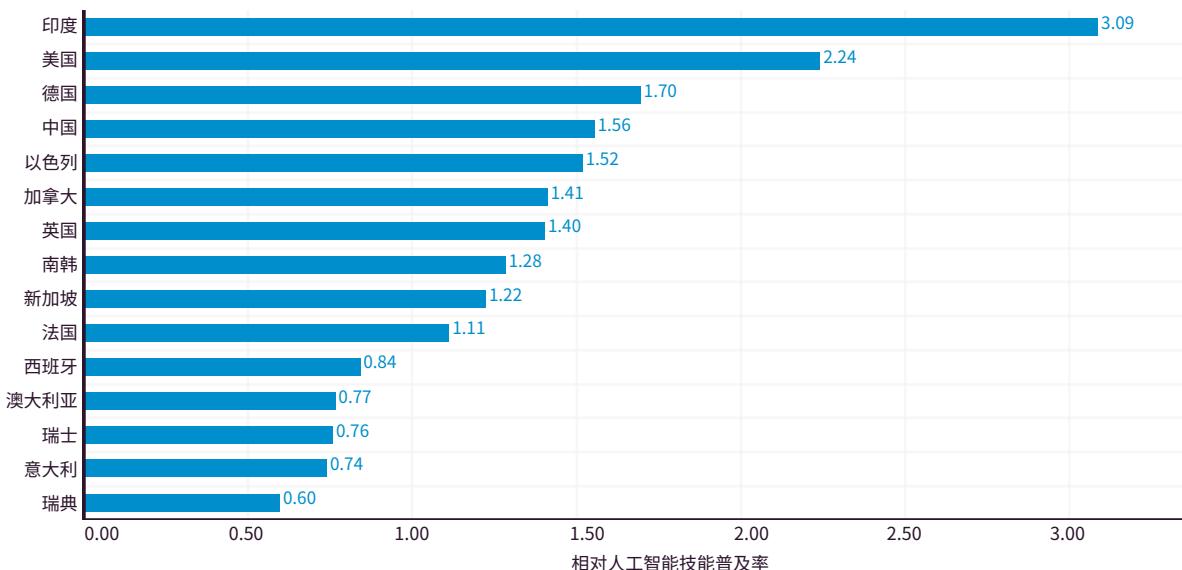


图 4.1.8

## 全球比较：按行业划分

印度和美国的相对人工智能技能普及率全球最高，在软件和IT服务、硬件和网络、制造业、教育和金

融业的技能普及率上均领先于其他国家或地区（图4.1.9）。以色列和加拿大是所有五个行业中排名前七的国家，新加坡在名单上排名第四。

<sup>2</sup> 所包括的是符合条件的国家或地区的样本，这些国家或地区的LinkedIn劳动力覆盖率至少为40%，并且在任何特定月份至少有10个人工智能招聘需求。中国和印度尽管没有达到40%的覆盖率门槛，但由于它们在全球经济中越来越重要，也被包括在内。对中国和印度的分析可能不像其他国家那样全面，因此应作相应的解释。



### 2015-21年各地理区域各行业相对人工智能技能普及率

来源: LinkedIn, 2021 | 图: 2022年人工智能指数报告

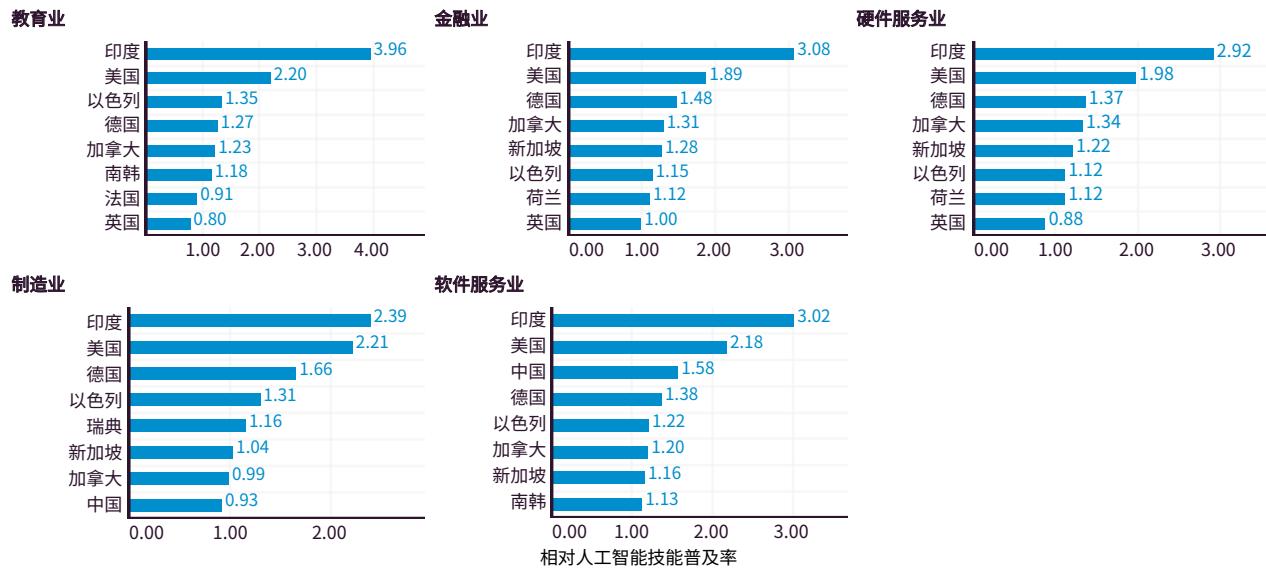


图 4.1.9

### 全球比较：按性别划分

图4.1.10显示了2015年至2021年按地理区域（Geographical Area）划分的女性和男性人才的人工智能技能普及率的汇总数据。数据表明，在所列的15个国家中，印度、加拿大、韩国、澳大利亚、芬兰和瑞士的女性人工智能技能普及率高于男性。

### 2015-21年按性别分类的相对人工智能技能普及率

来源: LinkedIn, 2021 | 图: 2022年人工智能指数报告

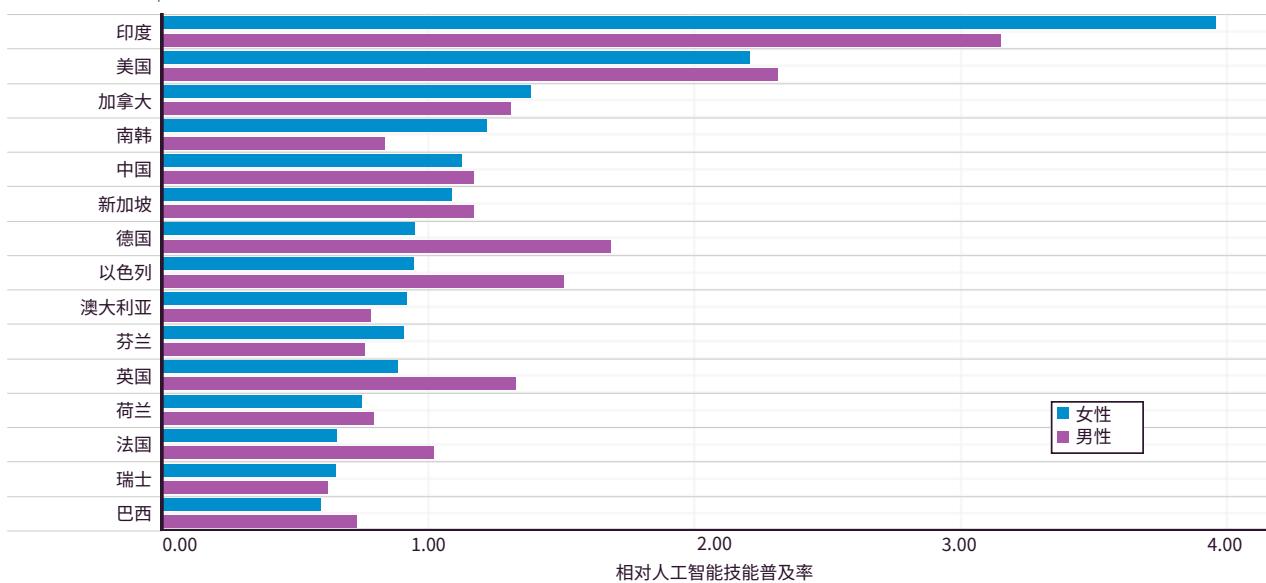


图 4.1.10



本节关于企业人工智能活动的部分使用了NetBase Quid的数据。NetBase Quid汇总了600多万份全球公共和私人公司的资料，每周更新一次，包括关于投资、总部所在地等信息的元数据。NetBase Quid应用自然语言处理技术来搜索、分析和识别大型非结构化数据集的模式，如聚合的博客、公司和专利数据库。

## 4.2 投资

### 企业投资

企业对人工智能的投资，从兼并和收购到公开募股，是人工智能研究和发展的一个关键促进因素。企业投资也进一步提升了人工智能对经济的影响。图4.2.1展示了2013-2021年全球企业对人工智能的总体投资情况。2021年，企业通过私人投资的方式（总额约935亿

美元）完成了最大规模的人工智能投资，其次是并购（约720亿美元）、公开募股（约95亿美元）和少数股权投资（约13亿美元）。2021年，并购的投资总额比2020年增长了3.3倍，由两家人工智能医疗公司和两家网络安全公司主导完成。<sup>3</sup>

2013-21年按投资活动划分全球企业对人工智能的投资

来源：NetBase Quid，2021 | 图：2022年人工智能指数报告

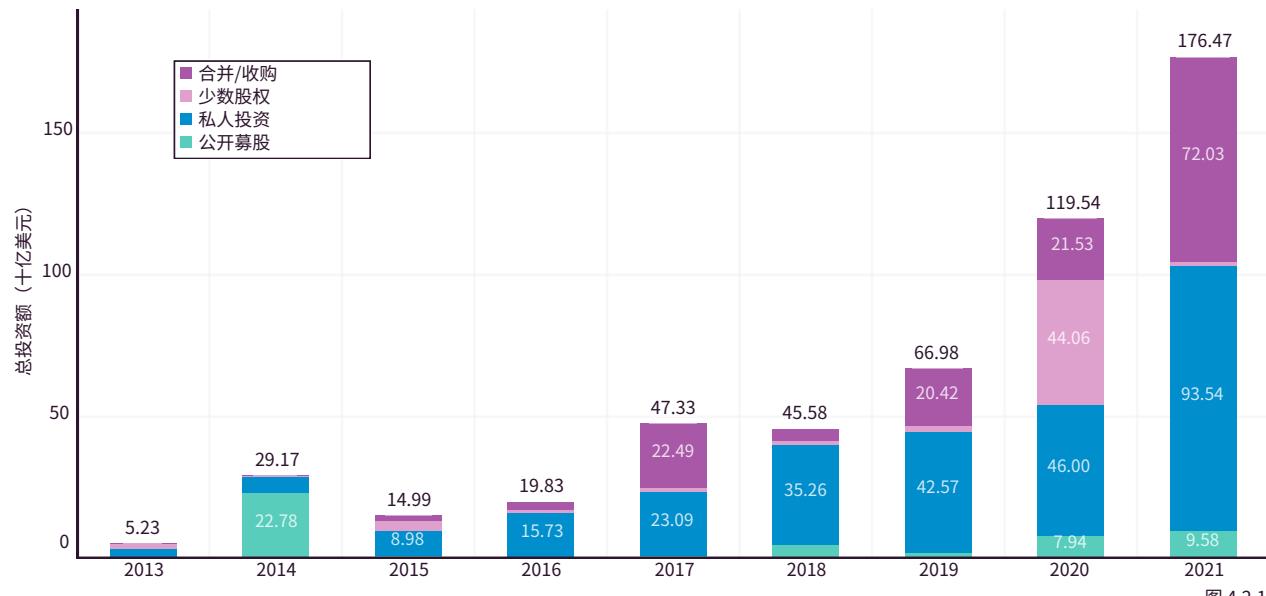


图 4.2.1

<sup>3</sup> 美国的Nuance Communications（由微软收购，198亿美元）、Varian Medical Systems（西门子，172亿美元）和Proofpoint（Thoma Bravo，124亿美元），其次是捷克共和国的Avast（NortonLifeLock，80亿美元）。



## 初创企业活动

以下章节内容分析了2013年至2021年全球获得超过150万美元投资的人工智能和机器学习公司。

## 全球趋势

2021年，全球人工智能的私人投资总额约为935亿美元，是2020年私人投资总额的两倍多（图4.2.2）。这标志着自2014年以来最大的同比增长（2013年至2014年的投资增加了一倍多）。

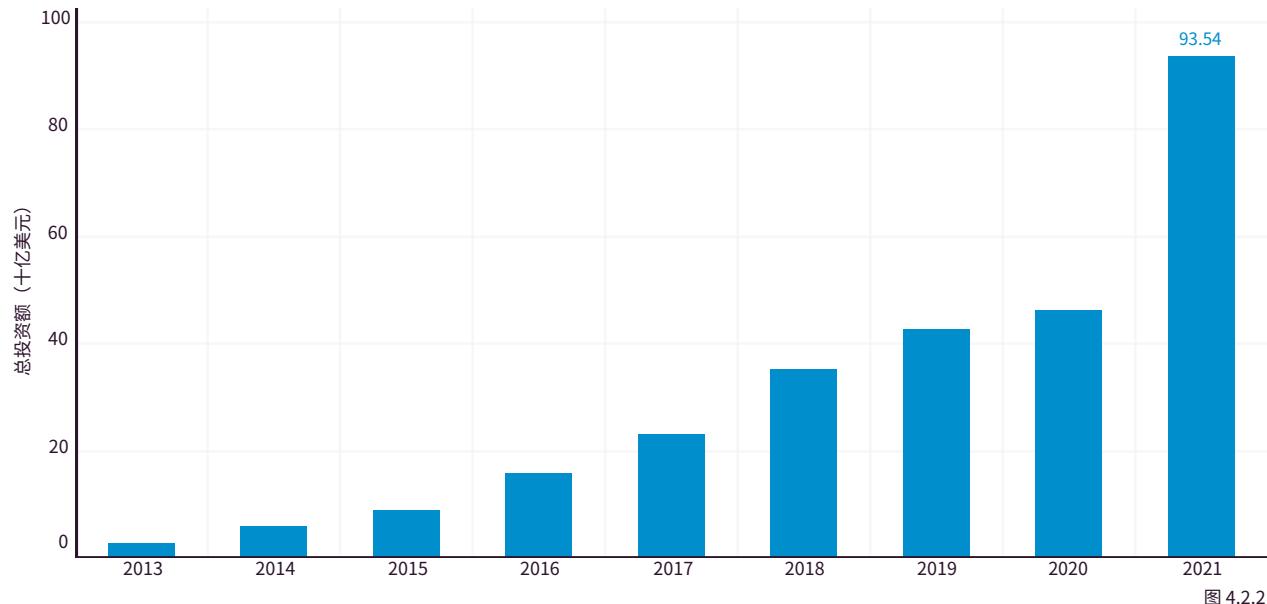
在披露融资金额的公司中，与2020年相比，2021年1亿至5亿美元的人工智能融资轮次数量增加了一倍多，而5000万至1亿美元的融资轮次也增加了一倍多（表4.2.1）。2020年，只有4轮价值5亿美元以上的融资；2021年，这一数字增长到15。2021年，公司吸引的投资明显增加，因为2021年的平均私人投资交易规模比2020年高出81.1%。

然而，图4.2.3显示，新获投资的人工智能公司的数量继续下降，从2020年的762家公司下降到2021年的746家公司----这也是从2018年开始连续第三年下降。2021年规模最大的私人投资由两家数据管理公司和两家机器人/自动驾驶公司主导完成。<sup>4</sup>

**2021年，全球人工智能私人投资总额约为935亿美元，是2020年私人投资总额的两倍以上。**

### 2013-21年人工智能领域的私人投资

来源: NetBase Quid, 2021 | 图: 2022年人工智能指数报告



<sup>4</sup> 最大的私人投资为Databricks（美国）、北京地平线机器人技术（中国）、Oxbotica有限公司（英国）和Celonis（德国）。



### 2013-21年全球新获投资的人工智能公司数量

来源: NetBase Quid, 2021 | 图: 2022年人工智能指数报告

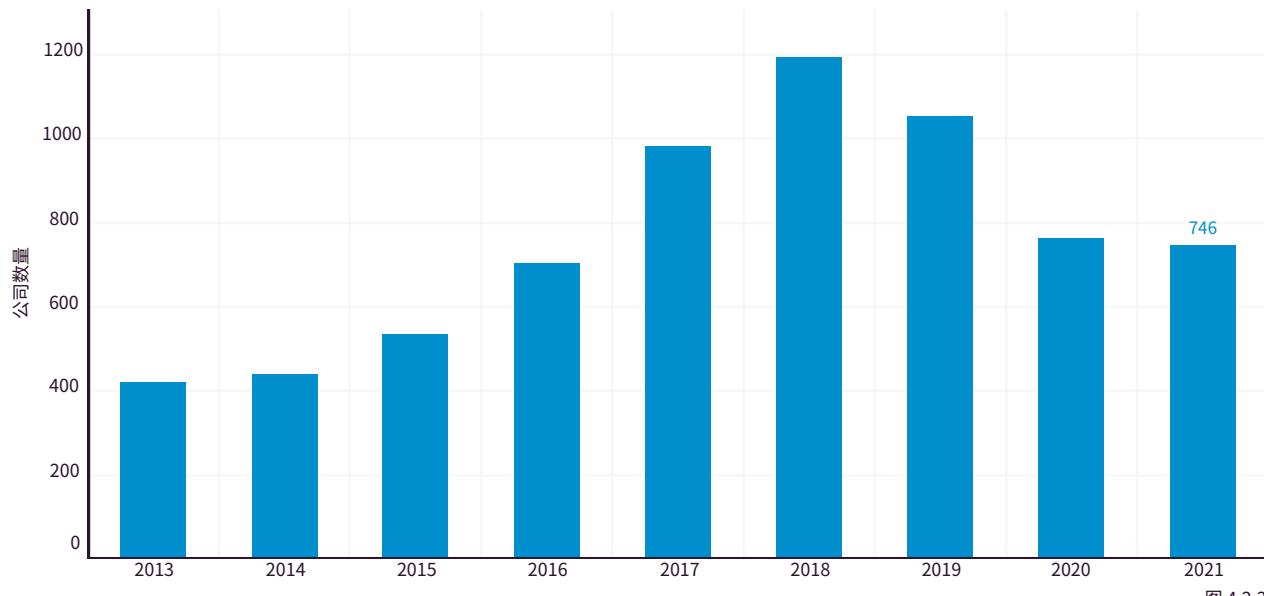


图 4.2.3

资金规模	2020	2021	总数
超10亿美元	3	5	8
5-10亿美元	1	10	11
1-5亿美元	93	235	328
5千万-1亿美元	85	194	279
低于5千万美元	2,102	2,120	4,222
未披露	354	395	749
<b>总数</b>	<b>2,638</b>	<b>2,959</b>	<b>5,597</b>

表 4.2.1



## 按资金数额的区域比较

2021年，如图4.2.4所示，美国在受投资的人工智能公司的总体私人投资方面居于世界领先地位，约为529亿美元，是排名第二的中国的三倍（172亿美元）。排名第三的是英国（46.5亿美元），其次是以色列（24亿美元）和德国（19.8亿美元）。图4.2.5显示，如果将2013年至2021年的私人投资总额加起来，排名情况完全相同。美国投资总额为1490亿美元，中国投资总额为619亿美元，其次是英国（108亿美元）、印度

（107.7亿美元）和以色列（61亿美元）。值得注意的是，2013-2021年美国对人工智能公司的私人投资总额是中国投资总额的两倍多，而中国的投资总额是英国同期投资总额的六倍。按地理区域（Geographical Area）划分，如图4.2.6所示，从2020年到2021年，美国、中国和欧盟的投资额都有所增长，其中美国的投资额分别是中国和欧盟的3.1倍和8.2倍，领先于其他国家。

2021年按地理区域划分的人工智能私人投资情况

来源: NetBase Quid, 2021 | 图: 2022年人工智能指数报告

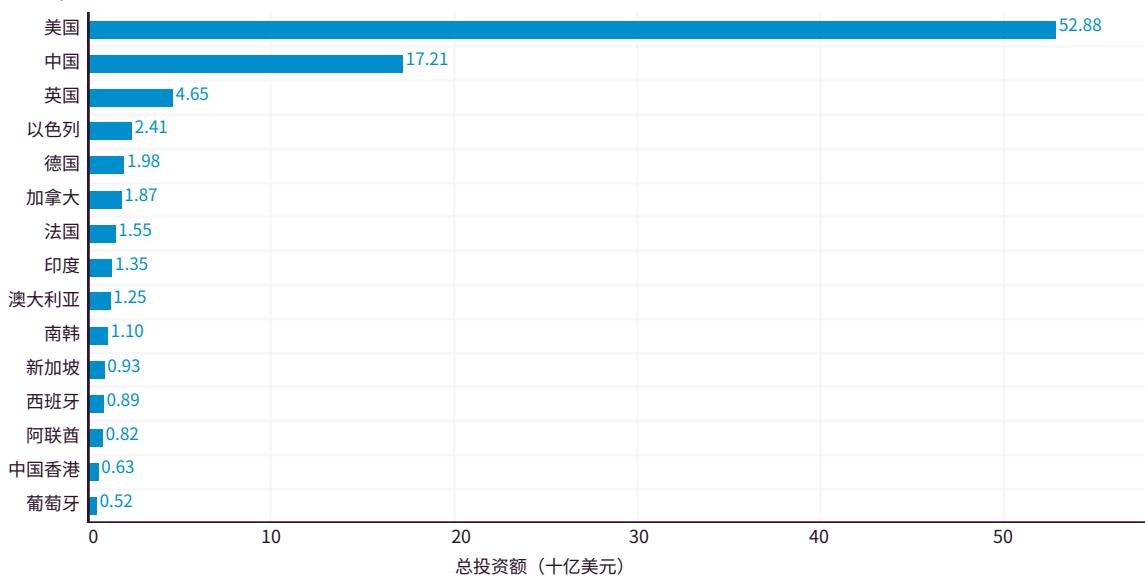


图 4.2.4



### 2013-21年按地理区域划分的人工智能私人投资情况

来源: NetBase Quid, 2021 | 图: 2022年人工智能指数报告

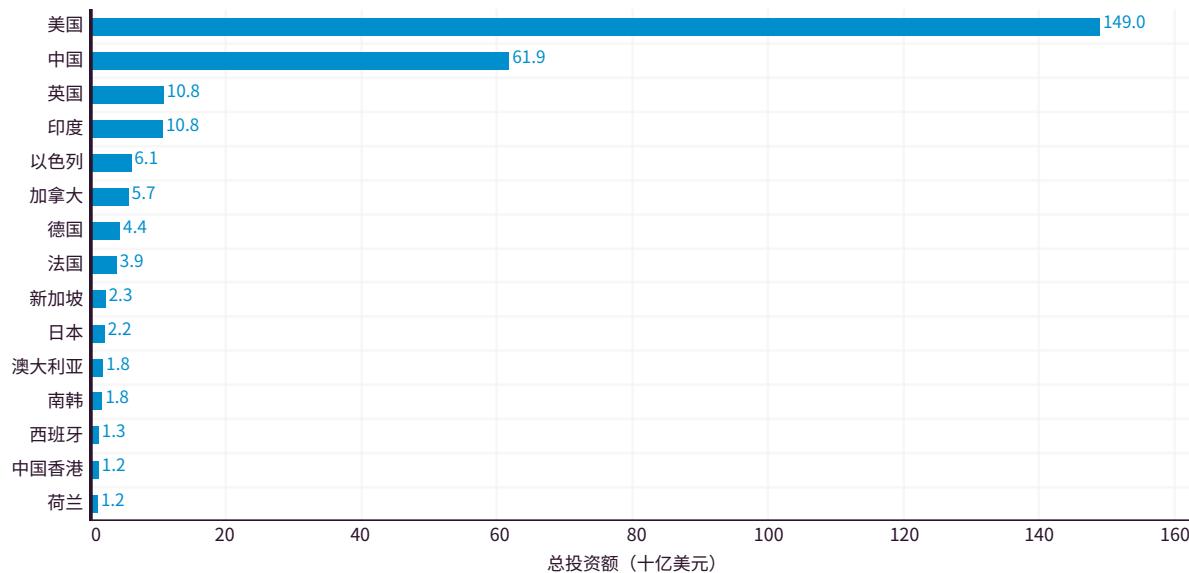


图 4.2.5

### 2013-21年按地理区域划分的人工智能私人投资情况

来源: NetBase Quid, 2021 | 图: 2022年人工智能指数报告

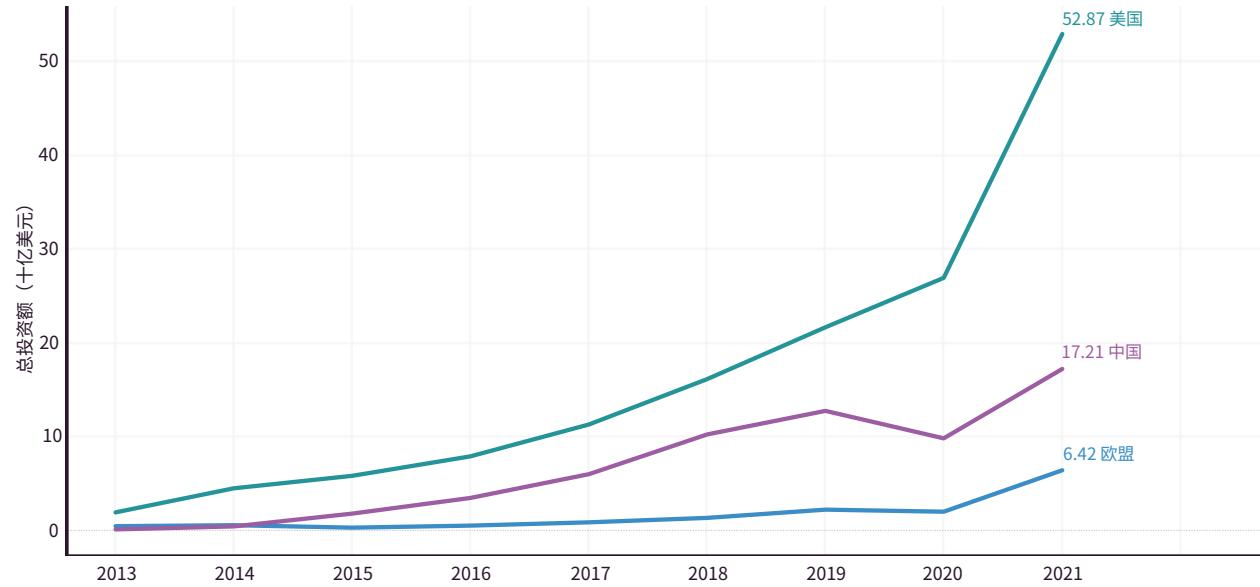


图 4.2.6



## 按新获投资的人工智能公司的区域比较

本节按各地区新获投资的人工智能公司的数量来细分投资数据。图4.2.7显示，2021年美国以299家公司领先，其次是中国的119家，英国的49家，以及以色列的28家。彼此之间的差距比较大。2013年至2021年的汇总数据呈现出类似的趋势（图4.2.8）。

然而，自2018年和2019年以来，美国和中国新获投资的人工智能公司的数量都有所下降（图4.2.9）。尽管有这种下降趋势，美国的新获投资公司数量仍然领先，2021年有299家公司新获投资，其次是中国（119家）和欧盟（96家）。

