



美国国家人工智能研究和发展战略计划

美国国家科学技术委员会

美国网络和信息技术研发小组委员会

2016 年 10 月



中国信息通信研究院政策与经济研究所编译组 整理

关于国家科学技术委员会

国家科学技术委员会（NSTC）是行政机构在构成联邦研发（R&D）企业的各种实体之间协调科学和技术政策的主要手段。NSTC 的主要目标之一是为联邦科学和技术投资制定明确的国家目标。NSTC 准备了旨在实现多个国家目标的研发包。NSTC 的工作在五个委员会下组建：环境委员会、自然资源和可持续发展委员会；国家与国土安全委员会；科学、技术、工程和数学（STEM）教育委员会；科学和技术委员会。每个委员会都负责监督专注于科学和技术不同方面的小组委员会和工作组。更多信息，请访问 www.whitehouse.gov/ostp/nstc。

关于白宫科技政策办公室

白宫科技政策办公室（OSTP）由 1976 年国家科学技术政策、组织和优先次序法案设立。OSTP 的使命有三方面：第一，就所有重要事项向总裁及其高级工作人员提供准确、相关和及时的科学和技术咨询；第二，确保行政机构的政策得到健全科学的通知；第三，确保行政机构的科学技术工作得到适当协调，为社会带来最大的益处。OSTP 的主任还担任科学和技术总裁助理，负责管理 NSTC。更多信息，请访问 www.whitehouse.gov/ostp。

关于网络和信息技术研究与开发小组委员会

网络和信息技术研发（NITRD）小组委员会是国家科学技术委员会（NSTC）技术委员会（CoT）下的一个机构。NITRD 小组委员会协调多机构研究和发展计划，以帮助确保美国在网络和信息技术方面的持续领导，满足联邦政府对先进网络和信息技术的需求，加快先进网络和信息技术的开发和部署。它还实施经 1998 年《下一代互联网研究法案》（P.L. 105-305）修订的 1991 年《高性能计算法》（PL 102-194）相关条款，以及 2007 年《美国创造机会，切实推进卓越技术、教育与科学（COMPETES）法》（P.L. 110-69）。更多信息，请参见 www.nitrd.gov。

致谢

本文档通过 NITRD 人工智能工作组的成员和工作人员的贡献而制定。特别感谢和感激帮助撰写、编辑和审查文档的其他贡献者：Chaitan Baru (NSF)、Eric Daimler (Presidential Innovation Fellow)、Ronald Ferguson (DoD)、Nancy Forbes (NITRD)、Eric Harder (DHS)、Erin Kenneally (DHS)、Dai Kim (DoD)、Tatiana Korelsky (NSF)、David Kuehn (DOT)、Terence Langendoen (NSF)、Peter Lyster (NITRD)、KC Morris (NIST)、Hector Munoz-Avila (NSF)、Thomas Rindfleisch (NIH)、Craig Schlenoff (NIST)、Donald Sofge (NRL)、和 Sylvia Spengler (NSF)。

版权信息

这是美国政府的一项工作，并属于公共领域。它可以自由分发、复制和翻译；感谢科学和技术政策办公室政策对出版的许可。任何翻译应包括一个免责声明，翻译是否准确为翻译者而不是 OSTP 的责任。请求将任何翻译的副本发送给 OSTP。根据知识共享 CC0 1.0 通用许可证的规定，此项工作可在全球范围内使用和重复使用。

2016 年 10 月 13 日

各位尊敬的同事：

我们很高兴通过这封信传递 NSTC 的国家人工智能研究与发展战略计划。该计划由人工智能工作组开发，人工智能工作组是应 NSTC 机器学习和人工智能小组委员会的要求，由 NSTC 的 NITRD 小组委员会任命的跨部门工作组。

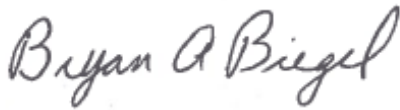
智能计算机系统一直是科幻小说的主题。现在，我们正在进入一个 AI 对我们的日常生活产生广泛和深远影响的时代，从精确医学到交通到教育和更多方面。作为回应，2016 年 5 月 3 日，白宫宣布了一系列行动，以促进 AI 的公共对话，确认与这种新兴技术相关的挑战和机遇，帮助政府更有效的利用 AI，并为 AI 的潜在利益和风险做好准备。作为这些行动的一部分，白宫指导创建人工智能研究和开发的国家战略。

由此产生的 AI 研发战略规划定义了一个高级框架，可用于确定 AI 中的科学和技术需求，跟踪研发投资的进展并最大限度地发挥其影响以满足这些需求。它还确立 AI 中联邦资助研发的优先事项，展望短期内 AI 对社会和世界长期变革影响的能力。

联邦政府的这种协调的 AI 研发尝试将帮助美国充分利用 AI 技术的全部潜力，以发展我们经济的同时改善我们的社会。然而，AI 研发战略规划并没有为个别联邦机构定义具体的研究议程。相反，机构将继续根据其使命、能力、权威和预算来追寻优先重点，同时进行协调，使整体研究组合与 AI 研发战略规划一致。

我们期待与联邦机构和其他关键合作伙伴继续这项重要的工作，并利用该计划指导未来在 AI 研发的决策。

诚挚的，



Bryan Biegel

网络和信息技术研发国家协调办公室
主任



James F. Kurose

国家科学基金会科学与工程部计算机与信息系副主任，

联合主席，网络和信息技术研发小组委员会

国家科学技术委员会

主席

John P. Holdren

科学技术委员会主席助理

兼科学技术政策办公室主任

成员

Afua Bruce

科学技术政策办公室

执行主任

机器学习和人工智能小组委员会

联合主席

Ed Felten

美国科学技术政策办公室

副首席技术官

联合主席

Michael Garriss

高级科学家

美国商务部

美国国家标准与技术协会

网络和信息技术研发小组委员会

联合主席

Bryan Biegel

网络和信息技术研发国家协调办公室主任

联合主席

James Kurose

国家科学基金会科学与工程部计算机与信息系副主任

网络和信息技术研发项目负责人工智能的任务组

联合主席

Lynne Parker

部门主任

信息与智能系统国家科学基金会成员

联合主席

Jason Matheny

主任

智力高级研究项目活动

Milton Corn

美国国家卫生研究所

Nikunj Oza

美国国家航空和航天局

William Ford

国家司法研究所

Robinson Pino

能源部

Michael Garriss

美国国家标准与技术协会

Gregory Shannon

科学技术政策办公室

Steven Knox

国家安全局

Scott Tousley

国土安全部

John Launchbury

国防部高级研究计划局

Faisal D'Souza网络和信息技术研发国家协调办公室
技术协调专员**Richard Linderman**

国防部长办公室

目 录

摘 要.....	7
一、 简介.....	9
（一） 《国家人工智能研究与发展战略计划》的目的.....	9
（二） 预期结果.....	11
（三） 利用人工智能推进国家优先事项的愿景.....	12
1、 促进经济发展.....	12
2、 改善教育机会和生活质量.....	13
3、 增强国家和国土安全.....	14
（四） 人工智能的现状.....	14
二、 研发战略.....	18
（一） 战略一：对人工智能研究进行长期投资.....	21
1、 提升基于数据发现知识的能力.....	21
2、 增强人工智能系统的感知能力.....	22
3、 了解人工智能的理论能力和局限性.....	22
4、 研究通用人工智能.....	23
5、 开发可扩展的人工智能系统.....	24
6、 促进类人的人工智能研究.....	24
7、 开发更强大和更可靠的机器人.....	25
8、 推动人工智能的硬件升级.....	26
9、 为改进的硬件创建人工智能.....	26
（二） 战略二：开发有效的人类与人工智能协作方法.....	28
1、 寻找人类感知人工智能的新算法.....	29
2、 开发增强人类能力的人工智能技术.....	30
3、 开发可视化和人机界面技术.....	30
4、 开发更高效的语言处理系统.....	31
（三） 战略三：了解并解决人工智能的伦理、法律和社会影响....	33
1、 改进公平性、透明度和设计责任机制.....	33
2、 建立符合伦理的人工智能.....	34

3、 设计符合伦理的人工智能架构.....	34
（四） 战略四：确保人工智能系统的安全可靠.....	36
1、 提高可解释性和透明度.....	36
2、 提高信任度.....	36
3、 增强可验证与可确认性.....	37
4、 保护免受攻击.....	38
5、 实现长期的人工智能安全和优化.....	38
（五） 战略五：开发用于人工智能培训及测试的公共数据集和环境	39
1、 开发满足多样化人工智能兴趣与应用的丰富数据集.....	39
2、 开放满足商业和公共利益的训练测试资源.....	40
3、 开发开源软件库和工具包.....	40
（六） 战略六：制定标准和基准以测量和评估人工智能技术.....	42
1、 开发广泛应用的人工智能标准.....	42
2、 制定人工智能技术的测试基准.....	42
3、 增加可用的人工智能测试平台.....	43
4、 促进人工智能社群参与标准和基准的制定.....	44
（七） 战略七：更好地了解国家人工智能人力需求.....	46
三、 建议.....	47
1、 建议一.....	47
2、 建议二.....	47
附录：首字母缩写词.....	48
译者注.....	50

摘要

人工智能（AI）是一种具有巨大社会和经济效益的革新性技术。人工智能有可能彻底改变我们的生活、工作、学习、发现和沟通的方式。人工智能研究可以推进美国的国家优先任务，包括增加经济繁荣、改善教育机会和生活质量，以及加强国家和国土安全。由于这些潜在的益处，美国政府已经对人工智能研究投资多年。然而，与联邦政府感兴趣的任何重要技术一样，指导人工智能领域联邦资助研发的总体方向时不仅具有巨大的机会，还必须考虑到一些注意事项。

2016 年 5 月 3 日，政府宣布成立一个新的国家科学技术委员会（NSTC）机器学习和人工智能小组委员会，以帮助协调联邦在人工智能领域的活动。¹ 该小组委员会于 2016 年 6 月 15 日，请求网络和信息技术研究和发展计划（NITRD）小组委员会编写《国家人工智能研究与发展战略计划》（以下简称“AI 研发战略计划”或《战略》）。之后成立了一个 NITRD 人工智能工作组，以确定人工智能研发为联邦的战略重大计划，特别关注产业不可能解决的领域。

这项《战略》为联邦资助的人工智能研究制定了一系列目标，既包括政府内部的研究，也包括联邦资助的政府外部研究，例如在学术界。这项研究的最终目标是产生新的人工智能知识和技术，为社会提供一系列积极效益，同时尽量减少负面影响。为实现这一目标，《战略》确定了联邦资助人工智能研究的以下重大计划：

战略一：对人工智能研究进行长期投资。 优先投资下一代人工智能，将促进新发现和洞察力，同时使美国在人工智能领域保持世界领先地位。

战略二：开发有效的人类与人工智能协作方法。 并非取代人类，大多数人工智能系统将与人类合作以实现最佳性能。需要研究来创建人类和人工智能系统之间的有效交互。

战略三：了解并解决人工智能的伦理、法律和社会影响。 我们期望人工智能技术根据我们持有人类同胞的正式和非正式规范表现。需要研究以了解人工智能的伦理、法律和社会影响，并开发设计符合伦理、法律和社会目标的人工智能系统的方法。

战略四：确保人工智能系统的安全可靠。 在人工智能系统广泛使用之前，

需要保证系统将以受控、充分定义和充分理解的方式安全地操作。需要进一步加强研究，以解决创建可靠、可信任和可信赖人工智能系统的挑战。

战略五：开发用于人工智能培训及测试的公共数据集和环境。训练数据集和资源的深度、质量和准确性显著影响人工智能性能。研究人员需要开发高质量的数据集和环境，并允许负责访问高质量数据集，以及测试和培训资源。

战略六：制定标准和基准以测量和评估人工智能技术。人工智能进步极其重要的是指导和评估人工智能进展的标准、测试基准、测试台和社区参与。需要进行额外的研究来开发广泛的评价技术。