2021年按地理区域划分的新获投资的人工智能公司数量

来源: NetBase Quid, 2021 | 图: 2022年人工智能指数报告

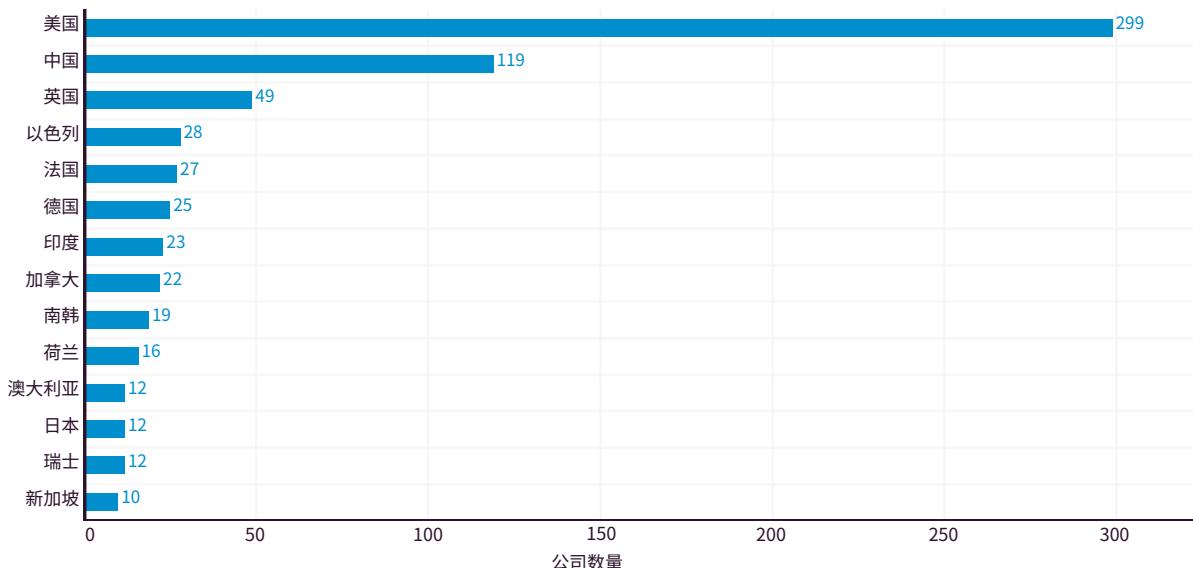


图 4.2.7



### 2013-21年按地理区域划分的新获投资的人工智能公司数量(总数)

来源: NetBase Quid, 2021 | 图: 2022年人工智能指数报告

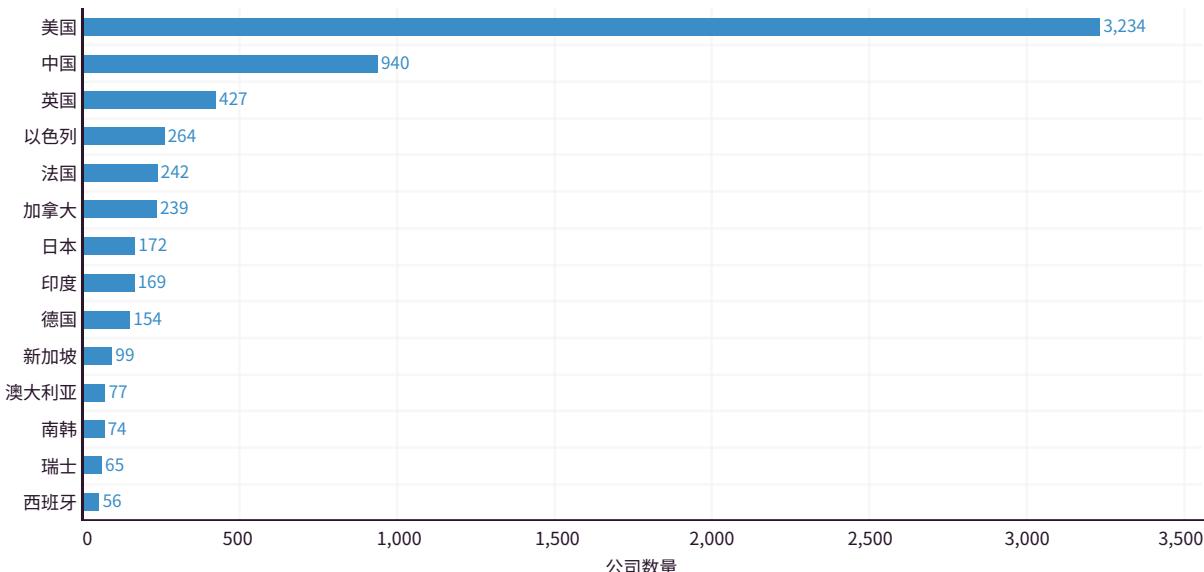


图 4.2.8

### 2013-21年按地理区域划分的新获投资的人工智能公司数量

来源: NetBase Quid, 2021 | 图: 2022年人工智能指数报告

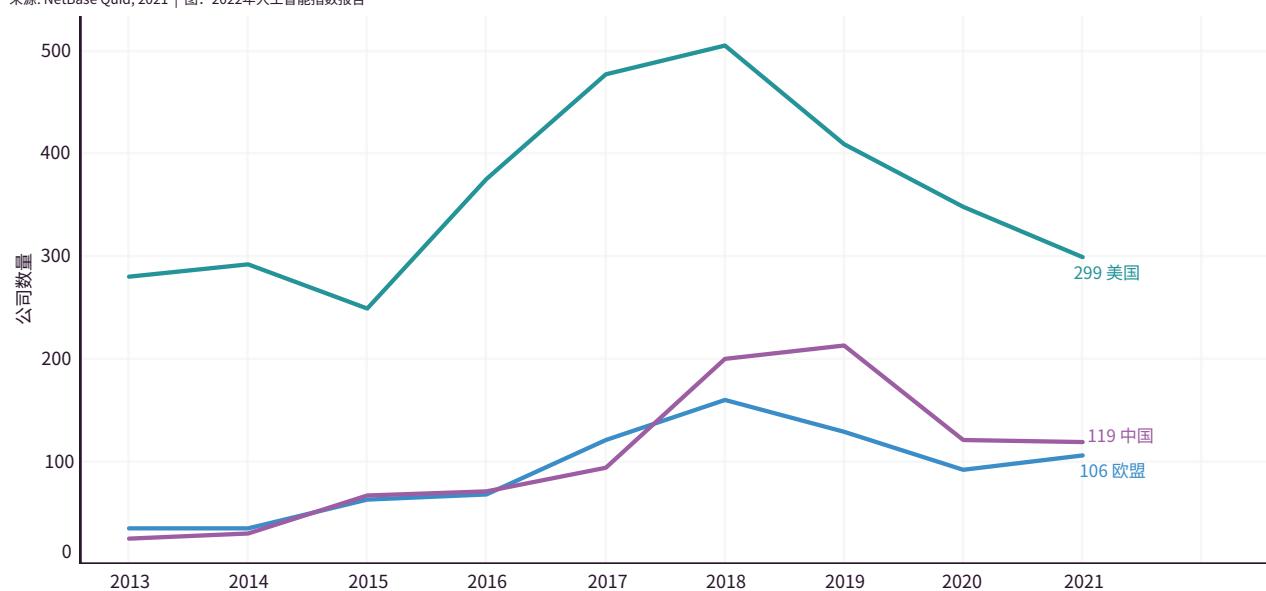


图 4.2.9



## 重点领域分析

私人人工智能投资也因重点领域而异。根据图4.2.10，2021年人工智能的最大私人投资集中在数据管理、处理和云领域（约122亿美元）。值得注意的是，这是2020年投资额（约46.9亿美元）的2.6倍，2021年进行的四项最大的私人投资中有两项是数据管理公司。排在第二位的是医疗和保健方面的私人投资（112.9亿美元），其次是金融技术（102.6亿美元）、影音（80.9亿美元）和半导体（60亿美元）。

按重点领域划分的人工智能私人投资，2020年vs 2021年

来源: NetBase Quid, 2021 | 图: 2022年人工智能指数报告

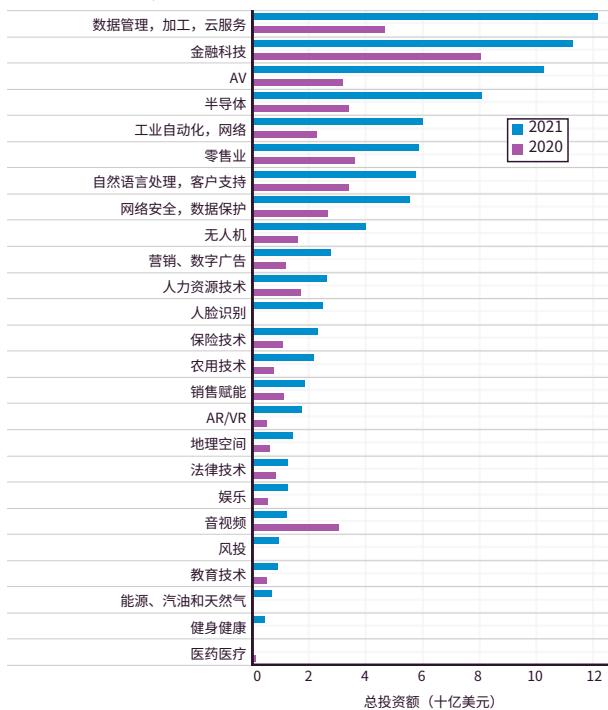


图 4.2.10

图4.2.11中的汇总数据显示，在过去五年中，医疗和保健类公司获得了全球最大的私人投资（289亿美元），其次是数据管理、处理和云（269亿美元），随后是金融技术（249亿美元）和零售（219.5亿美元）。此外，图4.2.12显示了2017-2021年各行业私人投资的总体趋势，其中，AV、网络安全和数据保护、健身和健康、医疗和保健以及半导体行业私人投资稳步增长。

2017-21年按重点领域划分的人工智能私人投资情况

来源: NetBase Quid, 2021 | 图: 2022年人工智能指数报告



图 4.2.11



### 2017-21年按重点领域划分的人工智能私人投资（总额）

来源: NetBase Quid, 2021 | 图: 2022年人工智能指数报告

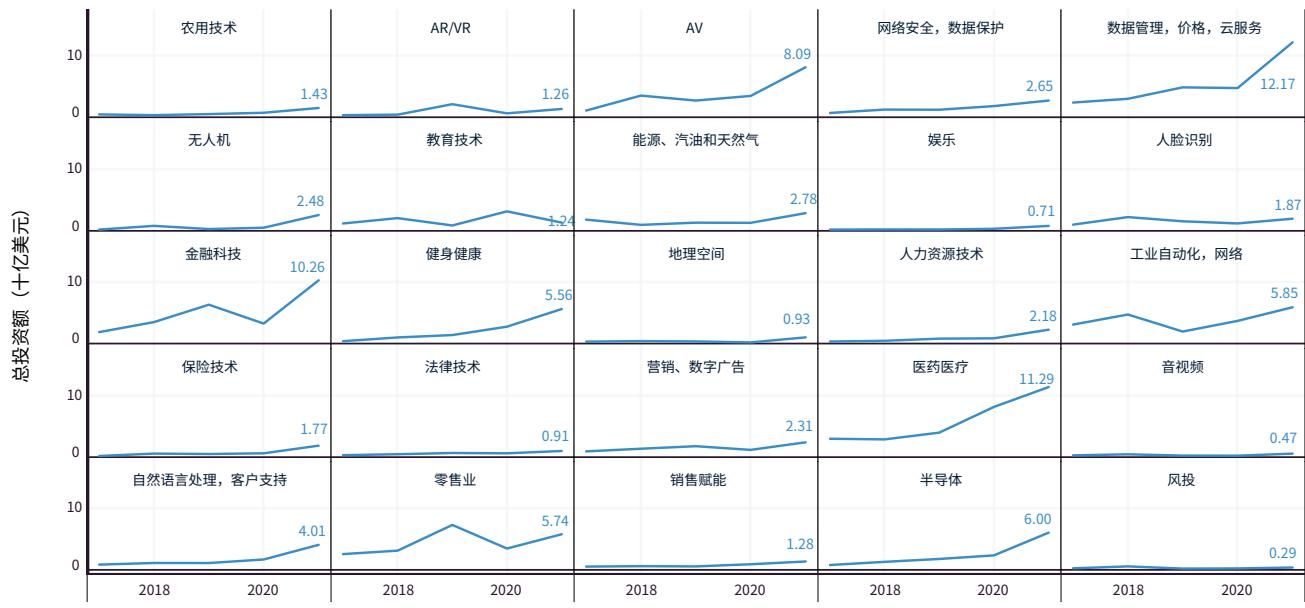


图 4.2.12



## 4.3 企业活动

### 产业界应用情况

本节关于企业人工智能活动的内容借鉴了麦肯锡2021年12月的《2021年人工智能状况》报告。该报告的结论基于2021年早些时候对1,843名参与者进行的全球在线调查。调查对象来自于不同的行业、公司、职能专业、任期和世界各地区，每个人都提供了关于当今人工智能状况的问题的答案。

### 全球应用人工智能的情况

图4.3.1显示了按地理区域（Geographical Area）划分的全球各组织应用人工智能的情况。2021年，印度以65%的应用率领先，其次是“亚太发达国家”（64%）、“发展中市场（包括中国、中东和北非）”（57%）和北美（55%）。所有地区的平均应用率为56%，比2020年增长了6%。值得注意的是，“发展中市场（包括中国、中东和北非）”比2020年增长了21%，而印度比2020年增长了8%。

2020-21年全球各组织应用人工智能的情况

来源: McKinsey & Company, 2021 | 图: 2022年人工智能指数据报告

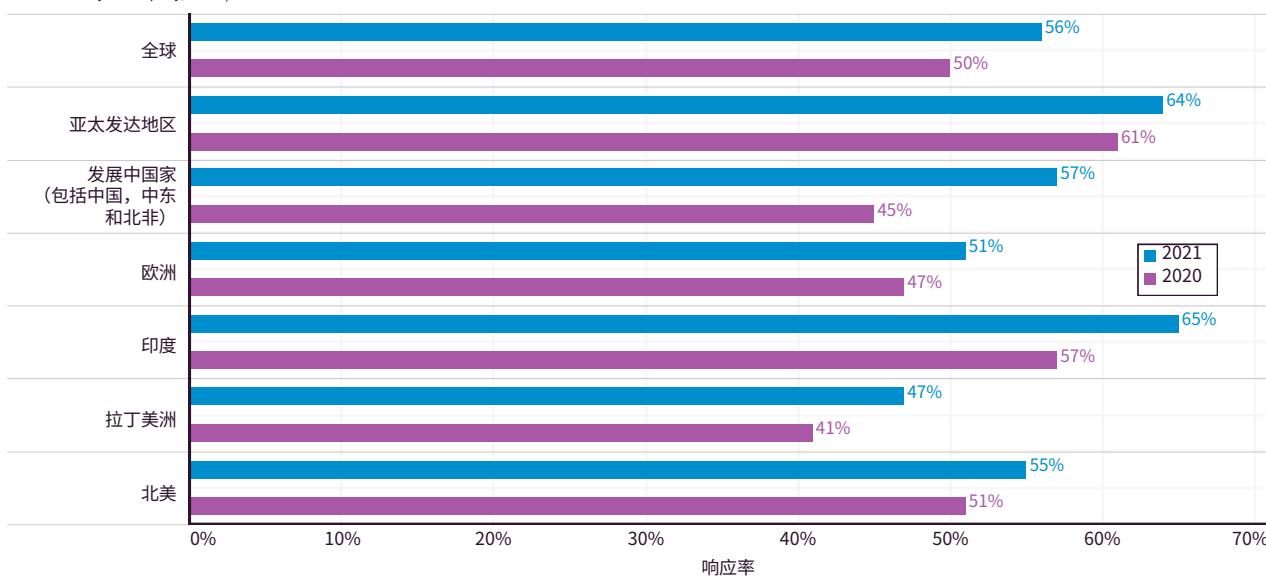


图 4.3.1



## 按行业和功能划分的人工智能应用情况

图4.3.2显示了2021年按行业和职能划分的人工智能应用情况。应用最多的是高科技/电信的产品和/或服务

开发（45%），其次是金融服务的服务运营（40%）、高科技/电信的服务运营（34%），以及金融服务的风险职能（32%）。

2021年按行业和功能划分的人工智能应用情况

来源: McKinsey & Company, 2021 | 图: 2022年人工智能指数报告

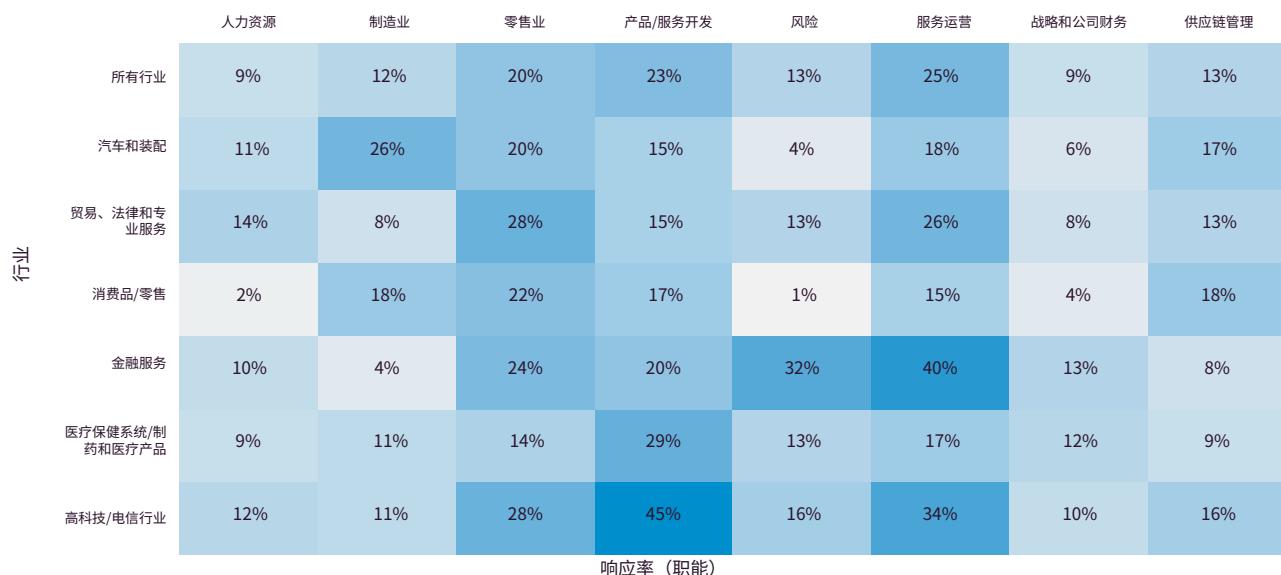


图 4.3.2



## 应用人工智能能力的类型

关于2021年嵌入标准业务流程的人工智能能力类型，如图4.3.3所示，嵌入率最高的是高科技/电信行业的自

然语言文本理解（34%），其次是金融服务和汽车及装配行业的机器人流程自动化（33%），以及金融服务的自然语言文本理解（32%）。

AI2021年标准业务流程中嵌入的人工智能能力

来源: McKinsey & Company, 2021 | 图: 2022年人工智能指数报告

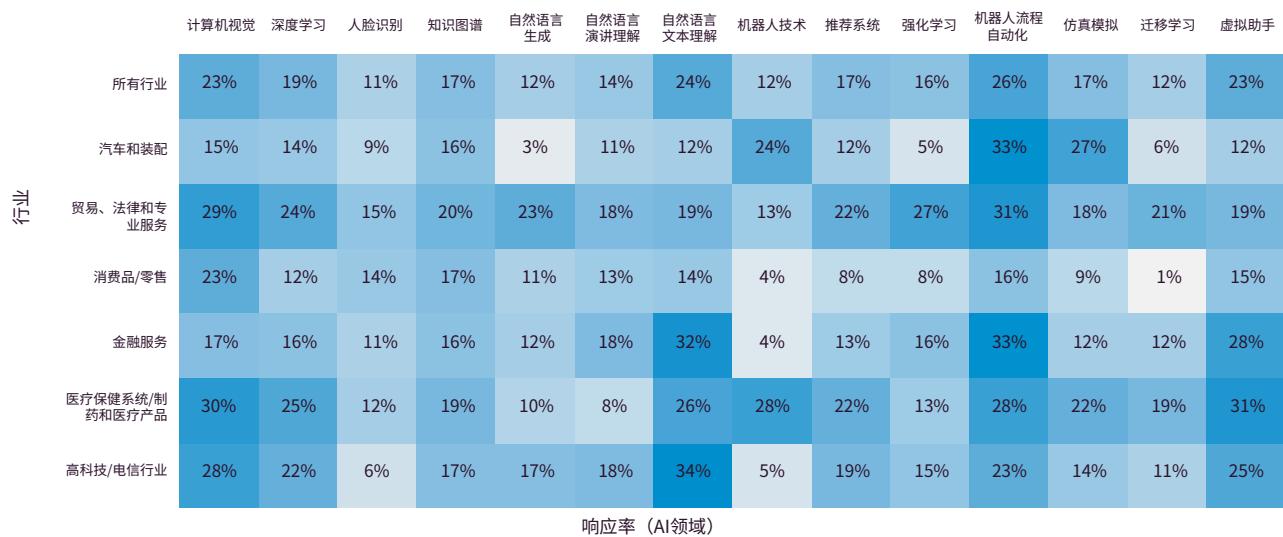


图 4.3.3



## 考虑和减轻应用人工智能的风险

调查对象认为2021年与应用人工智能技术最相关的风险是网络安全（55%的受访者），其次是合规性（48%）、可解释性（41%）和个人/私人隐私（41%）（图4.3.4）。与2020年相比，2021年关注网络安全的人工智能风险的组织减少，从2020年略高于60%的关注度下降到2021年的55%。同时，对人工智能监管合规风险的关注与2020年相比几乎没有变化。

图4.3.5显示了各组织正在采取措施防范的来自人工

智能的风险。调查对象给出最多的回答是网络安全（47%的受访者），其次是监管合规（36%）、个人/私人隐私（28%）和可解释性（27%）。值得注意的是，这些组织认为相关的风险和他们采取措施防范的风险之间的差距为：公平和公正的差距为10个百分点（29%到19%），监管合规的差距为12个百分点（48%到36%），个人/私人隐私的差距为13个百分点（41%到28%），可解释性的差距为14个百分点（41%到27%）。

### 2019-21年各组织认为相关的应用人工智能的风险

来源: McKinsey & Company, 2021 | 图: 2022年人工智能指数报告

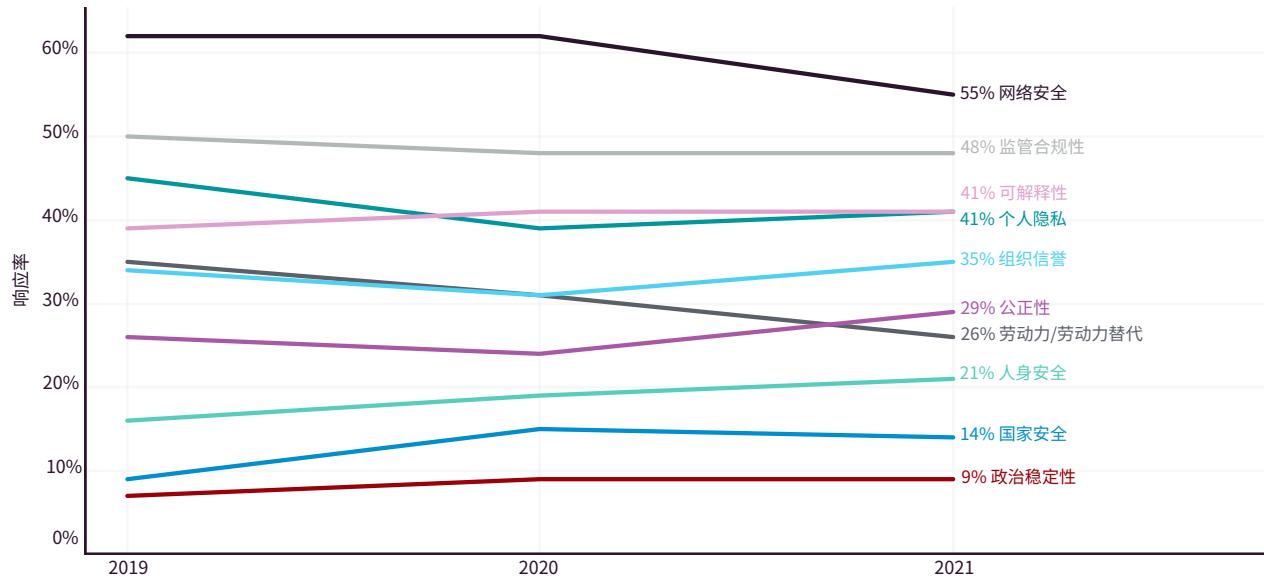


图 4.3.4



### 2019-21年各组织采取措施防范的应用人工智能的风险

来源: McKinsey & Company, 2021 | 图: 2022年人工智能指数报告

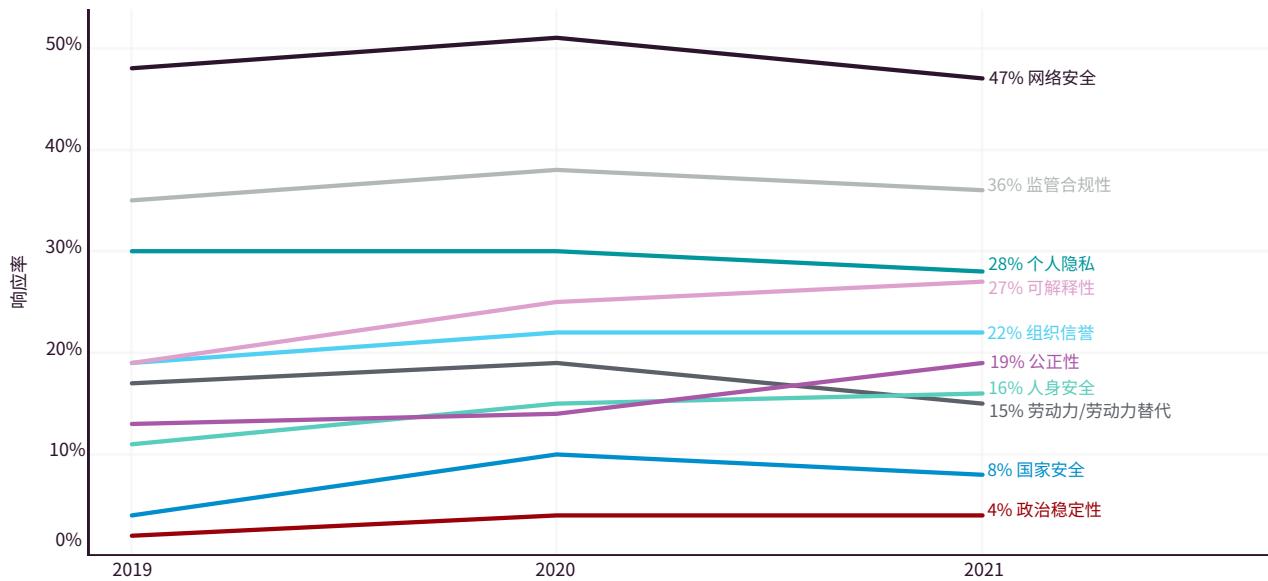


图 4.3.5



以下部分借鉴了计算机研究协会（Computing Research Association, CRA）Taulbee年度调查的数据。对于本节所介绍的最新调查，CRA在2020年秋季通过联系美国和加拿大的200多个具有博士学位授予资质的部门来收集数据。调查结果于2021年5月公布。CRA调查记录了美国和加拿大授予计算机科学(CS)、计算机工程(CE)或信息学(I)博士学位的学术单位在学生注册、学位授予、毕业生就业和教师工资方面的变化趋势。学术单位包括CS和CE专业的系，以及信息学或计算专业的学院或学校。

## 4.4 人工智能教育

### 北美地区的计算机科学本科毕业生

北美地区大多数人工智能相关的课程都是作为本科阶段计算机科学专业课程的一部分。从2010年到2020年，北美地区博士院校应届计算机科学本科生数量增

长了3.5倍（图4.4.1）。2020年有超过31,000名本科生完成了计算机科学专业本科阶段的学习--比2019年的人数增加了11.60%。

2010-20年北美地区博士院校计算机科学专业本科生毕业人数

来源：CRA Taulbee Survey, 2021 | 图：2022年人工智能指数报告

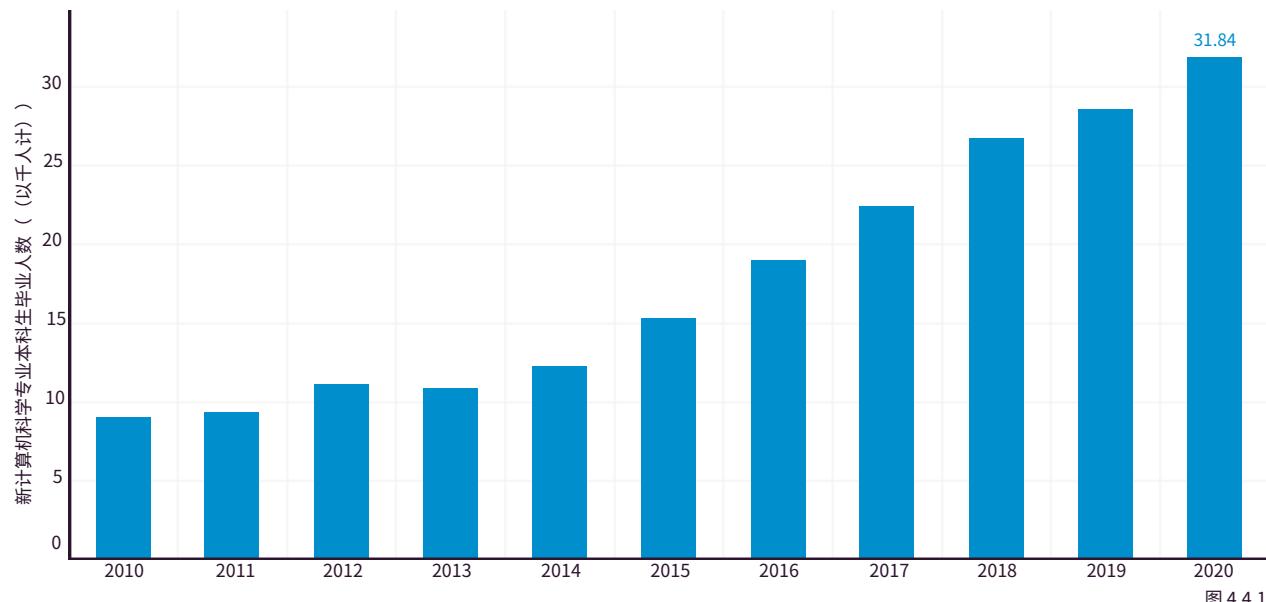


图 4.4.1



## 北美地区应届计算机科学博士的情况

以下小节显示了北美地区计算机科学博士毕业生的趋势，重点是与人工智能相关的专业。<sup>5</sup> CRA的调查共包括20个专业，其中两个专业与人工智能领域直接相关：人工智能/机器学习（AI/ML）和机器人/视觉。

### 按专业划分的应届计算机科学博士

2020年，每5个获得博士学位的应届计算机科学专业学生中就有1人是AI/ML专业的，AI/ML是过去十年中最受欢迎的专业（图4.4.2）。相对于其他18个专业，AI/ML也是2010年到2021年应届博士数量增长最明显的专业（图4.4.3）。机器人/视觉也是2020年博士毕业生中最受欢迎的计算机科学专业之一，在过去11年里，在应届计算机科学专业博士总数中的份额有1.4个百分点的变化。

**2020年，每5个获得博士学位的应届计算机科学专业学生中就有1人是AI/ML专业的，这是过去十年中最受欢迎的专业。**

### 2020年美国各专业应届计算机科学专业博士（占总数的%）情况

来源：CRA Taulbee Survey, 2021 | 图：2022年人工智能指数报告

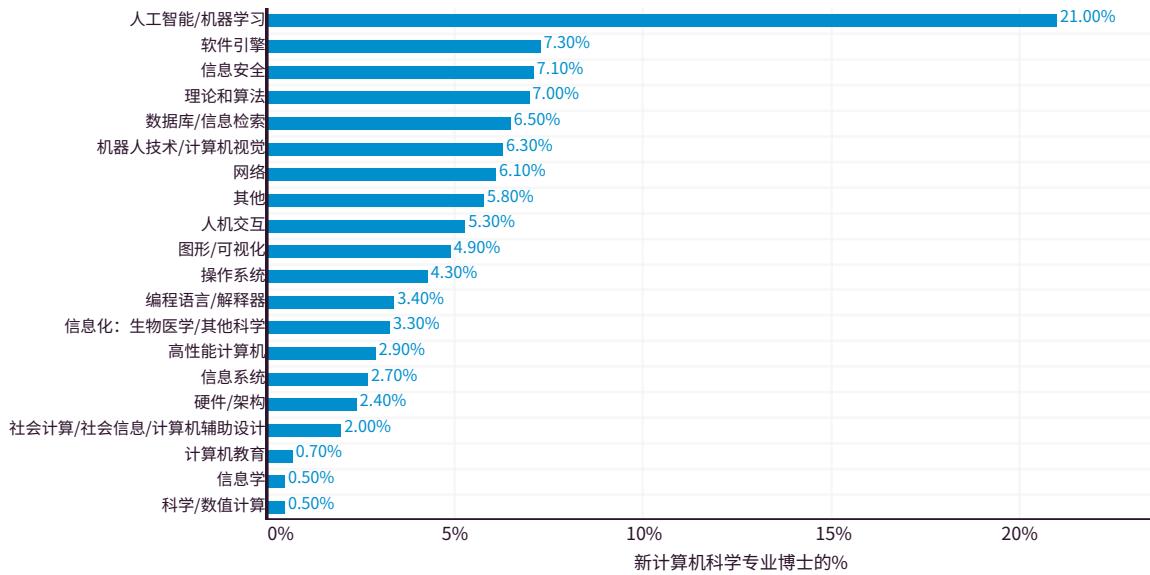


图 4.4.2

<sup>5</sup> 本节的应届计算机科学博士包括来自美国计算机科学学术单位（大学中的系、学院或学校）的博士毕业生。



### 2010-20年按专业划分的美国应届计算机科学专业博士的百分点变化

来源: CRA Taulbee Survey, 2021 | 图: 2022年人工智能指数报告

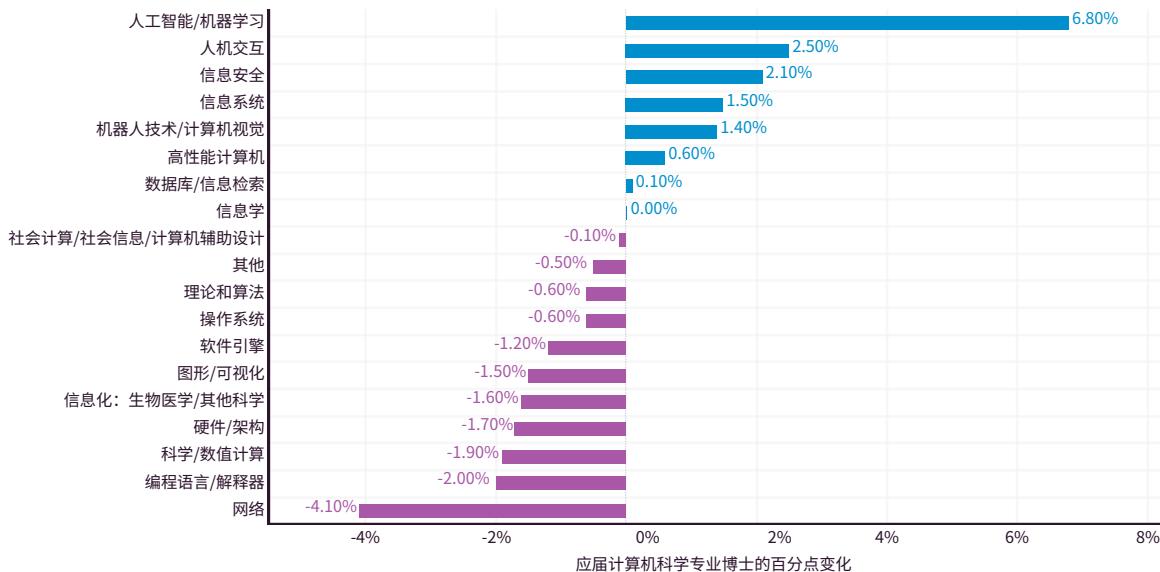


图 4.4.3

### 人工智能/机器学习和机器人/视觉专业的应届计算机科学博士

2010年至2020年期间，人工智能/机器学习和机器人/视觉专业的计算机科学博士毕业生人数分别增长了72.05%和50.91%。2019年至2020年，这两个专业应届博士的总人数略有减少，可能是受到了COVID-19的影响。

#### 2010-20年美国应届人工智能/机器学习和机器人/视觉专业的计算机科学博士

来源: CRA Taulbee Survey, 2021 | 图: 2022年人工智能指数报告



#### 2010-20年美国应届人工智能/机器学习和机器人/视觉专业的计算机科学博士(占总数的%)

来源: CRA Taulbee Survey, 2021 | 图: 2022年人工智能指数报告

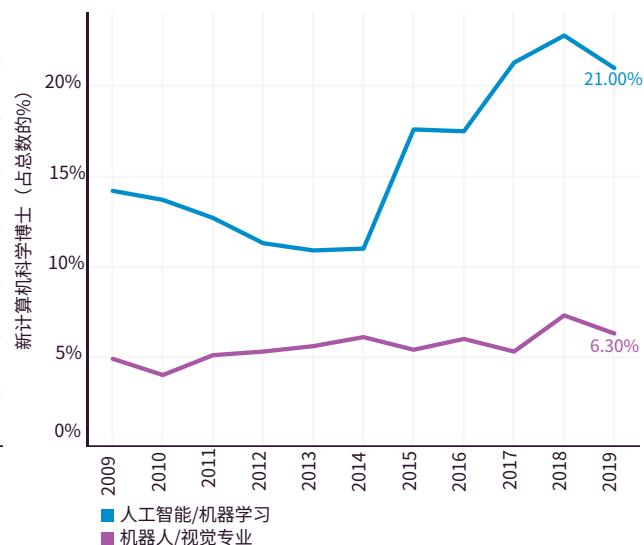


图 4.4.4a

图 4.4.4b



**应届人工智能博士在北美地区的就业情况**  
应届的人工智能博士毕业后选择在哪里工作？本节分析了北美地区的应届人工智能博士在学术界、产业界和政府部门的就业趋势。<sup>6</sup>

### 学术界vs.产业界vs.政府部门

2020年，北美地区应届人工智能博士选择在产业界工

作的比例略有下降，其份额从2019年的65.7%下降到2020年的60.2%，而进入学术界和政府部门的应届人工智能博士的份额变化不大（图4.4.5a和图4.4.5b）。请注意，2020年的数据可能受到越来越多的应届人工智能博士在毕业后即出国的影响，这个数字从2019年的19人增长到2020年的32人。

2010-20年北美地区学术界、政府部门或产业界新聘用人工智能专业博士的就业情况

来源：CRA Taulbee Survey, 2021 | 图：2022年人工智能指数报告

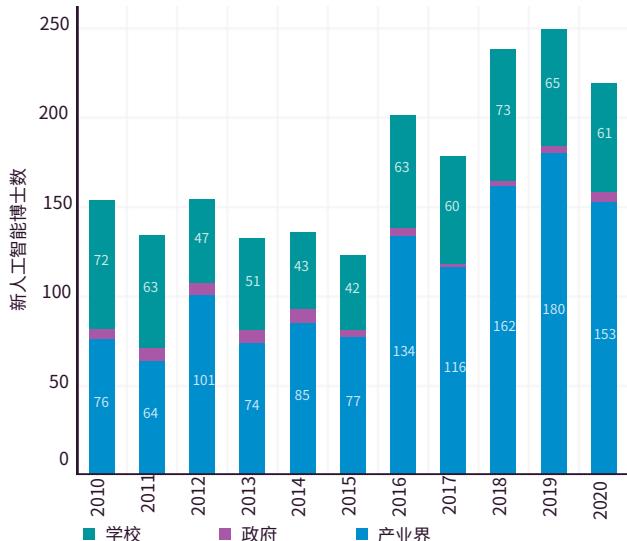


图 4.4.5a

2010-20年北美地区学术界、政府部门或产业界新聘用人工智能专业博士的就业情况(占总数的%)

来源：CRA Taulbee Survey, 2021 | 图：2022年人工智能指数报告

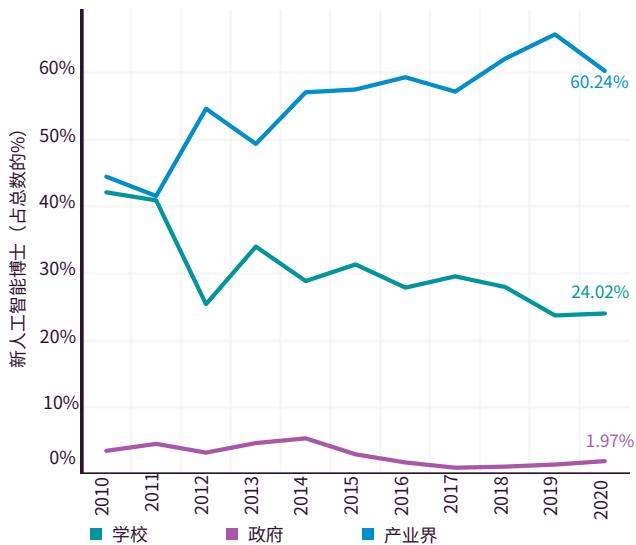


图 4.4.5b

<sup>6</sup> 本节中提及的应届人工智能博士包括美国和加拿大的计算机科学、计算机工程和信息学的学术单位（大学内的系、学院或学校）中专门从事人工智能专业研究的博士毕业生。



## 北美地区应届人工智能博士的多样性

### 按性别划分

图4.4.6显示，北美地区应届女性人工智能和计算机科学专业博士的比例仍然很低，从2010年到2020年变化不大。

2010-20年北美地区应届女性人工智能和计算机科学专业博士（占应届人工智能和计算机科学博士总数的%）

来源：CRA Taulbee Survey, 2021 | 图：2022年人工智能指数报告



图 4.4.6



## 按种族/族裔划分

根据图4.4.7, 从2010年到2020年, 在美国本土居民的应届人工智能博士中, 非西班牙裔白人和亚裔所占比例最大, 平均为65.2%和18.8%。相比之下, 在过去11年中, 平均约有1.5%的黑人或非裔美国人(非西班牙裔)和2.9%的西班牙裔毕业。图4.4.8显示了2010年至2020年期间,

美国计算机科学、计算机工程和信息学等专业授予美国本土居民的所有博士学位情况。在过去的11年中, 新毕业的白人(非西班牙裔)博士的比例变化不大, 而新毕业的黑人或非裔美国人(非西班牙裔)和西班牙裔计算机博士的比例明显较低。

2010-20年按种族/族裔划分的美国本土居民应届人工智能专业博士 (占总数的%)

来源: CRA Taulbee Survey, 2021 | 图: 2022年人工智能指数报告

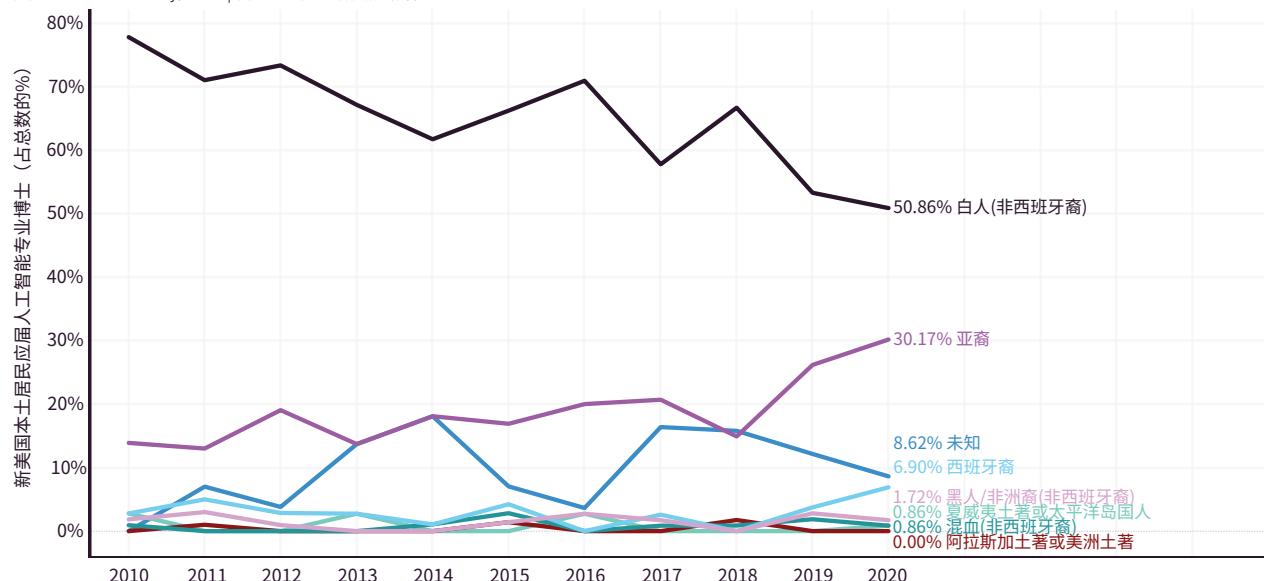


图 4.4.7

2010-20年按种族/族裔划分的美国本土居民应届计算机专业博士 (占总数的%)

来源: CRA Taulbee Survey, 2021 | 图: 2022年人工智能指数报告

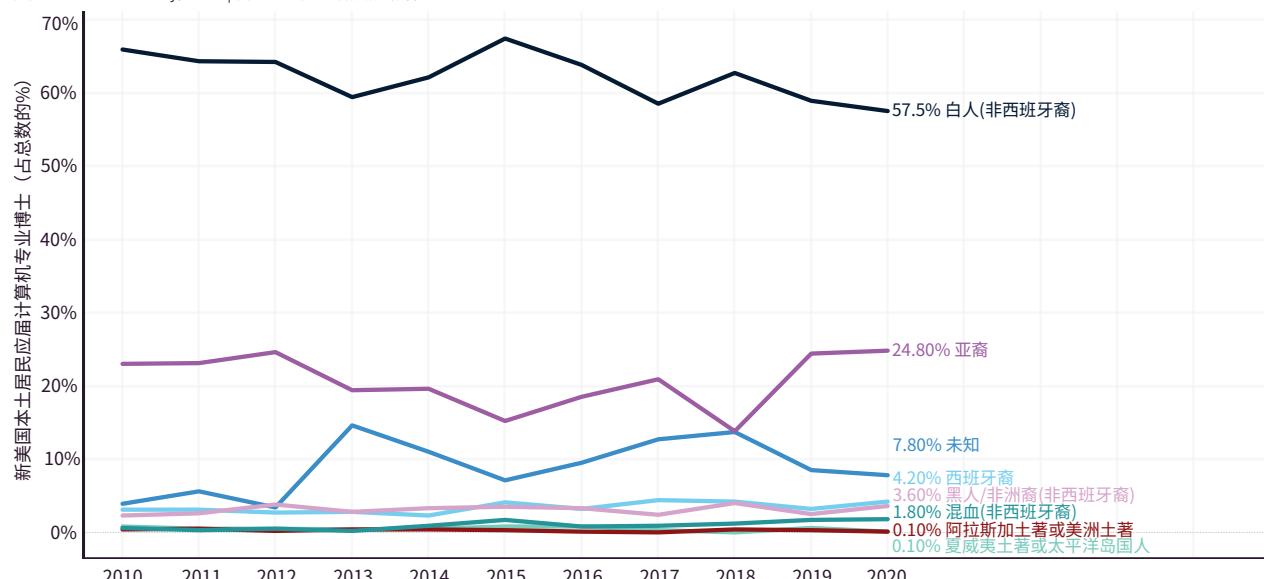


图 4.4.8



## 北美地区的应届国际人工智能博士

2020年北美地区应届人工智能博士中，国际学生的比例从2019年的64.3%略微下降到2020年的60.5%（图4.4.9）。作为比较，2022年毕业的所有计算机博士

中，有65.1%是国际学生。此外，2020年，更多的国际学生--占所有应届人工智能博士的14.0%--在美国以外的地方工作，而2019年这一数字为8.6%（图4.4.10）。

2010-20年北美地区应届国际人工智能博士（占应届人工智能博士总数的%）

来源：CRA Taulbee Survey, 2021 | 图：2022年人工智能指数报告

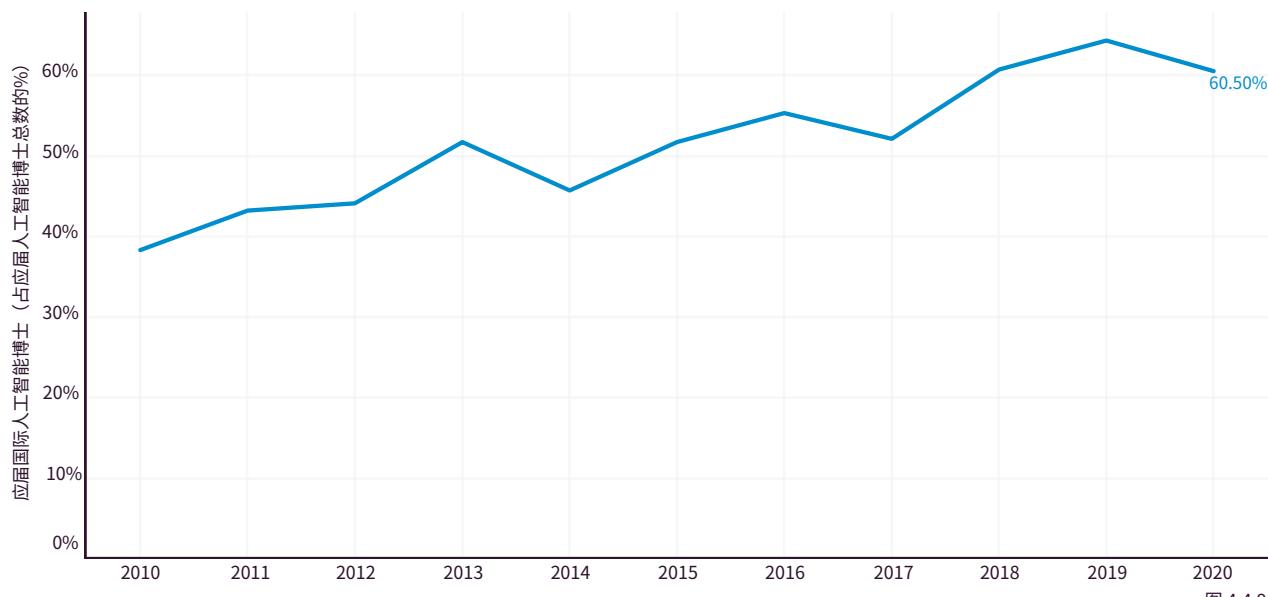


图 4.4.9

2020年按就业地点划分的美国应届国际人工智能博士（占总数的%）

来源：CRA Taulbee Survey, 2021 | 图：2022年人工智能指数报告

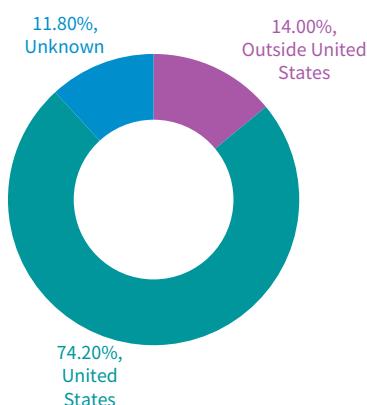


图 4.4.10



2022年  
人工智能指数报告

## 章节五 人工智能政策和治理



## 章节五 章节预览

概述	174
章节要点	175
<b>5.1 人工智能和政策制定</b>	<b>176</b>
关于人工智能的全球立法记录	176
按地理区域划分	177
美国联邦人工智能立法	178
对立法的进一步了解	179
美国州级人工智能立法	180
按州划分	181
政党赞助	182
立法记录中的人工智能提及情况	183
美国国会记录中的人工智能提及情况	183
全球立法程序中的人工智能提及情况	184
按地理区域划分	185
美国人工智能政策文件	186
按主题划分	187
<b>5.2 美国在人工智能方面的公共投资</b>	<b>188</b>
非国防性人工智能研发的联邦预算	188
美国国防部预算需求	189
国防部资金最多的五个项目	190
按部门划分的国防部人工智能研发支出	191
美国政府与人工智能有关的合同支出	192
合同支出总额	192
按部门和机构划分的合同支出情况	193
2021年五个开支最大的部门的最大合同	195

访问公开数据



# 概述

人工智能在过去十年中成为一个越来越普遍的话题，政府间、国家和区域组织一直致力于制定围绕人工智能治理的政策和战略。这些举措的动力来自于这样一种认识：必须找到方法来解决围绕人工智能的道德和社会问题，同时最大限度地发挥其效益。对人工智能技术进行积极和知情治理已成为世界上许多政府的优先事项。

本章探讨了人工智能技术和治理的交叉点，研究了不同国家、地区和美国各州政府如何努力管理人工智能技术。本章首先考察了全球和美国的人工智能政策制定情况，探讨了哪些国家和政治人物最热衷于推动人工智能立法，以及从隐私到伦理的哪类人工智能子主题是立法关注的重点。然后，这一章深入研究了全球人工智能领域最大的公共部门投资者之一--美国，并分析美国各个政府部门在过去五年中在人工智能方面的支出情况。



## 章节要点

- 人工智能指数对25个国家的人工智能立法记录的分析显示，包含 "人工智能 "的法案被通过成为法律的数量**从2016年的1个增长到2021年的18个**。西班牙、英国和美国在2021年通过的人工智能相关法案数量最多，各通过了三项。
- 美国的联邦立法记录显示，从2015年到2021年，与人工智能有关提案总数急剧增加，**而通过的法案数量仍然很少，只有2%最终确定立法**。
- 2021年，美国各州立法者**通过了每50个包含人工智能条款的提案中的一个**，而此类提案数量**从2012年的2个增长到2021年的131个**。
- 在美国，本届国会会议（第117届）有望创下自2001年以来与人工智能有关的最多提及次数，到2021年底，**即本届会议过半时，将有295次提及，而上届（第116届）则有506次**。



在过去的十年中，围绕人工智能治理规制的讨论已经加速，各种立法机构都提出了大量政策建议。本节首先研究了不同国家和地区已经提出或已经通过成为法律的人工智能相关立法，重点分析了美国的州级立法。然后，仔细研究了世界各地关于人工智能的国会和议会记录，最后是关于美国发表的政策文件数量的数据。

## 5.1 人工智能和政策制定

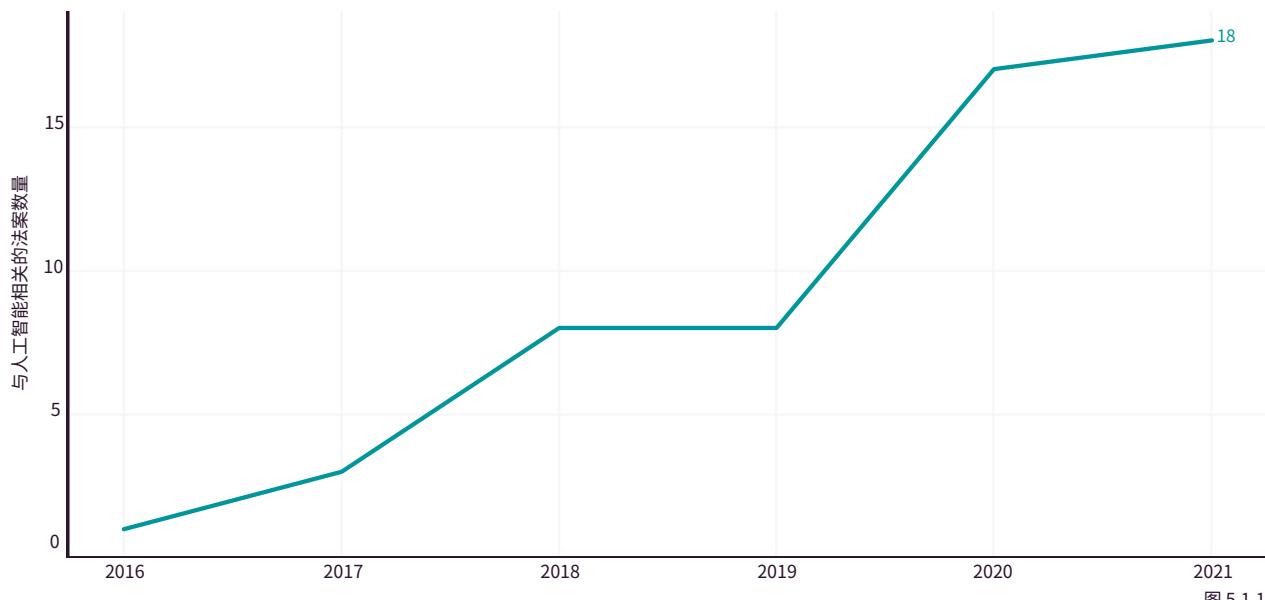
### 关于人工智能的全球立法记录

世界各地的政府和立法机构正越来越多地寻求通过法律为人工智能的发展和创新提供资金，同时促进以人为本的价值观的融合。人工智能指数对25个国家的立法机构在2016年至2021年期间通过的含有“人工智能”字样的法律进行了分析。