战略七：更好地了解国家人工智能人力需求。人工智能的进步将需要一个强大的人工智能研究人员社区。需要更好地了解人工智能当前和未来研发人员需求，以帮助确保有足够的人工智能专家能够应对本计划中概述的战略研发领域。

《战略》最后提出了两方面建议：

建议一：开发一个人工智能研发实施框架，以抓住科技机遇，并支持人工智能研发投资的有效协调，与本计划的第一至六项战略保持一致。

建议二：研究创建和维持一个健全的人工智能研发队伍的国家愿景，与本计划的战略第七项保持一致。

一、简介

（一）《国家人工智能研究与发展战略计划》的目的

1956 年，来自美国的计算机科学研究人员在新罕布什尔州的达特茅斯学院会面，讨论一个新兴的计算分支，即人工智能或 AI 的开创性思想。他们想象了一个世界，“机器使用语言，构成抽象和概念，解决现在人类的问题，并改善自己”。² 这次历史性会议为 AI 的政府和行业研究设置了几十年阶段，包括感知、自动推理/规划、认知系统、机器学习、自然语言处理、机器人和相关领域的进展。今天，这些研究进展已经产生影响我们日常生活的新兴经济部门，从制图技术到语音辅助智能手机，到邮件传递的手写识别，到金融交易，到智能物流，到垃圾邮件过滤，语言翻译，甚至更多。AI 进展也为精准医学、环境可持续性、教育和公共福利等领域的社会福利带来巨大的益处。³

过去 25 年来，AI 方法的显著增加在很大程度上得益于统计和概率方法的采用，大量数据的可用性以及计算机处理能力的提高。在过去十年中，机器学习的 AI 子领域，使计算机能够从经验或例子中学习，已经表现出越来越准确的结果，引起了人们对 AI 近期前景更多的兴趣。虽然最近注意到例如深度学习等统计方法的重要性，⁴ 但在其他各种领域 AI 也已经取得了影响深远的进展，例如：感知、自然语言处理、形式逻辑、知识展示、机器人技术、控制理论、认知系统架构、搜索和优化技术以及其他更多方面。

（注：深度学习是指使用多层神经网络的一系列方法的汇总，这些方法支持快速完成一度被认为无法自动化完成的任务。）

AI 的最近成就对这些技术的最终方向和影响已经产生了重要问题：当前 AI 技术的重要科学和技术瓶颈是什么？新的 AI 进展将提供什么积极，需要的经济和社会影响？如何继续安全和有益地使用 AI 技术？如何设计 AI 系统以符合伦理、法律和社会原则？这些进步对 AI 研发人员的影响是什么？

AI 研发的情况变得越来越复杂。虽然政府过去和现在的投资造就了 AI 的突破性方法，但其他部门也已成为 AI 的重要贡献者，包括广泛的行业和非营利组织。这种投资环境提出了关于联邦投资在 AI 技术发展中适当作用的重要问题。联邦对 AI 投资的正确优先事项是什么，特别是在行业不可能投资的领域和时间框架方面？是否有机会有产业和国际研发合作，推动美国的优先事项？

2015 年，美国政府对 AI 相关技术的未分类研发投资约为 11 亿美元。虽然这些投资已经产生了重要的新科学和技术，但是仍有机会在联邦政府之间进一步协调，使这些投资能够充分发挥潜力。⁵

认识到 AI 的革新性影响，2016 年 5 月，白宫科学和技术政策办公室（OSTP）宣布了一个新的跨部门工作组，以探讨 AI 的利益和风险。⁶ OSTP 还宣布了一系列四个研讨会，举办于 2016 年 5 月至 7 月的一段时间，旨在促进 AI 的公众对话，并确定其所带来的挑战和机遇。研讨会的结果是伴随公共报告《为人工智能的未来准备》的一部分，与该计划一起发布。

在 2016 年 6 月，新的 NSTC 机器学习和人工智能小组委员会 - 它被特许在联邦政府、私营部门和国际上与 AI 的进展保持同步，并帮助协调联邦在 AI 的活动，任命 NITRD 国家协调办公室（NCO）创建《国家人工智能研究与发展战略计划》。小组委员会指示本计划应传达一系列明确的研究优先重点，以解决战略研究目标，将联邦投资重点放在行业不太可能投资的领域，并解决扩大和维持 AI 研发人才渠道的需求。

本 AI 研发战略计划的输入来自广泛的来源，包括联邦机构、AI 相关会议的公开讨论、投资于 IT 相关研发的所有联邦机构的 OMB 数据呼叫、投资 IT 相关研发，OSTP 信息请求 RFI），该信息请求向公众征询了有关美国如何为未来的 AI⁷ 做出最佳准备的意见，以及 AI 公开出版物的信息。

该计划对 AI⁸ 的未来做出多个假想。首先，假设 AI 技术将继续发展至复杂巧妙并无所不在，而这多亏了政府和行业对 AI 研发的投资。第二，本计划假设 AI 对社会的影响将继续增加，其中包括就业、教育、公共安全和国家安全，以及对美国经济增长的影响。第三，假设行业对 AI 的投资将继续增加，因为最近的商业成就已增加了研发投资的预期回报。同时，本计划假设一些重要的研究领域不太可能获得来自行业的足够投资，因为它们受制于典型的公共物品投资不足问题。最后，本计划假设对 AI 专业的需求将继续在行业、学术界和政府内部增长，从而对公共和私人造成劳动力压力。

与 AI 研发战略计划相关的其他研发战略计划和方案包括《联邦大数据研究和发展战略计划》、⁹《联邦网络安全研究和发展战略计划》、¹⁰《国家隐私研究和发展战略》、¹¹《国家纳米技术倡议战略计划》、¹²《国家战略计算计划》、¹³《推进创新神经技术脑研究计划》¹⁴ 与《国家机器人方案》。¹⁵ 涉及某

些 AI 子领域的其他战略研发计划和战略框架处于发展阶段，其中包括视频和图像分析、健康信息技术、机器人和智能系统。这些额外计划和框架将提供补助和详细叙述本 AI 研发战略计划的协同建议。

（二）预期结果

本 AI 研发战略计划超越了近期的 AI 功能，着眼于 AI 对社会和世界的长期变革影响。AI 的最新研究进展让 AI 的潜力更为乐观，使行业得到迅猛发展，并让 AI 方法变得商业化。然而，虽然联邦政府可以利用 AI 的行业投资，但许多应用领域和长期研究挑战不会存在明确的近期利润驱动因素，因此不可能完全由行业进行解决。联邦政府是长期高风险研究计划以及近期发展工作的主要资金来源，以实现部门或机构的具体要求，或解决私营企业并不从事的重要社会问题。因此，联邦政府应该强调重大社会重要性领域内的 AI 投资，这不针对消费市场的领域，如用于公共卫生、城市系统与智慧社区、社会福利、刑事司法、环境可持续性和国家安全的 AI，以及加速 AI 知识和技术生成的长期研究。

跨联邦政府的 AI 协调研发工作将增加这些技术的积极影响，并为决策者提供用于解决与使用 AI 相关的复杂政策挑战的所需知识。此外，协调方法将有助于美国利用 AI 技术的全部潜力来改善社会。

本 AI 研发战略计划定义了一个高级框架，该框架可用于确定 AI 的科学和技术差距，并跟踪用于填补这些差距的联邦研发投资。AI 研发战略计划确定了 AI 短期和长期支持的战略优先事项，以此来解决重要的技术和社会挑战。然而，AI 研发战略计划并未为个别联邦机构定义具体的研究议程。相反，其为行政部门设定了目标，在这些目标中，各机构可以根据其任务、能力、权威和预算来决定优先顺序，以便整个研究组合能与 AI 研发战略计划保持一致。

AI 研发战略计划也并未制定 AI 的研究或使用政策，亦未就 AI 对就业和经济的潜在影响作更广泛的探讨。虽然这些议题对国家至关重要，但它们在题为“人工智能的机遇和挑战，这次会有所不同吗？”⁸ 的经济顾问委员会报告中进行了讨论。

AI 研发战略计划侧重于有助定义和推进确保 AI 责任、安全和权益用途的政策的研究投资。

（三）利用人工智能推进国家优先事项的愿景

推动此 AI 研发战略计划是未来世界充满希望的愿景，AI 将给所有社会成员带来显著益处。人工智能的进一步进展可以提升社会中几乎所有部门的福利，¹⁶ 让国家优先事项获得进展，其中包括促进经济发展、改善生活质量和加强国家安全。这种潜在利益的例子包括：

1、促进经济发展

新产品和服务可以创造新市场，并提高多个行业现有商品和服务的质量和效率。通过专业决策系统创造更有效的物流和供应链。¹⁷ 通过基于视觉的驾驶员辅助和自动/机器人系统，¹⁸ 能更有效地运输产品。通过用于控制制造工艺和调度工作流程的新方法来改善制造业。¹⁹

如何促进经济发展？

（1）**制造业：**技术进步能在制造业，包括整个工程产品生命周期内引发新工业革命。更多使用机器人技术能使制造业回归陆上。²⁰ AI 可以通过更可靠的需求预测、提升运营和供应链灵活性，以及对改变制造业营运的影响进行更好的预测来加速生产能力。AI 可以创造更智能、更快、更便宜和更环保的生产流程，这能提高工人的生产率、提高产品质量、降低成本并改善工人的健康和安全。²¹ 机器学习算法可以改善制造流程的调度并减少库存要求。²² 消费者可以从现时的商业级 3-D 打印中获利。²³

（2）**物流：**私营部门制造商和托运人可以使用 AI，通过适配调度和路线来改进供应链管理。²⁴ 通过自动调整天气、交通和意外事件的预期影响，让供应链更加牢固难以中断。²⁵

（3）**金融：**工业和政府可以使用 AI 提供多种规模的异常金融风险早期检测。²⁶ 安全控制可以确保金融系统自动减少恶意行为的机会，例如市场操纵、欺诈和异常交易。²⁷ 他们可以进一步提高效率并降低波动性和交易成本，同时预防系统性失效，例如定价泡沫和低估信用风险。²⁸

（4）**交通：**AI 可以增强所有交通方式，实质上影响所有类型的旅途的安全。²⁹ 它可以用于结构安全监测和基础设施资产管理，提高公众信任，降低维修和重建成本。³⁰ AI 可用于乘客和货运车辆，从而通过增强情景意识来提高安全性，并为司机和其他旅客提供实时路线信息。³¹ AI 应用还可以改善网络级移动

性并减少整个系统的能源使用和运输相关的排放。³²

（5）**农业：**AI 系统可以创建通往可持续农业的途径，使农业产品的生产、加工、储存、分配和消费更灵活。AI 和机器人能收集有关作物的特定场所和时间数据，仅在它们需要的时间和地点才应用所需的投入（例如水、化学品和化肥），并填补农业劳动力的紧迫缺口。³³

（6）**营销：**AI 方法能使商业实体更好地配合供应与需求，增加用来资助进行中资助私营部门发展的税收。³⁴ 其能预测和识别消费者需求³⁵，使他们以更低的成本获得更好的产品和服务。

（7）**通信：**AI 技术可以最大限度地有效利用带宽和信息存储和检索的自动化。³⁶ AI 可以改进数字通信的过滤、搜索、语言翻译和摘要，积极影响商业和我们的生活方式。³⁷

（8）**科学和技术：**AI 系统可以协助科学家和工程师阅读出版物和专利，使理论与之前的观察值更一致，使用机器人系统和模拟、进行实验，并设计新的设备和软件。³⁸

2、改善教育机会和生活质量

通过用于制定专有学习计划的虚拟导师来实现终身学习，以此根据每个人的兴趣、能力和教育需求进行自我挑战和参与其中。通过为每个人定做和调整的个性化健康信息，让人们能过上更健康 and 更积极的生活。智能家居和个人虚拟助手可以节省人们的时间，并减少每日重复任务所损失的时间。