总的来说，在所分析的25个国家中，共有55项与人工智能相关的法案最终获得了通过。图5.2.1显示，在过去六年中，与人工智能相关并最终通过成为法律的法案数量急剧增加。<sup>1</sup>

#### 2016-21年，25个国家通过的与人工智能相关的法案数量

来源：人工智能指数，2021 | 图：2022年人工智能指数报告



<sup>1</sup>请注意，本分析仅包括国家立法机构（如国会、议会）通过的、在法案文本的标题或正文中包含各种语言的关键词“人工智能”的法律。方法见附录。包括的国家有：澳大利亚、比利时、巴西、加拿大、中国、丹麦、芬兰、法国、德国、印度、爱尔兰、意大利、日本、荷兰、新西兰、挪威、俄罗斯、新加坡、南非、韩国、西班牙、瑞典、瑞士、英国和美国。



## 按地理区域划分

图5.1.2a显示了2021年颁布的包含人工智能关键词字样的法律数量。西班牙、英国和美国领先，各通过了三项。图5.1.2b显示了过去六年中的立法总数。美国以13项法案占据榜首，从2017年开始，之后的每年都有3项新法律通过。其次是俄罗斯、比利时、西班牙和英国。

**美国以13项法案占据榜首，从2017年开始，之后的每年都有3项新法律通过，其次是俄罗斯、比利时、西班牙和英国。**

2021年部分国家通过成为法律的与人工智能相关的法案数量

来源：人工智能指数，2021 | 图：2022年人工智能指数报告

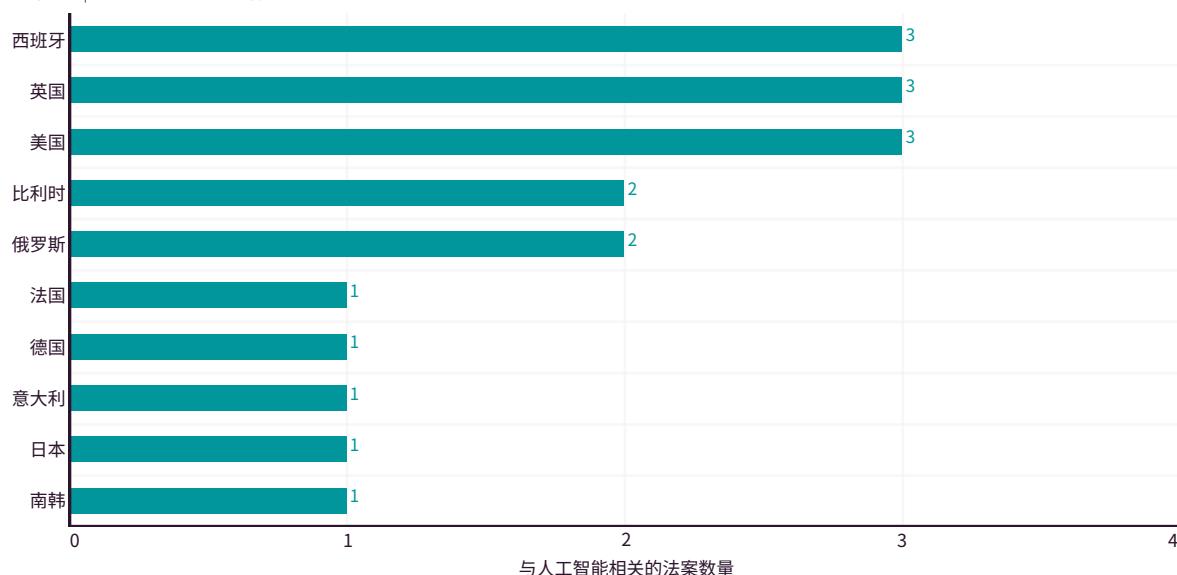


图 5.1.2a



### 2016-21年部分国家通过成为法律的与人工智能有关的法案数量（总数）

来源：人工智能指数，2021 | 图：2022年人工智能指数报告

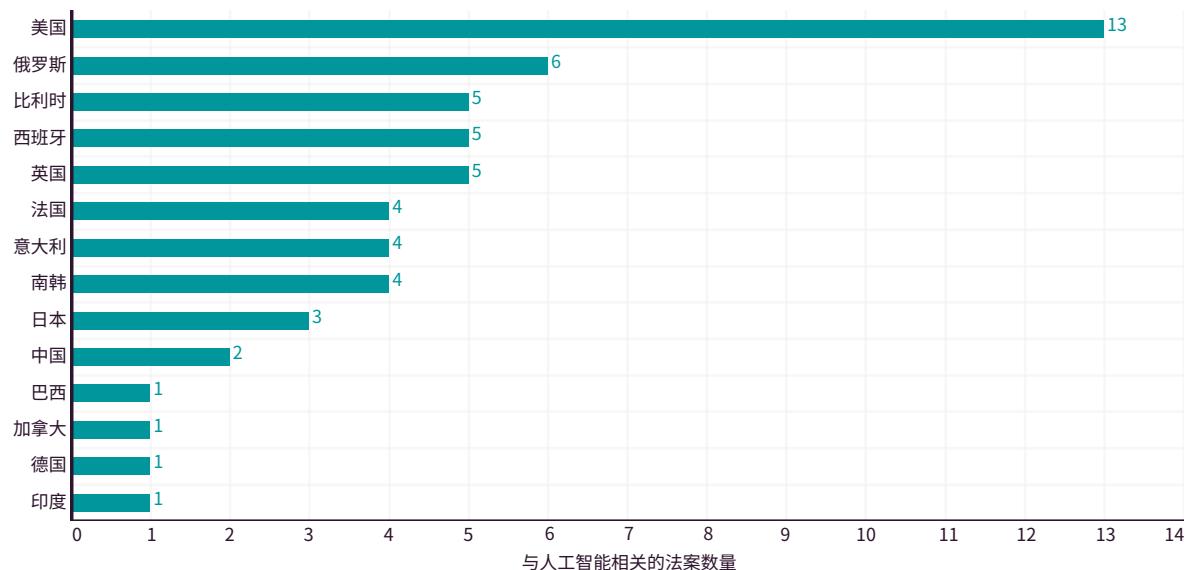


图 5.1.2b

### 美国联邦人工智能立法

仔细观察美国的联邦立法记录，我们会发现与人工智能有关的提案总数急剧增加（图5.1.3）。在2015年，只提出了一项联邦法案，而在2021年，提出了130项。虽然这一激增趋势很明显，但与人工智能有关的法案

最终通过的数量并没有跟上人工智能相关法案提出的增长速度。这种差距在2021年最为明显，所提出的联邦层面的人工智能相关法案中只有2%最终通过成为法律。

### 2015-21年美国与人工智能相关的法案数量（提出vs通过）

来源：人工智能指数，2021 | 图：2022年人工智能指数报告

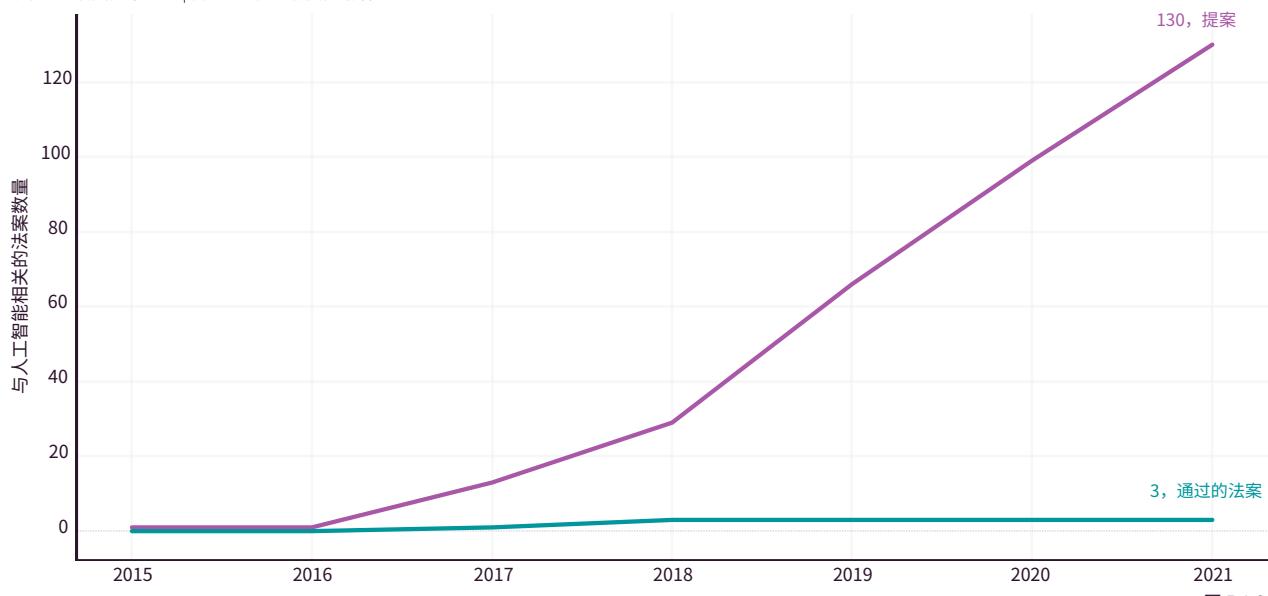


图 5.1.3



## 对立法的进一步了解

下面的小节深入探讨了自2016年以来最终通过成为法律的一些与人工智能相关的立法情况。表5.1.1给出了已经引起政策制定者关注的广泛的人工智能相关问题。

国家	通过年份	法案名称	描述
加拿大	2017	2017年预算执行法，第1号	该法案的一项规定授权加拿大政府向加拿大高级研究所支付1.25亿美元，以支持泛加拿大人工智能战略的发展。
中国	2019	中华人民共和国基本医疗卫生与健康促进法	该法案的一项规定旨在促进大数据和人工智能在健康和医疗领域的应用和发展，同时加快医疗和保健信息基础设施建设，制定关于医疗和健康数据的收集、存储、分析和应用的技术标准。
俄罗斯	2020	2020年4月24日第123-FZ号联邦法，关于为在俄罗斯联邦地区--联邦莫斯科市发展和实施人工智能技术创造必要条件而制定特别法规的试验，以及对《联邦个人数据法》第6条和第10条的修订。	这项法律为人工智能的发展和实施建立了一个实验框架，作为一项为期五年的实验，于2020年7月1日在莫斯科启动，包括允许人工智能系统为政府和某些商业活动处理匿名的个人数据的内容。
英国	2020	供应和拨款（主要估算）法案2020, c.13	该法案的一项规定授权资格和考试监管办公室（the Office of Qualifications and Examination Regulation）探索引入人工智能技术，以改善高风险资质的评分和管理。
美国	2020	IOGAN ACT: 识别生成式对抗网络输出的法案	该法案指示美国国家科学基金会支持致力于生成式对抗网络（deepfakes）和其他类似技术产出的研究。
比利时	2021	关于对求职者的指导和以解决方案为导向的支持的法令，N.327	该法案的一项规定指示政府建立一个名为道德委员会的咨询小组，如果要将人工智能工具用于数字化活动，则由该小组负责提交建议。
法国	2021	Law N:2021-1485 of November 15, 2021: 旨在减少法国数字技术对环境的影响	该法案建立了一个监测系统，以评估新出现的数字技术，特别是人工智能技术对环境的影响。

表 5.1.1



## 美国州级人工智能立法

根据Bloomberg Government自2012年以来提供的数据，美国各州近期提出了大量与人工智能有关的法案，这也看出人们对人工智能政策越来越关注。Bloomberg Government将包含人工智能、机器学习或算法偏见等人工智能相关关键词的法案归类为与人工智能有关。

与联邦层面的情况一样，在过去十年中，在州级提出的人工智能法案的数量有了大幅增加（图5.1.4）。2012年，在新泽西州议会议员安妮特·基哈诺（Annette Quijano）指示新泽西州机动车委员会为自动驾驶汽车建立驾驶执照背书时，提出了头两项与人工智能有关的立法。在过去的10年里，相关法案数量的增长幅度很大，从2012年的2项法案到2021年的131项。

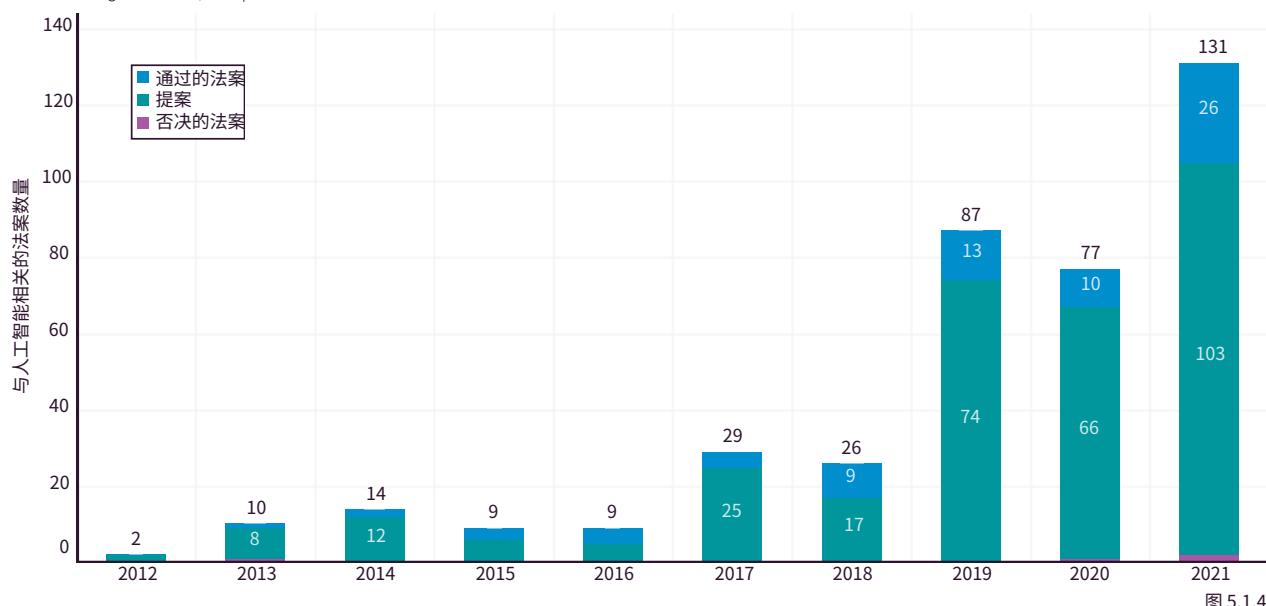
美国联邦与州级的人工智能相关立法之间的一个明显区别是，州级的人工智能法案提案中，有较大比例的

法案能够最终通过。2021年，在131项提出的州级法案中，有26项最终通过成为法律（20%），即每5项提出的法案中就有1项通过成为法律。与联邦层面相比，这一比例明显较高，2021年，联邦层面每50个提案中只有1个通过成为法律。

**美国联邦与州级的人工智能相关立法之间的一个明显区别是，州级的人工智能法案提案中，有较大比例的法案能够最终通过。**

2012-21年美国州级人工智能相关法案的数量

来源: Bloomberg Government, 2021 | 图: 2022年人工智能指数报告





## 按州划分

在美国，人工智能法律的制定在所有州都比较普遍。截至2021年，50个州中有41个州至少提出了一项与人工智能有关的法案，其中某些州在制定人工智能立法方面特别活跃。图5.1.5显示，马萨诸塞州提出的人工智能法案最多，自2012年以来有40项，其次是夏威夷州（35项）和新泽西州（32项）。在图5.1.6中我们重点关注2021年，马萨诸塞州是提出人工智能相关法案最多的州，有20项，其次是伊利诺伊州（15）和阿巴拉马州（12）。

2012-21年美国各州提出的与人工智能有关的法案数量（总数）

来源: Bloomberg Government, 2021 | 图: 2022年人工智能指数报告



图 5.1.5

2021年按州划分的美国各州提出的与人工智能相关的法案数量

来源: Bloomberg Government, 2021 | 图: 2022年人工智能指数报告



图 5.1.6



## 政党赞助

州级的人工智能立法数据显示，人工智能立法呈现了党派动态趋势。图5.1.7显示了民主党和共和党立法者在州一级提出的人工智能相关法案的数量。尽管自2012年以来，两党成员提出的人工智能法案数量都

在增加，但在过去的四年里，数据显示民主党人更有可能提出与人工智能相关的立法。在2018年，民主党人提出的人工智能法案只比共和党人多了两项，而在2021年，他们提出了39项。

2012-21年按政党划分，美国各州提出的与人工智能有关的法案数量

来源: Bloomberg Government, 2021 | 图: 2022年人工智能指数报告

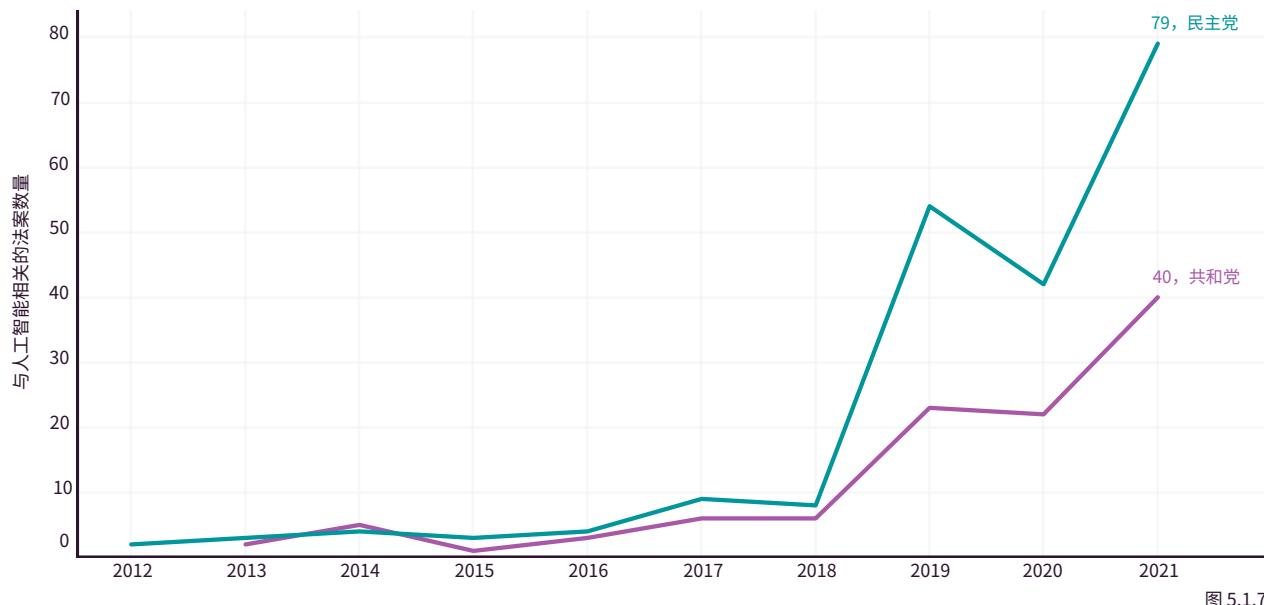


图 5.1.7



## 立法记录中对人工智能的提及情况

立法机构对人工智能兴趣的另一个晴雨表是政府和议会程序中提到 "人工智能" 的次数。本小节根据人工智能指数和Bloomberg Government数据，分析了美国国会记录和其他国家的议会程序中提及人工智能的数据。

### 美国国会记录中的人工智能提及情况

在过去的五年里，特别是在2021年，美国国会议程投入了越来越多的时间讨论人工智能问题。本节介绍了 Bloomberg Government 提供的数据，涉及国会议

中提到的人工智能相关关键词，按立法、国会委员会报告和国会研究服务报告分类展示。

根据图5.1.8，本届国会（第117届）有望（截至2021年底）创下2001年以来最大的与人工智能相关的提及次数。最近结束的国会议会，即第116届会议（2019-2020年），提到了506次人工智能，是第115届会议（2017-2018年）期间提到次数的近3.4倍，是第114届会议（2015-2016年）的30倍。

#### 2001-21年按立法会届次划分的美国国会记录中对人工智能的提及情况

来源: Bloomberg Government, 2021 | 图: 2022年人工智能指数报告

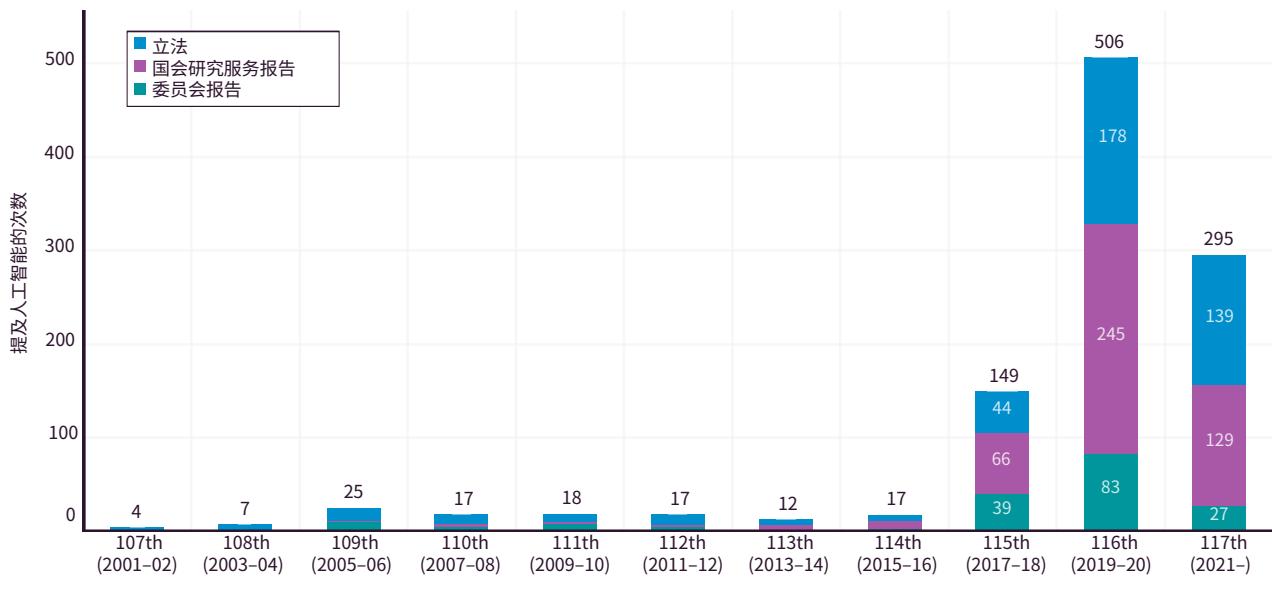


图 5.1.8



## 全球立法程序中的人工智能提及情况

AI mentions in governmental proceedings are on the rise. Not only in the United States, but also in many other countries around the world, the number of mentions of AI in government proceedings has increased. The AI Index found that from 2016 to 2021,

2021年25个国家包含 "人工智能"关键词的立法会议记录或议事录进行了分析。图5.1.9显示，在过去六年中，25个选定国家的立法会议记录中提及人工智能的次数增长了7.7倍。<sup>2</sup>

2016-21年25个选定国家的立法程序中提及人工智能的次数

来源: 人工智能指数, 2021 | 图: 2022年人工智能指数报告

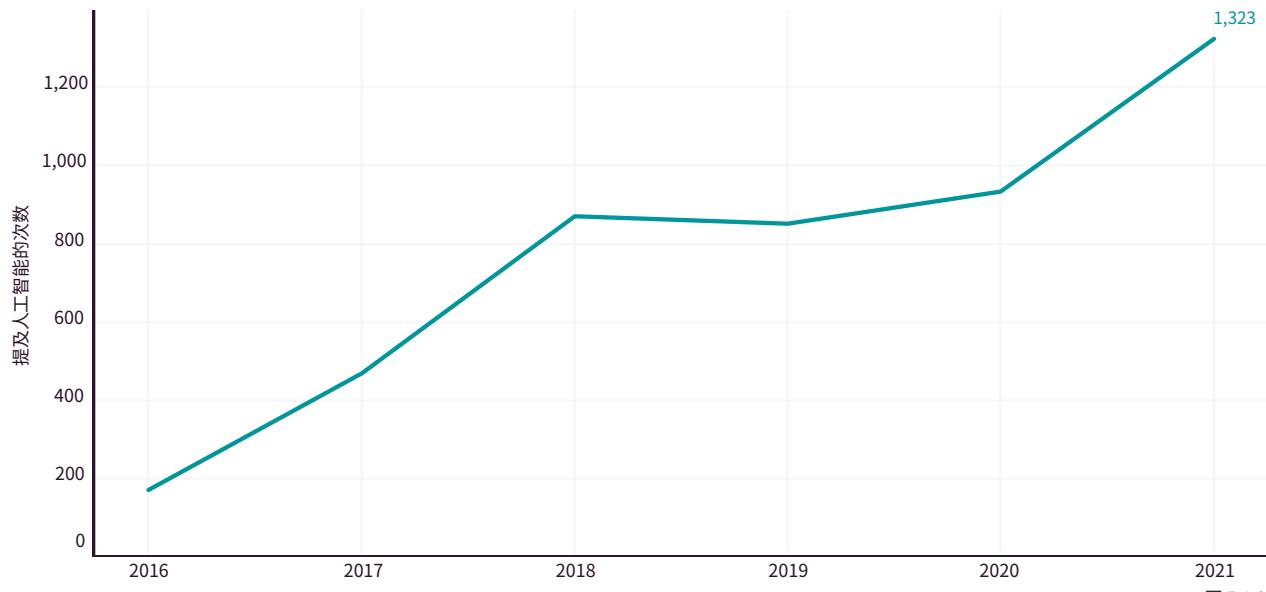


图 5.1.9

<sup>2</sup> 方法见附录。包括的国家有：澳大利亚、比利时、巴西、加拿大、中国、丹麦、芬兰、法国、德国、印度、爱尔兰、意大利、日本、荷兰、新西兰、挪威、俄罗斯、新加坡、南非、韩国、西班牙、瑞典、瑞士、英国和美国。