AI 将如何改善教育机会和社会福利？

（1）**教育：**AI-增强的学习型学校随处可见，通过其自动化辅导能衡量学生的发展¹⁶。AI 辅导员可补充面授教师，还可以因材施教。¹⁶ AI 工具可以促进终身学习并让所有社会成员获取新技能。¹⁶

（2）**医学：**AI 能支持从大规模基因组研究（如全基因组关联研究，排序研究）中识别出遗传风险的生物信息学系统，并预测新药物的安全性和有效性。³⁹ AI 技术允许进行多维度的数据评估，以研究公共卫生问题，并为医疗诊断和处方治疗提供决策支持系统。⁴⁰ AI 技术为个人提供药物定制；由此可提高医疗效果、患者舒适度和减少浪费。⁴¹

（3）**法律：**通过机器对法律个案史进行分析会变为普遍。⁴² 这些变得越来

越复杂的过程使辅助取证过程的分析水平得以提高。⁴²法律取证工具能识别和总结相关证据；这些系统甚至能以越来越复杂的方式制定法律条例。⁴²

（4）**个人服务：**AI 软件可以利用多来源的知识，为多种用途提供更准确的信息。⁴³自然语言系统可以为真实世界、嘈杂环境中的技术系统提供直观界面。⁴⁴个性化工具能为个人和群体调度提供自动辅助设备。⁴⁵文本能从多个搜索结果进行自动汇总，并在多种媒体中得以增强。⁴⁶AI 可以实现实时口语多语言翻译。⁴⁷

3、增强国家和国土安全

机器学习代理可以处理大量智能数据，并使用快速变化的战术来识别敌人的相关生存规律模式。⁴⁸这些代理还能向易受攻击的关键基础设施和主要经济部门提供保护。⁴⁹数字防御系统可以大大降低战场风险和伤亡。⁵⁰

如何增强国家和国土安全？

（1）**安全和执法：**执法和安全官员可以通过使用模式检测来检测个人行为者的异常行为或预测危险人群的行为，从而帮助建立一个更安全的社会。⁴⁸智能感知系统可以保护关键基础设施，如机场和发电厂。⁴⁹

（2）**安全和预测：**正常条件的分布式传感器系统和模式理解技术可以检测主要基础设施中断的概率何时会显著上升，无论是由自然原因还是人为原因引起。⁵¹这种预测能力能有助于指示出问题将在哪里发生，以防止或甚至在发生之前阻止中断。⁵¹

然而，这种积极使用 AI 的愿景，需要大量研发进展。许多关键和困难的技术挑战仍存在于所有 AI 子领域中，包括基础科学和应用领域。AI 技术还存在风险，例如随着人类因自动化系统而增长或被替代，劳动力市场可能会中断，以及 AI 系统安全性和可靠性的不确定性。本 AI 研发战略计划的后续章节讨论了 AI 研发投资的高级优先和战略领域，这些将支持这一愿景的同时将降低潜在破坏和风险。

（四）人工智能的现状

自其开始后，AI 研究已经历了三次技术浪潮。第一次浪潮集中于手工知识，20 世纪 80 年代着重于明确定义域的基于规则的专家系统，其中知识是从人类专家中采集，以“if-then”的规则进行运算，然后在硬件中执行。此类系

统推理可成功应用于狭义问题，但其并没有学习或处理不确定性的能力。然而，他们仍然能产生重要解决方案，并且今天的技术发展仍然非常活跃。

AI 研究的第二次浪潮起始于 21 世纪直到现在，其表现特点是机器学习的崛起。当应用于诸如图像和书写识别、语音理解和人类语言翻译的任务时，极为大量的数字数据可用性、相对便宜的大规模并行计算能力和经改良的学习技术让 AI 变得更为进步。这些进步的成果无处不在：智能手机执行语音识别、ATM 在书面支票执行手写识别、电子邮件应用程序执行垃圾邮件过滤、以及免费在线服务执行机器翻译。其中一些成就的关键是深度学习的发展。

如今，AI 系统在专业任务上的表现经常胜于人类。AI 首次超越人类表现的主要里程碑包括：国际象棋（1997 年）、⁵² trivia（2011 年）、⁵³ Atari 游戏（2013 年）、⁵⁴ 图像识别（2015 年）、⁵⁵ 语音识别（2015 年）、⁵⁶ 和 Go（译者注：即 AlphaGo 围棋。2016 年）。⁵⁷ 这类里程碑的步伐似乎越来越快，按现状，最佳表现系统是基于机器学习，而非手编码规则集合。

AI 的这些成就已得到了强大的基础研究基础的支持。这项研究正在扩大，可能会刺激未来的发展。作为一项指标，2013 年到 2015 年，以科学为索引提到“深度学习”的期刊文章的网站数量增加了六倍（图 1）。这一趋势也表明研究日益全球化，美国的出版物，或至少被引用过一次的出版物的数量不再领先世界（图 2）。

美国政府在人工智能研究中发挥了关键作用，尽管商业部门也积极参与人工智能相关研发。⁵⁸ 使用“深度学习”或“深度神经网络”这一术语的专利数量急剧增加（图 3）。从 2013 年到 2014 年，投入人工智能创业公司的风险资本增加了四倍。⁵⁹ 人工智能的应用现正为大型企业创造可观收益。⁶⁰ 人工智能对金融系统的影响甚至更大——全球金融交易近乎一半，即数万亿美元的交易，属于自动化（“算法”）交易。⁶¹

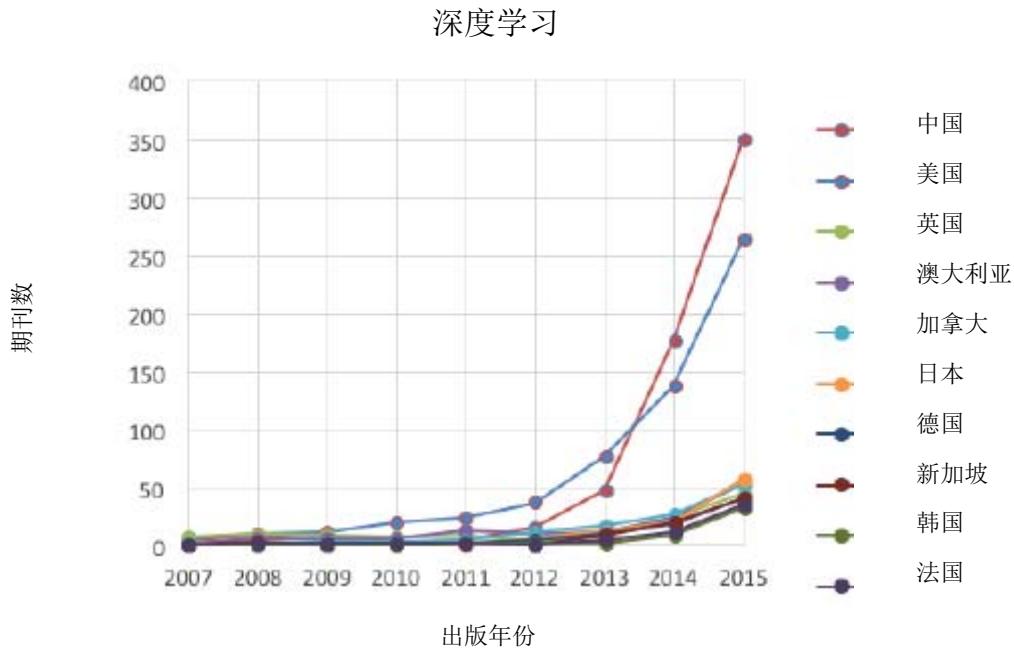


图 1：各国提到“深度学习”或“深度神经网络”的期刊文章⁶²



图 2：各国至少被引用过一次的提到“深度学习”或“深度神经网络”的期刊文章⁶³

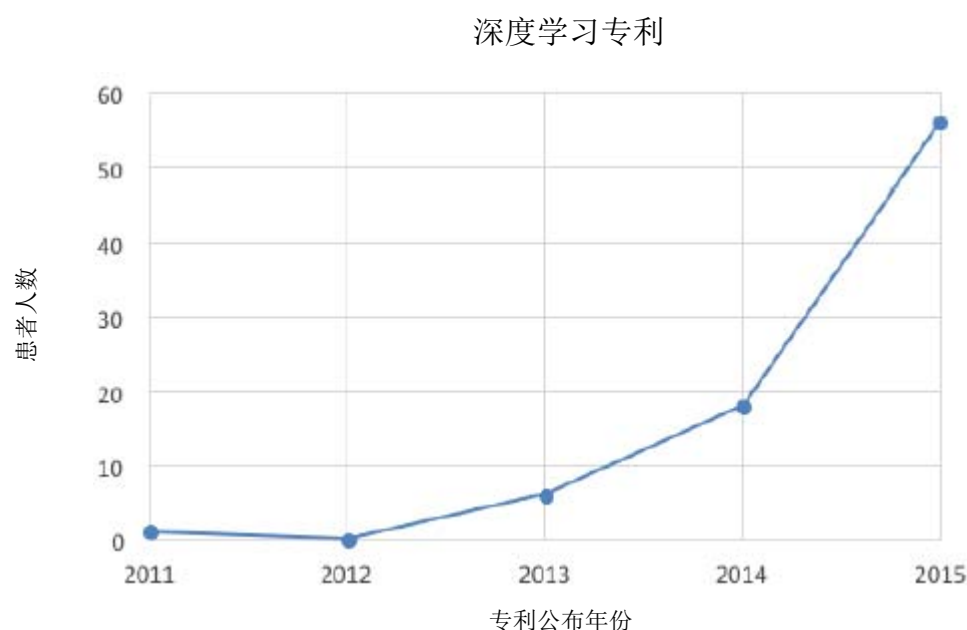


图 3：使用术语“深度学习”或“深度神经网络”的专利数量分析⁶⁴

尽管取得了进步，人工智能系统仍然有其局限性。几乎所有进步都是在能够有效完成专业任务的“狭义人工智能”方面取得；而在各种认知领域能够发挥有效作用的“广义人工智能”方面取得的进步很小。即使是在狭义人工智能方面，进步也不均衡。用于图像识别的人工智能系统需大量人力来标记数以千计的示例答案。⁶⁵相反，大多数人只需通过几个示例便可“一次性”掌握学习方法。虽然大多数机器视觉系统容易被具有重叠事物的复杂场景混杂，但是儿童可轻松进行“场景解析”。人容易理解的场景对于机器来说仍然很困难。

人工智能领域现处于第三次浪潮的开始阶段，注重解释性和通用人工智能技术。这些方法的目标是通过说明和界面修正加强学习模型，明确输出的基础和可靠性，以高透明度运作，超越狭义人工智能，获得可在更广泛任务领域中通用的功能。如果成功的话，工程师可创建系统，构建现实世界现象类的解释性模型，与人进行自然交流，在遇到新任务和情况时学习和思考，并通过总结过去的经验解决新问题。人工智能系统的解释性模型可通过先进方法自动构建。这些模型可实现人工智能系统的快速学习，可以向人工智能系统提供“含义”或“理解”，使人工智能系统获得更多通用功能。

二、研发战略

《战略》中所述研究重点侧重于行业不能解决的领域，因此这些领域最有可能从联邦投资中受益。这些重点研究涉及人工智能领域所有需求，包括感知、自动推理/规划、认知系统、机器学习、自然语言处理、机器人等人工智能子领域和相关领域中的常见需求。由于人工智能的广度，这些研究重点跨越整个领域，而不仅针对各子领域具体的单个研究难题。为了实施该规划，应制定详细的路线图，明确与规划一致的功能缺口。

战略一中所述最重要的联邦研究重点之一是对人工智能的持续长期研究，获得发现和深刻见解。美国联邦政府许多对高风险、高回报基础研究的投资已带来了今天赖以生存的革命性的技术进步，包括互联网、GPS、智能手机语音识别、心脏监视器、太阳能电池板、先进电池、癌症治疗等。人工智能的前景几乎可涉及社会的每一方面，带来显著积极的社会效益和经济效益。因此，为了在这一领域保持世界领先地位，美国必须重视对人工智能基础和长期研究的投资。