## 按地理区域划分

图5.1.10a显示了2021年部分国家立法程序中提及人工智能的次数。与通过成为法律的法案中提及人工智能的次数趋势类似，西班牙、英国和美国位居榜首。

2021年部分国家立法程序中提及人工智能的次数

来源: 人工智能指数, 2021 | 图: 2022年人工智能指数报告

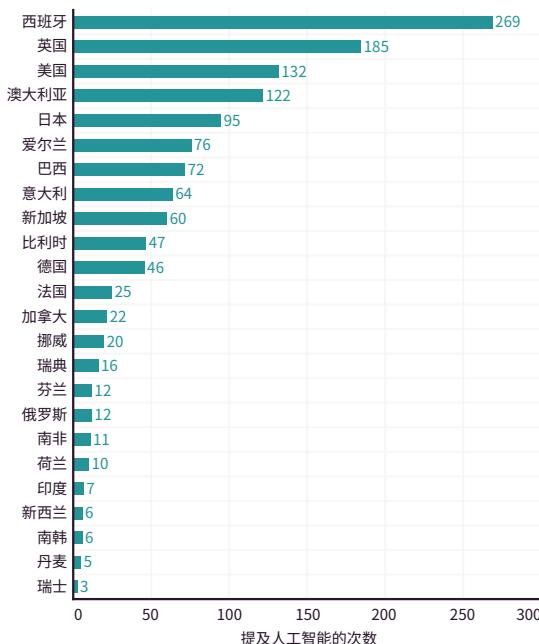


图 5.1.10a

图5.1.10b显示了过去六年中部分国家立法程序中提及人工智能的次数。英国以939次提及占据榜首，其次是西班牙、日本、美国和澳大利亚。

2021年部分国家立法程序中提及人工智能的次数

来源: 人工智能指数, 2021 | 图: 2022年人工智能指数报告

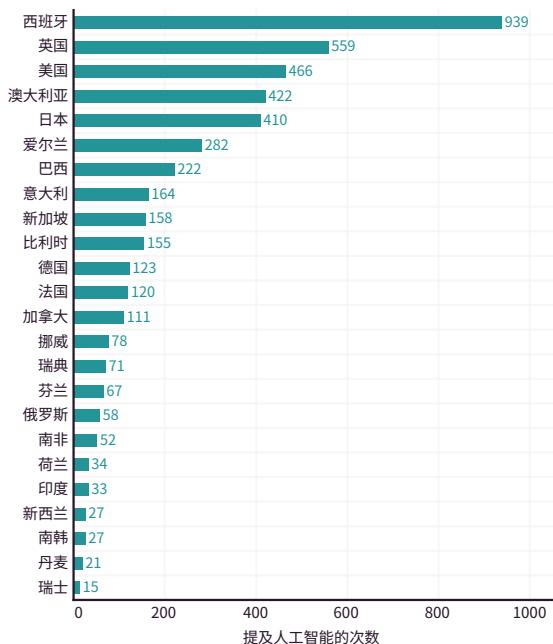


图 5.1.10b



## 美国人工智能政策文件

为了估计国家政府以外的活动，这些活动也在为人工智能相关规则的制定提供信息，人工智能指数追踪了过去四年中发表政策文件的55个美国组织。这些组织包括：智囊团和政策研究所（19）；大学研究所和研究项目（14）；公民社会组织、协会和联盟（9）；行业和咨询组织（9）；政府机构（4）。<sup>3</sup> 本节中的政策文件包括：研究论文、研究报告、简报或博客文章。这些政策文件涉及与人工智能有关的问题，并向政策制定者提出具体建议。将这些文件的主题划分为

主要和次要类别，其中主要主题是文件的主要焦点，而次要主题则是文件的一个副主题或简单探讨的问题。

图5.1.11显示了2018年至2021年发表的美国人工智能相关政策文件的数量，这一数据表明美国政策制定领域对人工智能普遍是感兴趣的。自2018年以来，政策文件的总数增加了两倍，在2020年达到顶峰，有273篇，在2021年略有下降，有210篇。

2018-21年美国本土组织的人工智能相关政策文件的数量

来源: 人工智能指数, 2021 | 图: 2022年人工智能指数报告

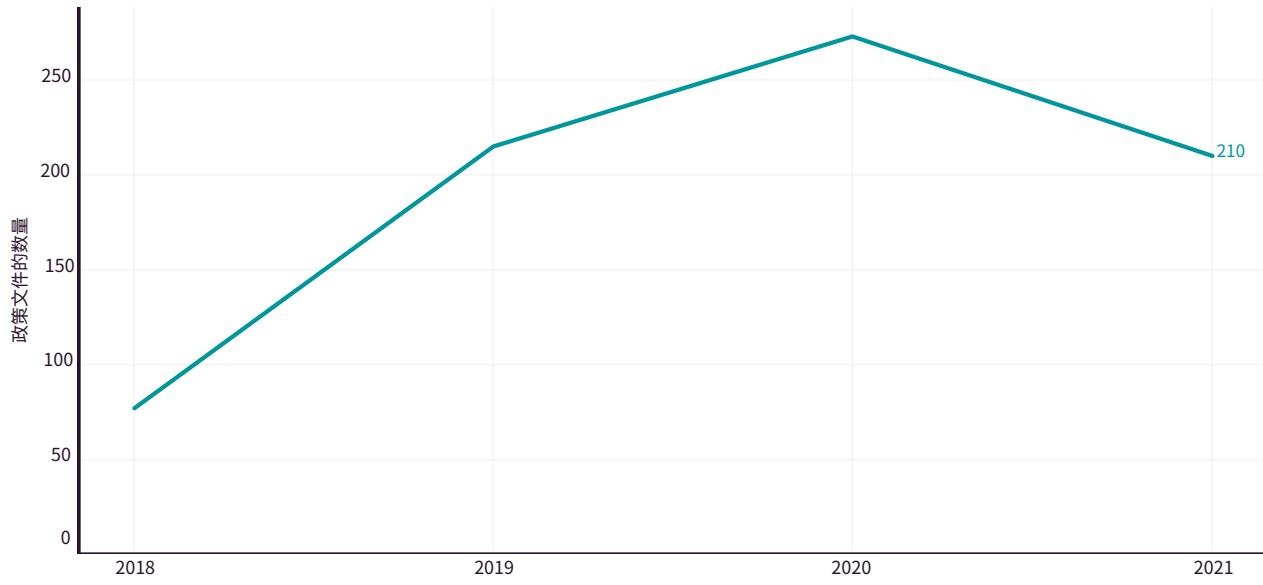


图 5.1.11

<sup>3</sup> 该指数所关注的组织的完整名单可在附录中找到。



## 按主题划分

2021年，隐私、安全和安保；创新和技术；以及道德（图5.1.12）是收到主要关注的主题。某些主题，如政府和公共管理、教育和技能，以及民主，并没有作

为主要主题突出，但它们作为次要主题被更频繁地提及。在人工智能专题中，相对较少关注的是那些与能源和环境、人文、物理科学、社会和行为科学有关的主题。

2021年按主题划分的美国本土组织与人工智能相关的政策文件数量

来源: 人工智能指数, 2021 | 图: 2022年人工智能指数报告

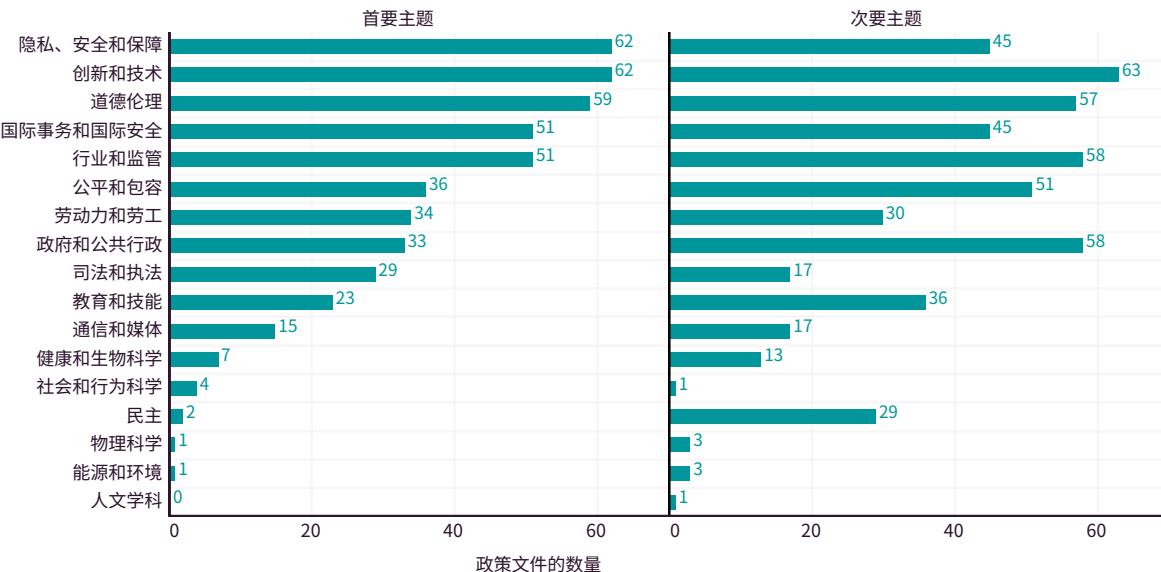


图 5.1.12



本节根据美国政府和Bloomberg Government的数据，分析了美国的公共人工智能投资情况。

## 5.2 美国在人工智能方面的公共投资

### 非国防性人工智能研发的联邦预算

2021年12月，美国国家科学技术委员会发布了一份报告，介绍了参与网络和信息技术研究与发展（Networking and Information Technology Research and Development，NITRD）计划和国家人工智能计划的各部门和机构的公共人工智能研发预算。该报告不包括国防和情报机构的机密人工智能研发投入信息。

在2021财年，非国防性的美国政府机构共拨出15.3亿美元用于人工智能研发支出，大约是2018财年的2.7倍（图5.2.1）。预计2022财年这一数字将上升8.8%，总额为16.7亿美元。<sup>4</sup> 非国防部门用于人工智能研发的金额不断增加，表明美国政府对公共部门资助人工智能研究和开发的兴趣持续浓厚，涉及众多联邦机构。

**非国防部门用于人工智能研发的金额不断增加，表明美国政府对公共部门资助人工智能研究和开发的兴趣持续浓厚，涉及众多联邦机构。**

#### 2018-22财年美国联邦政府对非国防部门人工智能研发的预算

来源：美国NITRD计划，2022 | 图：2022年人工智能指数报告

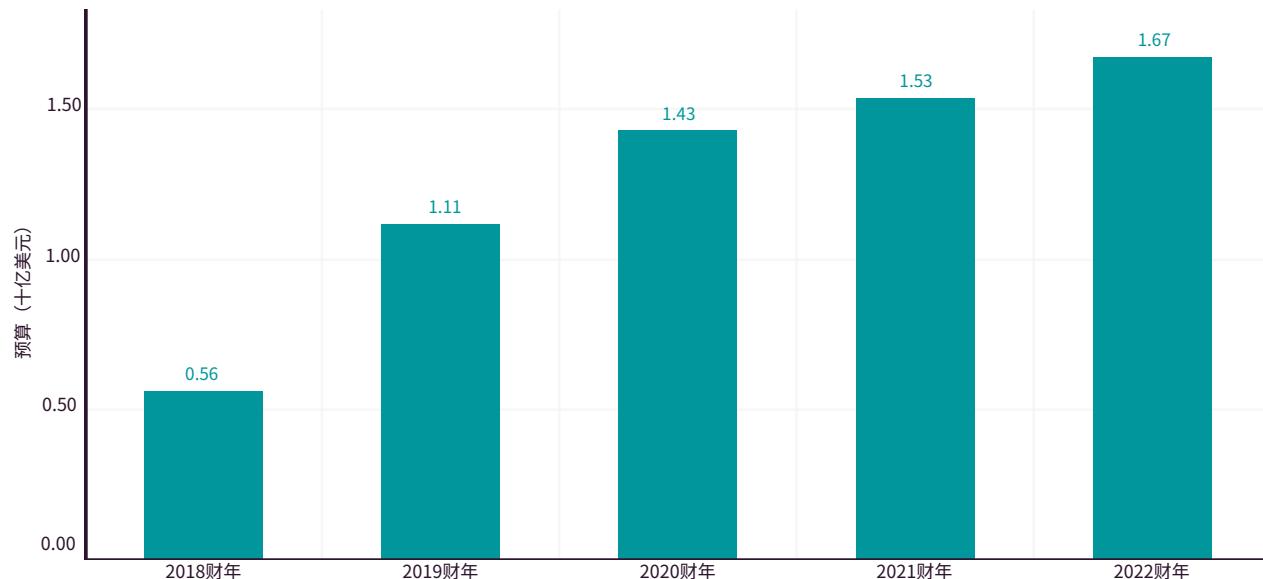


图 5.2.1

<sup>4</sup> 有关2018-22财年人工智能研发投入的详情，请参见NITRD网站，其中包括核心人工智能与人工智能交叉领域的细分。请注意，2018财年没有提供人工智能交叉预算数据。



## 美国国防部预算需求

美国国防部（DOD）在人工智能方面的支出可以通过查看国防部公开的与人工智能有关的研究、开发、测试和评估（RDT&E）的需求来获取。2021财年，国防部为500个人工智能研发项目拨款92.6亿美元（图

5.2.2），比2020年的86.8亿美元增加了6.68%。对于2022财年，到目前为止，该部门已经申请了100亿美元，考虑到额外申请和国会的拨款，这个数字可能还会增加。

### 2020-22财年美国国防部人工智能专项研究、开发、测试和评估（RDT&E）预算

来源：Bloomberg Government和美国国防部，2021 | 图：2022年人工智能指数报告

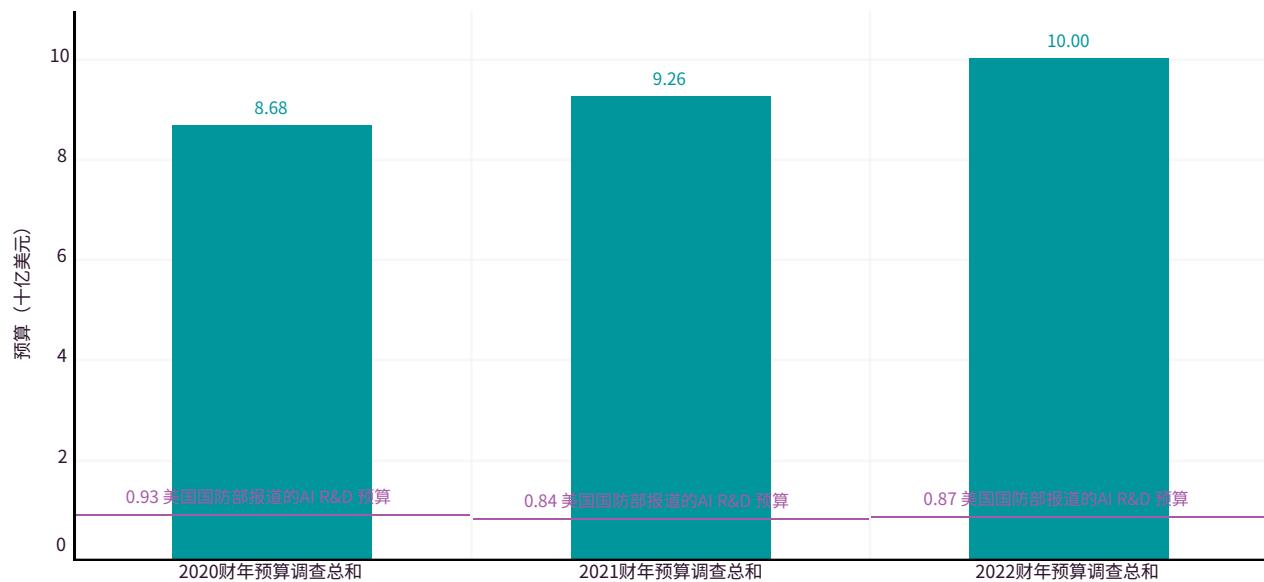


图 5.2.2

**重要的数据注意事项：**上图表明了量化公共人工智能支出的一项挑战。Bloomberg Government在国防部预算中搜索人工智能相关关键词的分析显示，该部门在2022财年提出了100亿美元用于人工智能特定的研发。然而，国防部自己测算的结果是一个较小的数

字，即8.74亿美元。这种差异可能是由于对人工智能相关预算项目的定义不同所造成的。例如，一个将人工智能用于网络防御的研究项目可能会将人力、硬件和操作相关的支出都纳入到人工智能相关的预算申请中，而其中的人工智能软件部分会小很多。



## 国防部资金最多的五个项目

本节重点对国防部优先考虑的一些人工智能相关研究项目进行了更加定性的分析。表5.2.1列出了2021年获得最多资金的五个国防部相关人工智能项目。在过去的一年里，国防部尝试将人工智能部署在一些专门的任务中，例如地理空间监测、减少大规模杀伤性武器造成的威胁等等。

项目名称	部门	资助金额（以百万计）	应用目的
1 快速能力开发和成熟化	陆军	257	资助可用于军事目的的各种人工智能相关技术原型的开发、工程、购置和运行。
2 反大规模杀伤性武器的先进技术发展	国防威胁署	254	开发可以 "拒绝、击败和破坏" 大规模杀伤性武器的技术。
3 算法战争跨职能团队--软件试点计划	美国国防部长办公室	230	加快国防部系统中人工智能技术的整合，以 "提高战争的速度和杀伤力"。
4 联合人工智能中心	国防信息系统局	137	开发、测试、制作原型和展示各种人工智能和机器学习能力，目的是将这些能力整合到众多领域中，其中包括 "供应链、个人恢复、基础设施评估、灾害期间的地理空间监测和网络感知。"
5 高性能计算现代化项目	陆军	96	调查、展示和孵化通用和特殊用途的超级计算环境，用于满足国防部的各类优先事项。

表 5.2.1



## 按部门划分的国防部人工智能研发支出

国防部在人工智能研发方面的支出可以在下级部门层面进一步细分，这体现了具体的国防部门--例如陆军和海军--在人工智能支出方面的对比（图5.2.3）。美国海军在2021财年是国防部开支最大的部门，且2022年

保持同样的趋势。他们在2022财年为人工智能相关项目申请了总计18.6亿美元的资金，其次是陆军（17.7亿美元）、国防部长办公室（11亿美元）和空军（8.83亿美元）。

2020-22财年美国国防部各部门用于人工智能研究、开发、测试和评估（RDT&E）的预算

来源：Bloomberg Government, 2021 | 图：2022年人工智能指数报告

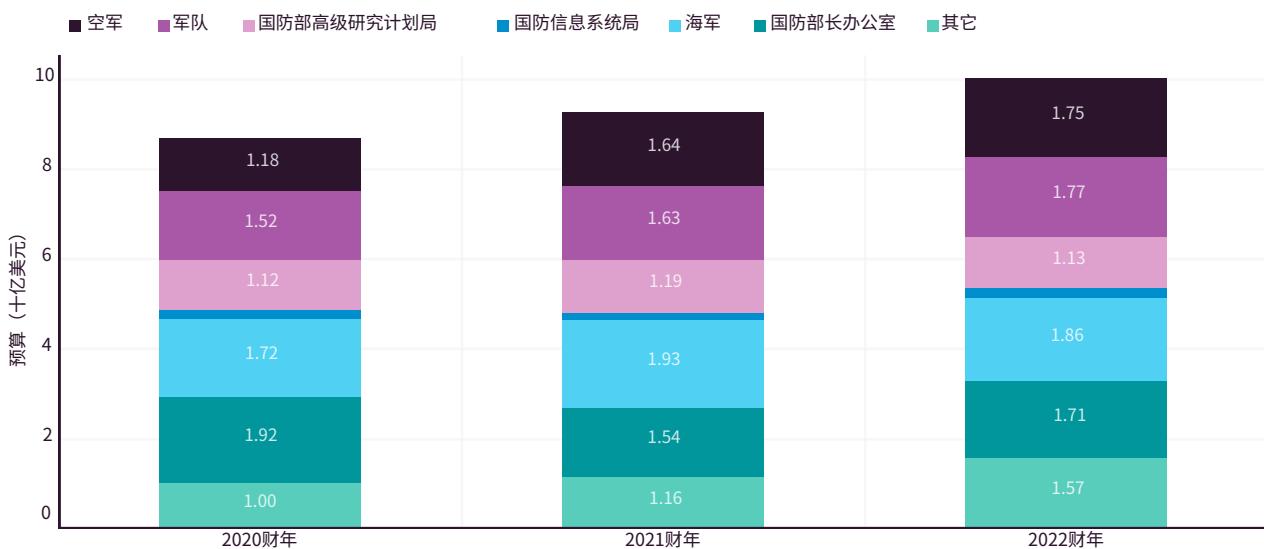


图 5.2.3



## 美国政府与人工智能有关的合同支出

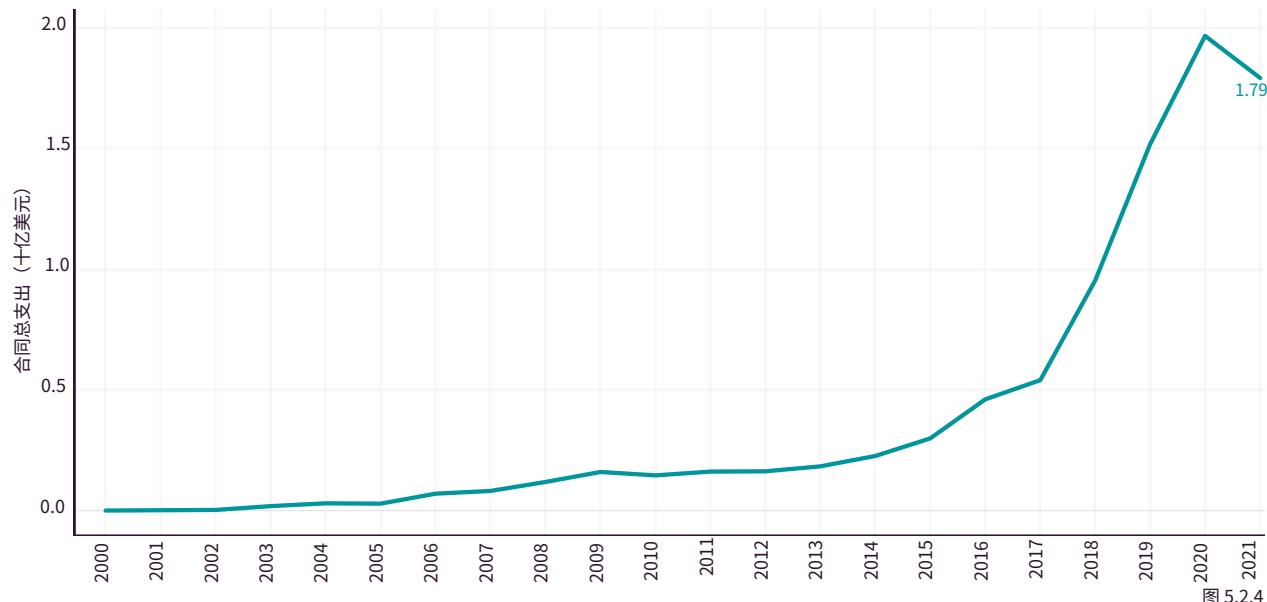
对人工智能的公共投资也可以通过联邦政府对人工智能相关合同的支出来衡量。美国政府机构经常将各种商品和服务的供应合同授予私营公司，这些商品和服务通常在一个机构的预算中占据最大份额。Bloomberg Government建立了一个模型，通过将所有在标题或描述中包含一组100多个人工智能特定关键词的合同交易加起来，来判断美国政府的合同是否是人工智能相关的。<sup>5</sup>

## 合同支出总额

2021年，联邦部门和机构共花费17.9亿美元用于人工智能相关合同支出。虽然这一数额几乎是2018年人工智能相关合同支出的两倍（大约9.2亿美元），但与2020年人工智能相关合同支出的数额相比略有下降，后者达到了19.7亿美元的峰值（图5.2.4）。

### 2000-21财年美国政府在人工智能方面的合同总支出

来源：Bloomberg Government, 2021 | 图：2022年人工智能指数报告



<sup>5</sup>请注意，承包商在采购过程中可能会在其应用程序中加入一些关键词，因此所包括的一些项目相对于其他部分的技术而言，人工智能的成分可能相对较少。



## 按部门和机构划分的合同支出情况

图5.2.5和5.2.6分别报告了2021年和2000年至2021年排名前10的联邦机构的人工智能相关合同支出。在这两个图表中，国防部的支出大大超过了美国政府的其他部门。2021年，国防部在人工智能相关合同上支出了11.4亿美元，大约是排名第二的部门--卫生和公共服务部（2.34亿美元）---的五倍。

在过去的四年里，人工智能合同的总支出情况类似。自2018年以来，国防部在人工智能合同上支出了52亿美元，大约是排名紧随其后的NASA（14.1亿美元）的七倍。事实上，自2018年以来，国防部在人工智能相关合同上的支出是所有其他政府机构总和的两倍。紧随国防部和NASA之后的是卫生和人类服务部（7亿美元）、国土安全部（3.62亿美元）和财政部（1.56亿美元）。

### 2021年美国政府各部门和机构在人工智能上的主要合同支出情况

来源：Bloomberg Government, 2021 | 图：2022年人工智能指数报告

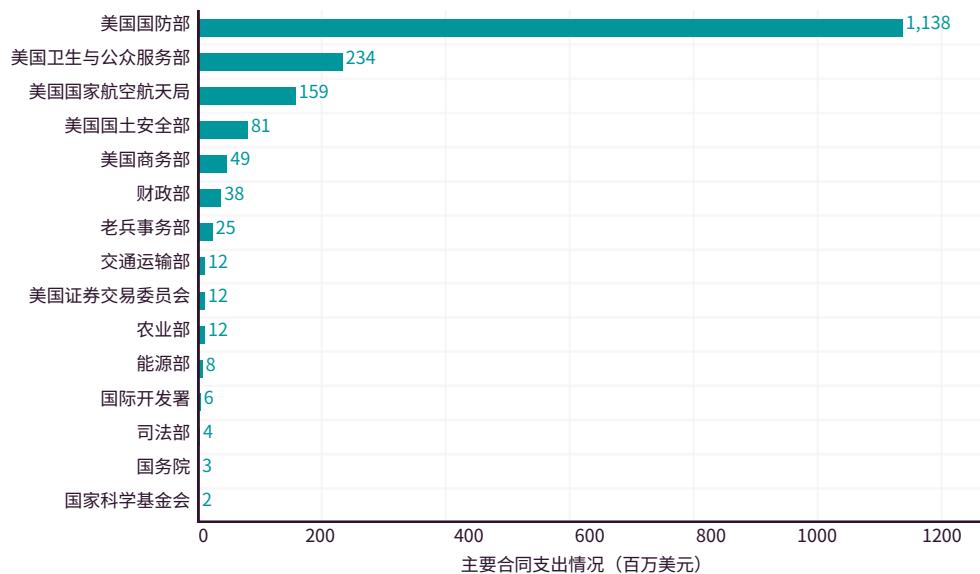


图 5.2.5



### 2000-21年按美国政府部门和机构划分的人工智能合同支出情况（总和）

来源：Bloomberg Government, 2021 | 图：2022年人工智能指数报告

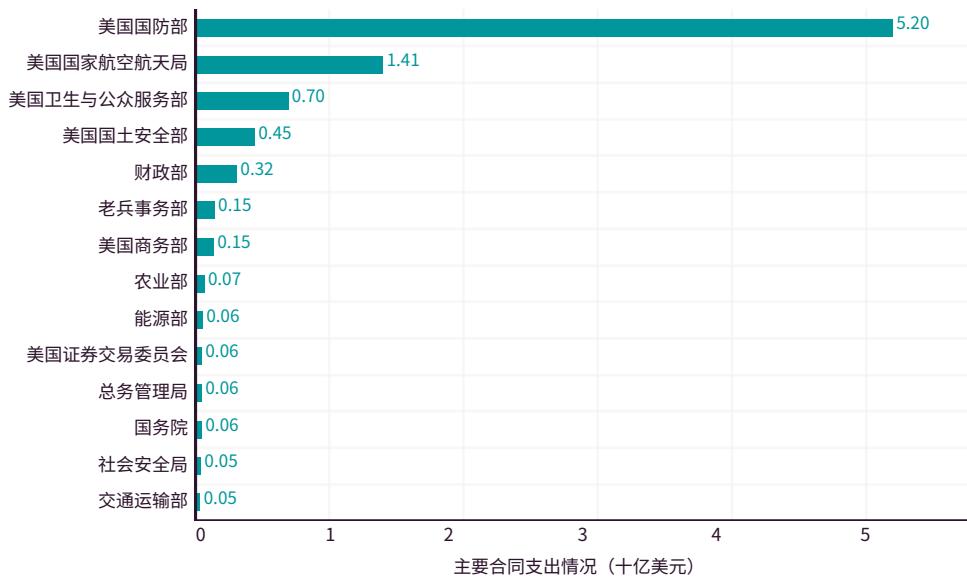


图 5.2.6



## 2021年五个开支最大的部门的最大合同

为了更好地展示不同的美国政府部门如何使用人工智能，表5.2.2给出了2021年五个人工智能相关开支最大的部门签署的金额最大的人工智能相关合同。去年，美国政府在人工智能方面的投资包括：建造自动驾驶汽车原型、开发可以协助烧伤分类的人工智能成像系统，以及创造能够实现更高层次月球导航的机器人。

合同名称	部门	总额(以百万计)	目的
在汽车网络安全、车辆安全技术、车辆轻量化、自动驾驶车辆和智能系统、互联车辆和先进能源储存技术等目标领域的原型服务	国防部	70	在汽车网络安全、车辆安全技术、自动驾驶车辆和智能系统领域获得原型。
生物医学高级研究与发展局(BARDA)	美国卫生署	20	开发光学成像设备和机器学习算法，协助对伤口和常规烧伤进行分类和治疗。
商用月球有效载荷服务	美国国家航空航天局	14	开发能够在月球南极航行的月球机器人，这些机器人的任务是获取月球资源和从事基于月球的科学活动。
SBIR-自主监控塔-交付订单	国土安全部	37	建造能够自主监控的塔台。
附表70：信息技术	国防部	13	利用人工智能技术开发一个可以改善专利搜索的原型。

表 5.2.2



2022年  
人工智能指数报告

## 附录



# 附录

<b>章节一</b>	研发 (R&D)	198
<b>章节二</b>	技术性能	200
<b>章节三</b>	技术AI伦理	212
<b>章节四</b>	经济和教育	216
<b>章节五</b>	人工智能政策和国家战略	222



# 章节一 研发 (R&D)

## 乔治敦大学安全与新兴技术中心

编写人：Sara Abdulla 和 James Dunham

安全与新兴技术中心 (The Center for Security and Emerging Technology, CSET) 是乔治城大学沃尔什外交学院的一个政策研究组织，主要在安全和技术的交叉领域进行数据驱动的研究，并向政界提供无党派分析结果。

## 来自CSET学术文献合并语料库的出版物

### 资料来源

CSET的学术文献合并语料库结合了来自Digital Science's Dimensions、Clarivate's Web of Science、Microsoft Academic Graph、China National Knowledge Infrastructure、arXiv以及Papers With Code的不同出版物。<sup>1</sup>

### 方法论

为了创建合并后的语料库，CSET使用出版物元数据对列出的来源进行了去重处理，然后对链接的出版物的元数据进行了合并。为了识别人工智能出版物，CSET使用了该语料库的一个英文子集：2010年以来与人工智能相关的出版物。<sup>2</sup> CSET的研究人员开发了一个分类器，通过利用arXiv文库来识别人工智能相关的出版物，在该文库中，作者和编辑按主题标记论文。

为了提供出版物的研究领域，CSET将分析语料库中的每篇出版物与微软学术图谱 (MAG) 的研究领域模型的预测相匹配，从而生成描述已发表的研究领域和相应分数的分层标签。<sup>3</sup> CSET的研究人员确定了2010年以来他们的人工智能相关出版物语料库中最常见的研究领域，并将所有其他领域的出版物记录为“其他人工

智能”。然后按照最高分的领域和出版年份对英语人工智能相关的出版物进行统计。

CSET还提供了与每个国家相关的人工智能相关工作的逐年引用情况。如果一份出版物的至少一位作者的组织关系属于某个国家，那么这份出版物就与该国家有关。并非所有出版物都有引文计数，那些没有计数的出版物不包括在引文分析中。在2010年至2020年期间发表的英文人工智能论文中，超过70%的论文有引文数据。

CSET将跨国合作计算为每份出版物的不同作者的国家对。每次合作只计算一次。例如，如果一个出版物有两个来自美国的作者和两个来自中国的作者，它就被算作一次中美合作。

此外，如果说有的话，还提供了按年份和出版类型（例如，学术期刊文章、会议论文）划分的出版物数量。如上所述，这些出版物类型按隶属国家进行分类。

CSET还提供了出版物的所属部门，与国家归属分析一样，部门与出版物通过作者的所属关系进行关联。并非所有的隶属关系都以部门为特征；CSET的研究人员主要依靠DigitalScience的GRID来实现这一目的，而并非所有的组织都能在GRID中找到或与之链接。<sup>4</sup> 当有附属部门时，论文按年份计入这些部门。学术出版物的跨部门合作采用与跨国合作分析中相同的方法计算。

## 来自CSET的AI专利数据集的专利

### 资料来源

CSET的人工智能专利数据集由CSET和1790 Analytics开发。它包括与人工智能的发展和应用有关的专利，由CPC/IPC代码和关键词来表示。

<sup>1</sup> 所有CNKI的内容都是由美国明尼苏达州明尼阿波利斯的East View Information Services为CSET提供的。

<sup>2</sup> 更多信息见James Dunham, Jennifer Melot, and Dewey Murdick, "Identifying the Development and Application of Artificial Intelligence in Scientific Text," arXiv [cs.DL], May 28, 2020, <https://arxiv.org/abs/2002.07143>.

<sup>3</sup> 这些分数是基于研究领域和论文嵌入之间的余弦相似度计算得到的。见Zhihong Shen, Hao Ma, and Kuansan Wang, "A Web-Scale System for Scientific KnowledgeExploration," arXiv [cs.CL], May 30, 2018, <https://arxiv.org/abs/1805.12216>.

<sup>4</sup> 关于数字科学的GRID数据集的更多信息，参见<https://www.grid.ac/>。



## 方法论

在本分析中，按年份和国家对专利进行分组，然后根据“专利族”层面进行统计。<sup>5</sup> CSET从一个专利族中的最新出版日期中提取年份值。这种方法的好处是可以捕捉到专利族内的更新（如修正案）。

将申请专利的第一个国家确定为专利的来源国<sup>6</sup>。

## GITHUB 星标

### 资料来源

GitHub：使用了星标历史（可在星标历史网站获得）

检索数据。

### 方法论

报告中的视觉图展示了不同的GitHub仓库在一段时间内的星标数量。这些仓库包括以下几个：

apache/cn/ailearning, apache/incubator-mxnet,  
Avik-Jain/100-Days-Of-ML-Code, aymericdamien/  
TensorFlow-Examples, BVLC/caffe, cafe2/caffe2,  
CorentinJ/Real-Time-Voice-Cloning, deepfakes /  
faceswap, dmlc/mxnet, exacity/deeplearningbook-  
chinese, fchollet/keras, floodsung/Deep-Learning-  
Papers-Reading-Roadmap, iperov/DeepFaceLab,  
Microsoft/CNTK, opencv/opencv, pytorch/pytorch,  
scikit-learn/scikit-learn, scutan90/DeepLearning-  
500-questions, tensorflow/tensorflow, Theano/  
Theano, Torch/Torch7.

### 误差

目前GitHub仓库没有提供计算用户何时从仓库中删除星标的方法。因此，报告的数据略微高估了星标的数量。与GitHub上的仓库的实际星标数量比较，可以发现数字相当接近，而且趋势没有改变。

5 专利是在“专利族”层面而不是“专利文件”层面进行分析的，因为专利族是一个专利文件的集合体，这个集合体中的全部文件与同一发明人/受让人的单一发明和/或创新相关。因此，在“专利族”层面上进行计算，可以在一个专利族中有多个专利文件或一个专利在多个司法管辖区申请的情况下减少人为的数字膨胀。

6 关于CSET分配国家价值的方法和实验的更多详情，请参见“专利和人工智能”的脚注26。关于CSET分配国家价值的方法和实验详情，见“专利和人工智能：入门”的脚注26，作者是Dewey Murdick和Patrick Thomas。 (Centerfor Security and Emerging Technology, September 2020), <https://doi.org/10.51593/20200038>。



## 章节二 技术性能

### ImageNet

关于ImageNet准确性的数据是通过详细的arXiv文献综述，并与 [Papers with Code](#) 上报告的技术进展相互参照而获得的。报告的日期对应于论文首次发表在arXiv上的年份，而报告的结果（top-1或top-5准确率）对应于每篇论文的最新版本中报告的结果。人类水平性能的估计来自 [Russakovsky et al., 2015](#)。了解更多关于LSVRC ImageNet竞赛和 [ImageNet](#) 数据集的信息。

#### ImageNet: Top-1 准确率

为了突出在不使用额外训练数据的情况下Top-1准确率的进展，分数取自以下论文：

[Adversarial Examples Improve Image Recognition](#)

[Billion-Scale Semi-Supervised Learning for Image Classification](#)

[Dual Path Networks](#)

[Densely Connected Convolutional Networks](#)

[EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks](#)

[Fixing the Train-Test Resolution Discrepancy: FixEfficientNet](#)

[ImageNet Classification with Deep Convolutional Neural Networks](#)

[Masked Autoencoders Are Scalable Vision Learners](#)

为了突出在使用额外训练数据的情况下top-1准确率的进展，分数取自以下论文：

[Big Transfer \(BiT\): General Visual Representation Learning](#)

[CoAtNet: Marrying Convolution and Attention for All Data Sizes](#)

[EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks](#)

[Self-Training with Noisy Student Improves ImageNet Classification](#)

[Sharpness-Aware Minimization for Efficiently Improving Generalization](#)

[Xception: Deep Learning with Depthwise Separable Convolutions](#)

#### ImageNet: Top-5准确率

为了突出在不使用额外训练数据的情况下top-5准确率的进展，分数取自以下论文：

[Adversarial Examples Improve Image Recognition](#)

[EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks](#)

[Exploring the Limits of Weakly Supervised Pretraining](#)

[Fixing the Train-Test Resolution Discrepancy: FixEfficientNet](#)

[GPipe: Efficient Training of Giant Neural Networks Using Pipeline Parallelism](#)

[High-Performance Large-Scale Image Recognition Without Normalization](#)

[ImageNet Classification with Deep Convolutional Neural Networks](#)

[Learning Transferable Architectures for Scalable Image Recognition](#)

[Squeeze-and-Excitation Networks](#)

为了突出在使用额外训练数据的情况下top-5准确率的进展，分数取自以下论文：

[Big Transfer \(BiT\): General Visual Representation Learning](#)

[Deep Residual Learning for Image Recognition](#)

[EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks](#)

[Florence: A New Foundation Model for Computer Vision](#)

[Self-Training with Noisy Student Improves ImageNet Classification](#)

[Xception: Deep Learning with Depthwise Separable Convolutions](#)



## STL-10

关于STL-10 FID分数的数据是通过详细的arXiv文献审查，并与 [Papers with Code](#) 上报告的技术进展相互参照而获得的。报告的日期对应于论文首次发表在arXiv上的年份，报告的结果（FID分数）对应于每篇论文的最新版本中报告的结果。关于 [STL-10 paper](#) 基准的详情可以在STL-10论文中找到。

为了突出STL-10的进展，分数取自以下论文：

[DEGAS: Differentiable Efficient Generator Search](#)  
[Dist-GAN: An Improved GAN Using Distance Constraints](#)  
[Off-Policy Reinforcement Learning for Efficient and Effective GAN Architecture](#)  
[Search Score Matching Model for Unbounded Data Score](#)

## CIFAR-10

CIFAR-10的FID得分数据是通过详细的arXiv文献审查，并与 [Papers with Code](#) 上报告的技术进展相互参照而获得的。报告的日期对应于论文首次发表在arXiv上的年份，而报告的结果（FID分数）对应于每篇论文的最新版本中报告的结果。关于CIFAR-10基准的详情可以在 [CIFAR-10 paper](#) 论文中找到。

为了突出CIFAR-10的进展，分数取自以下论文：

[AutoGAN: Neural Architecture Search for Generative Adversarial Networks](#)  
[Denoising Diffusion Probabilistic Models](#)  
[Improved Training of Wasserstein GANs](#)  
[Large Scale GAN Training for High Fidelity Natural Image Synthesis](#)  
[Score-Based Generative Modeling in Latent Space](#)

## FaceForensics++

关于FaceForensics++准确性的数据是通过详细的arXiv文献回顾检索出来的。报告的日期对应于论文首次发表在arXiv上或方法被引入的年份。通过FaceForensics，研究人员测试了先前的deepfake检

测方法。一个方法即使后来又被测试过，但是仍将它的引入年份确定为它被列入报告的年份。报告的结果（准确性）对应于每篇论文的最新版本中所报告的结果。关于FaceForensics++基准的详情可以在 [FaceForensics++](#) 论文中找到。

为了突出FaceForensics++的进展，分数取自以下论文：

[A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional Layer](#)  
[Detection of Deepfake Videos Using Long Distance Attention](#)  
[FakeCatcher: Detection of Synthetic Portrait Videos Using Biological Signals](#)  
[FaceForensics++: Learning to Detect Manipulated Facial Images](#)  
[Learning Spatiotemporal Features with 3D Convolutional Networks](#)  
[Recasting Residual-Based Local Descriptors as Convolutional Neural Networks](#)  
[Rich Models for Steganalysis of Digital Images](#)  
[Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues](#)  
[Xception: Deep Learning with Depthwise Separable Convolutions](#)

## Celeb-DF

关于Celeb-DF AUC的数据是通过详细的arXiv文献回顾检索出来的。报告的日期对应于论文首次发表在arXiv上或方法被引入的年份。通过Celeb-DF，研究人员测试了先前的deepfake检测方法。一个方法即使后来又被测试过，但是仍将它的引入年份确定为它被列入报告的年份。报告的结果（AUC）对应于每篇论文的最新版本中报告的结果。关于Celeb-DF基准的详情可以在 [Celeb-DF](#) 论文中找到。

为了突出Celeb-DF的进展，分数取自以下论文：

[Exposing DeepFake Videos by Detecting Face](#)



[Warping Artifacts](#)

[FaceForensics++: Learning to Detect Manipulated Facial Images](#)

[Face X-Ray for More General Face Forgery Detection](#)

[Spatial-Phase Shallow Learning: Rethinking Face Forgery Detection in Frequency Domain](#)

## Leeds Sports Poses

Leeds Sports Poses的正确关键点百分比（percentage of correct keypoints, PCK）的数据是通过详细的arXiv文献审查，并与[Papers with Code](#)上报告的技术进展相互参照而获得的。报告的日期对应于论文首次发表在arXiv上的年份，而报告的结果（PCK）对应于每篇论文的最新版本中报告的结果。关于Leeds Sports Poses基准的详情可以在[Leeds Sports Poses](#)论文中找到。

为了突出Leeds Sports Poses的进展，分数取自以下文件：

[Articulated Pose Estimation by a Graphical Model with Image Dependent Pairwise Relations](#)

[Human Pose Estimation via Convolutional Part Heatmap Regression](#)

[Jointly Optimize Data Augmentation and Network Training: Adversarial Data Augmentation in Human Pose Estimation](#)

[Knowledge-Guided Deep Fractal Neural Networks for Human Pose Estimation](#)

[OmniPose: A Multi-Scale Framework for Multi-Person Pose Estimation](#)

[Toward Fast and Accurate Human Pose Estimation via Soft-Gated Skip Connections](#)

## Human 3.6M

于Human 3.6M平均（每）关节位置误差的数据是通过详细的arXiv文献审查，并与[Papers with Code](#)上报告的技术进展相互参照而获得的。报告的日期对应于论文首次发表在arXiv上的年份，而报告的结果（MPJPE）对应于每篇论文的最新版本中报告的结

果。关于Human 3.6M基准的详情可以在[Human3.6M](#)论文中找到。

为了突出在不使用额外训练数据的情况下Human 3.6M的进展，分数取自以下论文：

[3D Human Pose Estimation in Video with Temporal Convolutions and Semi-Supervised Training](#)

[Conditional Directed Graph Convolution for 3D Human Pose Estimation](#)

[Cross View Fusion for 3D Human Pose Estimation Epipolar Transformers](#)

[Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments](#)

[Learning 3D Human Pose from Structure and Motion Robust Estimation of 3D Human Poses from a Single Image](#)

为了突出在使用额外训练数据的情况下Human 3.6M的进展，分数取自以下论文：

[Epipolar Transformers](#)

[Learnable Triangulation of Human Pose](#)

[TesseTrack: End-to-End Learnable Multi-Person Articulated 3D Pose Tracking](#)

## 城市景观挑战赛，像素级语义标签任务

城市景观挑战赛，像素级语义标签任务的平均IoU数据取自城市景观数据集，更具体地说，取自其像素级语义标签排行榜。关于Cityscapes数据集和其他相应的语义分割挑战的更多详情，可以访问[Cityscapes](#)数据集网页。

## CVC-ClinicDB 和 Kvasir-SEG

CVC-ClinicDB和Kvasir-SEG的mean dice数据是通过详细的arXiv文献审查，并与[Papers with Code](#)（[CVC-ClinicDB](#) 和 [Kvasir-SEG](#)）上报告的技术进展相互参照而获得的。报告的日期对应于论文首次发表在arXiv上的年份，而报告的结果（mean dice）对应于每篇论文的最新版本中报告的结果。关于CVC-ClinicDB基准的



详情可在 [CVC-ClinicDB](#) 数据库页面找到。关于Kvasir-SEG基准的详情可在 [Kvasir-SEG](#) 论文中找到。

为了突出CVC-ClinicDB的进展，分数取自以下论文：  
[DoubleU-Net: A Deep Convolutional Neural Network for Medical Image Segmentation](#)  
[Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation](#)  
[MSRF-Net: A Multi-Scale Residual Fusion Network for Biomedical Image Segmentation](#)  
[ResUNet++: An Advanced Architecture for Medical Image Segmentation](#)  
[U-Net: Convolutional Networks for Biomedical Image Segmentation](#)

为了突出Kvasir-SEG的进展，分数取自以下论文：  
[Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation](#)  
[MSRF-Net: A Multi-Scale Residual Fusion Network for Biomedical Image Segmentation](#)  
[PraNet: Parallel Reverse Attention Network for Polyp Segmentation](#)  
[U-Net: Convolutional Networks for Biomedical Image Segmentation](#)

**美国国家标准与技术研究所（NIST）人脸识别供应商测试（FRVT）和NIST FRVT遮挡效果**  
NIST FRVT 1:1验证精度的数据来自 [FRVT 1:1 验证排行榜](#)。关于NIST FRVT遮挡效果的数据来自 [FRVT 遮挡效果排行榜](#)。遮挡效果排行榜包含了在2020年3月中旬（即COVID-19开始的时间）之前和之后提交给FRVT的319种人脸识别算法的测试结果。

## 视觉问答（VQA）

关于VQA的数据取自最近几次VQA挑战赛。要了解更多关于VQA挑战赛的信息，请查阅以下链接。要了解更多关于2021年VQA挑战赛的信息，请参考以下链接。更具体地说，人工智能指数报告利用了来自以下几轮VQA挑战赛的数据。

[VQA Challenge 2016](#)  
[VQA Challenge 2017](#)  
[VQA Challenge 2018](#)  
[VQA Challenge 2019](#)  
[VQA Challenge 2020](#)  
[VQA Challenge 2021](#)

## Kinetics-400, Kinetics-600, and Kinetics-700

关于Kinetics-400、Kinetics-600和Kinetics-700的数据是通过详细的arXiv文献审查，并与Papers with Code ([Kinetics-400](#)、[Kinetics-600](#) 和 [Kinetics-700](#)) 上报告的技术进展相互参照而获得的。报告的日期对应于论文首次发表在arXiv上的年份，报告的结果（准确性）对应于每篇论文的最新版本中报告的结果。关于Kinetics-400基准详情可以在 [Kinetics-400](#) 论文中找到。关于Kinetics-600基准详情，可在 [Kinetics-600](#) 论文中找到。关于Kinetics-700基准详情可在 [Kinetics-700](#) 论文中找到。

为了突出Kinetics-400的进展，分数取自以下论文：  
[Co-Training Transformer with Videos and Images Improves Action Recognition](#)  
[Large-Scale Weakly-Supervised Pre-training for Video Action Recognition](#)  
[Multiview Transformers for Video Recognition](#)  
[Non-Local Neural Networks](#)  
[Omni-Sourced Webly-Supervised Learning for Video Recognition](#)  
[SlowFast Networks for Video Recognition](#)  
[Temporal Segment Networks: Towards Good Practices for Deep Action Recognition](#)

为了突出Kinetics-600的进展，分数取自以下论文：  
[Masked Feature Prediction for Self-Supervised Visual Pre-Training](#)  
[Multiview Transformers for Video Recognition](#)  
[Learning Spatio-Temporal Representation with Local and Global Diffusion](#)



## Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-offs in Video Classification SlowFast Networks for Video Recognition

为了突出Kinetics-700的进展，分数取自以下论文：  
Learn to Cycle: Time-Consistent Feature Discovery for Action Recognition  
Masked Feature Prediction for Self-Supervised Visual Pre-Training  
Multiview Transformers for Video Recognition

## ActivityNet：时间动作定位任务

在这项挑战中，有三个独立的任务，但都集中在一个主要问题上，即在ActivityNet基准的未修剪的视频中从时间上定位活动发生的位置。为了获得关于TALT最先进成果的信息，除了每年的ActivityNet挑战赛结果报告外，人工智能指数报告还对arXiv论文进行了详细调查。更具体地说，指数报告利用了以下信息来源：

TALT 2016  
TALT 2017  
TALT 2018  
TALT 2019  
TALT 2020  
TALT 2021

## 语境中的常见对象（COCO）

关于COCO平均精度（mAP50）的数据是通过详细的arXiv文献审查，并与Papers with Code上报告的技术进展相互参照而获得的。报告的日期对应于论文首次发表在arXiv上的年份，报告的结果（mAP50）对应于每篇论文的最新版本中报告的结果。关于COCO基准详情可在COCO论文中找到。

为了突出在不使用额外训练数据的情况下COCO的进展，分数取自以下论文：

An Analysis of Scale Invariance in Object Detection – SNIP  
Deformable ConvNets v2: More Deformable, Better

## Results

Dynamic Head: Unifying Object Detection Heads with Attentions  
Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks  
Mish: A Self Regularized Non-Monotonic Activation Function  
Scaled-YOLOv4: Scaling Cross Stage Partial Network

为了突出在使用额外训练数据的情况下COCO的进展，分数取自以下论文：

EfficientDet: Scalable and Efficient Object Detection  
Grounded Language-Image Pre-Training

## 你只看一次（YOLO）

关于YOLO平均精度（mAP50）的数据是通过详细的arXiv文献回顾和对GitHub仓库的调查获得的。报告的日期对应于论文首次发表在arXiv上或方法被引入的年份。更具体地说，指数报告利用了以下信息来源：

YOLO 2016  
YOLO 2018  
YOLO 2020  
YOLO 2021

2017年和2019年的YOLO成果没有纳入指数报告中，因为在文献回顾和GitHub仓库的调查中没有发现这几年YOLO的最新改进。

## 视觉常识推理（VCR）

VCR的技术进展来自于VCR排行榜；VCR排行榜的网页进一步描述了VCR挑战背后的方法。人类在VCR上的表现取自Zellers et al. (2018)。关于VCR基准详情可以在VCR论文中找到。



## SuperGLUE

SuperGLUE的基准数据来自于 SuperGLUE 排行榜。关于SuperGLUE基准详情，可见 SuperGLUE 论文和 SuperGLUE 软件工具箱。SuperGLUE的任务和评估指标是：

名称	识别器	指标
覆盖面广的诊断法	Ax-b	Matthew's Carr
CommitmentBank	CB	Avg.F1/Accuracy
选择合理的替代方案	COPA	Accuracy
多句子阅读理解	MultiRC	F1a/EM
识别文本关联性	RTE	Accuracy
语境中的单词	WiC	Accuracy
Winograd模式挑战	WSC	Accuracy
BoolQ	BoolQ	Accuracy
用常识推理的方式进行阅读理解	ReCoRD	F1/Accuracy
Winograd模式诊断	AX-g	Gender Parity/Accuracy

## SQuAD 1.1 和 SQuAD 2.0

SQuAD 1.1的性能数据取自 [Papers with Code](#)。SQuAD 2.0的性能数据来自于 [SQuAD 2.0 排行榜](#)。关于SQuAD1.1基准的详细信息，请参见 [SQuAD 1.1 论文](#)。关于SQuAD 2.0基准的详细信息，请参见 [SQuAD 2.0 论文](#)。

## 要求逻辑推理的阅读理解数据集（Reading Comprehension Dataset Requiring Logical Reasoning, ReClor）

关于ReClor成绩的数据取自 [ReClor 排行榜](#)。有关ReClor基准的详情，请见 [ReClor 论文](#)。

## arXiv

通过详细的arXiv文献审查，并与 [Papers with Code](#) 上报告的技术进展相互参照，检索到了面向arXiv召回的gisting评估研究（ROUGE-1）数据。报告的日期对应于论文首次发表在arXiv上的年份，报告的结果（ROUGE-1）对应于每篇论文的最新版本中报告的结果。有关arXiv基准详情，请见 [arXiv 数据集网页](#)。

为了突出在不使用额外训练数据的情况下arXiv的进展，分数取自以下论文：

[A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents](#)  
[Extractive Summarization of Long Documents by Combining Global and Local Context](#)  
[Get to the Point: Summarization with Pointer-Generator Networks](#)  
[Systematically Exploring Redundancy Reduction in Summarizing Long Documents](#)  
[Sparsifying Transformer Models with Trainable Representation Pooling](#)

为了突出在使用额外训练数据的情况下arXiv的进展，分数取自以下论文：

[Big Bird: Transformers for Longer Sequences](#)  
[Hierarchical Learning for Generation with Long](#)



### Source Sequences

[PEGASUS: Pre-Training with Extracted Gap-Sentences for Abstractive Summarization](#)

### PubMed

通过详细的arXiv文献审查，并与 [Papers with Code](#) 上报告的技术进展相互参照，检索到了面向PubMed召回的gisting评估研究（ROUGE-1）的数据。报告的日期对应于论文首次发表在arXiv上的年份，报告的结果（ROUGE-1）对于每篇论文的最新版本中报告的结果。有关PubMed基准详情，请见 [PubMed](#) 论文。

为了突出在不使用额外训练数据的情况下PubMed的进展，分数取自以下论文：

[A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents](#)  
[Extractive Summarization of Long Documents by Combining Global and Local Context](#)  
[Get to the Point: Summarization with Pointer-Generator Networks](#)  
[Sparsifying Transformer Models with Trainable Representation Pooling](#)

为了突出在使用额外训练数据的情况下PubMed的进展，分数取自以下论文：

[A Divide-and-Conquer Approach to the Summarization of Long Documents](#)  
[Hierarchical Learning for Generation with Long Source Sequences](#)  
[PEGASUS: Pre-Training with Extracted Gap-Sentences for Abstractive Summarization](#)

### 斯坦福自然语言推理（SNLI）

斯坦福自然语言推理（SNLI）的准确性数据是通过详细的arXiv文献回顾，并与 [Papers with Code](#) 上报告的技术进展相互参照而获得的。报告的日期对应于论文首次发表在arXiv上的年份，而报告的结果（准确性）对于每篇论文的最新版本中报告的结果。关于SNLI基准详情可以在 [SNLI](#) 论文中找到。

为了突出SNLI的进展，分数取自以下文件：

[Compare, Compress and Propagate: Enhancing Neural Architectures with Alignment Factorization for Natural Language Inference](#)  
[Convolutional Neural Networks for Sentence Classification](#)  
[Enhanced LSTM for Natural Language Inference](#)  
[Entailment as Few-Shot Learner](#)  
[Explicit Contextual Semantics for Text Comprehension](#)  
[Semantics-Aware BERT for Language Understanding](#)  
[Self-Explaining Structures Improve NLP Models](#)

### 归纳自然语言推理（aNLI）

归纳自然语言推理（aNLI）的数据来自Allen Institute for AI的 [aNLI](#) 排行榜。关于aNLI基准详情可以在 [aNLI](#) 论文中找到。

### SemEval 2014 Task 4 Sub Task 2

关于SemEval 2014Task 4Sub Task2准确性的数据是通过详细的arXiv文献审查和 [Papers with Code](#) 上报告的技术进展相互参照而获得的。报告的日期对应于论文首次发表在arXiv上的年份，报告的结果（准确性）对于每篇论文的最新版本中报告的结果。关于SemEval基准详情可以在 [SemEval 2014](#) 论文中找到。

为了突出SemEval的进展，分数取自以下文件：

[A Multi-Task Learning Model for Chinese-Oriented Aspect Polarity Classification and Aspect Term Extraction](#)  
[Aspect Level Sentiment Classification with Deep Memory Network](#)  
[Back to Reality: Leveraging Pattern-Driven Modeling to Enable Affordable Sentiment Dependency Learning](#)  
[Effective LSTMs for Target-Dependent Sentiment Classification](#)  
[Hierarchical Attention Based Position-Aware Network for Aspect-Level Sentiment Analysis](#)



[Investigating Typed Syntactic Dependencies for Targeted Sentiment Classification Using Graph Attention Neural Network](#)  
[Recurrent Attention Network on Memory for Aspect Sentiment Analysis](#)