许多人工智能技术将与人一起工作，¹⁶因此如何最好地建立主动帮助人工作的人工智能系统成为重要挑战¹⁶。人和人工智能系统之间的壁垒慢慢被打破，人工智能系统不断增强人的能力。如战略二所述，需开展基础研究，以研究有效的人机交互和协作方法。

人工智能的进步为社会带来了许多积极的益处，并提高了美国的国家竞争力。⁸然而，与大多数变革性技术一样，人工智能也为某些领域带来了风险，包括就业、经济、安全、伦理和法律问题。因此，随着人工智能科技的发展，联邦政府也必须投资研究更好地理解人工智能对所有这些领域的影响，并通过研发如战略三所述的符合伦理、法律和社会目标的人工智能系统以解决这些影响。

当前人工智能技术的一个关键缺口是缺乏确保人工智能系统安全性和可预测性的方法。确保人工智能系统的安全性是一项挑战，因为这些系统异常复杂并不断演变。面对这一安全挑战进行了几项重点研究。首先，战略四强调需建立可解释和透明的系统，这些系统受到用户信任，按用户可接受的方式运行，并确保可按用户想要的方式运行。人工智能系统的潜在功能和复杂性以及与人

类用户和环境的广泛相互作用，使增加对人工智能技术安全性和控制性研究的投资至关重要。战略五要求联邦政府对用于人工智能培训和测试的共享公共数据集进行投资，以推进人工智能的研究并对替代解决方案进行更有效的比较。战略六讨论了评估研发的标准和基准，以确定进展，缩小差距，为具体问题和挑战提供创新解决方案。标准和基准对于测量和评估人工智能系统以及确保人工智能技术符合设计功能和互用性的关键目标至关重要。

最后，人工智能技术在社会各领域的日益普及为人工智能研发专家带来了新的压力。⁶⁶ 人工智能领域的核心科学家和工程师将拥有大量机会，他们熟练掌握技术，具有新想法，能够拓展该领域的知识边界。国家应采取措施，确保熟练掌握人工智能技术的人才充足。战略七阐述了这一挑战。

以下图 4 提供了本《战略》总体结构的图形说明。底部一行（红色行）是影响所有人工智能系统开发的跨领域根本基础；这些基础在战略三-7 中进行了说明。中间第二行（浅蓝色和中蓝色行）包括推动人工智能进步所需许多领域的研究。这些基础研究领域（包括应用型基础研究）在战略一-2 中进行了概述。⁶⁷ 图形的顶部一行（深蓝色行）是预期受益于 AI 进步的应用示例，如本文件前文愿景部分所述。总之，本《战略》的这些组成部分为联邦投资定义了一个高层框架，可为该领域带来有效进展和积极的社会效益。

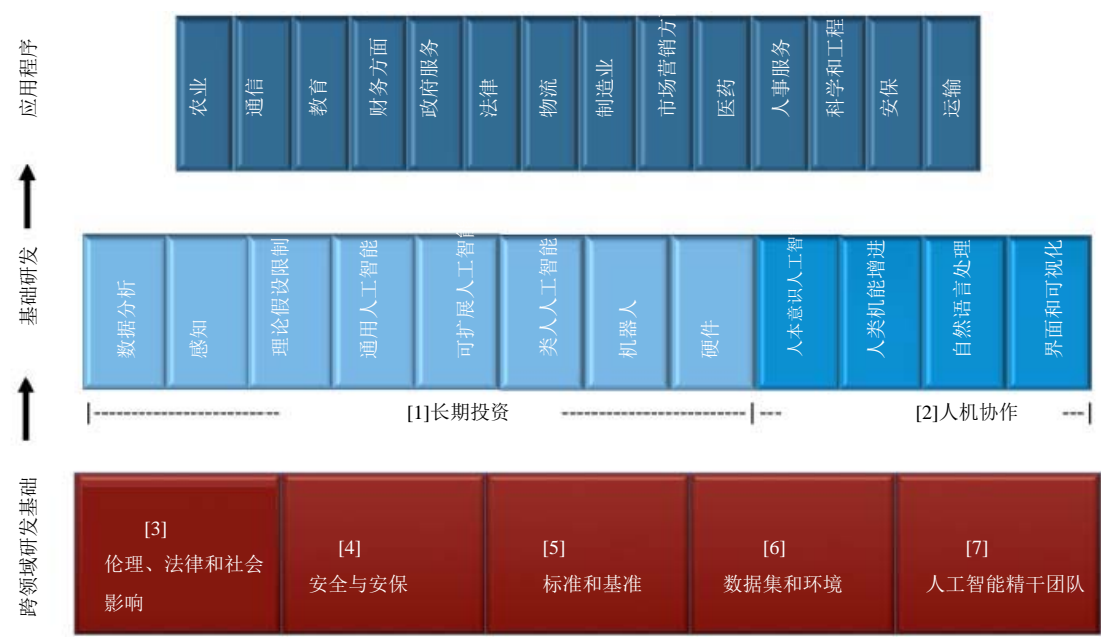


图 4. 人工智能研发战略规划的结构

跨领域研发基础（底部红色行）对于所有人工智能的研究都很重要。人工智能研发的

许多基础领域（浅蓝色和中蓝色行）可建立在这些跨领域基础上，以影响各种社会应用（顶部深蓝色行）。（括号中的小数字表示在本规划中的战略编号，战略有待进一步拓展。战略顺序并不表示重要性的优先次序。）

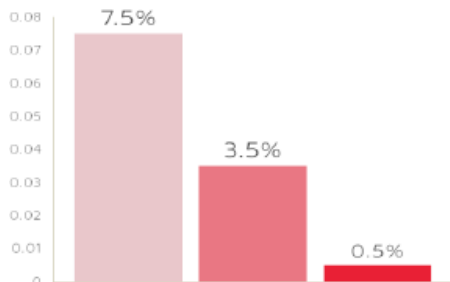
（一）战略一：对人工智能研究进行长期投资

需对具有潜在长期回报领域的人工智能研究进行投资。长期研究包含可预测结果的渐进式研究，对高风险研究的长期持续投资可带来高回报。

这些回报可在 5 年、10 年或更长时间内兑现。国家科学研究委员会最新报告强调了联邦投资在长期研究中的关键作用，指出“初步探索 and 商业化应用之间长期不可预测的孕育期需稳定的人力和财力”。⁶⁸ 报告进一步指出“从初步概念到成功上市通常需要几十年”。⁶⁸ 持续的基础研究工作带来高回报的有力证据包括万维网和深度学习。这两种案例的基础研究均始于 20 世纪 60 年代；历经 30 多年的持续研究，那些想法才转换成今天在各种人工智能中所见到的变革性技术。

例 1：国立卫生研究院（NIH）资助研究计算机病理学人工智能

影像诊断在癌症的病理诊断中具有关键作用。自 19 世纪后期以来，病理学家用于确定癌症诊断的主要工具是显微镜。病理学家通过手动检查癌症组织的染色切片来诊断癌症，确定癌症子类型。使用常规方法的病理诊断需要大量人力，不易重复，导致质量问题。



顶尖 AI 系统 病理学家 顶尖 AI 病理学家
人工智能在通过前哨淋巴结活检确定转移性乳腺癌方面能够有效降低病理学家的出错率。

新方法利用人工智能基础研究建立工具，使病理分析更有效、更准确、更具可预测性。在 2016 年转移性癌症检测 Camelyon 挑战大赛中，⁶⁹ 表现最佳的是基于人工智能的计算系统，出错率仅为 7.5%。⁷⁰ 辨识同组评估图像的一位病理学家出错率仅为 3.5%。结合人工智能系统的预测与病理学家的分析可将出错率降低至 0.5%，表示可减少 85% 的错误（见图像）。⁷¹

下文强调了上述某些领域。其他类型人工智能的重要研究在战略二至六中进行了讨论。

1、提升基于数据发现知识的能力

正如《联邦大数据研究和发展战略规划》中所述，⁹ 实现智能数据理解和知识发现需许多新的基础工具和技术。开发挖掘大数据中所有有用信息的更先进的机器学习算法中需取得进一步进展。

许多开放性研究问题围绕数据的创建和使用，包括对人工智能系统培训的准确性和适当性。当处理大量数据时，数据的准确性尤其具有挑战性，使人难以对其进行评估并从中提取信息。虽然许多研究通过数据质量保证方法确保数据清理和知识发现的准确性，但需进一步研究以提高数据清理技术的效率，建立发现数据不一致和异常的方法并使其可与人的反馈相结合。研究人员需探索新方法，以便同时挖掘数据和相关元数据。

许多人工智能的应用具有跨学科性质并利用异构数据。需对多模态机器学习进行进一步研究，以实现从不同类型数据（如离散数据、连续数据、文本数据、空间数据、时间数据、时空数据、图形数据）中获得知识发现。人工智能研究人员必须确定培训所需的数据量，并正确处理大型数据和长尾数据需求。他们也必须确定如何识别和处理纯统计学方法之外的小概率事件；在工作中利用知识源（即解释世界的任何类型的信息，如重力定律或社会规范的知识）和数据源，在学习过程中结合模型和本体；并且当大数据源不可用时，可利用有限数据获得有效的学习效果。

2、增强人工智能系统的感知能力

感知是智能系统面对世界的窗口。感知来自（可能为分布式）传感器数据，具有多种形态和形式，如系统本身的状态或环境的相关信息。传感器数据常与先验知识和模型一起进行处理和整合，以提取与人工智能系统任务相关的信息，如几何特征、属性、位置和速度。来自感知的综合数据形成环境感知，为人工智能系统提供综合知识和世界状态模型，有助于有效、安全地规划和执行任务。人工智能系统将极大地受益于硬件和算法方面的进步，获得更稳定和可靠的感知。传感器必须能够长距离捕获具有高分辨率的实时数据。感知系统需能够综合来自各种传感器和其他来源（包括计算云）的数据，以确定人工智能系统当前的感知对象并对其未来状态进行预测。对物体的检测、分类、识别和确认仍具有挑战性，特别是在杂乱和动态条件下。此外，通过使用传感器和算法的适当组合来大大改进人类的感知，使得 AI 系统可以更有效地与人类一起工作。¹⁶ 在整个感知过程中，需要一套用于计算和传播不确定性的框架来量化 AI 系统在其情境意识中的置信水平并提高准确性。

3、了解人工智能的理论能力和局限性

虽然许多 AI 算法的最终目标是使用仿人类的方案来解决开放式挑战，但我们不能很好地理解 AI 的理论能力和限制，以及这种仿人类的解决方案与 AI 算法一起使用的可行性能达到何种程度。需要理论工作来更好地理解为什么 AI 技术（特别是机器学习）通常在实践中起到良好作用。虽然不同的学科（包括数学，控制科学和计算机科学）都在研究这个问题，但该领域目前缺乏统一的理论模型或框架来理解 AI 系统性能。需要对计算可求解性进行额外的研究，这是指针对 AI 算法在理论上能够解决的问题类别的理解性，同样地，还针对了它们无法解决的问题。这种理解必须在现有硬件的背景下形成，以便了解硬件是如何影响这些算法的性能。

理解哪些问题在理论上是不可解决的，这样可以引导研究人员开发出这些问题的近似解决方案，甚至开辟出关于 AI 系统新型硬件的新研究路线。例如，人工神经网络（ANNs）于 20 世纪 60 年代被发明出来时，其只能被用于解决非常简单的问题。只是在硬件改进（例如并行化）之后使用 ANNs 来解决复杂的问题才变得可行，而且算法也被调整以利用新型硬件。这些发展是使得现今在深度学习方面取得重大进步的关键因素。