### WMT2014英法和英德

WMT2014英法和英德BLEU得分的数据是通过详细的arXiv文献综述，并与Papers with Code（[英法](#) 和 [英德](#)）上报告的技术进展相互参照而获得的。报告的日期对应于论文首次发表在arXiv上的年份，报告的结果（BLEU得分）对应于每篇论文的最新版本中报告的结果。关于WMT2014英法和英德基准详情可以在[WMT2014](#) 论文中找到。

为了突出在不使用额外训练数据的情况下WMT2014英法的进展，分数取自以下论文：

[Addressing the Rare Word Problem in Neural Machine Translation](#)  
[Google’s Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation](#)  
[MUSE: Parallel Multi-Scale Attention for Sequence to Sequence Learning](#)  
[R-Drop: Regularized Dropout for Neural Networks Scaling Neural Machine Translation](#)  
[Understanding the Difficulty of Training Transformers](#)  
[Weighted Transformer Network for Machine Translation](#)

为了突出在使用额外训练数据的情况下WMT2014英法的进展，分数取自以下论文：

[Understanding Back-Translation at Scale](#)  
[Very Deep Transformers for Neural Machine Translation](#)

为了突出在不使用额外训练数据的情况下WMT2014英德的进展，分数取自以下论文：

[BERT, mBERT, or BiBERT? A Study on Contextualized Embeddings for Neural Machine Translation](#)  
[Effective Approaches to Attention-based Neural Machine Translation](#)  
[Data Diversification: A Simple Strategy for Neural Machine Translation](#)  
[Fast and Simple Mixture of Softmaxes with BPE and Hybrid-LightRNN for Language Generation](#)  
[Google’s Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation](#)  
[Incorporating BERT into Neural Machine Translation](#)  
[Weighted Transformer Network for Machine Translation](#)

为了突出在使用额外训练数据的情况下WMT2014英德的进展，分数取自以下论文：

[Lessons on Parameter Sharing across Layers in Transformers](#)  
[Understanding Back-Translation at Scale](#)

### 商业上可用的MT系统的数量

关于商业上可用的MT系统数量的详细信息来自Intento的报告《2021年机器翻译的现状》。Intento是一家位于旧金山的创业公司，负责分析商业上可用的MT服务。

### LibriSpeech (Test-Clean and Other Dataset)

关于LibriSpeech (Test-Clean and Other)单词错误率的数据是通过详细的arXiv文献审查，并与Papers with Code（[Test-Clean](#) 和 [Other](#)）上报告的技术进展相互参照而获得的。报告的日期与论文首次发表在arXiv上的年份相对应，报告的结果（单词错误率）与每篇论文的最新版本中的结果相对应。关于LibriSpeech Test-Clean和Test-Other基准详情可以在[LibriSpeech](#) 论文中找到。

为了突出在不使用额外训练数据的情况下LibriSpeech



Test-Clean的进展，分数取自以下论文：

[ASAPP-ASR: Multistream CNN and Self-Attentive SRU for SOTA Speech Recognition](#)  
[Letter-Based Speech Recognition with Gated ConvNets](#)

[Neural Network Language Modeling with Letter-Based Features and Importance Sampling](#)

[SpeechStew: Simply Mix All Available Speech Recognition Data to Train One Large Neural Network](#)  
[State-of-the-Art Speech Recognition Using Multi-Stream Self-Attention With Dilated 1D Convolutions](#)

为了突出在使用额外训练数据的情况下LibriSpeech Test-Clean的进展，分数取自以下论文：

[Deep Speech 2: End-to-End Speech Recognition in English and Mandarin](#)  
[End-to-End ASR: From Supervised to Semi-Supervised Learning with Modern Architectures](#)  
[Pushing the Limits of Semi-Supervised Learning for Automatic Speech Recognition](#)

为了突出在不使用额外训练数据的情况下LibriSpeech Test-Other的进展，分数取自以下论文：

[Conformer: Convolution-Augmented Transformer for Speech Recognition](#)  
[Neural Network Language Modeling with Letter-Based Features and Importance Sampling](#)  
[SpeechStew: Simply Mix All Available Speech Recognition Data to Train One Large Neural Network](#)  
[Transformer-Based Acoustic Modeling for Hybrid Speech Recognition](#)

为了突出在使用额外训练数据的情况下LibriSpeech Test-Clean的进展，分数取自以下论文：

[Deep Speech 2: End-to-End Speech Recognition in English and Mandarin](#)  
[End-to-End ASR: From Supervised to Semi-Supervised Learning with Modern Architectures](#)

[Pushing the Limits of Semi-Supervised Learning for Automatic Speech Recognition](#)  
[W2v-BERT: Combining Contrastive Learning and Masked Language Modeling for Self-Supervised Speech Pre-Training](#)

## VoxCeleb

VoxCeleb 是一个视听数据集，由从上传至YouTube 的采访视频中提取的人类语音片段组成。VoxCeleb 包含了7000多个说话人的讲话，涵盖了广泛的种族、口音、职业和年龄--总计超过100万句话（人脸追踪是在“野外”捕获的，包括背景聊天、笑声、重叠的讲话、姿势变化和不同的照明条件），记录时间长达2000 小时（包括音频和视频）。每个片段至少有三秒长。该数据包含一个基于名人声音、短剧、电影和对话作品（如脱口秀）的音频数据集。最初的VoxCeleb 1（100,000句话，取自YouTube上的1,251位名人）被扩展到VoxCeleb 2（100万句话，来自6,112位名人）。

为了保持一致性，人工智能指数报告了最初的 VoxCeleb 数据集的分数。具体来说，指数报告利用了以下信息来源：

[The IDLAB VoxSRC-20 Submission: Large Margin Fine-Tuning and Quality-Aware Score Calibration in DNN Based Speaker Verification](#)  
[The SpeakIn System for VoxCeleb Speaker Recognition Challenge 2021](#)  
[VoxCeleb: A Large-Scale Speaker Identification Dataset](#)  
[VoxCeleb2: Deep Speaker Recognition](#)  
[VoxCeleb: Large-Scale Speaker Verification in the Wild](#)

## MovieLens 20M

关于MovieLens 20M的归一化折扣累积增益@100 (nDCG@100) 的数据是通过详细的arXiv文献审查，并与 [Papers with Code](#) 上报告的技术进展相互参照而获得的。报告的日期对应于论文首次发表在arXiv上的年份，而报告的结果 (nDCG@100) 对应于每篇论文



的最新版本中报告的结果。关于MovieLens系列基准详情可以在 [Harper et al. 2015](#) 中找到。

为了突出MovieLens 20M的进展，分数取自以下论文：  
[Deep Variational Autoencoder with Shallow Parallel Path for Top-N Recommendation \(VASP\)](#)  
[Enhancing VAEs for Collaborative Filtering: Flexible Priors & Gating Mechanisms](#)  
[RaCT: Toward Amortized Ranking-Critical Training for Collaborative Filtering](#)  
[Variational Autoencoders for Collaborative Filtering](#)

## Criteo

Criteo的曲线下面积得分 (AUC) 数据是通过详细的arXiv文献审查，并与 [Papers with Code](#) 上报告的技术进展相互参照而获得的。报告的日期对应于论文首次发表在arXiv上的年份，而报告的结果 (AUC) 对应于每篇论文的最新版本中报告的结果。关于Criteo基准详情，可以在 [Criteo Kaggle 挑战赛](#)页面上找到。

为了突出Criteo的进展，分数取自以下文件：  
[AutoInt: Automatic Feature Interaction Learning via Self-Attentive Neural Networks](#)  
[DeepFM: A Factorization-Machine Based Neural Network for CTR Prediction](#)  
[DeepLight: Deep Lightweight Feature Interactions for Accelerating CTR Predictions in Ad Serving](#)  
[FAT-DeepFFM: Field Attentive Deep Field-aware Factorization Machine](#)  
[MaskNet: Introducing Feature-Wise Multiplication to CTR Ranking Models by Instance-Guided Mask](#)  
[Product-Based Neural Networks for User Response Prediction](#)

## Arcade Learning Environment: Atari-57

Arcade Learning Environment: Atari-57的人类归一化平均分数是通过详细的arXiv文献审查，并与 [Papers with Code](#) 上报告的技术进展相互参照而获得的。报告的日期与论文首次发表在arXiv上的年份相对应，报告的结果（人类归一化平均分数）与每篇论文的最

新版本中报告的结果相对应。关于Arcade Learning Environment: Atari-57基准详情可以在 [Arcade Learning Environment](#) 论文中找到。

为了突出Arcade Learning Environment:Atari-57的进展，分数取自以下论文：  
[Dueling Network Architectures for Deep Reinforcement Learning](#)  
[Distributional Reinforcement Learning with Quantile Regression](#)  
[GDI: Rethinking What Makes Reinforcement Learning Different From Supervised Learning](#)  
[Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model](#)  
[Recurrent Experience Replay in Distributed Reinforcement Learning](#)

## Progen

关于Progen平均归一化分数的数据是通过详细的arXiv文献审查获得的。报告的日期对应于论文首次发表在arXiv上的年份，报告的结果（平均归一化分数）对应于每篇论文的最新版本中报告的结果。关于 Progen 基准详情可以在 [Progen](#) 论文中找到。

为了突出Progen的进展，分数取自以下论文：  
[Automatic Data Augmentation for Generalization in Reinforcement Learning](#)  
[Leveraging Procedural Generation to Benchmark Reinforcement Learning](#)  
[Procedural Generalization by Planning with Self-Supervised World Models](#)

## 国际象棋

关于国际象棋软件引擎性能的数据取自瑞典国际象棋计算机协会的顶级国际象棋软件引擎排名。瑞典国际象棋计算机协会对计算机国际象棋软件系统进行了相互测试，并发布了一份表现最好的系统排名表。瑞典国际象棋计算机协会制作的排名表是对国际象棋引擎性能的统计学意义上的衡量，因为引擎是在数千场类似锦标赛的比赛中相互竞争的，并且每个引擎都采用



相同的基础硬件。Magnus Carlsen的最高ELO分数的数据来自国际象棋联合会。

## 训练时间和加速器数量

关于人工智能系统的训练时间和加速器数量的数据来自MLPerf训练基准竞赛。更具体地说，人工智能指数报告利用了以下MLPerf训练比赛的数据：

[MLPerf Training v0.5, 2018](#)

[MLPerf Training v0.6, 2019](#)

[MLPerf Training v0.7, 2020](#)

[MLPerf Training v1.0, 2021](#)

[MLPerf Training v1.1, 2021](#)

关于MLPerf训练基准详情可以在[训练基准论文](#)中找到。关于当前基准类别详情以及提交和竞赛分项的技术信息，可以在[MLPerf Training 网页](#)上找到。

## ImageNet 训练成本

关于ImageNet训练成本的数据是基于DAWNench的研究和Deepak Narayanan的个人研究。DAWNBench是一个用于端到端深度学习训练和推理的基准套件。DAWNBench提供了一个常见的深度学习工作负载的参考集，用于量化不同的优化策略、模型架构、软件框架、云和硬件的训练时间、训练成本、推理延迟和推理成本。更多细节请见[DAWNBench](#).

由于DAWNbench在2020年3月后被废弃，最新一轮MLPerf提交的训练成本数据是由Deepak Narayanan手动收集的。



## 人工智能指数机器人学调查

该调查从2021年12月至2022年2月分三波向67所大学的509名机器人专业的教授在线发放。大学的选择是基于2021年世界大学排名来完成的，在全球范围内具有地理代表性。原始数据请见此文件夹。来自43所大学的101名教授完成了调查，其中包括：

奥尔堡大学, 丹麦	南非斯泰伦博斯大学
埃及艾因沙姆斯大学	瑞士洛桑联邦理工学院
美国卡内基梅隆大学	日本东京工业大学
美国哥伦比亚大学	加拿大不列颠哥伦比亚大学
美国康奈尔大学	美国加州大学伯克利分校
荷兰代尔夫特科技大学	美国加州大学洛杉矶分校
瑞士苏黎世联邦理工学院	美国加州大学圣地亚哥分校
香港科学与技术大学, 香港	英国剑桥大学
韩国科学与技术高级研究所, 韩国	南非开普敦大学
比利时鲁汶大学	英国伦敦大学学院
美国马萨诸塞理工学院	香港大学, 香港
新加坡南洋理工大学	美国伊利诺伊大学厄巴纳-香槟分校
墨西哥国家理工学院	马来亚大学, 马来西亚
新加坡国立大学, 新加坡	英国曼彻斯特大学
中国北京大学	美国密歇根大学
意大利米兰理工大学	西班牙加泰罗尼亚政治大学
智利, 智利天主教大学	美国德克萨斯大学奥斯汀分校
美国普林斯顿大学	日本东京大学
美国普渡大学	加拿大多伦多大学
德国亚琛工业大学	加拿大滑铁卢大学
韩国首尔国立大学	中国浙江大学
美国斯坦福大学	



## 章节三 技术AI伦理

### FACCT和NEURIPS的AI伦理学趋势

为了了解ACM公平、问责制和透明度会议的趋势，本节追踪了2018年至2021年在会议记录中发表的FAccT论文。我们将作者的隶属关系分为学术界、产业界、非营利机构、政府部门和独立类别，同时也跟踪其隶属机构的位置。有多个隶属机构的作者在每个类别（学术界和产业界）中只计算一次，但同一类型的多个隶属机构（即属于两个学术机构的作者）在该类别中计算一次。

在对NeurIPS出版物进行分析时，我们确定了以现实世界的影响为主题的研讨会，并在“医疗保健”、“气候”、“金融”、“发展中世界”或“其他”中为论文贴上一个单一的主要类别标签，其中“其他”表示与现实世界的使用案例有关，但不属于其他类别的论文。

我们统计了每个类别的论文数量，得出了图3.3.3中的数字。我们不会在多个类别中重复计算论文。我们注意到，这一计算对于2018年之前的数据可能并不准确，因为NeurIPS的社会影响方面的工作历史上一直被归入“人工智能对社会影响（AI for social impact）”下<sup>7</sup>，但最近已经被分割成更细化的研究领域，例如，专门为健康<sup>8</sup>、气候<sup>9</sup>、政策和治理<sup>10</sup>、灾难应对<sup>11</sup>和发展中世界的机器学习而举办的研讨会。<sup>12</sup>

为了跟踪图3.3.4-3.3.7中NeurIPS特定技术主题的趋势，我们统计了NeurIPS主赛道接收的标题中含有关键词的论文数量（例如，用于跟踪与因果效应有关的论文的关键词：“反事实”或“因果”），以及提交给相关研讨会的论文。这里给出了考虑分析的研讨会名单。

### 公平性和偏见指标的元分析

在对人工智能中的公平性和偏见指标进行分析时，我们确定并报告了在学术界被持续引用的基准和诊断指标，这些指标被报告在公共排行榜上，或者被报告在公开可用的基线模型上（例如，GPT-3、BERT、ALBERT）。我们注意到，研究论文的引用是一个滞后的指标，最近被采用的指标可能不会反映在2021年的数据中。

对于图3.1.1和3.1.2，我们跟踪来自以下论文和项目的指标：

[Aligning AI with Shared Human Values](#)  
[Assessing Social and Intersectional Biases in Contextualized Word Representations](#)  
[Bias in Bios: A Case Study of Semantic Representation](#)  
[Bias in a High-Stakes Setting](#)  
[BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation](#)  
[Certifying and Removing Disparate Impact](#)  
[CivilComments: Jigsaw Unintended Bias in Toxicity Classification](#)  
[CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models](#)  
[Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-Like Biases](#)  
[Equality of Opportunity in Supervised Learning](#)  
[Evaluating Gender Bias in Machine Translation](#)  
[Evaluating Gender Bias in Natural Language Inference](#)

7 见2018年人工智能的伦理、社会和治理问题研讨会，2018年人工智能促进社会公益研讨会，2019年人工智能促进社会公益联合研讨会，2020年抵制人工智能研讨会，2020年浏览人工智能研究的更广泛影响研讨会。

8 见2014年《用于临床数据分析、健康和基因组学的机器学习》，2015年《用于健康的机器学习》，2016年《用于健康的机器学习》，2017年《用于健康的机器学习》。

9 见2013年机器学习促进可持续发展，2020年人工智能促进地球科学，2019年，2020年，2021年用机器学习解决气候变化问题。

10 见2016年人与机器，2019年人工智能促进社会公益-公共政策联合研讨会，2021年以人为本的人工智能。

11 见2019年人工智能用于人道主义援助和灾害应对，2020年第二次人工智能用于人道主义援助和灾害应对研讨会，2021年第三次人工智能用于人道主义援助和灾害应对研讨会。

12 见2017-2021年发展中世界的机器学习研讨会。



[Examining Gender Bias in Languages with Grammatical Gender](#)  
[Fairness Through Awareness](#)  
[Gender Bias in Coreference Resolution](#)  
[Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods](#)  
[Gender Bias in Multilingual Embeddings and Cross-Lingual Transfer](#)  
[Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification](#)  
[Image Representations Learned with Unsupervised Pretraining Contain Human-Like Biases](#)  
[Measuring and Reducing Gendered Correlations in Pretrained Models](#)  
[Measuring Bias in Contextualized Word Representations](#)  
[Measuring Bias with Wasserstein Distance](#)  
[Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification](#)  
[On Formalizing Fairness in Prediction with Machine Learning](#)  
[On Measuring Social Biases in Sentence Encoders Perspective API](#)  
[RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models](#)  
[Scaling Language Models: Methods, Analysis & Insights from Training Gopher](#)  
[Semantics Derived Automatically from Language Corpora Contain Human-Like Biases](#)  
[StereоСet: Measuring Stereotypical Bias in Pretrained Language Models](#)  
[The Woman Worked as a Babysitter: On Biases in Language Generation](#)  
[When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness](#)

## 自然语言处理的衡量标准

在第3.2节中，我们追踪了Google的Jigsaw所创建的Perspective API的引用情况。Perspective API已经被自然语言处理领域的研究人员和工程师广泛采用。它的创造者将毒性定义为“粗鲁的、不尊重的或不合理的评论，有可能使人离开讨论。”该工具由在维基百科和新闻网站评论的专有数据集上训练的机器学习模型驱动。我们的分析中包括以下论文：

[#ContextMatters: Advantages and Limitations of Using Machine Learning to Support Women in Politics](#)  
[A General Language Assistant as a Laboratory for Alignment](#)  
[A Machine Learning Approach to Comment Toxicity Classification](#)  
[A Novel Preprocessing Technique for Toxic Comment Classification](#)  
[Adversarial Text Generation for Google's Perspective API](#)  
[Avoiding Unintended Bias in Toxicity Classification with Neural Networks](#)  
[Bad Characters: Imperceptible NLP Attacks](#)  
[Challenges in Detoxifying Language Models](#)  
[Classification of Online Toxic Comments Using Machine Learning Algorithms](#)  
[Context Aware Text Classification and Recommendation Model for Toxic Comments Using Logistic Regression](#)  
[Detecting Cross-Geographic Biases in Toxicity Modeling on Social Media](#)  
[Detoxifying Language Models Risks Marginalizing Minority Voices](#)  
[Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online](#)  
[HATEMOJI: A Test Suite and Adversarially Generated Dataset for Benchmarking and Detecting Emoji-Based Hate](#)  
[HotFlip: White-Box Adversarial Examples for Text](#)



## Classification

- Identifying Latent Toxic Features on YouTube Using Non-Negative Matrix Factorization
- Interpreting Social Respect: A Normative Lens for ML Models
- Knowledge-Based Neural Framework for Sexism Detection and Classification
- Large Pretrained Language Models Contain Human-Like Biases of What Is Right and Wrong to Do
- Leveraging Multilingual Transformers for Hate Speech Detection
- Limitations of Pinned AUC for Measuring Unintended Bias
- Machine Learning Suites for Online Toxicity Detection
- Mitigating Harm in Language Models with Conditional-Likelihood Filtration
- On-the-Fly Controlled Text Generation with Experts and Anti-Experts
- Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets
- Racial Bias in Hate Speech and Abusive Language Detection Datasets
- RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models
- Scaling Language Models: Methods, Analysis & Insights from Training Gopher
- Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP
- Social Bias Frames: Reasoning About Social and Power Implications of Language
- Social Biases in NLP Models as Barriers for Persons with Disabilities
- Stereotypical Bias Removal for Hate Speech Detection Task Using Knowledge-Based Generalizations
- The Risk of Racial Bias in Hate Speech Detection Towards Measuring Adversarial Twitter Interactions

## Against Candidates in the US Midterm Elections

- Toxic Comment Classification Using Hybrid Deep Learning Model
- Toxicity-Associated News Classification: The Impact of Metadata and Content Features
- Understanding BERT Performance in Propaganda Analysis
- White-to-Black: Efficient Distillation of Black-Box Adversarial Attacks
- Women, Politics and Twitter: Using Machine Learning to Change the Discourse

虽然Perspective API在机器学习研究中被广泛使用，也被用于测量在线毒性，但用于训练Perspective基础模型的特定领域（如新闻、维基百科）的毒性可能并不能广泛代表所有形式的毒性（如trolling）。其他已知的注意事项还包括对少数人所写的文本存在偏见：事实证明，Perspective API会对含有少数人身份的文字不成比例地赋予高毒性分数（如“我是一个同性恋者”）。因此，用来自Perspective API的标签建立的解毒技术会导致模型对少数人群体使用的语言建模能力较差，而且它们会避免提及少数人群体的身份。

我们注意到，文字嵌入关联测试（WEAT）部分报告的效果大小指标对稀有词高度敏感，因为已经证明，删除语料库中不到1%的相关文档会显著影响WEAT的效果。虽然我们报告的是测量沿性别和种族轴的偏见的嵌入关联任务的一个子集，但这些嵌入关联测试已经被扩展到量化跨交叉轴的效果（例如，欧洲裔美国人+男性，非洲裔美国人+男性，非洲裔美国人+女性）。

在对100多年的美国人口普查数据的嵌入分析中，通过计算平均嵌入距离的差异来衡量嵌入偏差。例如，性别偏见的计算方法是：与女性相关的词（如她、女性）的嵌入和与职业相关的词（如教师、律师）的嵌入相比的平均距离，减去与男性相关的词所计算的相同平均距离。



## 实事求是和真实性

### 定义

事实性、事实正确性、事实准确性和真实性等概念都是指符合事实或真相。最近人工智能方面的工作旨在评估语言模型内的事实正确性，并描述其局限性。

虽然人类的真实性是一个比较好理解的概念，但在人工智能的背景下，真实性并不是一个很好描述的概念。研究人员提出了系统真实性的框架--例如，一个广泛的真实性系统应该避免撒谎或使用真实的陈述来误导或诱导；在对话中应该是清晰的、信息丰富的、（大部分）合作的；并且应该具备良好的校准性、自我意识和开放对知识的限制。狭义的真实性定义可能只是指那些避免陈述错误的系统。TruthfulQA 的作者将系统定义为：它避免断言虚假的陈述；拒绝回答问题，表示不确定，或给出一个真实但不相关的答案，可能被认为是真实的，但不是有信息性的。

真实性与一致性有关：一个真实的系统是与人类的价值观和目标相一致的。根据 alignment 的定义，一个 alignment 的系统是一个有帮助的、诚实的、无害的系统。由于我们还不能衡量一个系统内的诚实度，所以真实性可以作为一种代理（proxy）。

一个诚实的系统是一个只会断言自己“相信”的系统，或者一个与自己的信念从不矛盾的系统。一个系统可以是诚实的，但不是真实的--例如，如果一个诚实的系统相信疫苗是不安全的，它就会诚实地宣称这一点，尽管这个声明在事实上是错误的。相反，一个系统可以是真实的，但不是诚实的：它可能认为疫苗是不安全的，但为了通过测试而断言它们是安全的。另一项工作提出，一个诚实的系统应该提供准确的信息，不误导用户，是经过校准的（例如，当它声称有80%的信心时，它应该有80%的正确率），并表达适当的不确定性水平。

Hallucination 是指语言模型捏造事实正确的支持性证据或输入文件中不存在的语句。在封闭式对话、总结或问答应用中，通常认为一个产生 Hallucination 的系统是不真实的。

### 训练数据中的语言多样性

训练数据中不平衡的语言分布会影响通用语言模型的性能。例如，Gopher 系列模型是在 MassiveText (10.5TB) 上训练的，这是一个99%为英语的数据集。同样地，GPT-3的训练数据中只有7.4%是非英语语言。相比之下，XGLM 是 Meta AI 最近推出的一个模型系列，在30种语言的训练数据上进行训练，并对低资源语言进行上采样，以构建一个更平衡的语言表征。参见 XGLM 论文中的图1，该图比较了 XGLM 和 GPT-3 的语言分布。

此外，XGLM 论文的图7通过比较 XGLM（一个多语言的语言模型）和 GPT-3（一个单语言模型）的性能，强调了语言模型能够有效存储事实知识的程度。使用 mLM 数据集对知识三元组的完成情况进行了评估，该数据集是用谷歌翻译从英语基准 LAMA 翻译过来的。GPT-3 在英语中表现优异，但 XGLM 在非英语语言中表现优异。进一步的结果显示，更多样化的语言表征可以提高语言模型在翻译等任务中的性能。

2021年，美国国会询问了社交媒体公司在非英语语言方面的内容审核方法，并强调了无论何种语言都要能平等获得真实和值得信赖的信息的重要性。在这些公司开始在其事实核查和内容审核过程中采用世界各地的语言模式的过程中，关键是要能够评估使用对非英语语言表现不佳的模型所产生的不成比例的负面影响。



# 章节四 经济和教育

## EMSI BURNING GLASS

编写人: Julia Nitschke, Summer Jasinski, Bledi Taskaand Rucha Vankudre

Emsi Burning Glass提供就业市场分析，帮助雇主、工人和教育工作者做出数据驱动的决策。该公司的人工智能技术分析了数以亿计的招聘信息和现实生活中的职业变化，以提供对劳动力市场模式的分析。这种实时战略情报提供了有效的关键信息，如什么工作最需要，雇主需要哪些具体技能，以及为劳动者提供最大潜力的职业方向。欲了解更多信息，请访问 [burning-glass.com](http://burning-glass.com)。

### 招聘信息数据

为了支持这些分析，Emsi Burning Glass挖掘了自2010年以来收集的数以百万计的招聘信息的数据集。Emsi Burning Glass收集了超过45000个在线招聘网站的招聘信息，以生成全面、实时的劳动力市场需求画像。它汇总了招聘信息，删除了重复信息，并从招聘信息文本中提取数据，包括关于职位名称、雇主、行业和地区的信息，以及所需的经验、教育和技能。

招聘信息对了解劳动力市场的趋势很有用，因为它们可以详细、实时地了解雇主计划招聘员工所需要的技能。为了评估招聘信息数据的代表性，Emsi Burning Glass进行了一些分析，将招聘信息的分布与美国官方政府和其他第三方来源的分布进行了比较。

美国政府关于职位发布的主要数据来源是劳工统计局进行的职位空缺和劳工流动调查（JOLTS）项目。根据JOLTS和Emsi Burning Glass之间的比较，Emsi Burning Glass数据所捕捉到的劳动力市场需求占总劳

动力需求的95%以上。没有在网上发布的工作通常是在由小企业发布（典型的例子是餐馆窗口的“招聘”标志）和或在工会的招聘大厅中发布。

### 衡量对人工智能的需求

为了衡量雇主对人工智能技能的需求，Emsi Burning Glass使用其超过17,000种技能的技能分类法。以下是Emsi Burning Glass数据中的人工智能技能清单，以及相关的技能集群。虽然有些技能被认为是专门属于人工智能集群的，在本报告中，以下所有技能都被视为人工智能技能。如果一个招聘启事要求具备这些技能中的一项或多项，则被认为是人工智能工作。

**人工智能:** 专家系统, IBM Watson, IPSoft Amelia, Ithink, 虚拟代理, 自主系统, 激光雷达, OpenCV, 路径规划, 遥感

**自然语言处理 (NLP) :** ANTLR, 自动语音识别 (ASR), 聊天机器人, 计算语言学, Distinguo, Latent Dirichlet Allocation, 潜在语义分析, Lexalytics, 词汇获取, 词汇语义, 机器翻译 (MT)。模块化音频识别框架 (MARF), MoSes, 自然语言处理, 自然语言工具包 (NLTK), 最近邻算法, OpenNLP, 情感分析/观点挖掘, 语音识别, 文本挖掘, 文本到语音 (TTS), 标记化, Word2Vec