4、研究通用人工智能

AI 方法可以分为“狭义 AI”和“通用 AI”。狭义的 AI 系统在特定的、定义明确的领域中执行单个任务，例如语音识别，图像识别和翻译。最近几项高度可见的狭义 AI 系统，包括 IBM Watson 和 DeepMind 的 AlphaGo，已经取得了重大成就。^{72, 73} 确实，这些特定的系统被赋予“超越人类”的标签，因为他们各自在 Jeopardy 和 Go 中的表现均胜过了最好的人类玩家。但这些系统只是举例说明了狭义 AI，因为它们只能应用于为他们专门设计的任务。将这些系统用在更广泛的问题上则需要大量的重新设计工作。相比之下，通用 AI 的长期目标是创造出能够在广泛的认知领域（包括学习，语言，知觉，推理，创造力和规划）中表现出人类智力之灵活性和多功能性的系统。广泛学习能力将为通用 AI 系统提供将知识从一个领域转移到另一个领域以及从经验和人类交互中学习的能力。通用 AI 自 AI 出现以来一直是研究者的远大目标，但目前的系统离实现这一目标仍然很远。目前正在探索狭义 AI 和通用 AI 之间的关系；从其中一个获得的经验教训也可能应用到改进另外一个上，反之亦然。虽然没有一致的共

识，大多数 AI 研究人员认为，通用 AI 距离现在仍有几十年的差距，需要长期、持续的研究投入来实现它。

5、开发可扩展的人工智能系统

AI 系统的不同分组和网络可以协调或自主协作以执行单个 AI 系统不可能完成的任务，并且还可以包括人类协作或带领团队。这种多 AI 系统的开发和使用该类系统的规划、协调、控制和可扩展性方面产生了重大的研究挑战。多 AI 系统的规划技术必须足够快，能够实时操作和适应环境的变化。它们应以流动的方式去适应可用通信带宽或系统退化的变化以及故障。许多先前的努力都是聚焦于集中规划和协调技术；然而，这些方法受制于单点故障，诸如规划器的信息丢失或者通向规划器的通信链接的信息丢失。分布式规划和控制技术难以通过算法实现，并且通常效率较低和不完整，但能潜在地为单点故障提供更强稳健性。未来的研究必须发现更有效、稳健和可扩展的技术，用于多个 AI 系统和人类群组的规划、控制和协作。

6、促进类人的人工智能研究

实现类人的人工智能，需要 AI 系统以人们能够理解的方式进行自我表达。这将引出新一代的智能系统，如智能教学系统、在执行任务时有效地协助人们的智能助手。然而，当前 AI 算法的工作方式与人们学习并执行任务的方式之间存在着巨大差距。人们能通过几个示例学习，或者通过接收正式指令和“提示”来执行任务，或者通过观察其他人执行任务。医学院采取这种方法，例如，医学院的学生通过观察有经验的医生完成复杂的手术来学习。即使在诸如世界冠军级围棋游戏中，大师级玩家只需几千场游戏来训练自己。相比之下，人类需要花费数百年的时间来完成训练 AlphaGo 所需的游戏次数。关于实现类人类 AI 新方法的更多基础研究将使得这些系统更接近这一目标。

例 2：NSF 资助的安全游戏理论框架

安全性是全世界的一个关键问题，无论是保护港口，机场和其他关键基础设施的挑战；保护濒危野生动植物，森林和渔业；压制城市犯罪；或网络空间的安全性。不幸的是，有限的安全资源在任何时候都需要保障完整的安全性能；我们必须优化有限安全资源的使用。



美国海岸警卫队



全球应用



洛杉矶警署



洛杉矶国际机场警察

许多类型的应用程序可能受益于实现安全性的游戏理论方式。

为此，“安全游戏”框架 - 基于计算机游戏理论的基础研究，同时也纳入了人类行为建模，不确定性下的 AI 规划以及机器学习的元素 - 在美国及世界各地为安全机构而建立和部署的决策辅助工具。⁷⁴ 例如，ARMOR 系统自 2008 年以来一直部署在 LAX 机场，美国联邦航空公司服务部门的 IRIS 系统自 2009 年以来一直在使用中，还有自 2011 年以来美国海岸警卫队一直在用的 PROTECT 系统。通常，给定有限的安全资源（例如，船只，空中巡警，警察）和大量不同值的目标（例如，不同的航班，机场处的不同终端），基于安全游戏的决策辅助会提供随机分配或巡逻计划，并考虑到不同目标的权重以及敌对目标对于不同安全态势的智能反应。

这些应用程序在不同安全机构（使用各种衡量指标，例如俘获率，红色警示队，巡逻计划随机性等）的工作表现中被证明显著改进了安全性能。⁷⁴

7、开发更强大和更可靠的机器人

在过去十年中机器人技术的重大进步对多种应用产生影响力，包括制造，物流，医药，医疗保健，国防和国家安全，农业和消费品。

虽然在历史上，机器人被设想用于静态工业环境，但最近的技术发展使得机器人与人类能够亲密合作。现今，机器人技术已如预期地展示出它们拥有补充、增加、增强或模拟人类身体能力或人类智力的能力。然而，科学家需要使这些机器人系统更强大、更可靠和更易使用。

研究人员需要更好地了解机器人的感知，以便从各种传感器中提取信息，并让机器人实时感知周围环境。机器人需要有更先进的认知和推理能力，这样机器人更好地理解物理世界并与之进行交互。适应和学习能力的提高将使机器人能够自我总结和自我评估，并从人类导师那里学习肢体运动。移动性和操作性是有待进一步研究的领域，这样一来机器人可以穿越崎岖的不确定性地形并

灵活应对各种物体。机器人需要学会以无缝方式进行团队合作，并以可信赖和可预测的方式与人类合作。

8、推动人工智能的硬件升级

虽然 AI 研究最常见的是与软件发展相关，但 AI 系统的性能在很大程度上取决于硬件。眼下深度机器学习的复兴与基于 GPU 的硬件技术进步及其改进的存储器、⁷⁵输入/输出、时钟速度、并行性和能量效率直接相关。针对 AI 算法而优化的开发硬件将实现比 GPU 更高的性能水平。其中一个实例是“神经形态”处理器，其受到大脑组织的自由启发，⁷⁶并且在一些情况下，针对神经网络的运行而对其进行优化。

硬件升级还可以提高数据高度密集型 AI 方法的性能。需要进一步研究在整个分布式系统中以受控方式打开和关闭数据通道的方法。还需要继续研究以使得机器学习算法能够高效地从高速数据中获取信息，包括从多个数据通道同时学习的分布式机器学习算法。更先进的基于机器学习的反馈方法将允许 AI 系统对来自大规模仿真、实验仪器和分布式传感器系统（如智能建筑和物联网（IoT））的数据进行智能采样或优先级排序。这样的方法可能需要动态 I/O 决策，其中基于重要性实时地作出选择来存储数据，而不是简单地以固定频率存储数据。

9、为改进的硬件创建人工智能

虽然改进的硬件可以产生更强大的 AI 系统，AI 系统也可以提高硬件的性能。⁷⁷这种互惠将引导硬件性能的进一步提高，因为解决计算的物理限制需要新的硬件设计方法。⁷⁸基于 AI 的方法可能对于改进高性能计算（HPC）系统的操作尤其重要。这样的系统消耗大量能量。

AI 用于预测 HPC 性能和资源使用，并进行在线优化决策以提高效率；更高级的 AI 技术可以进一步提高系统性能。AI 还可用于创建可自动重新配置的 HPC 系统，其可以在无人干预的情况下处理发生的系统故障⁷⁹

改进的 AI 算法可以通过减少处理器和存储器之间的数据移动来提高多核系统的性能。这是通向百亿次级计算系统的主要障碍，而这种系统比现在的超级计算机运行速度要快 10 倍。⁸⁰实际上，HPC 系统中的执行配置从来不相同，并且同时执行不同的应用，其中每个不同软件代码的状态随着时间独立地演进。

设计 AI 算法需要使其能够为 HPC 系统在线运行和大规模运行。

（注：进入百万兆级运算是指计算系统可实现至少每秒 10 亿次运算。）

（二）战略二：开发有效的人类与人工智能协作方法

虽然完全自主的 AI 系统在一些应用领域（例如，水下或深空探测）中将是比较重要的，但是许多其它应用领域（例如，灾难恢复和医学诊断）是通过人类和 AI 系统的组合协作得到最有效地解决而实现应用目标的。这种协作互动利用了人类和 AI 系统的互补性质。虽然针对人类- AI 协作的有效方法已经存在，但大多数是“单点解决方案”，只能奏效于特定环境中，并使用特定平台来实现特定目标。针对每个可能的应用程序实例来生成点解决方案是不能量化的；因此，需要更多的工作来越过这些点解决方案，并倾向于人类-AI 协作中更具一般性的方法。需要在设计“通用系统”和建立大量的“专用系统”之间进行权衡。前者能够在所有类型的问题中起到作用，构建所需人力较少且能更加便利地不同应用之间进行切换；而后者可以针对每个问题更有效率地发挥作用。

未来应用程序将在人类和 AI 系统之间的功能作用划分，人与 AI 系统之间的相互作用性质，人类和其他 AI 系统协同工作的数量以及人类和 AI 系统交流及共享情境意识的方式上存在大大地不同。人类和 AI 系统之间的功能作用划分通常属于以下类别之一：

（1）AI 执行辅助人类的功能：AI 系统执行支持人类决策者的外围任务。例如，AI 可以帮助人类检索工作记忆，短期或长期记忆，以及预测任务。

（2）AI 执行分担人类高认知负荷的功能：当人类需要帮助时，AI 系统执行复杂的监视功能（例如飞机中的地面接近警告系统），决策和自动医疗诊断。

（3）AI 执行代替人类的功能：AI 系统执行人类对其能力非常有限的任务，例如用于复杂的数学运算，用于争议性操作环境中的动态系统控制引导，用于有害或有毒环境中的自动化系统的控制方面，以及用于系统应非常快速地响应的情况下（例如，在核反应堆控制室中）。

实现人类与 AI 系统之间的有效交互需要额外的研发，以确保系统设计不会导致过度的复杂性，置信不足或置信过度。可以通过培训和体验来增加人类对人工智能系统的熟悉程度，以确保人类能很好的了解人工智能系统的功能以及

人工智能系统能够和不能够做什么。为了解决这些问题，在设计和开发这些系统时应使用某些以人为中心的自动化原则：⁸¹

（1）对人工智能系统的界面、控制和显示，采用直观、人性化的设计。

（2）保持操作人员对信息的及时了解。显示关键信息，人工智能系统的状态以及对这些状态的更改。

（3）对操作人员进行培训。参与一般性知识、技能和能力（KSA）的定期复训，以及进行人工智能系统采用的算法和逻辑和系统的预期故障模式的培训。

（4）确保自动化设备灵活。对于希望是否决定使用人工智能系统的操作人员而言，部署人工智能系统应被视为计划选项。同样重要的是，在过度工作负荷或疲劳期间，用于支持人类操作人员的自适应人工智能系统的设计和运用也是非常重要的。^{82, 83}

当创建能与人类进行有效合作的系统时，研究人员会遇到许多基本挑战。以下小节中列出了其中一些重要的挑战。

1、寻找人类感知人工智能的新算法

多年以来，人工智能算法已经能够解决日益复杂的问题。然而，这些算法的功能和人类对这些系统的可用性之间存在一定的差距。人类感知智能系统需要能够直观地与用户进行互动，并且能够实现无缝人机协作。直观的互动包括浅层互动，例如当用户舍弃由系统推荐的选项时；基于模型方法，考虑用户过去的行为；或甚至基于准确的人类认知模型的用户意图的深层模型。必须开发人为干预模型，允许智能系统仅在必要和适当时介入人类操作。智能系统还应该具有增强人类认知的能力，使用户在需要时知道需要检索哪些信息，即使他们没有明确地向系统提示该信息。未来的智能系统必须能够解释人类社会的行为规范，并相应地采取行动。如果智能系统具有一定程度的情感智能，则智能系统可以更有效地与人类一起协作，使得他们可以识别用户的情绪并做出适当地响应。另一个研究目标是超越单人和单台机器的互动，朝向“系统的系统（systems-of-systems）”方向发展，即由多个机器与多个人互动的协同工作。