**神经网络:** Caffe深度学习框架, 卷积神经网络 (CNN), 深度学习, Deep learning4j, Keras, 长短期记忆 (LSTM), MXNet, 神经网络, Pybrain, 循环神经网络 (RNN), TensorFlow

**机器学习:** AdaBoost算法, Boosting (Machine Learning), Chi Square自动交互检测 (CHAID), 分类算法, 聚类算法, 决策树, 降维, 谷歌云机器学习平台, 梯度



提升, H2O(软件), Libsvm, 机器学习, Madlib, Mahout, 微软认知工具包, MLPACK (C++库), Mlpy, 随机森林, 推荐系统, Scikit-learn, 半监督学习, 监督学习(机器学习), 支持向量机(SVM), 语义驱动减法聚类法(SDSCM), Torch (机器学习), 无监督学习, Vowpal, Xgboost

**机器人学:** Blue Prism, 机电系统, 运动规划, Motoman机器人编程, 机器人框架, 机器人系统, 机器人操作系统 (ROS) , 机器人编程, 伺服驱动器/电机, 同步定位和绘图 (SLAM) 。

**视觉图像识别:** 计算机视觉、图像处理、图像识别、机器视觉、物体识别

## LINKEDIN

编写人: Akash Kaura 和 Murat Erer

### 国家样本

纳入的本报告分析的国家是一个符合条件的国家样本, 这些国家的LinkedIn劳动力覆盖率至少为40%, 并且在任何特定月份至少有10个人工智能招聘需求。中国和印度尽管没有达到40%的覆盖率门槛, 但由于它们在全球经济中越来越重要, 也被包括在内。对中国和印度的分析可能不像其他国家那样全面, 因此应作相应的解释。

### 技能 (和AI技能)

LinkedIn会员在他们的LinkedIn档案中自我报告他们的技能。目前, LinkedIn确定了超过38,000种不同的、标准化的技能。这些技能已经被LinkedIn的分类学家编码并分类为249个技能组, 这就是数据集中所代表的技能组。构成人工智能技能组的顶级技能是机器学习、自然语言处理、数据结构、人工智能、计算机视觉、图像处理、深度学习、TensorFlow、Pandas (软件) 和OpenCV等等。

技能分组是由专家级分类学家通过相似度指数方法得出的, 以衡量行业层面的技能构成。行业是根据ISIC 4行业分类法来划分的 (Zhu et al., 2018) 。

### 技能基因组

对于任何实体 (职业或工作、国家、部门等), 技能基因组是该实体的50种 "最具特色的技能 "的有序列表 (矢量)。这些最具特色的技能是通过TF-IDF算法来确定的目标实体最具代表性的技能, 同时对那些对该特定实体没有增加多少信息的泛在技能 (如Microsoft Word) 进行降级。

TF-IDF是一种统计措施, 评估一个词 (在这里是指技能) 对一个选定的实体的代表性。这是由两个指标相乘得到的:

1. 一项技能在一个实体中的词频 (TF) 。
  2. 该技能在一组实体中的对数反实体频率 (IDF) 。
- 这表明一个词在整个实体集合中的常见或罕见程度。IDF越接近于0, 一个词就越常见。

因此, 如果该技能在LinkedIn实体中非常常见, 并出现在许多工作或会员描述中, IDF将接近于0。另一方面, 如果该技能是特定实体所独有的, IDF将接近于1。详情请见LinkedIn 的技能基因组和LinkedIn-世界银行方法学说明。

### AI技能普及率

本指标的目的是通过以下方法衡量一个实体 (在特定的国家、行业、性别等) 的人工智能技能的程度。

- 计算2015-2021年LinkedIn会员在特定实体 (职业、行业等) 中所有自我添加技能的频率。
- 使用TF-IDF模型对技能频率进行重新加权, 得到该实体中最具代表性的前50项技能。这50项技能构成了该实体的 "技能基因组"。
- 计算属于人工智能技能组的技能在所选实体的顶级技能中的份额。



解释：人工智能技能普及率表明人工智能技能在各职业中的普及程度，或LinkedIn会员在工作中使用人工智能技能的程度。例如，工程师职业的前50项技能是根据它们在LinkedIn会员资料中出现的加权频率计算的。如果工程师拥有的技能中有四项属于人工智能技能组，则表明，人工智能技能在工程师中的普及率估计为8%（例如，4/50）。

## 工作或职业

LinkedIn会员的职业title是标准化的，被归入大约15,000个职业。这些职业并不针对特定的部门或国家。这些职业被进一步标准化为大约3600个职业代表（Occupation representatives）。职业代表在不考虑资历的情况下将具有共同作用和专长的职业进行分组。

## 人工智能工作或职业

一个人工智能工作（技术上是职业代表）是一个需要人工智能技能来完成工作的职业代表。技能普及率被用来作为一个信号，表明人工智能技能是否在一个职业代表中普遍存在（在该职业代表可能存在的任何部门中）。这种职业的例子包括（但不限于）：机器学习工程师、人工智能专家、数据科学家、计算机视觉工程师等。

## 人工智能人才

如果一个LinkedIn会员在他们的个人资料中明确添加了人工智能技能和/或他们在人工智能职业代表中的任职，则认为他们是人工智能人才。使用人工智能人才的数量计算人才集中度指标（例如，为了计算国家层面的人工智能人才集中度，我们使用国家层面的人工智能人才的数量与各自国家的LinkedIn会员的数量相比较）。

## 相对人工智能技能普及率

为了对各国的技能普及率进行比较，要对技能基因组进行计算，并选择一个相关的基准（如全球平均水平）。然后在控制职业的情况下，在一个国家和基准的人工智能技能普及率之间构建一个比率。

解释。一个国家的相对人工智能技能普及率为1.5，表明人工智能技能在一组重叠的职业中的频率是基准的1.5倍。

## 全球比较

为了进行跨国比较，我们提出了人工智能技能的相对普及率，衡量标准是每个人工智能技能在特定国家职业中的普及率之和，除以人工智能技能在国家样本中重叠职业的全球平均普及率。

解释：相对普及率为2意味着该国人工智能技能的平均普及率是全球同一组职业平均水平的2倍。

## 全球比较：按行业分类

按国家划分的行业的相对人工智能技能普及率提供了人工智能技能普及率在各行业和样本国家的深度部门分解。

解释：一个国家在教育部门的相对人工智能技能普及率为2，意味着该国人工智能技能的平均普及率是该部门同一组职业的全球平均值的2倍。



## 相对人工智能招聘指数

- LinkedIn招聘率或整体招聘率是一个按LinkedIn会员数标准化的招聘措施。它的计算方法是，在工作开始的同一时期，增加新雇主的LinkedIn会员的百分比，除以相应地点的LinkedIn会员总数。
- 人工智能雇用率是按照整体雇用率方法计算的，但只考虑归类为人工智能人才的成员。
- 相对人工智能招聘指数是将人工智能招聘率的变化速度与总体招聘率的变化速度进行归一化，提供了一个市场上人工智能人才招聘的增长速度是高于、等于还是低于总体招聘的情况展示。当人工智能招聘和整体招聘以相同的速度逐年增长时，相对人工智能招聘指数等于1.0。

解释：相对人工智能招聘指数显示每个国家在人工智能人才招聘方面的增长速度相对于该国整体招聘的增长速度。比率为1.2意味着人工智能人才招聘的增长超过了整体招聘增长的20%。

## NETBASE QUID

编写人：Julie Kim和Tejas Sirohi

NetBase Quid提供以人工智能为动力的消费者和市场情报，以便在一个噪声环境和不可预测的世界中实现业务重塑。该软件应用人工智能来揭示大型非结构化数据集的模式，并生成可视化效果，使用户能够准确、快速、有效地做出由数据驱动的智能决策。NetBase Quid使用布尔查询来搜索社会、新闻、论坛和博客、公司和专利数据源以及其他自定义数据集中的重点领域、主题和关键词。然后，NetBase Quid根据语义相似度将这些数据点可视化。

## 搜索、数据来源和范围

NetBase Quid将来自多个数据源的600多万份全球公共和私营公司的资料编入索引，以便在公司描述中进行搜索，同时过滤并包括从投资信息到公司信息的元数据，如成立年份、总部所在地等。NetBase Quid每周都会更新公司信息。NetBase Quid算法从每个文件中读取大量的文本数据，根据相似语言在不同的文件之间建立联系。这个过程以一个巨大的规模重复进行，从而产生一个具有不同集群的网络（包含不同的主题或重点领域）。我们可以根据NetBase Quid识别的关键词、短语、人物、公司、机构以及其他输入软件的元数据来确定趋势。

## 数据

组织数据来自Capital IQ和Crunchbase。这些公司包括世界各地的所有类型的公司（私营、上市、经营、作为子公司经营、停业）。投资数据包括私人投资、并购、公开募股、PE/VCs、企业风险投资部门、政府和美国境内外机构的少数股权。有些数据根本无法获取--例如，当投资者或投资者的资金数额未被披露时。NetBase Quid还嵌入了公司的地理信息，如成立年份和总部所在地。

NetBase Quid默认嵌入了Capital IQ的数据，并为Capital IQ未收录的数据添加了Crunchbase的数据。这不仅生成了所有全球组织的全面和准确的数据，还捕捉到了早期的初创企业和融资事件数据。公司信息每周都会上传。

## 搜索参数

NetBase Quid使用布尔查询在存档的公司数据库中，在其业务描述和网站中搜索重点领域、主题和关键词。我们可以通过总部地区、投资金额、运营状况、



组织类型（私营/公营）和成立年份来过滤搜索结果。然后，NetBase Quid根据语义相似度将这些公司可视化。如果搜索结果中有超过7000家公司，NetBase Quid将根据语言算法选择7000家最相关的公司进行可视化。

布尔搜索：“人工智能”或“AI”或“机器学习”或“深度学习”

公司：

- 图4.2.1：从2012年1月1日到2021年12月31日，全球人工智能和机器学习公司的投资（私人、IPO、并购）。
- 图4.2.2-4.2.12：过去10年（2012年1月1日至2021年12月31日）投资超过150万美元的全球人工智能和机器学习公司--通过Quid的相关性算法，在7500家公司中选出7000家公司。

## 目标事件的定义

- 私人投资：私募是指公司向选定的投资者或选定的投资者群体私下出售新发行的证券（股权或债务）。买方在私募中获得的股权通常是少数股权（低于50%），尽管也有可能通过私募获得公司的控制权，而在这种情况下，私募将成为多数股权投资。
- 少数派投资：这些指的是在Quid的少数股权收购，当买方收购实体、资产产品和业务部门的现有所有股权少于50%时，就会发生这种情况。
- 并购：这是指买方收购实体、资产产品和业务部门现有所有权的50%以上的股权。

## 计算研究协会（CRA TAULBEE调查）

编写人：Betsy Bizot (CRA高级研究助理)

### 资料来源

计算研究协会(CRA)的成员是活跃在计算研究领域的200多个北美地区的组织：计算机科学和计算机工程的学术部门；产业、政府和学术界的实验室和中心；以及附属的专业协会（AAAI、ACM、CACS/AIC、IEEE计算机协会、SIAM USENIX）。CRA的使命是通过与产业界、政府部门和学术界联合，加强计算机领域的研究和高等教育，从而提高创新能力。[在这里了解更多关于CRA的信息。](#)

### 调查方法

CRA Taulbee调查在每个学年的秋季通过联系200多个博士学位授予部门来收集调查数据。关于Taulbee调查的细节可以[在这里找到](#)。Taulbee并不直接调查学生。该部门确定了每个应届博士的专业领域，以及他们的就业类型。每学年的9月至1月收集上一学年授予学位的博士的数据。结果在数据收集结束后的5月公布。因此，2020年的数据是在前一年春天得到的，而为2021年提供的数字是在2020年5月得到的。

CRA Taulbee调查只发给计算机科学、计算机工程和信息科学/系统的博士系。从历史上看，(a) Taulbee涵盖了美国1/4到1/3的计算机科学学士学位获得者；(b) 在Taulbee学校获得学士学位的女性比例低于整体水平；(c) Taulbee跟踪整体计算机科学生产的趋势。



## 差异

- 在博士就业市场趋势中特别值得关注的是关于人工智能博士专业领域的指标。专业领域的分类在2008年发生变化，在2016年澄清。从2004-2007年，人工智能和机器人被归为一类；从2008年至今，人工智能都是独立的专业领域；2016年向受访者澄清了人工智能是包括机器学习的。
- 关于新的终身职位聘用趋势的说明（总体上，特别是在AAU学校）。在2018年的Taulbee调查中，我们首次询问了有多少新员工来自以下来源：应届博士、博士后、产业界和其他学术机构。结果显示，29%的新晋助理教授来自于其他学术机构。
- 有些人可能是教学或研究人员，而不是终身教职，但可能有一些机构之间的流动，这意味着雇用的总人数夸大了实际的新员工总数。



# 章节五 人工智能政策和国家战略

## BLOOMBERG GOVERNMENT

编写人：Amanda-Allen

BLOOMBERG GOVERNMENT是一项基于订阅的高级服务，为与政府沟通或受政府影响的专业人士提供全面的信息和分析。BLOOMBERG GOVERNMENT的数字工作空间提供新闻、分析和数据驱动的决策工具，为政府事务和签约专业人士提供智能优势。欲了解更多信息或演示，请访问 [about.bgov.com](http://about.bgov.com)。

**立法文件：**BLOOMBERG GOVERNMENT维护了一个国会文件库，包括法案、国会预算办公室的评估，以及由国会委员会、国会研究处和其他办公室发布的报告。BLOOMBERG GOVERNMENT还收录了各州的立法议案。对于 "人工智能政策和治理"一节，BLOOMBERG GOVERNMENT的分析师确定了所有立法、国会委员会报告和CRS报告，这些报告均提及了一个或多个人工智能的关键词。

## 方法论

**合同支出：**BLOOMBERG GOVERNMENT的合同智能工具（CIT）从 [www.fpds.gov](http://www.fpds.gov) 获取合同数据。CIT包括一个政府在人工智能相关合同上的支出模型，该模型是基于政府定义的产品服务代码和100多个人工智能相关关键词的组合。对于 "美国政府合同支出" 部分，BLOOMBERG GOVERNMENT的分析师使用了从2000财年到2021财年的合同支出数据。

**国防部RDT&E预算：**BLOOMBERG GOVERNMENT整理了国防部审计长提供的所有RDT&E预算申请项目。对于 "美国国防部（DOD）预算" 部分，BLOOMBERG GOVERNMENT使用一组人工智能特定的关键词来识别与人工智能和机器学习相关的500个专门的预算活动，在2021财年的价值合计59亿美元。



## 关于人工智能的全球立法记录

人工智能指数在25个国家的国会或议会的网站上用各自的语言对关键词 "人工智能" 进行了法案全文的搜索，以找到通过并最终称为法律的人工智能相关法案。请注意，仅包括2015年至2021年由国家级立法机构通过并签署成为法律（即由总统或获得皇家同意）的法律文件。未来的人工智能指数报告希望能够引入对其他类型的法律文件的分析，例如由国家或超国家层面的立法机构、政府机构等通过的条例和标准。

### 澳大利亚

网址: [www.legislation.gov.au](http://www.legislation.gov.au)

关键词: artificial Intelligence

筛选:

- 立法类型: 法案
- 文件夹: 众议院部, 参议院部

注: 解释性备忘录中的文本不计入。

### 比利时

网址: <http://www.ejustice.just.fgov.be/loi/loi.htm>

关键词: intelligence artificielle

### 巴西

网址: <https://www.camara.leg.br/legislacao>

关键词: inteligência artificial

筛选:

- 联邦立法
- 类型: 法律

### 加拿大

网址: <https://www.parl.ca/legisinfo/>

关键词: artificial Intelligence

注: 对结果进行了进一步分析，以确定所提出的法案中有多少最终获得通过（即获得皇家同意），并记录了法案的状态。

### 中国

网址: <https://flk.npc.gov.cn/>

关键词: 人工智能

筛选:

- 立法机构: 全国人民代表大会常务委员会

### 丹麦

网址: <https://www.retsinformation.dk/>

关键词: kunstig intelligen

筛选:

- 文件类型: 法律

### 芬兰

网址: <https://www.finlex.fi/>

关键词: tekoäly

注: 在现行立法部分注意到

### 法国

网址: <https://www.legifrance.gouv.fr/>

关键词: intelligence artificielle

筛选:

- 综合文本
- 文件类型: 法律

### 德国

网址: <http://www.gesetze-im-internet.de/index.html>

关键词: künstliche Intelligenz

筛选:

- 所有目前有效的联邦法典、法规和条例
- Volltextsuche (全文)
- Und-Verknüpfung der Wörter (全字)

### 印度

网址: <https://www.indiacode.nic.in>

关键词: artificial intelligence

注: 所用的网站允许搜索合法化标题中的关键词，但不允许搜索全文，因此它对这个特定的研究没有用处。使



用 "网站" 功能进行谷歌搜索，以 "人工智能" 为关键词搜索该网站。

## 爱尔兰

网址: [www.irishstatutebook.ie](http://www.irishstatutebook.ie)  
关键词: artificial intelligence

## 意大利

网址: <https://www.normattiva.it/>  
关键词: intelligenza artificiale  
筛选:

- 文件类型: 法律

## 日本

网址: <https://elaws.e-gov.go.jp/>  
关键词: 人工知能  
筛选:

- 全文
- 法律

## 荷兰

网址: <https://www.overheid.nl/>  
关键词: kunstmatige intelligentie  
筛选:

- 文件类型: Wetten

## 新西兰

网址: [www.legislation.govt.nz](http://www.legislation.govt.nz)  
关键词: Artificial intelligence  
筛选:

- 文件类型: 法案
- 状态选项: 对于状态的选项 (例如: 有效的行为、当前的法案等)

## 挪威

网址: <https://lovdata.no/>  
关键词: kunstig intelligens

## 俄罗斯

网址: <http://graph.garant.ru:8080/SESSION/PILOT/main.htm> (俄罗斯联邦联邦会议联邦委员会官方网站中的 "The Federal Laws" 数据库)  
关键词: искусственный интеллект  
筛选:

- 文本中的字词

## 新加坡

网址: <https://sso.agc.gov.sg/>  
关键词: artificial intelligence  
筛选:

- 文件类型: 现行法案和附属立法

## 南非

网址: [www.gov.za](http://www.gov.za)  
关键词: artificial intelligence  
筛选:

- 文件: 法案

注: 这个搜索功能似乎没有在全文的范围内搜索，所以没有返回任何结果。因此，使用 "网站" 功能进行谷歌搜索，以 "人工智能" 为关键词搜索网站。

## 韩国

网址: <https://law.go.kr/eng/>; <https://elaw.klri.re.kr/>  
关键词: artificial Intelligence or 인공 지능  
筛选:

- 类型: 法案

注: 不能搜索组合词，所以要进行单独分析。

## 西班牙

网址: <https://www.boe.es/>  
关键词: inteligencia artificial  
筛选:

- 类型: 法律
- 国家首脑 (针对已通过的法律)



## 瑞典

网址: <https://www.riksdagen.se/>

关键词: artificiell intelligens

筛选: 瑞典法规法则

## 瑞士

网址: <https://www.fedlex.admin.ch/>

关键词: intelligence artificielle

筛选:

- 文本类别: 联邦宪法、联邦法案和联邦法令、各种文本、命令和其他形式的立法。
- 立法的公布期限定为2015-2021年。

## 英国

网址: <https://www.legislation.gov.uk/>

关键词: artificial intelligence

筛选:

- 立法类型: 英国公共法令和英国法定文书

## 美国

网址: <https://www.congress.gov/>

关键词: artificial intelligence

筛选:

- 资料来源: 立法

立法状态: 已成为法律



## 在与人工智能有关的立法程序中提及人工智能的情况

对于世界各地与人工智能有关的立法程序中提到的人工智能，人工智能指数在25个国家的国会或议会的网站上用各自的语言对关键词 "人工智能" 进行了搜索，通常是在名为 "会议记录"、"hansard" 等栏目下。

### 澳大利亚

网址: [https://www.aph.gov.au/Parliamentary\\_Business/Hansard](https://www.aph.gov.au/Parliamentary_Business/Hansard)

关键词: artificial intelligence

### 比利时

网址: [http://www.parlement.brussels/search\\_form\\_fr/](http://www.parlement.brussels/search_form_fr/)

关键词: intelligence artificielle

Filter

- 文件类型: 全部

### 巴西

网址: <https://www2.camara.leg.br/atividade-legislativa/discursos-e-notas-taquigraficas>

关键词: inteligência artificial

筛选:

- 联邦立法
- 类型: 法律

### 加拿大

网址: <https://www.ourcommons.ca/PublicationSearch/en/?PubType=37>

关键词: artificial Intelligence

### 中国

网址: Various reports on the work of the government

关键词: 人工智能

注: 全国人民代表大会每年举行一次，不提供完整的立法程序。因此，分析中包含的计数只搜索了大会会议发布的唯一公开文件中提到的人工智能，即总理所做的《政府工作报告》。

### 丹麦

网址: <https://www.retsinformation.dk/>

关键词: kunstig intelligens

筛选:

- 会议记录

### 芬兰

网址: <https://www.eduskunta.fi/>

关键词: tiedot

筛选:

- 议会事务和文件
- 公开文件: 会议记录
- 行为类型: 全体会议

### 法国

网址: <https://www.assemblee-nationale.fr/>

关键词: intelligence artificielle

筛选:

- 会议的辩论报告

注: 此类文件只在2017年开始编制。

### 德国

网址: <https://dip.bundestag.de/>

关键词: künstliche Intelligenz

筛选:

- 演讲, 要求在全体会议上发言

### 印度

网址: <http://loksabhap.nic.in/>

关键词: artificial intelligence

筛选:

- 精确的词/短语



## 爱尔兰

网址: <https://www.oireachtas.ie/>

关键词: artificial intelligence

筛选: 议会辩论的内容

## 意大利

网址: <https://aic.camera.it/aic/search.html>

关键词: intelligenza artificiale

筛选:

- 类型: 全部
- 按精确的短语搜索

## 日本

网址: <https://kokkai.ndl.go.jp/#/>

关键词: 人工知能

筛选:

- Full text
- Law

## 荷兰

网址: [https://www.tweedekeamer.nl/kamerstukken?pk\\_campaign=breadcrumb](https://www.tweedekeamer.nl/kamerstukken?pk_campaign=breadcrumb)

关键词: kunstmatige intelligentie

筛选:

- 议会文件-全体会议报告

## 新西兰

网址: <https://www.parliament.nz/en/pb/hansard-debates/>

关键词: artificial intelligence

## 挪威

网址: <https://www.stortinget.no/no/Saker-og-publikasjoner/Publikasjoner/Referater/>

关键词: kunstig intelligens

注: 该搜索功能不允许直接在几分钟内输入关键词。因此, 使用“网站”功能进行谷歌搜索, 以“人工智能”为关键词搜索网站。

## 俄罗斯

网址: <http://transcript.duma.gov.ru/>

关键词: искусственный интеллект

筛选:

- 文本中的字词

## 新加坡

网址: <https://sprs.parl.gov.sg/search/home>

关键词: artificial intelligence

## 南非

网址: <https://www.parliament.gov.za/hansard>

关键词: artificial intelligence

注: 此搜索功能不会在全文范围内进行搜索, 因此没有返回任何结果。使用“网站”功能进行谷歌搜索, 以“artificial intelligence”为关键词在<https://www.parliament.gov.za/storage/app/media/Docs/hansard/>中搜索。

## 韩国

网址: <http://likms.assembly.go.kr/>

关键词: 인공 지능

筛选:

- 会议类型: 全部

## 西班牙

网址: <https://www.congreso.es/>

关键词: inteligencia artificial

筛选:

- 议会程序的官方出版物

## 瑞士

网址: <https://www.parlament.ch/>

关键词: intelligence artificielle

筛选:

- 议会程序



## 瑞典

网址: [https://www.riksdagen.se/sv/global/  
sok/?q=&doktyp=prot](https://www.riksdagen.se/sv/global/sok/?q=&doktyp=prot)

关键词: artificiell intelligens

筛选:

- 会议记录

## 英国

网址: <https://hansard.parliament.uk/>

关键词: artificial intelligence

筛选:

- 参考文献

## 美国

网址: <https://www.congress.gov/>

关键词: artificial intelligence

筛选:

- 资料来源: 国会记录
- 国会记录部分: 参议院、众议院和扩展备注

- 私营企业: Google AI, Microsoft AI, Nvidia, OpenAI
- 智囊团和政策研究所: 美国企业研究所、阿斯彭研究所、大西洋理事会、布鲁金斯研究所、卡内基国际和平基金会、卡托研究所、新美国安全中心、战略与国际研究中心、对外关系委员会、传统基金会、哈德逊研究所、MacroPolo、国家安全研究所、新美国基金会、兰德公司、洛克菲勒基金会、史汀生中心、城市研究所、威尔逊中心
- 大学机构和研究项目: 人工智能与人类 康奈尔大学；纽约大学AI Now研究所；加州大学洛杉矶分校法学院AI Pulse；哈佛大学贝尔弗科学与国际事务中心；哈佛大学Berkman Klein中心；普林斯顿大学信息技术政策中心；加州大学伯克利分校长期网络安全中心。乔治敦大学安全和新兴技术中心；加州大学伯克利分校CITRUS政策实验室；胡佛研究所；斯坦福大学以人为本的人工智能研究所；麻省理工学院互联网政策研究倡议；麻省理工学院林肯实验室；普林斯顿公共和国际事务学院

## 美国人工智能政策文件

### 机构

为了对推动人工智能政策的思想领导力有更细致的了解，我们追踪了美国境内或在美国有重要影响力的55个组织（比去年的36个组织名单有所扩大）发表的政策文件，涉及四个大类。

- 民间社会、协会和财团: 算法正义联盟, 医疗领域人工智能联盟, Amnesty International, EFF, 未来隐私论坛, 人权观察, IJIS研究所, 电气和电子工程师协会, 人工智能伙伴关系
- 咨询业: 埃森哲、贝恩公司、波士顿咨询公司、德勤公司、麦肯锡公司
- 政府机构: 国会研究处、国会图书馆、国防技术信息中心、政府问责 国防技术信息中心、政府问责办公室, 五角大楼图书馆

### 方法论

每个广泛的主题领域都是基于描述具体论文内容的基础关键词的集合。我们包括17个主题，这些主题代表了2018-2021年间与人工智能相关的大部分内容。这些主题领域和相关的关键词具体如下。

- 健康和生物科学: 医学、医疗保健系统、药物发现、护理、生物医学研究、保险、健康行为、COVID-19、全球健康。
- 物理科学: 化学、物理学、天文学、地球科学
- 能源与环境: 能源成本、气候变化、能源市场、污染、保护、石油和天然气、替代能源
- 国际事务和国际安全: 国际关系、国际贸易、发展中国家、人道主义援助、战争、区域安全、国家安全、自主武器
- 司法与执法: 民事司法、刑事司法、社会司法、警察、公共安全、法院



- 通讯与媒体：社交媒体、虚假信息、媒体市场、deepfake
- 政府与公共管理：联邦政府、州政府、地方政府、公共部门的效率、公共部门的有效性、政府服务、政府福利、政府项目、公共工程、公共交通
- 民主：选举、权利、freedoms, liberties,、个人自由
- 产业与监管：经济、反垄断、并购、竞争、金融、管理、供应链、电信、经济监管、技术标准、自动驾驶汽车产业和监管
- 创新与技术：人工智能技术的进步和改进、研发、知识产权、专利、创业、创新生态系统、初创企业、计算机科学、工程
- 教育与技能：儿童早期教育、K-12教育、高等教育、STEM、学校、课堂、再培训
- 劳动力与劳动：劳动力供需、人才、移民、移徙、人事经济学、工作前景
- 社会和行为科学：社会学、语言学、人类学、民族研究、人口学、地理学、心理学、认知科学
- 人文科学：艺术、音乐、文学、语言、表演、戏剧、古典文学、历史、哲学、宗教、文化研究
- 公平与包容：偏见、歧视、性别、种族、社会经济不平等、残疾、弱势人群
- 隐私、安全和保障：匿名、GDPR、消费者保护、人身安全、人为控制、网络安全、加密、黑客行为
- 伦理：透明度、问责制、人类价值、人权、可持续性、可解释性、可说明性、决策规范



人工智能指数  
2022年度报告

 斯坦福大学  
以人为本人工智能研究院  
(斯坦福HAI)