人类人工智能系统的互动具有广泛的目标。人工智能系统需要能够代表多

个目标，他们为达到这些目标可以采取的行动，对这些行动的约束和其他因素，以及容易适应目标的修改。此外，人类和人工智能系统必须共享共同目标，并相互了解他们和他们当前状态的相关方面。需要进一步研究来概括人类人工智能系统的这些方面，以开发需要较少人体工程学的系统。

2、开发增强人类能力的人工智能技术

尽管人工智能研究先前的大多数焦点是关于匹配或胜过人类执行狭窄任务的算法，但是需要额外的工作来开发在许多领域增强人类能力的系统。人类增强研究包括固定设备（例如计算机）上使用的算法；可穿戴设备（如智能眼镜）；植入装置（如大脑连接）；以及特定的用户环境（例如特制的手术室）。例如，增强的人类意识可以使医疗助手，基于多个设备组合的数据读数指出医疗过程中的错误。其他系统可以通过帮助用户回忆，适用于用户当前状况的过去经验来增强人类的认知。

人类和人工智能系统之间的另一种类型的协作，涉及对智能数据理解的主动学习。在主动学习中，从领域专家处寻求输入，并且当学习算法不确定时仅对数据进行学习。这是减少需要首先生成的训练数据量或需要学习的量的重要技术。主动学习也是获得领域专家输入和提高学习算法信任度的关键方法。主动学习迄今为止只在监督式学习中使用——需要进一步的研究将主动学习纳入无监督学习（例如集中、异常检测）和强化学习中去。⁸⁴ 随机网络允许领域知识以先验概率分布的形式包含在内。必须寻求允许机器学习算法并入领域知识的一般方法，无论是以数学模型、文本或其他形式。

（注：84. 监督学习需要人类提供正确的结果，而强化学习和无监督学习则不需要。）

3、开发可视化和人机界面技术

更好的可视化和用户界面是需要更多开发，以帮助人们了解大量现代数据集和来自不同信息来源的额外领域。可视化和用户界面必须以人类可理解的方式，清楚地呈现来自他们的日益复杂的数据和信息。在安全危急操作中提供实时结果是重要的，并且可以通过增加计算能力和互连系统来实现。在这些类型的情况下，用户需要可以为实时响应而快速传达正确信息的可视化和用户界面。

人类人工智能系统的协作可以应用于各种环境中，并且对通信存在约束。

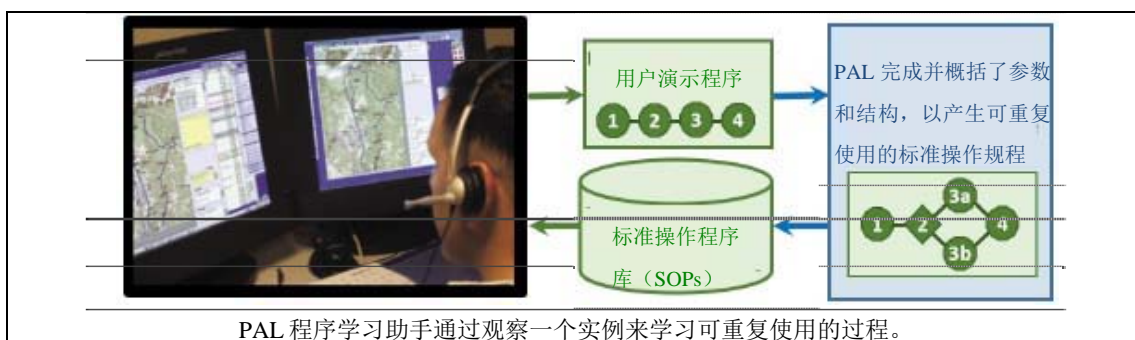
在一些领域，人类人工智能通信延迟低，通信快速可靠。在其他领域（例如，美国航空航天局布置在火星上的勇气号和机遇号探测器），人与人工智能系统之间的远程通信的延迟时间非常长（例如，地球和火星之间的往返时间为 5-20 分钟），因此需要部署能在很大程度上自主操作，且只向其传送高级别战略目标的平台。这些通信要求和约束是研发用户界面的重要考虑因素。

4、开发更高效的语言处理系统

使人们能通过口语和书面语言与人工智能系统进行互动一直是人工智能研究人员的目标。虽然已经取得了重大进展，但是在人与人工智能系统能像人与人之间进行有效沟通之前，必须在语言处理中解决相当大的开放性研究挑战。语言处理的最新进展已经归功于使用数据驱动的机器学习方法，其产生了成功的系统，例如，在安静的环境中成功的实时识别流利的英语语音。然而，这些成就只是实现更长期目标的第一步。当前系统不能处理现实世界的挑战，例如在嘈杂环境中的语音、带浓重口音的语音、儿童语音、受损的语音和手语语音。还需要开发能够与人进行实时对话的语言处理系统。这种系统需要推断人类对话者的目标和意图，使用针对当前情况适当的语境、风格和修辞，并在对话产生误解的情况下对其进行修复。需要进一步研究更易于普及不同语言的系统。此外，需要更多的以语言处理系统更容易访问的形式研究，来获得有用的结构化领域知识。

例 3：DARPA 的可学习个性化助理（PAL）项目创建了苹果商业化技术，如 Siri

计算技术对现代生活的每一个方面都至关重要，但我们每天使用的信息系统缺乏人类认知的一般的、灵活的能力。在 PAL 计划中，⁸⁵ DARPA 着手创建可以从经验、推论中学习的认知助手，并通过语音界面告诉他们该做什么。DARPA 设想了 PAL 技术，使得信息系统对于用户更加高效和有效。DARPA 和 PAL 的执行者与军事运营商合作，将 PAL 技术应用于命令和控制问题，PAL 程序学习技术被整合到未来版本的战斗通信和 10（见图）的美国军队指挥所中，并在世界各地使用。



DARPA 还非常了解 PAL 技术的商业潜力，特别是需要基于语音的智能手机交互的移动应用。DARPA 强烈鼓励 PAL 商业化，并且为了响应 DARPA 的鼓励，在 2007 年创建了 Siri 公司，以便在通过基于语音界面管理信息和自动化任务来帮助用户的系统中使 PAL 技术商业化。2010 年 4 月，Siri 公司被苹果收购，该公司进一步开发了这些技术，使其成为用于 iPhone 和 iPad 的 Apple 移动操作系统的一个组成部分，也是其定义的特征。

在许多其他领域的语言处理研究方面的进展，也需要使人类和人工智能系统之间的互动更加自然和直观。必须为语言和书面语言的模式建立稳健的计算模型，其为情绪状态、情感和立场提供根据，并确定语言和文本中隐含的信息。人工智能系统环境文本中的基础语言，需要能物理世界中操作的新的语言处理技术，例如在机器人中。最后，由于人们在线互动交流的方式，可能与语音互动完全不同，因此完善这些文本中使用的语言模式，使得社交人工智能系统可以更有效的与人进行互动。

（三）战略三：了解并解决人工智能的伦理、法律和社会影响

当人工智能自主代理行为时，我们期望他们根据我们人类正式和非正式的规范行事。作为基本的社会秩序的力量、法律和道德，既能告知也能判断人工智能系统的行为。首要研究需要涉及对道德、法律的理解，以及人工智能的社会影响，且人工智能设计的开发方式必须符合道德规范、法律和社会原则。还必须考虑隐私问题；有关这一问题的进一步信息可以在《国家隐私研究和发展战略》中找到。

与任何技术一样，法律和道德原则将告知人工智能可接受的用途；是如何将这些原则应用于这项新技术所面临的挑战，特别是那些涉及自主性、代理和控制的技术。如《强健和有益的人工智能研究重点》（Research Priorities for Robust and Beneficial Artificial Intelligence）所示：“为了构建稳健的表现良好的系统，我们当然需要在每个应用领域中确定什么是良好的行为方式。这种伦理维度与工程技术可用的问题，这些技术如何可靠，以及作出了怎样的取舍——所有计算机科学领域、机器学习和更广泛可用的人工智能专业知识是紧密相关的。”⁸⁶

该领域的研究可以受益于来自涉及计算机科学、社会和行为科学、伦理、生物医学科学、心理学、经济学、法律和政策研究专家的多学科视角。需要对 ITRD 相关 IT 领域（即信息技术领域以及上述学科）的内部和外部领域进行进一步的调查研究，以便为人工智能系统的研发和使用及其对社会的影响提供信息。以下小节探讨了该领域中关键信息技术研究的挑战。

1、改进公平性、透明度和设计责任机制

人们对数据密集型人工智能算法出错和滥用的敏感性，以及对性别、年龄、种族或经济类可能产生的影响表示了许多关注。在这方面，适当收集和使用人工智能系统的数据是一个重要挑战。然而，除了纯粹的数据相关问题，出现在人工智能设计上的更大的问题本质上是公正、公平、透明和负责。研究人员必须学会如何设计这些系统，以使他们的行动和决策是透明且是容易被人解释的，因此可以检查其可能包含的任何偏差，而不仅仅是学习和重复这些偏差。如何表示和“编码”人类价值和信仰体系是重要的研究课题。科学家们还

必须研究可以在什么程度上将正义和公平的考虑设计到系统中去，以及如何在当前工程技术范围内实现这一点。

2、建立符合伦理的人工智能

除了正义和公平的基本假设之外，还有人担心人工智能系统是否能够表现出一般伦理原则所容忍的行为。如何改进人工智能框架在道德伦理中新的“机器相关”问题，或什么用途的人工智能可能被认为是不道德的？伦理在本质上是一个哲学问题，而人工智能技术依赖于并受到工程的限制。因此，在技术可行的范围内，研究人员必须努力开发与现有法律、社会规范和道德伦理一致或相符的算法和架构——这显然是一项非常具有挑战性的任务。伦理原则通常有不同程度的模糊性，并且难以转化为精确的系统和算法设计。当人工智能系统，尤其是新的自主决策算法，面临基于独立和可能冲突的价值体系道德困境时，还有一些混乱。伦理问题因文化、宗教和信仰而异。然而，可以制定可接受的道德参考框架，以指导人工智能系统进行推理和决策，以解释和证明其结论和行为。需要一种多学科方法来生成反映适当价值体系的训练数据集，包括当存在道德问题或冲突价值困难时，显示首选行为的实例。这些实例可以包括法律或道德的“少见或极端案件（corner cases）”，由对用户透明的结果或判断进行标记。⁸⁷ 人工智能需要适当的方法来解决价值冲突，在严格规则行不通的地方，该系统结合的原则可以解决复杂情况下的实际情况。

3、设计符合伦理的人工智能架构

必须在基础研究方面取得额外的进展，以确定如何最好地设计包含道德推理的人工智能系统架构。已提出了各种方法，例如两层监控架构，其将操作 AI 从负任何操作行动的伦理或法律评估的监视代理中分离。⁸⁷ 另一种观点是倾向于选择安全工程——使用用于 AI 代理体系结构的精确概念框架来确保 AI 行为是安全且对人无害。⁸⁸ 第三种方法是使用集合理论原则来构成伦理体系结构，这结合了限制行动以符合道德原则的 AI 系统行为逻辑限制。⁸⁹ 随着 AI 系统变得更加普遍，他们的架构或许包括能在多个级别承担道德问题的子系统，其中包括：⁹⁰ 匹配规则的快速响应模式、用于放慢描述和辩护行为反应的审议推理、显示用户可信度的社会信号，以及社会历程——其会运作超过甚至更长的时间尺度以便系统能遵守文化规范。研究人员需要关注如何能最好地处理符合道德、

法律和社会目标的人工智能系统的整体设计。

（四）战略四：确保人工智能系统的安全可靠

在广泛使用 AI 系统之前，需要确保系统以受控的方式安全和可靠地进行操作。需要进行研究来解决这一挑战：创建可靠、真实和可信赖的 AI 系统。正如其他复杂的系统，AI 系统面临着重要的安全和安保挑战：⁹¹

（1）**复杂及不确定的环境**：在许多情况下，AI 系统是用于在复杂环境中进行操作，并存在不能进行详尽检查或测试的大量潜在状态。系统可能面临在其设计期间从未考虑过的条件。

（2）**紧急行为**：对于在部署后进行学习的 AI 系统，系统的行为可能主要由无监督条件下的学习阶段来决定。在这种情况下，可能难以预测系统的行为。

（3）**目标设定的偏差**：由于将人类目标转换为计算机指令极为困难，对 AI 系统编程的目标可能未必符合程序员预期目标。

（4）**人机交互限制**：在许多情况下，人工智能系统的性能会极大地受到人类交互的影响。在这些情况下，人类反应的变化可能会影响系统的安全性。⁹²

为了解决这些和其他问题，需要进行额外投资来提高人工智能的安全性和可靠性，⁹³其中包括可解释性与透明度、信任、验证与确认、抵御攻击的安全性以及长期的 AI 安全和数值调整。

1、提高可解释性和透明度

一项关键的研究挑战是增加 AI 的“可解释性”或“透明度”。包括基于深度学习在内的多种算法，对于用户来说是不透明的，只有很少的现有机制对它们的结果进行解释。这尤其会对诸如医疗保健等领域造成问题，在该领域，医生需要解释来验证特定诊断或治疗过程。AI 技术如决策树感应提供内在的解释，但通常不太准确。因此，研究人员必须开发透明的系统，并能在本质上向用户解释其结果的原因。

2、提高信任度

为了取得信任，AI 系统设计人员需要创建具有信息性和用户友好界面的准确、可靠的系统，而操作人员必须花时间进行充分的培训，以了解系统操作和性能限制。用户广泛信任的复杂系统（例如车辆手动控制）将趋于透明（系统

以用户可见的方式进行操作）、可信（用户接受系统的输出）、可审计（可以评估系统）、可靠（系统按用户期望行动）和可恢复（用户可以在需要时恢复控制）。当前和未来 AI 系统的重大挑战仍是不稳定的软件生产技术质量。随着发展在人类与 AI 系统之间架起更强的连接，信任领域内的挑战让变化和与日俱增的能力并驾齐驱，预期会采用和长期使用技术进步，并为研究设计、建造和使用的最佳实施制定管理原则和政策，包括为安全操作对操作人员进行适当培训。

3、增强可验证与可确认性

需要新的方法来验证和确认 AI 系统。“可验证”确定系统满足正式规范，而“可确认性”确定系统满足用户的操作需求。安全 AI 系统可能需要新的评估方法（确定系统是否发生故障，可能在预期参数以外运行时）、诊断方法（确定故障原因）和维修方式（调整系统以解决故障）。对于超过时间延长期进行自主操作的系统，系统设计者可能未考虑系统会遇到的每种条件。这种系统可能需要拥有自我评估、自我诊断和自我修复的能力，以变得稳健和可靠。

例 4：美国宇航局艾姆斯研究中心——在出现故障前预测出故障

因为基于模型的异常检测方式的不足，所以美国宇航局艾姆斯研究中心在 2003 年开发了一种数据驱动的异常检测方法，称为感应监测系统（IMS）。从那时起，其便已用于美国宇航局内的系统健康监测应用程序，包括监测航天飞机和国际空间站（ISS）以及非美国宇航局应用程序。



在 2014 年对猎户座载人飞船进行发射测试，在此期间 IMS 用于监测电气系统。



C-130 大力士军用运输机的预测性软件用来预测阀门出现的故障，该阀门用于切换引擎间的气流。

2012 年，综合工程管理解决方案（CEMSol）授权和增强了 IMS，并与美国宇航局艾姆斯研究中心和洛克希德·马丁公司合作将 IMS 作为洛克希德 C-130 大力士运输机的集成系统健康监测系统进行测试。洛克希德·马丁公司在该项测试中投资了 70,000 美元，并随即在降低的维护成本和任务延迟中收获了将近 10 倍的利润。⁹⁴

4、保护免受攻击

为了处理事故，嵌入关键系统中的 AI 必须耐用，但也应该安全，以应付大范围的蓄意网络攻击。安全工程包括了解系统的漏洞，以及有意对其进行攻击的行动者的行动。虽然在 NITRD 网络安全研发战略中心详细论述了网络安全研发需求，但一些网络安全风险是针对 AI 系统的。例如，一个关键的研究领域是“对抗机器学习”，其通过“污染”训练数据、修改算法或通过对阻碍其得到正确识别的某一目标进行微小变化（例如，欺骗面部识别系统的假肢），以此来探索对 AI 系统危害会到多大程度。在需要高度自主性的网络安全系统中实施 AI 也是一个需要进一步研究的领域。在此领域的一个最近的工作实例是 DARPA 的网络大挑战，其涉及 AI 代理自主分析和抵制网络攻击。⁹⁵

5、实现长期的人工智能安全和优化

AI 系统可能最终能“循环自我改进”，当中大量软件修改会由软件自身进行，而非由人类程序员进行。为了确保自我修改系统的安全性，需要进行额外研究来进行开发：自我监测架构通过人类设计者的原始目标来检查用于行为一致性的系统；限制策略用于防止系统在评估期间进行释放；在价值学习中，用户值、目标或意图可以由系统进行推断；并且可证明价值架构能抵抗自我修改。

（五）战略五：开发用于人工智能培训及测试的公共数据集和环境

AI 的益处将持续增加，但仅限于其培训和测试资源得以发展到可利用的阶段。训练数据集和其他资源的多样性、深度、质量和准确性大大影响 AI 性能。许多不同的 AI 技术要求培训和测试使用高质量的数据，以及动态的交互式测试平台和模拟环境。如果 AI 培训和测试仅限于已经拥有有价值数据集和资源的几个实体，那么这不只是一个技术问题，还是发展会经历的一个重大“公共事业”挑战，但我们必须同时尊重商业和个人权利和数据利益。需要进行研究来为各种 AI 应用开发出高质量的数据集和环境，并负责获取良好的数据集和测试和培训资源。还需要额外的开源软件库和工具包来加速 AI 研发的前进。以下小节概述了这些关键的重要领域。

1、开发满足多样化人工智能兴趣与应用的丰富数据集

AI 培训和测试数据集的完整性和可用性对确保科学的可靠结果至关重要。支持数字领域可再现研究所需的技术和社会技术基础设施已视为是一个重要挑战，并且对 AI 技术也是至关重要的。缺乏具有确认来源的经审查和可用公开数据集来保障再现性，是影响 AI 自信发展的关键因素。⁹⁶ 如在其他数据密集型科学，捕获数据源是至关重要的。研究人员必须能使用相同和不同的数据集来重现结果。数据集必须能代表具挑战性的真实应用，而不仅仅是简化版本。

为了快速取得进展，应集中提供政府持有的现有数据集，那些可以通过联邦资金得以开发的数据，并尽可能提供由工业界持有的数据集。

机器学习方面的 AI 挑战通常与“大数据”分析有关。考虑到各种相关数据集，其仍是一个需要适当陈述、使用 and 进行非结构化或半结构化数据分析的日益突出的挑战。如何使用绝对和相对术语来表示数据（上下文相关的）？当前真实世界的数据库可能极易受到不一致、不完整和嘈杂数据的影响。因此，许多数据预处理技术（例如，数据清理、集成变换、简化和表达）对建立 AI 应用的有用数据集是非常重要的。数据预处理如何影响数据质量，特别是在执行附加分析时？

鼓励分享专用于政府资助研究的 AI 数据集，可能会促成创新的人工智能方

法和解决方案。然而，需要技术来确保数据的安全共享，因为当数据所有者与研究团体共享数据时会承担风险。数据集的开发和共享还必须遵守适用的法律法规，并以合乎道德的方式进行。风险会以各种方式产生：不当使用数据集、不准确或不当披露、以及限制数据认同技术来确保隐私和机密性保护。

2、开放满足商业和公共利益的训练测试资源

随着全球数据、数据源和信息技术的不断扩大，数据集的数量和大小也在不断增加。分析数据的技术和科技未能迎合大量的原始信息源。数据捕获、管理、分析和可视化都是重要的研究挑战，而从大量数据中提取科学所需的有价值知识正处于落后。虽然存在数据存储库，但它们通常无法处理数据集的扩展，它们具有有限的数据源信息，且不支持语义丰富的数据搜索。需要动态和灵活的资源库。

用于支持 AI 研究需求的公开/共享基础设施计划中的一个例子是由国土安全部（DHS）开发的 IMPACT 计划（网络风险和信托政策和分析信息市场）。⁹⁷ 此计划通过协调和开发真实数据和信息共享能力（包括工具，模型和方法）来支持全球网络安全风险研究工作。IMPACT 还支持国际网络安全研发社区、关键基础设施提供商及其政府支持者之间的实证数据共享。AI 研发将得益于所有 AI 应用程序中的类似程序。

3、开发开源软件库和工具包

增加开源软件库和工具包的可用性，以此通过网络连接为任何开发人员提供最先进的 AI 技术接入。其他资源中的诸如 Weka 工具包、⁹⁸ MALLET、⁹⁹ 和 OpenNLP、¹⁰⁰ 等资源，加速了 AI 的开发和应用。开发工具，包括免费或低成本的代码存储库和版本控制系统，以及免费或低成本开发语言（例如 R、Octave 和 Python）为使用和扩展这些储存库提供了低准入。此外，对于那些可能不想直接集成这些储存库的人来说，任何数量的云基础机器学习服务，可以通过极少或无需使用编程的低延时网络协议来执行诸如图像分类之类的任务。最后，众多这种 Web 服务还提供专用硬件，包括基于 GPU 的系统。可以合理地假设用于 AI 算法的专用硬件，包括神经形态处理器，也将通过这些服务而变得随手可得。

这些资源共同提供了一种 AI 技术基础设施，该基础设施通过允许企业家开

发解决狭窄领域问题的解决方案来促进市场创新，且无需昂贵的硬件或软件，无需高水平的 AI 专业知识，并允许按需快速扩展系统。对于狭窄的 AI 领域，相对于众多其他技术领域，市场创新的界限极低。

为了支持这一领域的持续高水平创新，美国政府会大力推进开发、支持和使用开放式 AI 技术。最为有益的是开源，其使用标准化或开放格式和开放标准来表示语义信息，其中包括域本体当可用时。

政府还可以通过在政府内部加快使用开放 AI 技术，鼓励开放 AI 资源更大范围的应用，从而帮助创新者维持进入的低门槛。政府应尽可能为开源项目提供算法和软件。因为政府存在具体顾虑，比如更加强调数据的隐私和安全，因此政府需要制定机制，以缓冲人工智能系统的采用。例如，创建一个可以跨政府机构执行“水平搜索”的特别小组，从而在部门间查找特定的 AI 应用程序区域，再确定需要解决的具体问题，以允许采用这些机构的此类技术，该方法可能很有用。

（六）战略六：制定标准和基准以测量和评估人工智能技术

标准、基准、测试平台以及 AI 社群应用对于指导和促进 AI 技术的研发至关重要。以下小节概述了必须取得更大进展的领域。

1、开发广泛应用的人工智能标准

必须加快制定标准，以跟上性能快速发展和 AI 应用领域不断扩大的步伐。标准可以提供能够持续使用的要求、规范、指导或特性，以确保 AI 技术实现功能和互操作性的关键目标，并使其执行可靠且安全。标准的采用能够为技术进步带来可信度，并帮助扩大互操作性市场。已制定的 AI 相关标准的一个示例是 P1872-2015（机器人和自动化的标准本体论），它由电气和电子工程师协会（IEEE）共同制定。该标准提供了知识表示的系统方法以及一套通用的术语和定义。这些使得人、机器人和其他人工系统之间能够进行明确的知识转移，并为机器人 AI 技术的应用奠定基础。AI 的所有子域都需要 AI 标准开发取得更大进步。

需要开发标准解决：

- （1）**软件工程**：管理系统复杂性、维持性、安全性、监测并控制紧急行为；
- （2）**性能**：确保准确性、可靠性、稳健性、可访问性和可扩展性；
- （3）**指标**：量化影响性能和遵守标准的因素；
- （4）**安全**：评估系统、人机交互、控制系统和法规遵从性的风险管理和危害分析；
- （5）**可用性**：确保接口和控件有用、高效和直观；
- （6）**互操作性**：通过标准和兼容性接口定义可互换组件、数据和事务模型；
- （7）**安全**：处理信息的保密性、完整性和可用性，以及网络安全；
- （8）**隐私**：在处理、传输或存储时控制信息保护；
- （9）**可追溯性**：提供事件记录（实施、测试和完成），以及数据管理；
- （10）**特定领域**：定义特定领域的标准词汇和相应框架

2、制定人工智能技术的测试基准

由测试和评估组成的基准为制定标准和评估标准遵从性提供了量化措施。基准通过促进旨在解决战略性方案精选的进展推动创新；它们还提供客观数据，以跟踪 AI 科学和技术的演变。为了有效评估 AI 技术，必须开发相关的有效测试方法和指标并使其标准化。标准测试方法将规定用于评估、比较和管理 AI 技术性能的协议和程序。需要标准指标来定义可量化的衡量标准，以便表征 AI 技术，其中包括但不限于：准确性、复杂性、信任和能力，风险与不确定性；可解释性；意外偏差；与人类表现相比较；以及经济影响。值得注意的是基准是由数据驱动的。策略五讨论了数据集在培训和测试中的重要性。

作为 AI 相关基准的成功例子，国家标准与技术研究所（NIST）已开发一套全面的标准测试方法和相关的性能指标，用于评估应急救援机器人的关键能力。其目的是通过使用通过标准测试方法获得的机器人能力的统计学显著数据来促进不同机器人模型的定量比较。这些比较可以指导购买决策，并帮助开发人员了解部署能力。通过用于机器人操作设备的国家安全应用的 ASTM 国际标准委员会而得到的测试方法正被标准化（被称为标准 E54.08.01）。测试方法的版本可通过 RoboCup 救援机器人组的竞争来挑战研究社区，¹⁰¹ 这强调的是自主能力。另一个例子是用于工业自动化竞争的灵活机器人（ARIAC）¹⁰²，这是 IEEE 和 NIST 之间共同努力的结晶，它通过利用人工智能和机器人规划的最新进展，以提高机器人灵活性。这次比赛的核心重点是测试工业机器人系统的灵活性，目标是使车间员工的工作更高效、更自主，并节省车间工人时间。

虽然这些工作为推动 AI 基准发展奠定了坚实的基础，但它们仍受限于特定领域。而更为广泛的领域需要额外的标准、测试平台和基准，以确保 AI 解决方案的广泛适用和采用。

3、增加可用的人工智能测试平台

《未来的网络实验》（Cyber Experimentation of the Future）报告中指出了测试平台的重要性：¹⁰³ “测试平台是至关重要的，因为研究人员可凭此使用实际操作数据和良好测试环境中的方案对现实世界系统建模并进行实验...。所有 AI 领域都需要具备足够的测试平台。政府拥有大量政府特有的任务敏感数据，但其中大部分数据不能分发给外部研究团体。为学术和工业研究人员建立适当的程序，用于在由特定机构建立的安全和策划测试平台环境中进行研究。

AI 模型和实验方法可通过访问这些测试环境，由研究团体进行共享和验证，除此之外无法为 AI 科学家、工程师和学生提供独特的研究机会。

4、促进人工智能社群参与标准和基准的制定

需要政府的领导和协调，促使其标准化并鼓励其在政府、学术界和行业中广泛使用。由用户、行业、学术界和政府组成的 AI 社群必须积极参与开发标准和基准程序。由于每个政府机构会根据其作用和使命以不同方式参与该社群，因此可协调利用社群互动，从而加强其影响。需要这种协调聚集用户驱动的需求，预测开发人员驱动的标准，并提升教育机会。用户驱动的需求塑造了挑战问题的目标和设计，使技术评估成为可能。集中研发社群基准，进而明确进度、缩小差距、并推动具体问题创新解决方案的发展。这些基准必须包括定义和设置地面实况的方法。基准仿真和分析工具的创建同样可加快 AI 的发展。这些基准结果同样有助于根据用户需要匹配正确技术，形成用于遵循标准、合格产品列表和潜在资源选择的客观准则。

行业和学术界是新兴 AI 技术的主要来源。这对于促进和协调其参与标准和基准活动是至关重要的。随着解决方案的出现，通过共享技术架构的共同愿景，很大机会可以预期实现开发者和用户驱动的标准，发展新标准的参考实施以说明其可行性，并通过竞争前测试，以确保解决方案的高质量和可操作性和开发技术应用的最佳实践。

高影响力、基于社区、与 AI 相关的基准程序的一个成功例子是文本检索会议（TREC）¹⁰⁴，它由 NIST 于 1992 年启动，用于为大规模评估信息检索方法提供必要的基础设施。参加 TREC 的团体已超过 250 个，其中包括大型和小型的学术和商业组织。

由 TREC 提出的标准、广泛适用、精心构造的数据集已被认为是振兴信息检索的研究。^{105, 106}第二个例子是在机器视觉识别领域用于生物统计学的 NIST 周期性基准程序，¹⁰⁷特别是面部识别技术（FERET）。其于 1993 年开始用于评估人脸识别技术，¹⁰⁸其提供了人脸照片的标准数据集，旨在支持面部识别算法开发和评估协议。这项工作经过多年的努力已经进入面部识别供应商测试（FRVT）阶段，¹⁰⁹其中包括数据集分布、大量挑战问题、以及分离技术评估。这个基准程序为面部识别技术的改进做出了重大贡献。TREC 和 FRVT 都可以作为 AI 相关

社区基准活动的有效示例，但在 AI 其他领域需要进行类似的工作。

值得注意的是，制定和采用标准以及参与基准活动都需要付出代价。当他们看到显著效益时，可使研发机构获得激励。更新机构间的采购流程以包括建议书中请求纳入的 AI 标准具体要求，这将激励社区进一步参与标准的制定和采用。基于社区的基准，如 TREC 和 FRVT，同样通过提供本来不可见的培训和测试数据类型，降低壁垒并加强激励措施，促进技术人员之间的健康竞争，从而获得最佳算法，并提供用于相关资源选择的目标和比较性能指标。

（七）战略七：更好地了解国家人工智能人力需求

要获得该战略中概述的所需 AI 研发进展，需要足够的 AI 研发人员。AI 研发领域最强的国家将在未来的自动化中确立领先地位。他们将成为能力要求（如算法创建和开发）、能力验证和商业化的领跑者。开发技术专长将为这些进步奠定基础。

虽然目前不存在官方的 AI 人力资源数据，但最近大量的商业和学术部报告表明，在 AI 领域中，专家的短缺现象越来越严峻。据报道，AI 专家短缺，¹¹⁰且这个需求预计将持续增加。⁶⁶据报道，高科技公司投入大量资源，用于招募具有 AI 专业知识的教师和学生。¹¹¹据报道，大学和工业进入招募和留用 AI 人才的激战中。¹¹²

需要开展更多研究，以更好了解 AI 研发在当前和未来的国家劳动力需求。需要数据说明 AI 研发人员的当前状况，包括学术界、政府和行业的需求。研究应探索 AI 工作场所的供求力量，从而帮助预测未来的劳动力需求。需要了解预计 AI 研发人员的来源。应将教育途径和潜在再培训机会考虑在内。

还应探讨多元化问题，因为研究表明，一个具备多元化信息技术的工作人员可以改进成果。¹¹³一旦更好了解当前和未来 AI 研发人员的需求，就可以考虑实施适当的计划和行动，以解决现有或预期的劳动力挑战。

三、建议

联邦政府可以全面支持本计划的七个战略重点，并通过支持以下建议实现其愿景：

1、建议一

开发一个人工智能研发实施框架，以确定机遇，并支持人工智能研发投资的有效协调，与本计划的第一至六项战略保持一致。

联邦机构应通过 NITRD 合作开发一个研发实施框架，促进本计划所述研发挑战的协调和进展。这将使各机构能够轻松规划、协调和合作支持该战略计划。实施框架应根据其任务、能力、权威和预算，将各机构的研发重点考虑在内。根据实施框架，可能需要建立融资计划，用于协调执行 AI 的国家研究议程。为了帮助实施这一战略计划，NITRD 应考虑成立一个以人工智能为重点、与现有工作组相协调的联合工作组。

2、建议二

研究创建和维持健康的人工智能研发队伍的国家图景，与本计划的战略第七项保持一致。

一个健康且充满活力的 AI 研发队伍对于解决本报告所述研发战略挑战是非常重要的。虽然一些报告表明 AI 研发专家可能越来越短缺，但未有官方劳动力数据说明 AI 研发人员的当前状态、预计的人力资源来源以及 AI 劳动力的供求力量。鉴于 AI 研发人员在解决本计划中确定的战略重点方面的作用，因此需要更好了解如何获得和/或维持一个健康的 AI 研发人员队伍。NITRD 应该研究如何最好地说明和定义当前和未来的 AI 研发人员需求，并发展其他研究或建议，从而确保得到足够的研发人员以解决国家的 AI 需求。如研究结果所示，适当的联邦机构应该采取措施，确保创建并维护一个健康的国家 AI 研发人员队伍。

附录：首字母缩写词

3-D	三维
AI	人工智能
ANNs	人工神经网络
ARIAC	用于工业自动化竞赛的敏捷机器人
ARMOR	路由随机监测助理
ASTM	美国测试和材料协会
ATM	自动取款机
BRAIN	美国创新神经技术脑研究计划
CEMSol	综合工程管理方案
COMPETES	《创造机遇，显著提升美国科技教育领域优势地位》法案
CoT	技术委员会
DARPA	美国国防部高级研究计划局
DHS	美国国土安全部
DoD	美国国防部
DOE	美国能源部
DOT	美国运输部
FERET	人脸识别技术
FRVT	人脸识别供应商测试
GPS	全球定位系统
GPU	图形处理单元
HPC	高性能计算
I/O	输入/输出
IBM	国际商业机器公司
IEEE	美国电气和电子工程师协会
IMPACT	用于政策和分析网络风险和信任的信息市场
IMS	感应式监测系统
IoT	物联网
IRIS	国际智能随机调度（系统）
ISS	国际空间站
IT	信息技术
KSA	知识、技能和能力
LAX	洛杉矶世界机场
MALLET	一种开源的用于机器学习的语言软件包
NASA	美国国家航空和航天局
NCO	NITRD 国家协调办公室
NIH	美国国家卫生研究院
NIST	美国国家标准与技术协会
NITRD	网络信息技术研究与开发（小组委员会）
NLP	自然语言处理
NRL	美国海军研究实验室
NSF	国家科学基金会
NSTC	国家科学技术委员会
OMB	管理和预算办公室
OSTP	美国白宫科学技术政策办公室

PAL	具有学习能力的个性化助理
PROTECT	打击恐怖主义的港口防御作战/战术的实施
R&D	研究和发展，研发
RFI	信息邀请书
S&T	科学和技术
STEM	科学、技术、工程和数学
TREC	文本检索会议
U.S.	美国

译者注

2016 年 10 月 13 日，美国白宫科技政策办公室（OSTP）下属国家科学技术委员会（NSTC）发布了《国家人工智能研究与发展战略计划》（National Artificial Intelligence Research and Development Strategic Plan）。本报告深入探讨了人工智能的发展现状、应用领域以及潜在的公共政策问题，并提出了美国优先发展的人工智能七大战略方向及两方面建议，对我国人工智能产业发展具有重要的借鉴意义。

根据原文的版权信息，本报告为美国政府的工作成果，属于公共领域，可自由翻译与传播。此中文译本由中国信息通信研究院政策与经济研究所编译组编译整理，受译者水平和翻译时间所限，文中编译错误与不足在所难免，仅供读者参阅，敬请指正。依据原文版权的公共性，此中文译本也可自由分发和复制。本文翻译的正确性由译者承担，不代表所在机构的立场。

编译组：刘铁志 施羽暇 尹昊智